Wiley Series in Probability and Statistics

THIRD EDITION

Nonparametric Statistical Methods

MYLES HOLLANDER
DOUGLAS A. WOLFE
ERIC CHICKEN

WILEY

Nonparametric Statistical Methods

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Nonparametric Statistical Methods

Third Edition

Myles Hollander

Department of Statistics
Florida State University
Tallahassee, Florida

Douglas A. Wolfe

Department of Statistics
Ohio State University
Columbus, Ohio

Eric Chicken

Department of Statistics
Florida State University
Tallahassee, Florida

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Hollander, Myles.
Nonparametric statistical methods. –Third edition / Myles Hollander, Department of Statistics, Florida State University, Tallahassee, Florida, Douglas A. Wolfe, Department of Statistics, Ohio State University, Columbus, Ohio, Eric Chicken, Department of Statistics, Florida State University, Tallahassee, Florida.
pages cm
Includes bibliographical references and indexes.
ISBN 978-0-470-38737-5 (hardback) -
1. Nonparametric statistics. I. Wolfe, Douglas A. II. Chicken, Eric, 1963– III. Title.
QA278.8.H65 2013
519.5–dc23

2013007061

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To our wives,
Glee, Marilyn, and Rebecca.

Contents

Preface

The nonparametric approach is the preferred methodology for statisticians and other scientists. We list some of its advantages in Section 1.1. In the third edition, we retain our emphasis on applications to real-world situations. We want our readers to learn how to apply nonparametric techniques in a variety of settings and to understand the assumptions underlying the methods.

In this third edition, we have improved the 11 chapters of the second edition and added five new chapters. The new chapters cover topics of recent and current interest, namely, density estimation, wavelets, smoothing, ranked set sampling, and Bayesian nonparametrics. R programs are now used to perform calculations. See Section 1.5 for a description of R.

The second edition was used primarily for a one-semester senior undergraduate/first-year graduate course for students having had a prior course in statistics. With the added coverage here, there is ample material for a two-semester course. Nevertheless, we expect most teachers will still opt for a one-semester course and choose specific chapters in accordance with their interests and those of their students.

Many friends and colleagues have helped us with this project.

Grant Schneider, a graduate student at the Ohio State University, provided invaluable support in the conversion from complete reliance on null distribution tables in our second edition to the exclusive use of R programs to obtain appropriate critical values and compute associated *P*-values in this third edition. He wrote new R programs for all of the statistical procedures in Chapter 15 and for a majority of the many procedures in Chapters 5–7, and modified existing programs for the other procedures in those three chapters, leading to significantly improved computational speed in most cases. He also organized all of the R programs used in this third edition into a documented collection that is formally registered as an R package specifically linked to this third edition. We owe Grant a special thanks for his leadership role in this important aspect of our new edition.

Rachel Becvarik wrote new R programs for Chapters 11 and 16 and provided a spark.

Jelani Wiltshire and Michael Rosenthal assisted with LaTeX typesetting.

James Stricherz provided computing support and Pamela McGhee and Marylou Tatis provided office support.

Our editors Steve Quigley, Susanne Steitz-Filler, and Sari Friedman were dedicated from the inception to the completion. Our production manager Melissa Yanuzzi carefully guided the manuscript through the production process.

To all these helpmates, we are very grateful.

Tallahassee, Florida
Columbus, Ohio
Tallahassee, Florida
August 2013

MYLES HOLLANDER
DOUGLAS A. WOLFE
ERIC CHICKEN

Chapter 1

Introduction

1.1 ADVANTAGES OF NONPARAMETRIC METHODS

Roughly speaking, a nonparametric procedure is a statistical procedure that has certain desirable properties that hold under relatively mild assumptions regarding the underlying populations from which the data are obtained. The rapid and continuous development of nonparametric statistical procedures over the past $7\frac{1}{2}$ decades is due to the following advantages enjoyed by nonparametric techniques:

1. Nonparametric methods require few assumptions about the underlying populations from which the data are obtained. In particular, nonparametric procedures forgo the traditional assumption that the underlying populations are normal.
2. Nonparametric procedures enable the user to obtain exact *P*-values for tests, exact coverage probabilities for confidence intervals, exact experimentwise error rates for multiple comparison procedures, and exact coverage probabilities for confidence bands without relying on assumptions that the underlying populations are normal.
3. Nonparametric techniques are often (although not always) easier to apply than their normal theory counterparts.
4. Nonparametric procedures are often quite easy to understand.
5. Although at first glance most nonparametric procedures seem to sacrifice too much of the basic information in the samples, theoretical efficiency investigations have shown that this is not the case. Usually, the nonparametric procedures are only slightly less efficient than their normal theory competitors when the underlying populations are normal (the home court of normal theory methods), and they can be mildly or wildly more efficient than these competitors when the underlying populations are not normal.
6. Nonparametric methods are relatively insensitive to outlying observations.
7. Nonparametric procedures are applicable in many situations where normal theory procedures cannot be utilized. Many nonparametric procedures require just the ranks of the observations, rather than the actual magnitude of the observations, whereas the parametric procedures require the magnitudes.
8. The Quenouille–Tukey jackknife (Quenouille (1949), Tukey (1958, 1962)) and Efron's computer-intensive (1979) bootstrap enable nonparametric approaches to be used in many complicated situations where the distribution theory

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

needed to support parametric methods is intractable. See Efron and Tibshirani (1994).

9. Ferguson's Dirichlet process (1973) paved the way to combine the advantages of nonparametric methods and the use of prior information to form a Bayesian nonparametric approach that does not require distributional assumptions.
10. The development of computer software has facilitated fast computation of exact and approximate P-values for conditional nonparametric tests.

1.2 THE DISTRIBUTION-FREE PROPERTY

The term *nonparametric*, introduced in Section 1.1, is imprecise. The related term *distribution-free* has a precise meaning. The distribution-free property is a key aspect of many nonparametric procedures. In this section, we informally introduce the concept of a distribution-free test statistic. The related notions of a distribution-free confidence interval, distribution-free multiple comparison procedure, distribution-free confidence band, asymptotically distribution-free test statistic, asymptotically distribution-free multiple comparison procedure, and asymptotically distribution-free confidence band are introduced at appropriate points in the text.

Distribution-Free Test Statistic

We introduce the concept of a distribution-free test statistic by referring to the two-sample Wilcoxon rank sum statistic, which you will encounter in Section 4.1.

The data consist of a random sample of m observations from a population with continuous probability distribution F_1 and an independent random sample of n observations from a second population with continuous probability distribution F_2. The null hypothesis to be tested is

$$H_0 : F_1 = F_2 = F, \ F \text{ unspecified.}$$

The null hypothesis asserts that the two random samples can be viewed as a single sample of size $N = m + n$ from a common population with unknown distribution F. The Wilcoxon (1945) statistic W is obtained by ranking the combined sample of N observations jointly from least to greatest. The test statistic is W, the sum of the ranks obtained by the Y's in the joint ranking.

When H_0 is true, the distribution of W does not depend on F; that is, when H_0 is true, for all a-values, the probability that $W \leq a$, denoted by $P_0(W \leq a)$, does not depend on F.

$$P_0(W \leq a) \text{ does not depend on } F. \tag{1.1}$$

The distribution-free property given by (1.1) enables one to obtain the distribution of W under H_0 without specifying the underlying F. It further enables one to exactly specify the type I error probability (the probability of rejecting H_0 when H_0 is true) without making distributional assumptions, such as the assumption that F is a normal distribution; this assumption is required by the parametric t-test.

The details concerning how to perform the Wilcoxon test are given in Section 4.1.

1.3 SOME REAL-WORLD APPLICATIONS

This book stresses the application of nonparametric techniques to real data. The following 10 examples are a sample of the type of problems you will learn to analyze using nonparametric methods.

EXAMPLE 1.1 ***Dose–Response Relationship.***

In many situations, a dose–response relationship may not be monotonic in the dosage. For example, with *in vitro* mutagenicity assays, experimental organisms may not survive the toxic side effects of high doses of the test agent, so there may be a reduction in the number of organisms at risk of mutation. This would lead to a downturn (i.e., an umbrella pattern) in the dose–response curve. The data in Table 6.10 were considered by Simpson and Margolin (1986) in a discussion of the analysis of the Ames test results. Plates containing *Salmonella* bacteria of strain TA98 were exposed to various doses of Acid Red 114. Table 6.10 gives the number of visible revertant colonies on the 18 plates in the study, three plates for each of the six doses (in μg/ml): 0, 100, 333, 1000, 3333, and 10,000. How can we test the hypothesis of equal population median numbers at each dose against the alternative that the peak of the dose–response curve occurs at 1000 μg/ml? How can we determine which particular pairs of doses, if any, significantly differ from one another in the number of revertant colonies? Which particular doses, out of 100, 333, 1000, 3333, and 10,000, differ significantly from the 0 dose in terms of the number of revertant colonies? For doses that significantly differ, how can we estimate the magnitude of the difference? How can we simultaneously estimate all 15 "contrasts," $\tau_1 - \tau_2, \tau_1 - \tau_3, \tau_1 - \tau_4, \tau_1 - \tau_5, \tau_1 - \tau_6, \tau_2 - \tau_3, \tau_2 - \tau_4, \tau_2 - \tau_5, \tau_2 - \tau_6, \tau_3 - \tau_4, \tau_3 - \tau_5, \tau_3 - \tau_6, \tau_4 - \tau_5, \tau_4 - \tau_6, \tau_5 - \tau_6$, where, for example, $\tau_1 - \tau_2$ denotes the difference between the population medians at dose 0 and dose 100. The methods in Chapter 6 can be used to answer these questions.

EXAMPLE 1.2 ***Shelterbelts.***

Shelterbelts are long rows of tree plantings across the direction of prevailing winds. They are used in developed countries to protect crops and livestock from the effects of the wind. A study was performed by Ujah and Adeoye (1984) to see if shelterbelts would limit severe losses from droughts regularly experienced in the arid and semiarid zones of Nigeria. Droughts are considered to be a leading factor in declining food production in Nigeria and in the neighboring countries. Ujah and Adeoye studied the effect of shelterbelts on a number of factors related to drought conditions, including wind velocity, air and soil temperatures, and soil moisture. Their experiment was conducted at two locations about $3\frac{1}{2}$ km apart, near Dambatta. Table 7.7 presents the wind velocity data, averaged over the two locations, at various distances leeward of the shelterbelt. The data are given as percent wind speed reduction relative to the wind velocity on the windward side of the shelterbelt. The data are given for 9 months (data were not available for July, November, and December) and five leeward distances, namely, 20, 40, 100, 150, and 250 m, from the shelterbelt. Does the percent reduction in average wind speed tend to decrease as the leeward distance from a shelterbelt increases? Which particular leeward distances, if any, significantly differ from one another in percent reduction in average wind speed? How can the difference in percent reduction for two leeward distances be

estimated? Chapter 7 presents nonparametric methods that will enable you to analyze the data and answer these questions.

EXAMPLE 1.3 ***Nasal Brushing.***

In order to study the effects of pharmaceutical and chemical agents on mucociliary clearance, doctors often use the ciliary beat frequency (CBF) as an index of ciliary activity. One accepted way to measure CBF in a subject is through the collection and analysis of an endobronchial forceps biopsy specimen. This technique is, however, a rather invasive method for measuring CBF. In a study designed to assess the effectiveness of less invasive procedures for measuring CBF, Low et al. (1984) considered the alternative technique of nasal brushing. The data in Table 8.10 are a subset of the data collected by Low et al. during their investigation. The subjects in the study were all men undergoing bronchoscopy for the diagnosis of a variety of pulmonary problems. The CBF values reported in Table 8.10 are averages of 10 consecutive measurements on each subject.

How can we test the hypothesis of independence versus the alternative that the CBF measurements corresponding to nasal brushing and endobronchial forceps are positively associated? If there is evidence that the alternative is true, this would support the notion that nasal brushing is an acceptable substitute to measure CBF for the more invasive endobronchial forceps biopsy technique. How can we obtain an estimate of a measure of the strength of association between the two techniques' CBF values? How can we compute confidence intervals for such a measure? These questions can be answered by the methods described in Chapter 8.

EXAMPLE 1.4 ***Coastal Sediments.***

Coastal sediments are an important reservoir for organic nitrogen (ON). The degradation of ON is bacterially mediated. The mineralization of ON involves several distinct steps, and it is possible to measure the rates of these processes at each step. During the first stage of ON remineralization, ammonium is generated by heterotrophic bacteria during a process called *ammonification*. Ammonium can then be released to the environment or can be microbially transformed to other nitrogenous species. The data in Table 9.4 are from the work by Mortazavi (1997) and are based on four sediment cores that were collected in Apalachicola Bay, Florida, in April 1995 and brought back to the main campus at the Florida State University for analysis. The flux of ammonium to the overlying water was measured in each core during a 6-h incubation period. It is desired to know if there is a significant difference in ammonium flux between the cores. This is a regression problem, and it can be studied using the methods in Chapter 9.

EXAMPLE 1.5 ***Care Patterns for Black and White Patients with Breast Cancer.***

Diehr et al. (1989) point out that it is well known that the survival rate of women with breast cancer tends to be lower in Blacks than Whites. Diehr and her colleagues sought to determine if these survival differences could be accounted for by differences in diagnostic methods and treatments. Diehr et al. reported on various breast cancer patterns; one pattern of interest was *liver scan*. Did patients with local or regional disease have a

liver scan or CT scan of the liver? The data are given in Table 10.14. The data are for the 19 hospitals (out of 107 hospitals participating in the study) that had enough Black patients for individual analysis. How can we determine, for a specific hospital, if there was a significant difference between the chance of a White patient receiving a scan and the chance of a Black patient receiving a scan? How can the data from the 19 hospitals be utilized to get an overall assessment? The methods in Chapter 10 provide the means to answer these questions.

EXAMPLE 1.6 ***Times to First Review.***

The data in Table 11.18, from Hollander, McKeague, and Yang (1997), relate to 432 manuscripts submitted to the Theory and Methods Section of the *Journal of the American Statistical Association* (*JASA*) in the period January 1, 1994, to December 13, 1994. Of interest is the time (in days) to first review. When the data were studied on December 13, 1994, 158 papers had not yet received the first review. For example, for a paper received by the *JASA* on November 1, 1994, and still awaiting the first review on December 13, 1994, we know on December 13 that its time to review is greater than 33 days, but at that point we do not know the actual time to review. The observation is said to be *censored*. How can we use the censored and uncensored observations (i.e., the ones for which we know the exact times to first review) to estimate the distribution of the time to first review? Chapter 11 shows how to estimate distributions when some of the data are censored.

EXAMPLE 1.7 ***Spatial Ability Scores of Students.***

In a study examining the relation between student mathematical performance and their preference for solving problems, Haciomeroglu and Chicken (2011) gathered data on a student's spatial ability using four tests of visualization. For each student, these four test scores were combined into a single score representing their overall measure of spatial ability. High scores are associated with students with strong spatialization skills, while low scores reflect weak spatialization. The spatial ability scores for 68 female and 82 male high school students enrolled in advanced placement calculus classes in Florida are given in Tables 12.1 and 12.3, respectively. What is the distribution of spatial ability scores for the population represented by this sample of data? Does the distribution for the male students appear to possess different characteristics than that of the female students? These questions are problems in density estimation. Methods on this are given in Chapter 12.

EXAMPLE 1.8 ***Sunspots.***

Andrews and Herzberg (1985) provide data on mean monthly sunspot observations collected at the Swiss Federal Observatory in Zurich and the Tokyo Astronomical Observatory from the years 1749 to 1983. The data display excessive variability over time, obscuring any underlying trend in the cycle of sunspot appearances. The data do not follow any apparent analytical form or simple parametric model so a general nonparametric regression setting is appropriate. A powerful method for obtaining the trend from a noisy set of observations in cases such as this is by the use of wavelet estimation and thresholding. Wavelet analysis will provide a smoothed and accurate estimate of

the noise-free trend underlying the observed data. Chapter 13 provides details on using wavelet methods for this type of problem.

EXAMPLE 1.9 ***Effective Auditing to Detect Fraud.***

Account auditing is one of the most important ways to ensure that a company's stated records accurately represent the true financial transactions of the business. Being able to detect fraudulent accounting practices is vital to the integrity of the business and its management. Statistical sampling is a well-established approach for conducting such audits, as in almost all settings, the number of accounts of interest is far too large for a complete census. One major concern with statistical audits is that assessing the true values of the accounts selected to be part of the statistical sample can be quite time-intensive and, hence, expensive. It is therefore of interest to limit the number of accounts sampled for audit, while still providing adequate assurance that we gather enough information to accurately assess the reliability of the company's financial records. A ranked set sampling approach to select representative observations from a population allows an auditor to formally audit fewer accounts while maintaining the desired level of precision in his or her assessment. This leads to time savings and overall cost reduction for the auditing process. Tackett (2012) provided a collection of sales invoice records data for an electrical/plumbing distribution center that contained some fraudulent accounts where the charges (stated book values) for transactions were larger than the audited values for the materials actually delivered in those transactions. These data are given in Table 15.1. The ranked set sampling techniques described in Chapter 15 provide an effective mechanism for minimizing the auditing expense in assessing the fraudulent nature of these sales invoice records.

EXAMPLE 1.10 ***Times to Death with Cardiovascular Disease.***

The Framingham Heart Study is a well-known ongoing longitudinal study of cardiovascular disease. The original study cohort consisted of a random sample of 5209 adults aged 28 through 62 years residing in Framingham, Massachusetts between 1948 and 1951. The data in Table 16.1 were provided by McGee (2010) and consist of an extinct cohort of 12 men who were 67 years and over at the time of the fourth exam. How can we estimate the survival distribution underlying this population? How can we incorporate expert opinion concerning the remaining life for men under those or similar situations? This is a survival problem that incorporates prior information. It can be studied using the methods of Chapter 16.

1.4 FORMAT AND ORGANIZATION

The basic data, assumptions, and procedures are described precisely in each chapter according to the following format. *Data* and *Assumptions* are specified before the group of particular procedures discussed. Then, for each technique, we include (when applicable) the following subsections: *Procedure*, *Large-Sample Approximation*, *Ties*, *Example*, *Comments*, *Properties*, and *Problems*. We now describe the purpose of each subsection.

Procedure

This subsection contains a description of how to apply the procedure under discussion.

Large-Sample Approximation

This subsection contains an approximation to the method described in *Procedure*. The approximation is intended for use when the sample size (or sample sizes, as the case may be) is large. Our R programs enable small-sample and large-sample applications.

Ties

A common assumption in the development of nonparametric procedures is that the underlying population(s) is (are) continuous. This assumption implies that the probability of obtaining tied observations is zero. Nevertheless, tied observations do occur in practice. These ties may arise when the underlying population is not continuous. They may even arise if the continuity assumption is valid. We simply may be unable, owing to inaccuracies in measurement, to distinguish between two very close observations (temperatures, lengths, etc.) that emanate from a continuous population. The *Ties* subsection contains a prescription to adjust the necessary steps in the *Procedure* in order that we may treat tied observations. The adjusted procedure should then be viewed as an approximation.

Example

This subsection is basic to our text. We present a problem in which the procedure under discussion is applicable. The reader has a set of data he or she may use to apply each step of the *Procedure*, to become familiar with our notation, and to gain familiarity in performing the method. In many examples, computations are done directly and using R commands. In addition to practice, the example provides the first step toward developing an appreciation for the simplicity (difficulty) of the procedure and toward developing an intuitive feeling of how the procedure summarizes the data. The enthusiastic reader can seek out the journal article on which the example is based to obtain a more detailed specification of the experiment (in some cases our descriptions of the experiments are simplified so that the examples can be easily explained) and to question whether the *Assumptions* underlying the nonparametric method are indeed satisfied.

Comments

The comments supplement the text. In the comments, we may discuss the underlying assumptions, give an intuitive motivation for the method being considered, relate the method to other procedures in different parts of the book, provide helpful computational hints, or single out certain references including historical references.

Properties

This subsection is primarily intended as a set of directions for the reader who wishes to probe the theoretical aspects of the subject and, in particular, the theory of the procedure

under discussion. No theory is presented, but the citations guide the reader to sources furnishing the basic properties and their derivations.

Problems

Typically, the first problem of each *Problems* subsection provides practice in applying the procedure just introduced. Some problems require a comparison of an exact procedure with its large-sample approximation. Other problems are more thought provoking. We sometimes ask the reader to find or create an example that illustrates a desirable or undesirable property of the procedure under discussion.

There are occasional deviations from the format. For example, in many of the sections devoted to estimators and confidence intervals, there is no need for a *Ties* subsection, because the procedures described are well defined even when ties observations occur. In some chapters, the *Assumptions* are given before the particular (group of) sections that contain procedures based on those *Assumptions.*

Efficiency

How do the nonparametric procedures we present compare with their classical competitors, which are based on parametric assumptions such as the assumption of normality for the underlying populations? The answer depends on the particular problem and procedures under consideration. When possible, we indicate a partial answer in an efficiency section at the end of each chapter.

1.5 COMPUTING WITH R

In many of our *Example* subsections, we not only illustrate the direct computation of the procedure but also provide the output obtained using various commands in the statistical computing package R. R is a general-purpose statistical package that provides a wide range of data analysis capabilities. It is an open source program that is available for a variety of computing platforms. Users may obtain the software free of charge through the Comprehensive R Archive Network (CRAN). CRAN is a network of ftp and web servers that provide all the necessary files and instructions for downloading and installing R. It also contains numerous manuals and FAQs to assist users.

One of the strengths of R is its openness. Individuals around the world may create packages of statistical commands and routines to be distributed to any other interested users through CRAN. The standard distribution of R contains the resources to perform many of the nonparametric methods described in this book. Additional packages are readily available that perform more specialized analyses such as the density estimation procedures and wavelet analyses in the book's later chapters. Whenever a command is referenced that is not a part of the standard installation of R but instead comes from an add-on package, we make a note of this and specify which package is needed to perform the analysis.

R is also a programming language. If one cannot find an existing statistical methodology within R that will perform a suitable analysis, it is possible to program unique commands to fill this void. This falls under the topic of programming, rather than statistical analysis. As such, programming within R is not covered. The main procedures

discussed in this book have specific sets of existing commands that will perform the appropriate actions.

Many analyses include graphical as well as numeric output. R has a significant number of built-in graphing functions and is very flexible in that it allows users to create unique and detailed graphs to suit their specific needs.

The results of statistical analyses performed using R may vary slightly from those presented in the text. When they exist, these differences will be minor and will depend on the hardware configuration of the machine used to run the analyses. We also note that, for large sample sizes, many of the programs will use Monte Carlo approximations by default. Specifying methods="Exact", while more computationally intensive, will ensure that the user's output matches the text.

1.6 HISTORICAL BACKGROUND

Binomial probability calculations were used early in the eighteenth century by the British physician Arbuthnott (1710) (see Comment 2.13). Nevertheless, Savage (1953) (also see Savage (1962)) designated 1936 as the true beginning of the subject of nonparametric statistics, marked by the publication of the Hotelling and Pabst (1936) paper on rank correlation. Scheffé (1943), in a survey paper, pointed to (among others) the articles by Pearson (1900, 1911) and the presence of the sign test in Fisher's first edition of "Statistical Methods for Research Workers" Fisher (1925). Other important papers, in the late 1930s, include those by Friedman (1937), Kendall (1938), and Smirnov (1939). Wilcoxon (1945), in a paper that is brief, yet elegant in its simplicity and usefulness, introduced his now-famous two-sample rank sum test and paired-sample signed rank test. The rank sum test was given by Wilcoxon only for equal sample sizes, but Mann and Whitney (1947) treated the general case. Wilcoxon's procedures played a major role in stimulating the development of rank-based procedures in the 1950s and 1960s, including rank procedures for multivariate situations. Further momentum was provided by Pitman (1948), Hodges and Lehmann (1956), and Chernoff and Savage (1958), who showed that nonparametric rank tests have desirable efficiency properties relative to parametric competitors. An important advance that enabled nonparametric methods to be used in a variety of situations was the jackknife, introduced by Quenouille (1949) as a bias-reduction technique and extended by Tukey (1958, 1962) to provide approximate significance tests and confidence intervals.

There was major nonparametric research in the 1960s, and the most important contribution was that of Hodges and Lehmann (1963). They showed how to derive estimators from rank tests and established that these estimators have desirable properties. Their work paved the way for the nonparametric approach to be used to derive estimators in experimental design settings and for nonparametric testing and estimation in regression. Two seminal papers in the 1970s are those by Cox (1972) and Ferguson (1973). Cox's paper sparked research on nonparametric models and methods for survival analysis. Ferguson (1973) presented an approach (based on his Dirichlet process prior) to nonparametric Bayesian methods that combines the advantages of the nonparametric approach and the use of prior information incorporated in Bayesian procedures. Susarla and van Ryzin (1976) used Ferguson's approach to derive nonparametric Bayesian estimators of survival curves. Dykstra and Laud (1981) used a different prior, the gamma process, to develop a Bayesian nonparametric approach to reliability. Hjort (1990b) proposed nonparametric Bayesian estimators based on using beta processes to model the cumulative hazard

function. In the late 1980s and the 1990s, there was a surge of activity in Bayesian methods due to the Markov chain Monte Carlo (MCMC) methods (see, for example, Gelfand and Smith (1990), Gamerman (1991), West (1992), Smith and Roberts (1993), and Arjas and Gasbarra (1994)). Gilks, Richardson, and Spiegelhalter (1996) give a practical review. Key algorithms for developing and implementing modern Bayesian methods include the Metropolis–Hastings–Green algorithm (see Metropolis et al. (1953), Hastings (1970), and Green (1995)) and the Tanner–Wong (1987) data augmentation algorithm.

One of the important advances in nonparametric statistics in the past $3\frac{1}{2}$ decades is Efron's (1979) bootstrap. Efron's computer-intensive method makes use of the (ever-increasing) computational power of computers to provide standard errors and confidence intervals in many settings, including complicated situations where it is difficult, if not impossible, to use a parametric approach (see Efron and Tibshirani (1994)).

In the new millennium, the development of nonparametric techniques continues at a vigorous pace. The *Journal of Nonparametric Statistics* is solely devoted to nonparametric methods and nonparametric articles are prevalent in most statistical journals. A special issue of *Statistical Science* (Randles, Hettmansperger, and Casella, 2004) contains papers written by nonparametric experts on a wide variety of topics. These include articles on robust analysis of linear models (McKean, 2004), comparing variances and other dispersion measures (Boos and Brownie, 2004), use of sign statistics in one-way layouts (Elmore, Hettmansperger, and Xuan, 2004), density estimation (Sheather, 2004), multivariate nonparametric tests (Oja and Randles, 2004), quantile–quantile (QQ) plots (Marden, 2004), survival analysis (Akritas, 2004), spatial statistics (Chang, 2004), ranked set sampling (Wolfe, 2004), reliability (Hollander and Peña, 2004), data modeling via quantile methods (Parzen, 2004), kernel smoothers (Schucany, 2004), permutation-based inference (Ernst, 2004), data depth tests for location and scale differences for multivariate distributions (Li and Liu, 2004), multivariate signed rank tests in time series problems (Hallin and Paindaveine, 2004), and rank-based analyses of crossover studies (Putt and Chinchilli, 2004).

Books dealing with certain topics in nonparametrics include those on survival analysis (Kalbfleisch and Prentice, 2002 and Klein and Moeschberger, 2003), density estimation, smoothers and wavelets (Wasserman, 2006), rank-based methods (Lehmann and D'Abrera, 2006), reliability (Gámiz, Kulasekera, Limnios, and Lindquist, 2011), and categorical data analysis (Agresti, 2013).

We delineated advantages of the nonparametric approach in Section 1.1. In addition to those practical advantages, the theory supporting nonparametric methods is elegant, and researchers find it challenging to advance the theory. The primary reasons for the success and use of nonparametric methods are the wide applicability and desirable efficiency properties of the procedures and the realization that it is sound statistical practice to use methods that do not depend on restrictive parametric assumptions because such assumptions often fail to be valid.

Chapter 2

The Dichotomous Data Problem

INTRODUCTION

In this chapter the primary focus is on the dichotomous data problem. The data consists of n independent repeated Bernoulli trials having constant probability of success p. On the basis of these outcomes, we wish to make inferences about p. Section 2.1 introduces the binomial distribution and presents a binomial test for the hypothesis $p = p_0$, where p_0 is a specified success probability. Section 2.2 gives a point estimator $\widehat{p}$ for p. Section 2.3 presents confidence intervals for p. Section 2.3 also contains the generalization of the binomial distribution to the multinomial distribution, confidence intervals for multinomial probabilities and a test that the multinomial probabilities are equal to specified values. Section 2.4 presents Bayesian competitors to the frequentist estimator $\widehat{p}$ of Section 2.2. The Bayesian estimators incorporate prior information.

Data. We observe the outcomes of n independent repeated Bernoulli trials.

Assumptions

A1. The outcome of each trial can be classified as a success or a failure.

A2. The probability of a success, denoted by p, remains constant from trial to trial.

A3. The n trials are independent.

2.1 A BINOMIAL TEST

Procedure

To test

$$H_0 : p = p_0, \tag{2.1}$$

where p_0 is some specified number, $0 < p_0 < 1$, set

$$B = \text{number of successes.} \tag{2.2}$$

a. *One-Sided Upper-Tail Test.* To test

$$H_0 : p = p_0$$

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

versus

$$H_1 : p > p_0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } B \geq b_\alpha; \text{ otherwise do not reject,} \tag{2.3}$$

where the constant b_α is chosen to make the type I error probability equal to α. The number b_α is the upper α percentile point of the binomial distribution with sample size n and success probability p_0. Due to the discreteness of the binomial distribution, not all values of α are available (unless one resorts to randomization). Comment 3 explains how to obtain the b_α values. See also Example 2.1.

b. *One-Sided Lower-Tail Test.* To test

$$H_0 : p = p_0$$

versus

$$H_2 : p < p_0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } B \leq c_\alpha; \text{ otherwise do not reject.} \tag{2.4}$$

Values of c_α can be determined as described in Comment 3. Here, c_α is the lower α percentile point of the binomial distribution with sample size n and success probability p_0. For the special case of testing $p = \frac{1}{2}$,

$$c_\alpha = n - b_\alpha. \tag{2.5}$$

Equation 2.5 is explained in Comment 4.

c. *Two-Sided Test.* To test

$$H_0 : p = p_0$$

versus

$$H_3 : p \neq p_0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } B \geq b_{\alpha_1} \text{ or } B \leq c_{\alpha_2}; \text{ otherwise do not reject,} \tag{2.6}$$

where b_{α_1} is the upper α_1 percentile point, c_{α_2} is the lower α_2 percentile point, and $\alpha_1 + \alpha_2 = \alpha$. See Comment 3.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of B, suitably standardized. To standardize, we need to know the mean and variance of B when the null hypothesis is true. When H_0 is true, the mean and variance of B are, respectively,

$$E_{p_0}(B) = np_0, \tag{2.7}$$

$$\text{var}_{p_0}(B) = np_0(1 - p_0). \tag{2.8}$$

Comment 8 gives the derivations for (2.7) and (2.8).

The standardized version of B is

$$B^* = \frac{B - E_{p_0}(B)}{\{\text{var}_{p_0}(B)\}^{1/2}} = \frac{B - np_0}{\{np_0(1 - p_0)\}^{1/2}}. \tag{2.9}$$

When H_0 is true, B^* has, as n tends to infinity, an asymptotic $N(0, 1)$ distribution. Let z_α denote the upper α percentile point of the $N(0, 1)$ distribution. To find z_α, we use the `qnorm`(1-α,0,1). For example, to find $z_{.05}$, we apply qnorm(.95,0,1) and obtain $z_{.05} = 1.645$.

The normal approximation to procedure (2.3) is

$$\text{Reject } H_0 \text{ if } B^* \geq z_\alpha; \text{ otherwise do not reject.} \tag{2.10}$$

The normal approximation to procedure (2.4) is

$$\text{Reject } H_0 \text{ if } B^* \leq -z_\alpha; \text{ otherwise do not reject.} \tag{2.11}$$

The normal approximation to procedure (2.6), with $\alpha_1 = \alpha_2 = \alpha/2$, is

$$\text{Reject } H_0 \text{ if } |B^*| \geq z_{\alpha/2}; \text{ otherwise do not reject.} \tag{2.12}$$

EXAMPLE 2.1 ***Canopy Gap Closure.***

Dickinson, Putz, and Canham (1993) investigated canopy gap closure in thickets of the clonal shrub *Cornus racemosa.* Shrubs often form dense clumps where tree abundance has been kept artificially low (e.g., on power-line right of ways). These shrub clumps then retard reinvasion of the sites by trees. Individual clumps may persist for many years. Clumps outlast the lives of the individual stems of which they are formed; stems die and leave temporary holes in the canopies of the clumps. Closure of the hope (gap) left by dead stems occurs in part by the lateral growth of stems that surround the hole. Opening of the gap often occurs when individual branches of hole-edge stems die. Between sample dates, more branches in six out of seven gaps in clumps, at a site with nutrient-poor and dry soil, died than lived. Let us say we have a success if more branches die than live in the gaps in clumps. Let p denote the corresponding probability of success. We suppose that the success probability for sites that are nutrient rich with moist soil has been established by previous studies to be 15%. Do the nutrient-poor and dry soil sites

have the same success probability as the nutrient-rich and moist soil sites or is it larger? This reduces to the hypothesis-testing problem

$$H_0 : p = .15$$

versus

$$H_1 : p > .15.$$

Our sample size is $n = 7$ and we observe $B = 6$ successes. From the R command `round(pbinom(0:7,7,.15,lower.tail=F),4)`, we obtain, rounded to four places, the probabilities $P_{.15}(B > x)$ for $x = 0, \ldots, 7$. (The notation $P_{.15}(B > x)$ is shorthand for the probability that $B > x$, computed under the assumption that the true success probability is .15.) The $P_{.15}(B > x)$ probabilities are

x	0	1	2	3	4	5	6	7
$P_{.15}(B > x)$	.6794	.2834	.0738	.0121	.0012	.0001	.0000	.0000

To find $P_{.15}(B \geq x)$ note $P_{.15}(B \geq x) = P_{.15}(B > x - 1)$. Reasonable possible choices for α are .0738, .0121, .0012, .0001. Suppose we choose to use $\alpha = .0121$. We note $P(B > 3) = P(B \geq 4) = .0121$ and thus we see $b_{.0121} = 4$. Thus the $\alpha = .0121$ test is

$$\text{Reject } H_0 \text{ if } B \geq 4; \text{ otherwise do not reject.}$$

Our observed value is $B = 6$ and thus we reject H_0 at $\alpha = .0121$. To find the P-value, which is $P_{.15}(B \geq 6)$, we can find $P_{.15}(B > 5)$ using the R command `pbinom(5,7,.15,lower.tail=F)`. Alternatively, we can find the P-value using the R command `binom.test(6,7,.15,"g")`. We find $P = .000069$, or rounded to four places, P is .0001. This is the smallest significance level at which we can reject H_0 (in favor of the alternative $p > .15$) with our observed value of B. We conclude that there is strong evidence against H_0 favoring the alternative. For more on the P-value, see Comment 9.

EXAMPLE 2.2 ***Sensory Difference Tests.***

Sensory difference tests are often used in quality control and quality evaluation. The triangle test (cf. Bradley, 1963) is a sensory difference test that provides a useful application of the binomial model. In its simplest form, the triangle procedure is as follows. To each of n panelists, three test samples are presented in a randomized order. Two of the samples are known to be identical; the third is different. The panelist is then supposed to select the odd sample, perhaps on the basis of a specified sensory attribute. If the panelists are homogeneous trained judges, the experiment can be viewed as n independent repeated Bernoulli trials, where a success corresponds to a correct identification of the odd sample. (If the panelists are not homogeneous trained judges, we may question the validity of Assumption A2.) Under the hypothesis that there is no basis for discrimination, the probability p of success is $\frac{1}{3}$, whereas a basis for discrimination would correspond to values of p that exceed $\frac{1}{3}$.

Byer and Abrams (1953) considered triangular bitterness tests in which each taster received three glasses, two containing the same quinine solution and the third a different

quinine solution. In their first bitterness test, the solutions contained .0075% and .0050%, respectively, of quinine sulfate. The six presentation orders, LHH, HLH, HHL, HLL, LHL, and LLH (L denotes the lower concentration, H the higher concentration), were randomly distributed among the tasters. Out of 50 trials, there were 25 correct selections and 25 incorrect selections.

We consider the binomial test of $H_0 : p = \frac{1}{3}$ versus the one-sided alternative $p > \frac{1}{3}$ and use the large-sample approximation to (2.3). We set $\alpha = .05$ for purposes of illustration. To find $z_{.05}$, the 95^{th} quantile of the $N(0,1)$ distribution, we use the `R` command `qnorm(.95,0,1)`, and find $z_{.05} = 1.645$. Thus approximation (2.10), at the $\alpha = .05$ level, reduces to

$$\text{Reject } H_0 \text{ if } B^* \geq 1.645; \text{ otherwise do not reject.}$$

From the data we have $n = 50$ and B (the number of correct identifications) $= 25$. Thus from (2.9), with $p_0 = \frac{1}{3}$, we obtain

$$B^* = \frac{25 - 50\left(\frac{1}{3}\right)}{\left\{50\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right\}^{1/2}} = 2.5.$$

The large sample approximation value $B^* = 2.5 > 1.645$ and thus we reject $H_0 : p = \frac{1}{3}$ in favor of $p > \frac{1}{3}$ at the approximate $\alpha = .05$ level. Thus there is evidence of a basis for discrimination in the taste bitterness test. To find the P-value corresponding to $B^* = 2.5$, one can use the `R` command `pnorm(2.5)`. The P-value is `1-pnorm(2.5)=.0062`. Thus, the smallest significance level at which we reject H_0 in favor of $p > \frac{1}{3}$ using the large-sample approximation is .0062. (Note the exact P-value in this case is given by R as `1-pbinom(24,50,1/3)=.0108`.)

Comments

1. *Binomial Test Procedures.* Assumptions A1–A3 are the general assumptions underlying a binomial experiment. Research problems possessing these assumptional underpinnings are common, and thus the binomial test procedures find frequent use. A particularly important special case in which procedures (2.3), (2.4), and (2.6) are applicable occurs when we wish to test hypotheses about the unknown median, θ, of a population. The application of binomial theory to this problem leads to a test statistic, B, that counts the number of sample observations larger than a specified null hypothesis value of θ, say θ_0. For this particular special case, the statistic B is referred to as the *sign statistic*, and the associated test procedures are referred to as sign test procedures. See Sections 3.4 and 3.8 for a more detailed discussion of the sign test procedures corresponding to (2.3), (2.4), and (2.6).
2. *Distribution-Free Test.* The critical constant b_α of (2.3) is chosen so that the probability of rejecting H_0, when H_0 is true, is α. We can control this type I error because Assumptions A1–A3 and a specification of p (the null hypothesis specifies p to be equal to p_0) determine, without further assumptions regarding the underlying populations from which the dichotomous data emanate, the probability distribution of B. Thus, under Assumptions A1–A3, the test given by (2.3) is

said to be a distribution-free test of H_0. The same statement can be made for tests (2.4) and (2.6).

3. *Illustration of Lower-Tail and Two-Tailed Tests.* Suppose $n = 8$ and we wish to test $H_0 : p = .4$ versus $p > .4$ via procedure (2.3). Using the methods illustrated in Example 2.1 to obtain binomial tail probabilities, we can find

b	0	1	2	3	4	5	6	7	8
$P_{.4}(B \geq b)$	1	.9832	.8936	.6846	.4059	.1737	.0498	.0085	.0007

(Recall that the $P_{.4}$ notation indicates that the probabilities are computed under the assumption that $p = .4$.) Hence, we can find constants b_α that satisfy the equation $P_{.4}\{B \geq b_\alpha\} = \alpha$ only for certain values of α. For $\alpha = .0085$, $b_{.0085} = 7$. For $\alpha = .0498$, $b_{.0498} = 6$. As α increases, the critical constant b_α decreases. Thus, when we increase α, it is easier to reject H_0; hence, we increase the power or, equivalently, decrease the probability of a type II error for our test (against a particular alternative). Similarly, if we lower α, we raise the probability of a type II error. This is illustrated in Comment 9.

Again consider the case $n = 8$ and suppose we want to test $p = .4$ versus the alternative $p < .4$. We can use the lower-tail test described by (2.4). For example, suppose we want $\alpha = .1064$. Then $P_{.4}\{B \geq 2\} = .8936$ and $P_{.4}\{B \leq 1\} = 1 - .8936 = .1064$. Thus, in (2.4), $c_{.1064} = 1$ and this yields the $\alpha = .1064$ test; namely, reject H_0 if $B \leq 1$ and accept H_0 if $B > 1$.

We close this comment with an example of the two-sided test described by (2.6). For convenience, we stay with the case $n = 8$ and test $H_0 : p = .4$. Note 6 is the upper .0498 percentile point of the null distribution of B and 1 is the lower .1064 percentile point. Thus the test that rejects H_0 when $B \geq 6$ or when $B \leq 1$ and accepts H_0 when $1 < B < 6$ is an $\alpha = .0498 + .1064 = .1562$ two-tailed test.

4. *Binomial Distribution.* The statistic B has been defined as the number of successes in n independent Bernoulli trials, each trial having a success probability equal to p. The distribution of the random variable B is known as the binomial distribution with parameters n and p.

For the special case when $p = \frac{1}{2}$, it can be shown that the distribution of B is symmetric about its mean $n/2$. This implies that

$$P_{.5}\{B \geq x\} = P_{.5}\{B \leq (n - x)\} \quad \text{for} \quad x = 0, \ldots, n. \tag{2.13}$$

Equation (2.13) implies that the lower α percentile point of the binomial distribution, with $p = .5$, is equal to n minus the upper α percentile point. This result was expressed by (2.5) after we introduced the lower-tail test given by (2.4).

5. *Motivation for the Test Based on B.* The statistic B/n is an estimator (see Section 2.2) of the true unknown parameter p. Thus, if $p > p_0$, B/n will tend to be larger than p_0. This suggests rejecting $H_0 : p = p_0$ in favor of $p > p_0$ for large values of B and serves as partial motivation for (2.3).

6. *An Example of the Exact Distribution of B.* The exact distribution of B can be obtained from the equation

$$B = \sum_{i=1}^{n} \psi_i, \tag{2.14}$$

where

$$\psi_i = \begin{cases} 1, & \text{if the } i\text{th Bernoulli trial is a success,} \\ 0, & \text{if the } i\text{th Bernoulli trial is failure.} \end{cases}$$

We consider the 2^n possible outcomes of the configurations $(\psi_1, \ldots, \psi_n)$ and use the fact that under H_0, any outcome with b 1's and $(n-b)$ 0's has probability $p^b(1-p)^{n-b}$. For example, in the case $n = 2$, $p = \frac{1}{4}$, the $2^2 = 4$ possible outcomes for (ψ_1, ψ_2) and associated values of B are as follows:

(ψ_1, ψ_2)	$P_{.25}\{(\psi_1, \psi_2)\}$	$B = \psi_1 + \psi_2$
$(0,0)$	$\left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^{2-0} = \frac{9}{16}$	0
$(0,1)$	$\left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^{2-1} = \frac{3}{16}$	1
$(1,0)$	$\left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^{2-1} = \frac{3}{16}$	1
$(1,1)$	$\left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^{2-2} = \frac{1}{16}$	2

Thus, for example, $P_{.25}\{B \geq 1\} = P_{.25}\{B = 1\} + P_{.25}\{B = 2\} = \frac{6}{16} + \frac{1}{16} = \frac{7}{16}$.

7. *The Exact Distribution of B.* By methods similar to the particular case illustrated in Comment 6, it can be shown that for each of the $n+1$ possible values of B (namely, $b = 0, \ldots, n$), we have

$$P_p\{B = b\} = \binom{n}{b} p^b (1-p)^{n-b}. \tag{2.15}$$

In (2.15), the symbol $\binom{n}{b}$ (read "binomial *n, b*") is given by

$$\binom{n}{b} = \frac{n!}{b!(n-b)!}, \tag{2.16}$$

where the symbol $m!$ (read "m factorial") is, for positive integers, defined as $m! = m(m-1)(m-2)\ldots(3)(2)(1)$, and 0! is defined to be equal to 1. The number $\binom{n}{b}$ is known as the number of combinations of n things taken b at a time. It is equal to the number of subsets of size b that may be formed from the members of a set of size n. The distribution given by (2.15) is known as the binomial distribution with parameters n and p.

8. *The Asymptotic Distribution of B.* Using representation (2.14), we find the mean B is

$$E_p(B) = E_p\left(\sum_{i=1}^{n} \psi_i\right) = \sum_{i=1}^{n} E_p(\psi_i) = np,$$

where we have used the calculation

$$E_p(\psi_i) = 1 \cdot P(\psi_i = 1) + 0 \cdot P(\psi_i = 0) = 1 \cdot p + 0 \cdot (1 - p) = p.$$

Then, using the fact that $\psi_1, \psi_2, \ldots, \psi_n$ are independent,

$$\text{var}_p(B) = \text{var}_p\left(\sum_{i=1}^{n} \psi_i\right) = \sum_{i=1}^{n} \text{var}_p(\psi_i). \tag{2.17}$$

The variance of any one of the indicator random variables ψ_i is determined as follows. Note $\psi_i^2 = \psi_i$ and thus

$$E_p(\psi_i^2) = E_p(\psi_i) = p,$$

and

$$\text{var}_p(\psi_i) = E_p(\psi_i^2) - \{E_p(\psi_i)\}^2 = p - p^2 = p(1 - p).$$

Hence, from (2.17),

$$\text{var}_p(B) = \sum_{i=1}^{n} p(1 - p) = np(1 - p).$$

The random variable B is a sum of independent and identically distributed random variables and hence the central limit theorem (cf. Casella and Berger, 2002, p. 236) establishes that, as $n \to \infty$, $(B - np)/\sqrt{np(1 - p)}$ has a limiting $N(0, 1)$ distribution.

9. *The P-Value.* Rather than specify an α level and report whether the test rejects at that specific α level, it is more informative to state the lowest significance level at which we can reject with the observed data. This is called the *P-value*. Consider the $\alpha = .0085$ test (test T_1, say) and the $\alpha = .0498$ test (T_2) of $H_0 : p = .4$ versus $p > .4$ for the case $n = 8$. Suppose in an actual experiment that our observed value of B is 7. Then with test T_2 we reject H_0 because the critical region for test T_2 consists of the values $\{B = 6, B = 7, B = 8\}$ and our observed value 7 is in the critical region. Thus, it is correct for us to state that the value $B = 7$ is significant at the $\alpha = .0498$ level. But the value $B = 7$ is also significant at the $\alpha = .0085$ level. If we simply state that we reject H_0 at the .0498 level, we do not convey the additional information that, with the value $B = 7$, we also can reject H_0 at the .0085 level. To remedy this, the following approach is suggested.

   Suppose, as in the previous example, large values of some statistic S (say) lead to rejection of the null hypothesis. Let s denote the observed value of S. Compute $P_0\{S \geq s\}$, the probability, under the null hypothesis, that S will be greater than or equal to s. This is the lowest level at which we can reject H_0. The observation $S = s$ will be significant at all levels greater than or equal to $P_0\{S \geq s\}$ and not significant at levels less than $P_0\{S \geq s\}$.

   To further illustrate this point, consider the test of $p = \frac{1}{3}$ versus $p > \frac{1}{3}$ of Example 2.2. We apply procedure (2.10), based on the large-sample approximation to the null distribution of B. The (approximate) $\alpha = .05$ test rejects if

$B^* \geq 1.645$ and accepts otherwise. The observed value of B^* is $B^* = 2.5$ and thus we can reject $p = \frac{1}{3}$ in favor of $p > \frac{1}{3}$ at the .05 level. In Example 2.2, we found $z_{.0062} = 2.5$. Thus, the smallest significance level at which we can reject is approximately .0062, and this statement is more informative than the statement that the .05 test leads to rejection.

10. *Calculating Power.* Take $n = 8$, and consider the following two tests of $H_0 : p = .4$ versus $p > .4$, based on (2.3). Test T_1, corresponding to $\alpha = .0085$, rejects H_0 if $B \geq 7$ and accepts otherwise. Test T_2, corresponding to $\alpha = .0498$, rejects H_0 if $B \geq 6$ and accepts otherwise. Suppose, in fact, that the alternative $p = .5$ is true. Let R_1 denote the power of the test T_1 (for this alternative) and let R_2 denote the power of the test T_2. Thus, R_1 is the probability of rejecting H_0 with test T_1 and R_2 is the probability of rejecting H_0 with test T_2. These powers are to be calculated when the alternative $p = .5$ is true. Using the R commands `pbinom(6,8,.5,lower.tail=F)` and `pbinom(5,8,.5,lower.tail=F)`, we obtain

$$R_1 = P_{.5}\{B \geq 7\} = P_{.5}\{B > 6\} = .0352$$

$$R_2 = P_{.5}\{B \geq 6\} = P_{.5}\{B > 5\} = .1445$$

For the alternative $p = .5$, let β_1 denote the probability of a type II error using test T_1 and let β_2 denote the probability of a type II error using test T_2. We find

$$\beta_1 = 1 - R_1 = .9648, \quad \beta_2 = 1 - R_2 = .8555.$$

Test T_1 has a lower probability of a type I error than test T_2, but the probability of a type II error for test T_1 exceeds that of test T_2. Incidentally, the reader should not be shocked at the very high values of β_1 and β_2. The alternative $p = .5$ is quite close to the null hypothesis value $p = .4$ and a sample of size 8 is simply not large enough to make a better (in terms of power) distinction between the hypothesis and alternative.

11. *More Power Calculations.* We return to Example 2.2 concerning sensory difference tests. Suppose we have $n = 50$ and we decide to employ the approximate $\alpha = .05$ level test of $H_0 : p = \frac{1}{3}$ versus $H_1 : p > \frac{1}{3}$. Recall that test rejects H_0 if

$$\frac{B - n\left(\frac{1}{3}\right)}{\{n\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\}^{1/2}} > 1.645$$

and accepts H_0 otherwise. What is the power of this test if in fact $p = .6$? We approximate the power using the asymptotic normality of B, suitably standardized. If $p = .6$, then

$$\frac{B - n(.6)}{\{n(.6)(.4)\}^{1/2}}$$

has an approximate $N(0, 1)$ distribution. Using this, we find

$$\begin{aligned}
\text{Power} &= P_{.6}\left\{\frac{B - n\left(\frac{1}{3}\right)}{\left\{n\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right\}^{1/2}} > 1.645\right\} \\
&= P_{.6}\left\{B > \left[(1.645)\left\{n\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right\}^{1/2}\right] + n\left(\frac{1}{3}\right)\right\} \\
&= P_{.6}\left\{\frac{B - n(.6)}{\{n(.6)(.4)\}^{1/2}} > \left[\frac{(1.645)\left\{n\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right\}^{1/2} + n\left(\frac{1}{3}\right) - n(.6)}{\{n(.6)(.4)\}^{1/2}}\right]\right\} \\
&\doteq P\{Z > -2.27\},
\end{aligned}$$

where $Z = \{B - n(.6)\}/\{n(.6)(.4)\}^{1/2}$ is approximately a $N(0, 1)$ random variable and -2.27 is the value, when $n = 50$, of the term in large square brackets. Using `1-pnorm(-2.27)`, we find power $\doteq P\{Z > -2.27\} = .9884$.

12. *Counting Failures Instead of Successes.* Define B^- to be the number of failures in the n Bernoulli trials. Note that B^- could be defined by (2.14) with ψ_i replaced by $(1 - \psi_i)$, for $i = 1, \ldots, n$. Test procedures (2.3), (2.4), and (2.6) could equivalently be based on B^-, because $B^- = n - B$.
13. *Some History*. The binomial distribution has been utilized for statistical inferences about dichotomous data for more than 300 years. Binomial probability calculations were used by the British physician Arbuthnott (1710) in the early eighteenth century as an argument for the sexual balance maintained by Divine Providence and against the practice of polygamy. Bernoulli trials are so named in honor of Jacques Bernoulli. His book "Ars Conjectandi" (1713) contains a profound study of such trials and is viewed as a milestone in the history of probability theory. (LeCam and Neyman (1965) reported that the original Latin edition was followed by several in modern languages; the last reproduction, in German, appeared in 1899 in No. 107 and No. 108 of the series *Ostwald's Klassiker der exakten Wissenschaften*, Wilhelm Engelman, Leipzig.) Today, the binomial procedures remain one of the easiest and most useful sets of procedures in the statistical catalog.

Properties

1. *Consistency.* Test procedures (2.3), (2.4), and (2.6) will be consistent against alternatives for which $p >, <$, and $\neq p_0$, respectively.

Problems

1. Stanton (1969) investigated the problem of paroling criminal offenders. He studied the behavior of all male criminals paroled from New York's correctional institutions to original parole supervision during 1958 and 1959 (exclusive of those released to other warrants or to deportation). The parolees were observed for 3 years following their releases or until they exhibited some delinquent parole behavior. In a study involving a very large number of subjects, Stanton considered criminals convicted of crimes other than first- or second-degree murder. He found that approximately 60% of these parolees did not have any delinquent behavior during the 3 years following their releases.

During the same period, Stanton found that 56 of the 65 paroled murderers (first- or second-degree murderers who were also original parolees) in the study had no delinquent parole behavior. Let a success correspond to a male murderer on original parole who does not exhibit any delinquent parole behavior in the 3-year observation period. Note that we could question Assumptions A2 in this context; parolees convicted of first-degree murder may have a different success probability than parolees convicted of second-degree murder. Even the parolees in the first-degree (or second-degree) group may have different individual success probabilities. For pedagogical purposes, we proceed as if Assumption A2 is valid and denote the common success probability by p.

It is of interest to investigate whether murderers are better risks as original parolees than are criminals convicted of lesser crimes. This suggests testing $H_0 : p = .6$ against the alternative $p > .6$. Perform this test using the large-sample approximation to procedure (2.3).

2. Describe a situation in which Assumptions A1 and A2 hold but Assumption A3 is violated.

3. Describe a situation in which Assumptions A1 and A3 hold but Assumption A2 is violated.

4. Suppose that 10 Bernoulli trials satisfying Assumptions A1–A3 result in 8 successes. Investigate the accuracy of the large-sample approximation by comparing the smallest significance level at which we would reject $H_0 : p = \frac{1}{2}$ in favor of $p > \frac{1}{2}$ when using procedure (2.3) with the corresponding smallest significance level for the large-sample approximation to procedure (2.3) given by (2.10).

5. Return to the $\alpha = .0121$ test of Example 2.1. Recall that the test of $H_0 : p = .15$ versus $H_1 : p > .15$ rejects H_0 if in $n = 7$ trials there are 4 or more successes and accepts H_0 if there are 3 or fewer successes. What is the power of that test when (a) $p = .4$, (b) $p = .6$, and (c) $p = .8$?

6. A standard surgical procedure has a success rate of .7. A surgeon claims a new technique improves the success rate. In 20 applications of the new technique, there are 18 successes. Is there evidence to support the surgeon's claim?

7. A multiple-choice quiz contains ten questions. For each question there are one correct answer and four incorrect answers. A student gets three correct answers on the quiz. Test the hypothesis that the student is guessing.

8. Return to Example 2.2 and, in the case of $n = 50$, approximate the power of the $\alpha = .05$ test when $p = .5$.

9. Forsman and Lindell (1993) studied swallowing performance of adders (snakes). Captive snakes were fed with dead field voles (rodents) of differing body masses and the number of successful swallowing attempts was recorded. Out of 67 runs resulting in swallowing attempts, 58 where successful and 9 failed. (A failure was easy to detect because the fur of a partly swallowed and regurgitated vole is slick and sticks to the anterior part of the body.) Test the hypothesis that $p = .6$ against the alternative $p > .6$.

10. Table 2.1 gives numbers of deaths in US airline accidents from 2000 to 2010. (The entry for 2001 does not include the death toll in the September 11, 2001 plane hijackings.) See the TODAY article by Levin (2011), which cites data from the National Transportation Board.

Suppose you view each trial year as a success if there are no U.S. Airline deaths and a failure otherwise. Discuss the validity of Assumptions A1 and A2. (Mann's test for trend, covered in Comment 8.14, can be used to obtain an approximate P-value for assessing the degree of trend in deaths.)

Table 2.1 Deaths in US Airlines Accidents

2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
89	266	0	22	13	22	50	0	0	50	0

Source: A. Levin (2011).

2.2 AN ESTIMATOR FOR THE PROBABILITY OF SUCCESS

Procedure

The estimator of the probability of success p, associated with the statistic B, is

$$\widehat{p} = \frac{B}{n}. \tag{2.18}$$

EXAMPLE 2.3 ***Example 2.2 Continued.***

Consider the triangle test data of Example 2.2. Then $\widehat{p} = B/n = (25/50) = .5$. Thus our point estimate of p, the probability of correctly identifying the odd sample, is $\widehat{p} = .5$.

Comments

14. *Observed Relative Frequency of Success.* The statistic $\widehat{p}$ is simply the observed relative frequency of success in n Bernoulli trials satisfying Assumptions A1-A3. Thus $\widehat{p}$ qualifies as a natural estimator of p, the unknown probability of success in a single Bernoulli trial. That is, we estimate the true unknown probability of success by the observed frequency of success.
15. *Standard Deviation of $\widehat{p}$.* We have shown in Comment 8 that the variance of B is $np(1-p)$, where p is the success probability. It follows that the variance of $\widehat{p}$ is

$$\text{var}(\widehat{p}) = \frac{p(1-p)}{n}. \tag{2.19}$$

The standard deviation of $\widehat{p}$ is

$$sd(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}}. \tag{2.20}$$

Note that $sd(\widehat{p})$ cannot be computed unless we know the value of p, but it can be estimated by substituting $\widehat{p}$ for p in (2.20). This quantity, which we denote as $\widehat{sd}(\widehat{p})$, is a consistent estimator of $sd(\widehat{p})$. The quantity $\widehat{sd}(\widehat{p})$ is also known as the *standard error* of $\widehat{p}$. We have

$$\widehat{sd}(\widehat{p}) = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}. \tag{2.21}$$

Rather than simply stating the value of $\widehat{p}$ when reporting an observed relative frequency of success, it is important to also report the value of $\widehat{sd}(\widehat{p})$, which (as does $\text{var}(\widehat{p})$) measures the variability of the estimate.

Thus, for the adder data of Problem 9, we could report

$$\widehat{p} = \frac{58}{67} = .87; \quad \widehat{sd}(\widehat{p}) = \sqrt{\frac{\left(\frac{58}{67}\right)\left(\frac{9}{67}\right)}{67}} = .04.$$

Alternatively, we could use a confidence interval for p (see Section 2.3).

16. *Sample Size Determination.* Suppose we want to choose the sample size n so that $\widehat{p}$ is within a distance D of p, with probability $1-\alpha$. That is, we want

$$P_p\left(-D < \widehat{p} - p < D\right) = 1 - \alpha.$$

This is equivalent to

$$P_p\left(\frac{-D}{\sqrt{\frac{p(1-p)}{n}}} < \frac{\widehat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{D}{\sqrt{\frac{p(1-p)}{n}}}\right) \doteq 1-\alpha.$$

The variable $(\widehat{p}-p)/\sqrt{p(1-p)/n}$ has an asymptotic $N(0,1)$ distribution and thus we know

$$P\left(-z_{\alpha/2} < \frac{\widehat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) \doteq 1-\alpha.$$

From the two previous equations, we see that

$$\frac{D}{\sqrt{\frac{p(1-p)}{n}}} \doteq z_{\alpha/2}.$$

Solving for n yields

$$n \doteq \frac{(z_{\alpha/2})^2 p(1-p)}{D^2} \tag{2.22}$$

Expression (2.22) requires a guess or estimate for p because p is not known. The function $p(1-p)$ is maximized at $p = \frac{1}{2}$ and decreases to zero as p approaches 0 or 1. Thus we obtain the most conservative sample size by substituting $\frac{1}{2}$ for p in (2.22). This yields

$$n = \frac{(z_{\alpha/2})^2}{4D^2} \tag{2.23}$$

17. *Competing Estimators.* Suppose you observe $B = 0$ in n trials. Depending on the situation, you may have little faith in the estimate $\widehat{p} = 0$. For example, you take a random sample of 10 smokers on a college campus and find no one in the sample smokes. You do not, however, believe that the probability is 0 that a randomly selected student is a smoker. A similar dilemma occurs when $B = n$. One alternative estimator of p is $\tilde{p}$ defined by (2.24) and presented in Section 2.3 on confidence intervals for p. Other alternative estimators use the Bayes estimators presented in Section 2.4.

Properties

1. *Maximum Likelihood Estimator.* The estimator $\widehat{p}$ is the maximum likelihood estimator.
2. *Standard Deviation.* For the standard deviation of $\widehat{p}$ see Comment 15.
3. *Asymptotic Normality.* For asymptotic normality, see Casella and Berger (2002, p 236).

Problems

11. Calculate $\widehat{p}$ for the parolee data of Problem 1 and obtain an estimate of the standard deviation of $\widehat{p}$.

12. Obtain an estimate for the standard deviation of the estimate $\widehat{p}$ calculated in Example 2.1.

13. Suppose $n = 7$. What are the possible values for $\widehat{p}$? When $\alpha = .05$, what are the possible values for $\tilde{p}$ defined by (2.24)?

14. Suppose you are designing a study to estimate a success probability p. Determine the sample size n so that $\widehat{p}$ is within a distance .05 of p with probability .99.

2.3 A CONFIDENCE INTERVAL FOR THE PROBABILITY OF SUCCESS (WILSON)

Procedure

Set

$$\tilde{p} = \widehat{p}\left(\frac{n}{n + z_{\alpha/2}^2}\right) + \frac{1}{2}\left(\frac{z_{\alpha/2}^2}{n + z_{\alpha/2}^2}\right), \tag{2.24}$$

$$p_{\mathrm{L}}^{W}(\alpha) = \tilde{p} - z_{\alpha/2}V^* \tag{2.25}$$

$$p_{\mathrm{U}}^{W}(\alpha) = \tilde{p} + z_{\alpha/2}V^*, \tag{2.26}$$

where

$$V^* = \left\{\frac{1}{n + z_{\alpha/2}^2}\left[\widehat{p}(1-\widehat{p})\left(\frac{n}{n + z_{\alpha/2}^2}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{z_{\alpha/2}^2}{n + z_{\alpha/2}^2}\right)\right]\right\}^{1/2}. \tag{2.27}$$

With $p_{\mathrm{L}}(\alpha)$ and $p_{\mathrm{U}}(\alpha)$ defined by (2.25) and (2.26),

$$P_p\{p_{\mathrm{L}}^{W}(\alpha) < p < p_{\mathrm{U}}^{W}(\alpha)\} \approx 1 - \alpha. \tag{2.28}$$

The classical large-sample confidence interval (see Comment 19) is centered at $\widehat{p}$. The Wilson confidence interval is centered at $\tilde{p}$ which is a weighted average of $\widehat{p}$ and 1/2 (see Comment 18 and (2.24)).

EXAMPLE 2.4 ***Tempting Fate.***

Risen and Gilovich (2008) conducted a number of studies designed to explore the notion that it is bad luck to tempt fate. In one study, participants were read a scenario in which a student had recently finished applying to graduate school and his top choice was Stanford University. In the scenario, the student's optimistic mother sent him a Stanford T-shirt in the mail. Risen and Gilovich asked a group of 20 participants to consider that the student could either stuff the T-shirt in a drawer while waiting for Stanford's admission decision or could wear the shirt the next day. The question asked of the 20 participants was would a person be more upset receiving a rejection from Stanford after having worn the Stanford shirt than after having stuffed the shirt in a drawer. Eighteen of the 20 participants thought the person would be more upset having worn the shirt. (The person when he wears the shirt "tempts fate" but it is more of a superstitious nature than, for example, tempting fate by walking outside in the middle of a storm replete with lightening. The latter actually increases your chance of a serious accident while wearing the shirt does not affect the chance of admission.) Let p denote the probability that a participant thought the person would be more upset having worn the shirt.

To directly find the Wilson interval for this tempting fate data, we can use the function `binom.confint` from the library `binom`. If we enter `binom.confint(x = 18, n = 20, conf.level = .95, methods = "all")` we obtain the Wilson interval along with a number of other confidence intervals including the Laplace–Wald interval of Comment 19, the Agresti–Coull interval of Comment 20, and the Clopper–Pearson interval of Comment 21. The Wilson 95% interval is (.699, .972).

The null hypothesis of no effect underlying the Risen and Glovich studies is that people are unconcerned about tempting fate, which, in terms of p, is $H_0 : p = 1/2$. With $B = 18$, $n = 20$, we find the one-sided P-value is $P_{1/2}(B \geq 18) = .0002$. Thus there is strong evidence that the participants feel people will avoid tempting fate. The P-value of .0002 can be obtained directly from the `R` function `pbinom(18,20,.5,lower.tail=F)` or equivalently from `1-pbinom(18,20,.5)`.

Comments

18. *The Wilson Confidence Interval.* In general, confidence intervals can be obtained by inverting tests. For a general parameter θ, a two-sided $100(1-\alpha)\%$ confidence interval consists of those θ_0 values for which the two-sided test of $\theta = \theta_0$ does not reject the null hypothesis $\theta = \theta_0$. The confidence interval given by (2.25) and (2.26) is due to Wilson (1927) (see also Agresti and Caffo (2000), Agresti and Coull (1998), Brown, Cai and DasGupta (2001), and Agresti (2013)). It is also called the score interval (see Agresti (2013). The interval is the set of p_0 values for which $|\widehat{p} - p_0|/\{(p_0(1-p_0)/n)\}^{1/2} < z_{\alpha/2}$. The midpoint $\tilde{p}$ of the interval is a weighted average of $\widehat{p}$ and $1/2$ with the weights $n/(n + z^2_{\alpha/2})$ and $z^2_{\alpha/2}(n + z^2_{\alpha/2})$, respectively. This midpoint equals the sample proportion obtained if $z^2_{\alpha/2}/2$ pseudo observations are added to the number of successes and $z^2_{\alpha/2}/2$ pseudo observations are added to the number of failures. We can write this midpoint $\tilde{p}$ as

$$\tilde{p} = \frac{B + z^2_{\alpha/2}/2}{n + z^2_{\alpha/2}},$$

which is equivalent to (2.24).

The quantity $(V^*)^2$ (see (2.27)) is a weighted average of the variance of a sample proportion when $p = \widehat{p}$ and the variance of a sample proportion when $p = 1/2$, where $n + z_{\alpha/2}^2$ is used in place of the sample size n.

Brown, Cai, and DasGupta (2001) studied various confidence intervals for p. For $n \leq 40$, they recommended the Wilson interval and an alternative interval due to Jeffreys. For $n > 40$ they found that the Wilson interval, the Jeffreys interval, and the Agresti–Coull interval (see Comment 20) are comparable.

19. *The Laplace–Wald Confidence Interval.* This interval can be obtained by inverting large-sample Wald tests (cf. Agresti, 2013). The approximate $100(1-\alpha)\%$ interval is the set of p_0 values for which $|\widehat{p} - p_0|/\{\widehat{p}(1-\widehat{p})/n\}^{1/2} < z_{\alpha/2}$. The interval is

$$p_{\mathrm{L}}^{\mathcal{L}W}(\alpha) = \widehat{p} - z_{\alpha/2}\{(\widehat{p}(1-\widehat{p}))/n\}^{1/2}, \tag{2.29}$$

$$p_{\mathrm{U}}^{\mathcal{L}W}(\alpha) = \widehat{p} + z_{\alpha/2}\{(\widehat{p}(1-\widehat{p}))/n\}^{1/2}, \tag{2.30}$$

where $\widehat{p} = B/n$. The interval was used by Laplace (1812), and here, we denote it as the $\mathcal{L}W$ interval.

Brown, Cai, and DasGupta (2001) highlight disadvantages of the $\mathcal{L}W$ interval. There exist pairs (p, n), which they call unlucky pairs, for which the coverage probability is much smaller than the nominal coverage probability $1-\alpha$. The phenomenon of oscillation occurs in n for fixed p and in p for fixed n. They also note that severe changes in the coverage occur in nearby p for fixed n and in nearby n for fixed p. Furthermore, even for large sample sizes, significant changes in coverage probabilities occur in nearby p and in many cases the coverage of the $\mathcal{L}W$ interval is strictly smaller than the nominal level. In particular, Brown, Cai, and DasGupta (2001) show for all $n \leq 45$, the actual coverage of the 99% $\mathcal{L}W$ interval is strictly less than the nominal level for all $0 < p < 1$. See their Examples 1–5.

The $\mathcal{L}W$ interval can be found directly using the `R` function `binom.confint`. For the tempting fate data, if we enter `binom.confint(x=18, n=20,conf.level=.95,methods="all")`, the output for the $\mathcal{L}W$ interval (labeled the "asymptotic" interval in the output) is (.769, 1.031). The parameter p must be between 0 and 1. Thus the upper value 1.031 should be changed to 1.

20. *The Agresti–Coull Confidence Interval.* The Agresti–Coull interval is also centered at $\tilde{p}$. Let $\tilde{q} = 1 - \tilde{p}$. The Agresti–Coull (1998) two-sided confidence interval for p with confidence coefficient approximately $1-\alpha$ is

$$p_{\mathrm{L}}^{AC}(\alpha) = \tilde{p} - z_{\alpha/2}(\tilde{p}\tilde{q})^{1/2}\tilde{n}^{-1/2} \tag{2.31}$$

$$p_{\mathrm{U}}^{AC}(\alpha) = \tilde{p} + z_{\alpha/2}(\tilde{p}\tilde{q})^{1/2}\tilde{n}^{-1/2} \tag{2.32}$$

With $p_{\mathrm{L}}(\alpha)$ and $p_{\mathrm{U}}(\alpha)$ defined by (2.31) and (2.32),

$$P_p\{p_{\mathrm{L}}^{AC}(\alpha) < p < p_{\mathrm{U}}^{AC}(\alpha)\} \approx 1 - \alpha. \tag{2.33}$$

The Agresti–Coull interval is an alternative to the classical Laplace–Wald interval; one with a better centering point ($\tilde{p}$ instead of $\widehat{p}$). For the case when

$\alpha = .05$, if you substitute "2" for $z_{.025} = 1.96$, it can be thought of as the "add two successes and two failures" interval. Brown, Cai, and DasGupta (2001) recommend the Agresti–Coull interval for practical use when $n \geq 40$, although it is never shorter than the Wilson interval. Its relative simplicity and ease of description make it particularly attractive for an introductory course. See Brown, Cai, and DasGupta (2001) for comparisons of various confidence intervals for the binomial parameter. The Agresti–Coull interval can be found directly from the `R` function `binom.confint`. For the tempting fate data if we enter `binom.confint(x=18, n=20, conf.level=.95, methods="all")`, we find the approximate 95% interval to be (.687, .984).

21. *The Clopper–Pearson Confidence Interval.* The Clopper–Pearson (1934) confidence interval is obtained by inverting the equal-tail binomial test. That is, if $B = b$ is observed, the Clopper–Pearson interval is defined by $p_{\mathrm{L}}^{CP}(\alpha)$, $p_{\mathrm{U}}^{CP}(\alpha)$, where $p_{\mathrm{L}}^{CP}(\alpha)$ and $p_{\mathrm{U}}^{CP}(\alpha)$ are, respectively, the solutions in p to the equations

$$P_p(B \geq b) = \alpha/2, \quad P_p(B \leq b) = \alpha/2.$$

The endpoints of the $100(1-\alpha)\%$ confidence interval are defined by the following equations:

$$p_{\mathrm{L}}^{CP}(\alpha) = \frac{B}{B + (n - B + 1) f_{\alpha/2,2(n-B+1),2B}} \tag{2.34}$$

$$p_{\mathrm{U}}^{CP}(\alpha) = 1 - p_{\mathrm{L}}^{\alpha}(n, n - B), \tag{2.35}$$

where B is the number of successes in the n Bernoulli trials and f_{γ,n_1,n_2} is the upper γth percentile for the F distribution with n_1 degrees of freedom in the numerator and n_2 degrees of freedom in the denominator.

The Clopper–Pearson interval is conservative,

$$P_p\{p_{\mathrm{L}}^{CP}(\alpha) < p < p_{\mathrm{U}}^{CP}(\alpha)\} \geq 1 - \alpha. \tag{2.36}$$

The conservativeness can be extreme in that for any fixed p, the true coverage probability can be much larger than $1 - \alpha$ unless n is quite large.

The Clopper–Pearson interval can be found directly from the `R` function `binom.confint`. For the tempting fate data we apply `binom.confint( x=18,n=20,conf.level=.95,methods="all")` and find the CP interval is (.683, .987). In the output the CP interval is labeled as "exact".

22. *Equivariance*. Binomial confidence interval procedures that satisfy (2.35) are said to be equivariant (Casella, 1986). The motivation for the term *equivariance* is that the binomial distribution is invariant under the transformations $B \to n - B$ and $p \to 1 - p$. See Casella (1986) for further details. The Clopper–Pearson intervals are equivariant but they are not the only ones that enjoy the equivariance property. Casella (1986) gives a method for refining equivariant binomial confidence intervals to obtain new intervals with uniformly shorter lengths for the same confidence coefficient.

23. *The Multinomial Distribution.* The binomial distribution given by (2.15) can be extended to situations where an experiment has k ($k \geq 2$) possible outcomes or categories, say $A_1, A_2, \ldots, A_k$, which are mutually exclusive and exhaustive.

We let $P(A_i) = p_i, i = 1, \ldots, k$, where $\sum_{i=1}^{k} p_i = 1$. Furthermore, let X_i be the number of times A_i occurs in the n trials. The k variables $X_1, X_2, \ldots, X_k$ are said to have the multinomial distribution with parameter $n, \mathbf{p}$, where $\mathbf{p} = (p_1, p_2, \ldots, p_k)$. The distribution is given by

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}, \quad (2.37)$$

where

$$\binom{n}{x_1, x_2, \ldots, x_k} = \frac{n!}{x_1! x_2! \ldots x_k!}. \quad (2.38)$$

The quantity $\binom{n}{x_1, x_2, \ldots, x_k}$ is known as the *multinomial coefficient*. It is equal to the number of distinguishable arrangements of x_1 A_1's, x_2 A_2's, ..., x_k A_k's. The mean and variance of X_i are

$$E(X_i) = np_i, \ \text{var}(X_i) = np_i(1 - p_i), \ i = 1, \ldots k.$$

The covariance between X_i and X_j is

$$\text{cov}(X_i, X_j) = -np_i p_j.$$

24. *Some Examples Where the Multinomial Arises.*

Example A: In Paradise Paved, Florida, 44% of the voters are Democrats, 42% are Republican, and 14% are in some other category (Independent, Tea Party, etc.). Suppose a random sample of 20 voters are polled yielding X_1 Democrats, X_2 Republicans, and X_3 in the other category. Then, (X_1, X_2, X_3) has a multinomial distribution with parameters $n = 20$ and $\mathbf{p} = (.44, .42, .14)$.

Example B: Cohen and Bloom (2010) report on data from the National Health Interview Survey, 2008. The distribution of health insurance status for male adults aged 20–29 years was 57.5% had private insurance, 5.6% were on Medicaid, 1.6% had some other form of insurance, and 35.3% were uninsured. Suppose a random sample of 40 was obtained from this population yielding X_1 on private insurance, X_2 on Medicaid, X_3 on some other coverage, and X_4 uninsured. Then, (X_1, X_2, X_3, X_4) has a multinomial distribution with parameters $n = 40$ and $\mathbf{p} = (.575, .056, .016, .353)$.

25. *Estimation for the Multinomial Distribution.* For the multinomial distribution, if we observe the frequencies $X_1, X_2, \ldots, X_k$, where X_i is the number of times the event A_i occurs in n experiments, the standard frequentist estimators of $p_1, p_2, \ldots, p_k$ are the sample proportions

$$\widehat{p_i} = X_i/n, \quad i = 1, \ldots, k. \quad (2.39)$$

With $k \geq 2$, asymptotic $100(1 - \alpha)\%$ simultaneous confidence intervals for the $M = k(k-1)/2$ pairwise differences $\{p_i - p_j\}, 1 \leq i < j \leq k$, can be obtained using Bonferroni's inequality. They are

$$p_{\text{L}}^{(i,j)} = \widehat{p_i} - \widehat{p_j} - z_{\alpha/2M}\{[\widehat{p_i} + \widehat{p_j} - (\widehat{p_i} - \widehat{p_j})^2]/n\}^{1/2}, \quad (2.40)$$

$$p_{\text{U}}^{(i,j)} = \widehat{p_i} - \widehat{p_j} + z_{\alpha/2M}\{[\widehat{p_i} + \widehat{p_j} - (\widehat{p_i} - \widehat{p_j})^2]/n\}^{1/2}, \quad (2.41)$$

where $\widehat{p_i} = X_i/n,\ i = 1, \ldots, k$. The M intervals given by (2.40) and (2.41) satisfy, for large n,

$$P\{p_{\mathrm{L}}^{(i,j)} < p_i - p_j < p_{\mathrm{U}}^{(i,j)}, 1 \le i < j \le k\} \approx 1 - \alpha. \tag{2.42}$$

That is, the probability is approximately $1 - \alpha$ that the M intervals simultaneously contain the M differences $\{p_i - p_j, i < j\}$.

Asymptotic $100(1 - \alpha)\%$ simultaneous confidence intervals for $p_i, i = 1, \ldots, k$, based on Bonferroni's inequality, are obtained by solving

$$(\widehat{p_i} - p_i)^2 = (z_{\alpha/2k})^2 p_i(1 - p_i)/n \tag{2.43}$$

for lower and upper limits $p_{\mathrm{L}}^{(i)}, p_{\mathrm{U}}^{(i)}$ (see Goodman (1965), Miller (1981a), Fitzpatrick and Scott (1987) and Agresti (2013)).

26. *Pearson's Chi-Squared Goodness-of-Fit Test for Specified Multinomial Probabilities.* Pearson's (1900) chi-squared statistic can be used to test, on the basis of n experiments with frequencies $X_1, X_2, \ldots, X_k$ corresponding to the k categories, the hypothesis that the multinomial probabilities $p_1, p_2, \ldots, p_k$ are equal to specified or known values $p_1^0, p_2^0, \ldots, p_k^0$. Pearson's chi-squared statistic is

$$\chi^2 = \sum_{i=1}^{k} \left\{ \frac{(X_i - np_i^0)^2}{np_i^0} \right\} \tag{2.44}$$

Note deviations of the X_i's from their hypothesis expected values (the np_i^0,s) are magnified by the $(X_i - np_i^0)^2$ terms leading to large values of χ^2. That is, significantly large values of χ^2 indicate a deviation from the hypothesis

$$H_0: \quad p_1 = p_1^0,\ p_2 = p_2^0,\ \ldots,\ p_k = p_k^0 \tag{2.45}$$

in favor of the alternative

$$H_1: \quad p_i \ne p_i^0 \text{ for at least on value of } i. \tag{2.46}$$

If one computes $\chi^2 = \chi^2_{obs}$ (the observed value), one can find the corresponding P-value, the probability under the null hypothesis that $\chi^2 \ge \chi^2_{obs}$, by summing the probabilities given by (2.37) over all possible multinomial outcomes yielding $\chi^2 \ge \chi^2_{obs}$. It is more convenient, however, to use a large-sample approximation.

Pearson showed that when H_0 is true, the distribution of χ^2 as $n \to \infty$, is that of a chi-squared distribution with $k - 1$ degrees of freedom. Thus, an approximate α-level test is

$$\text{Reject } H_0 \text{ if } \chi^2 \ge \chi^2_{\alpha,k-1};\ \text{otherwise do not reject.} \tag{2.47}$$

The P-value is found by referring the observed value of χ^2 to the χ^2_{k-1} distribution.

The χ^2 approximation is good when each of the np_i^0's is not too small. A general foot rule is it's good if $np_i^0 \ge 5$ for each value of i.

Table 2.2 Outcomes of Pea Plant Experiments

Trait	Dominant		Recessive		χ^2	P-value	Expected Ratio
Seed shape	Round	5474	Angular	1850	.2629	.6081	3:1
Cotyledon color	Yellow	6022	Green	2001	.015	.9025	3:1
Seed coat color	Colored	705	White	224	.3907	.5319	3:1
Pod shape	Inflated	882	Constricted	299	.0635	.801	3:1
Pod color	Green	428	Yellow	152	.4506	.5021	3:1
Flower position	Axial	651	Terminal	207	.3497	.5543	3:1
Stem length	Long	787	Short	277	.6065	.4361	3:1

Source: D.J. Fairbanks and B. Rytting (2001).

27. *Checking for Data Fudging: The Fit May be Too Good.* The chi-squared statistic rejects the goodness-of-fit null hypothesis (2.45) if χ^2 is too large. Small values in the lower tail of the null distribution of χ^2, however, can give an indication that the fit is too good and that perhaps the data have been "cooked" so that they would appear to support the hypothesized model values.

    A classic example of the use of Pearson's χ^2 involves Gregor Mendel's famous genetics experiments on pea plants. Mendel, a European monk whom many biologists regard as the father of genetics, cross-pollinated purebred plants with specific traits and observed and recorded the results over many generations. Table 2.2, based on data in Fairbanks and Rytting (2001), gives the f_2 generation (the second offspring of cross-pollinated purebred plants) pertaining to seven pea characteristics: (1) seed shape (round or angular), (2) cotyledon (part of the embryo within the seed) color (yellow or green), (3) seed coat color (colored or white), (4) pod shape (inflated or constricted), (5) pod color (green or yellow), (6) flower position (axial or terminal), (7) stem length (long or short).

    All of the χ^2 statistics in Table 2.2 are based on one degree of freedom (df). The sum of seven independent χ^2 statistics with one df follows a χ^2 distribution with $df = 7$. (More generally, the sum $\sum_i \chi^2$ of independent χ^2 statistics, with the i^{th} having $df = m_i$, follows a χ^2 distribution with $df = \sum m_i$.) Summing the seven χ^2 values in column 6 of Table 2.1 yields $\sum_{i=1}^{7} \chi^2 = 2.1389$ and $P(\sum_{i=1}^{7} \chi^2 \leq 2.1389) = .0482$. Thus the value 2.1389 falls in the lower tail of the distribution and arouses suspicion that the fit is too good.

    Mendel did many more experiments than those represented in Table 2.2. Agresti (2013) and Fisher (1936) summarized Mendel's experiments and obtained a χ^2 value of 42 based on $df = 84$. We find $P(\chi^2_{84} \leq 42) = .000035$. This chi-square value is extremely small and it is smaller than would be expected when the model fits. Fisher suspected that an overzealous assistant might have biased the data. Other possibilities include the "several left-in-the drawer" theory in which Mendel may have only reported the "best" results and omitted the results of other experiments. Nevertheless, over time the works of Mendel and many others have led to acceptance of Mendel's genetic theories. Pires and Branco investigate a model that may alleviate the controversy. See their paper and Box (1978) for more details about the history of the Mendel-Fisher difficulty.

    The χ^2 values can be readily obtained from R functions. For example, the $\chi^2 = .2629$ value given in the first row of Table 2.2 is obtained from `chisq.test(c(5474,1850), p=c(.75,.25))` yielding the output $\chi^2 = .2629$, $df = 1$, P-value=.6081. The lower-tail probability $P(\chi^2_{84} \leq 42) =$

.000035 corresponding to the value $\chi^2_{84} = 42$ is obtained by the R function `pchisq(42,84)`.

28. *Testing Equal Probabilities.* In the case when the multinomial probabilities are specified to be equal, that is H_0 is taken to be $p_1 = 1/k, \ldots, p_k = 1/k$, the chi-squared statistic reduces to

$$\chi^2 = \left\{ (k/n) \sum_{i=1}^{k} X_i^2 \right\} - n. \tag{2.48}$$

29. *The Case k = 2.* When $k = 2$, the multinomial setting reduces to the binomial setting and Pearson's χ^2 test is a test of $p = p_0$. In this case, the approximate test defined by (2.47) is equivalent to the approximate two-sided test of $p = p_0$ versus the alternative $p \neq p_0$ given by (2.12).

Properties

1. *Distribution-Freeness.* For Bernoulli trails satisfying Assumptions A1–A3, (2.36) holds. Thus, $(p_{\mathrm{L}}^{CP}(\alpha), p_{\mathrm{U}}^{CP}(\alpha))$ is a confidence interval for p with confidence coefficient at least $1 - \alpha$.

Problems

15. For the parolee data of Problem 1, obtain the Wilson, Laplace–Wald, Agresti–Coull, and Clopper–Pearson confidence intervals for p, each with an approximate confidence coefficient of .96. Compare the four intervals.

16. Shlafer and Karow (1971) considered some of the problems involved with cardiac preservation. In particular, they were interested in the morphological and physiological injury occurring in hearts that had been frozen to various temperatures without the benefit of a cryoprotectant. Hearts from adult rats were perfused with a balanced salt solution *in vitro* for 20 min, and during this time, contractions were noted. After disconnection from the perfusion apparatus, each heart, surrounded by a plastic shield, was inserted into a metal canister and chilled by an acetone bath (maintained at -20°C by addition of dry ice) until the lowest desired temperature was attained. The individual hearts were then thawed (in 1 min or less) by removing the metal canister and running 35°C tap water over the plastic shields, being careful to prevent water from flowing directly over the hearts. After thawing, the hearts were again perfused with the balanced salt solution. Hearts spontaneously resuming coordinated atrioventricular contractions within 20 min of thawing were considered to be "survivors" of the freeze–thawing process.

The authors conducted experiments where the lowest attained temperatures were -10, -12, -17, and -20°C. We focus here on the data for the -12°C investigation, in which the authors found that of six hearts frozen to -12°C, three were survivors. If we let success denote survival, then p represents the probability that a rat heart frozen to -12°C will spontaneously resume coordinated atrioventricular contractions within 20 min after thawing and perfusion with a balanced salt solution. Obtain the Wilson, Laplace–Wald, Agresti–Coull, and Clopper–Pearson confidence intervals for p, each with an approximate confidence coefficient of .90. Compare the intervals.

17. Many materials that are satisfactory for use in air will ignite and burn in pure oxygen when subjected to mechanical impact. This problem is of vital concern to the aerospace industry, which uses enormous amounts of both liquid and gaseous oxygen. In particular, there is a need for guidelines to aid the designer in selecting materials to be employed in pressurized oxygen

systems. In order to provide an appropriate method for determining gaseous oxygen-material compatibility, the Kennedy Space Center developed a gaseous oxygen impact test procedure. Jamison (1971) reported on the use of this testing scheme to analyze the gaseous oxygen-material compatibility for 33 Apollo spacecraft test materials. One such material tested, silicone elastomer 342, failed the gaseous oxygen impact test (i.e., the material ignited) in 4 out of 20 trials. Let p denote the probability of ignition for silicone elastomer 342 when subjected to the conditions employed in the gaseous oxygen impact test. Obtain the Wilson, Laplace–Wald, Agresti–Coull, and Clopper–Pearson confidence intervals for p, each with an approximate confidence coefficient .95. Compare the intervals.

18. Ehlers (1995) performed a 1-year follow-up study of panic disorder. As partial motivation for the study, the author offered the following quote from Wolfe and Maser (1994, 241): "Little is known about the long-term course of disorder. The limited findings to date suggest that in most cases it is a chronic disorder that waxes and wanes in severity. However, some people may have a limited period of dysfunction that never recurs, while others tend to have a more severe and complicated course." In this study, diagnoses were made by trained interviewers (either the author, Ehlers, or trained graduate students) according to the criteria of the revised third edition of the "Diagnostic and Statistical Manual of Mental Disorders" (DSM-III-R; American Psychiatric Association, 1987). One year after initial assessments, participants were mailed a questionnaire for the purpose of assessing their current symptoms. In this problem, we discuss one small portion of the data that were obtained. Out of 46 people who were initially diagnosed as "infrequent panickers," 23 experienced panic attacks during follow-up. Obtain the approximate 95% Wilson, Laplace–Wald, Agresti–Coull, and Clopper–Pearson confidence intervals for p, the probability that an infrequent panicker will experience panic attacks during a 1-year follow-up period. Compare the intervals.

19. For the triangle bitterness tests data of Example 2.2, obtain the Wilson, Laplace–Wald, Agresti–Coull, and Clopper–Pearson confidence intervals for p, each at an approximate confidence coefficient of .90. Compare the intervals.

20. Consider the study in Problem 17 and describe the inherent sources for error when one uses a mailed questionnaire.

21. Consider the study in Problem 18 and discuss the possibility of unintentional bias entering the study, because some of the diagnoses were made by the author of the study.

22. A table of random numbers contains randomly generated digits that have been generated so that the following two properties are satisfied.

(I) *Equal Likelihood Property.* Focus on any particular spot in the table, such as the second digit in the fourth row. The probability it will be a 0 equals the probability it will be a 1 ... equals the probability it will be a 9. More succinctly, $P(0) = P(1) = \ldots = P(9) = 1/10$.

(II) *Mutual Independence Property.* Focus on any particular spot in the table, such as the third digit in the second row. Knowing some or even all of the digits in the table, except for the one you are considering, does not change the probability it will be a 0, or a 1, ..., or a 9. This probability remains a 1/10.

For the 400 randomly generated digits in Table 2.3, test the equal likelihood property.

Table 2.3 Four Hundred Random Digits

24253	39427	80642	36718	92164	77732	69754	01291	53704	33054
34302	60309	27186	22418	59962	13934	67591	17476	21559	73437
76809	84341	74012	50947	83214	19967	44219	75929	13182	34858
85183	35958	04301	49628	91493	66103	65699	04241	82441	38112
27541	79187	99777	22894	83283	56218	86183	74497	21070	78935
74188	09083	54938	79920	27158	24864	31116	33173	43032	52000
13270	57457	30968	65978	67679	91216	47969	39204	46030	93954
89150	53922	40537	23169	46948	05519	72171	85417	31580	98102

23. Devore (1991) gives an example from the paper "Linkage Studies of the Tomato" (*Tran. Royal Canadian Inst.*, 1931, pp 1-19) on phenotypes from a dihybrid cross of tall, cut-leaf tomatoes with dwarf, potato-leaf tomatoes. Dihybrid crosses arise as follows. If two different characteristics of an organism are each controlled by a single gene, and a pure strain having genotype AABB is crossed with a pure strain having genotype aabb (A,B are dominant alleles; a,b are recessive), the resulting genotype is AaBb. A dihybrid cross occurs when the first-generation organisms are crossed among themselves. The data for a dihybrid cross of tall, cut-leaf tomatoes with dwarf, potato-leaf tomatoes, based on a sample of n=1611, are as follows: Tall, cut-leaf (926); tall, potato-leaf (288); dwarf, cut-leaf (293); dwarf, potato-leaf (104) (tall is dominant for size, cut-leaf is dominant for leaf type). Mendelian inheritance theory predicts that there are four categories of probabilities occurring, 9/16, 3/16, 3/16, 1/16. Test if the data support Mendelian theory.

24. Consider the National Health Interview Survey of Example B of Comment 24. Suppose a random sample of 100 from that population yielded the results: 60 males on private insurance, 5 on Medicaid, 4 on some other form of insurance, and 31 uninsured. Is such a sample consistent with the specified population probabilities?

25. For the data of Example B of Comment 24, find approximate 90% confidence intervals for the six pairwise differences $p_1 - p_2$, $p_1 - p_3$, $p_1 - p_4$, $p_2 - p_3$, $p_2 - p_4$, $p_3 - p_4$.

26. Establish the expression for χ^2 given by (2.48), Comment 28.

27. Show the equivalence of the two tests described in Comment 29.

2.4 BAYES ESTIMATORS FOR THE PROBABILITY OF SUCCESS

The estimator $\widehat{p}$ of Section 2.2 is a frequentist estimator of p; it does not utilize prior information. In this section, we consider Bayes estimators of p that make use of prior information. The estimators are based on the Beta class $Beta(r, s)$ $(r > 0,\ s > 0)$ of prior distributions for p. In the Bayesian approach p is considered a random variable. Guidance for the choice of the prior distribution is presented in Comments 30–33.

Procedure

For squared-error loss, the Bayes estimator of p when using the prior distribution $Beta(r, s)$, is the mean of the posterior distribution of p, given $B = b$. This mean can be denoted as

$$E(p|B = b) = \frac{b + r}{r + s + n} \tag{2.49}$$

Note that the Bayes estimator can be rewritten as

$$E(p|B = b) = \left(\frac{n}{n + r + s}\right)\frac{b}{n} + \left(\frac{r + s}{n + r + s}\right)\frac{r}{r + s}. \tag{2.50}$$

In the form (2.50), we can see that the Bayes estimator is a weighted average of the observed proportion of successes b/n and the prior guess at p, namely $E(p) = r/(r + s)$. The weights are $n/(n + r + s)$ and $(r + s)/(n + r + s)$. Note that as n gets large, the second term in (2.50) tends to 0 and the first term tends to b/n. This is a reflection of the fact that as the sample size gets large, the observed data dominate the prior information.

EXAMPLE 2.5 ***Percentage of Smokers.***

In May 2010, E. Chicken polled his two statistics courses to investigate the percentage of students who smoke cigarettes. The classes were STA 4321 Introduction to Mathematical Statistics and STA 4502-5507 Applied Nonparametric Statistics. The 4321 class had 27 students comprising 5 females and 22 males. The 5507 class had 17 students comprising 5 females and 12 males. In STA 4321, there was one smoker, a female. In STA 5507, there were five smokers, three females and two males. To have a large sample size, the two classes were combined. Thus, out of a total of 44 students, 6 were smokers. The classical frequentist estimator for the true proportion p of smokers in the college population is $\widehat{p} = 6/44 = .136$ or 13.6%.

A Bayesian approach can effectively be employed because there are many studies concerning smoking rates. For example, a National Health Interview Survey (NHIS) (2008) estimated the smoking rates to be 23.1% for men and 18.3% for women. We will illustrate Bayesian approaches using a noninformative prior (see Comment 31) and an informative prior (see Comments 32 and 33). Using the noninformative Bayes–Laplace prior $Beta(1, 1)$ (see Comment 31), we find from (2.49) with $n = 44$, $b = 6$, $r = 1$, $s = 1$,

$$E(p|B = 6) = \frac{6 + 1}{2 + 44} = \frac{7}{46} = .152$$

or 15.2%.

We can also use an informative prior such as the one mentioned in Comment 32. From the NHIS (2008) results, it is reasonable to take $p^* = .20$ as a good guess at the percentage of smokers. Setting

$$p^* = .20 = r/r + s$$

as in (2.56), and taking $\sigma^* = .05$, we have from (2.57),

$$(.05)^2 = \frac{rs}{(r + s)^2(r + s + 1)}$$

Solving the previous two equations for r and s yields

$$r = 12.6, \quad s = 50.4,$$

and from (2.49)

$$E(p|B = 6) = \frac{6 + 12.6}{12.6 + 50.4 + 44} = .174$$

or 17.4%.

Note that in this example, both Bayesian estimators, one based on a noninformative prior and the other on an informative prior, are closer to what might have been expected, considering the results of the NHIS survey. Yet times change, you are not reading this in 2008, and there are strong efforts in the United States to reduce the incidence of smoking. Furthermore, statisticians have played a prominent role in discovering and assessing the increased risks of various health problems (e.g., lung cancer, heart disease, and emphysema) associated with smoking, so it is not surprising that in statistics classes in a college population the incidence may be less than that in broader populations.

Comments

30. *Bayes Estimators.* The Bayesian approach incorporates prior information into the estimation procedure. One starts with a prior density for p, which is viewed as a random variable. After observing the data $B = b$, the prior and the data are used to compute the posterior density of p. The conjugate prior for p is the beta distribution, $Beta(r, s)$. A random variable Y has a beta distribution with parameters r, s ($r > 0$, $s > 0$) if Y has the density function

$$\begin{aligned} f(y) &= \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} y^{r-1}(1-y)^{s-1}, \quad 0 < y < 1 \\ &= 0 \qquad\qquad\qquad\qquad\qquad\quad \text{otherwise.} \end{aligned} \tag{2.51}$$

The mean of the distribution is

$$E(y) = \frac{r}{r+s} \tag{2.52}$$

and the variance is

$$\text{var}(Y) = \frac{rs}{(r+s)^2(r+s+1)}. \tag{2.53}$$

Figure 2.1 shows various beta densities.

The posterior distribution of p, given $B = b$, is readily shown to be $Beta(b + r, n - b + s)$, that is, a beta distribution with parameters $b + r$ and $n - b + s$. For squared-error loss, the Bayes estimator of p is the mean of this posterior distribution as given in (2.49), namely,

$$E(p|B = b) = \frac{b+r}{b+r+n-b+s} = \frac{b+r}{r+s+n} \tag{2.54}$$

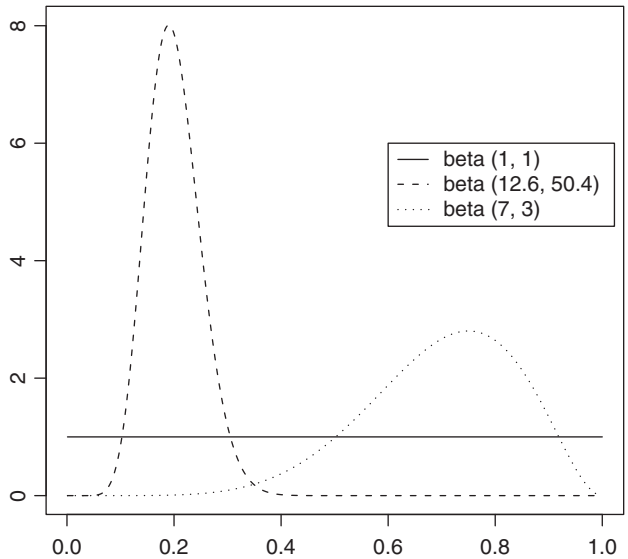

Figure 2.1 Three beta densities.

As we have noted in the Procedure subsection, the Bayes estimator can be rewritten as

$$E(p|B=b) = \left(\frac{n}{n+r+s}\right)\frac{b}{n} + \left(\frac{r+s}{n+r+s}\right)\frac{r}{r+s} \tag{2.55}$$

In the form (2.55), we see that the Bayes estimator is a weighted average of the observed proportion of successes b/n and the prior guess $E(p) = r/(r+s)$.

31. *Choice of Prior When Minimal Information is Available.* When there is little information about the parameter of interest, Bayesians, who still want to employ the Bayesian structure and machinery, favor using a noninformative prior. A noninformative prior, roughly speaking, contains little information about the parameter of interest, or favors no possible value of that parameter over other possible values (see Berger (1985, p. 82)). In the case we are discussing in this chapter, namely, where the parameter of interest is p, the probability of success, Berger (1985, p. 89) considers the following priors to be reasonable noninformative priors (i) $f_1(p) = 1$, the uniform prior corresponding to $Beta(1,1)$, is often referred to as the Bayes–Laplace prior. (ii) $f_2(p) = p^{-1}(1-p)^{-1}$ is known as the *Haldane prior*. It can be roughly viewed as $Beta(0,0)$ and yields as the Bayes estimator $\widehat{p} = B/n$, (iii) $f_3(p)$ proportional to $[p(1-p)]^{-1/2}$ is the $Beta(1/2,1/2)$ prior due to Jeffreys (1961), and (iv) $f_4(p)$ proportional to $p^p(1-p)^{1-p}$, a prior not in the Beta family and one that arises from an approach due to Zellner (1977). Berger points out that f_1 is a proper density, as are f_3 and f_4 suitably normalized, whereas f_2 is improper.

32. *Choice of Prior when Prior Information is Available.* The Beta family is often used and is quite flexible for characterizing prior information. If one has a good guess, p^* say, from a prior experiment perhaps, about the location of p, it is reasonable to set p^* equal to the mean of the Beta distribution, namely,

$$p^* = \frac{r}{r+s} \tag{2.56}$$

If a reasonable choice for the standard deviation, say σ^*, is also available then set

$$(\sigma^*)^2 = \frac{rs}{(r+s)^2(r+s+1)}, \tag{2.57}$$

the Beta variance, and solve (2.56) and (2.57) for r and s, to obtain r^*, s^* say, then use $Beta(r^*, s^*)$. This is illustrated in Example 2.5.

33. *Choice of Prior in the Case of Zero Events.* Tuyl, Gerlach, and Mengerson (2008) consider the case of zero observed events, that is, $B = 0$ (or equivalently $B = n$). Their paper considers the four noninformative priors discussed by Berger (1985, p. 89) and mentioned in Comment 31. They recommend the Bayes–Laplace prior as a consensus noninformative prior. They also note that the use of a Beta prior with small Beta parameters, namely, $r, s, < 1$ should be avoided, both for noninformative and informative priors. One of their examples suggests that when p is known to be very small, an informative prior from $Beta(1,s)$ with $s > 1$ seems appropriate but a $Beta(r,s)$ with $r < 1$ can be too informative.

34. *The Prior Should Have Support on the Entire Parameter Space.* A prior distribution should not be so restrictive that it prevents the data from telling the true story. For example, in our problem of estimating p, if one chooses a prior that puts all of its probability on a proper subset of $[0, 1]$ and the true value of p is outside of that subset, the resulting Bayes estimator will not converge to the true value as the sample size grows. Thus, for example, if you put a uniform prior on the interval $[0, 1/2]$ and the true value of p is greater than $1/2$, the Bayes procedure will not converge to the true value.

    We want, for large samples, the data to have most of the influence. If you have a spread out prior that covers the parameter space, then for large samples the data will dominate and your posterior distribution will not be too influenced by your prior distribution.

35. *Bayesian Updating.* Suppose your prior distribution is $Beta(r, s)$, and suppose that you observe b successes in n binomial trials. Then, as noted in Comment 30, your posterior distribution is $Beta(b + r, n - b + s)$ and the Bayes estimator is $(b + r)/(r + s + n)$. Now suppose you obtain a second sample of size m and, in that second sample, you have c successes and $m - c$ failures. From your first sample, your new prior is $Beta(b + r, n - b + s)$, and after observing c successes, your new posterior is $Beta(b + r + c, s + m + n - b - c)$ and your Bayes estimator is $(b + r + c)/(b + r + c + s + m + n - b - c)$ or $(b + r + c)/(r + s + m + n)$. This agrees with what you would obtain by pooling the two samples of n and m with successes b and c, respectively. For if you start with a $Beta(r, s)$ prior, then obtain $b + c$ successes in a combined sample of $n + m$, the posterior based on the pooled sample is $Beta(r + b + c, s + m + n - b - c)$ with corresponding Bayes estimator $(r + b + c)/(r + b + c + s + m + n - b - c)$ or $(r + b + c)/(r + s + m + n)$.

36. *Bayes Estimation for the Multinomial Distribution.* For the Bayesian approach, the conjugate density is the Dirichlet distribution with parameters $\beta_1, \ldots, \beta_k$; letting $\mathbf{p} = (p_1, \ldots, p_k)$, the density is

$$f(\mathbf{p}) = \frac{\Gamma(\beta_0)}{\prod_{i=1}^{k} \Gamma(\beta_i)} \cdot \prod_{i=1}^{k} p_i^{\beta_i - 1}, \quad 0 < p_i < 1, \sum_{i=1}^{k} p_i = 1,$$

    where $\beta_i > 0$ and $\beta_0 = \sum_{i=1}^{k} \beta_i$. We continue with this density in Chapter 16, where the Dirichlet distribution is generalized to the Dirichlet process.

    The mean and variance of the Dirichlet are

$$E(p_i) = \frac{\beta_i}{\beta_0}, \quad \text{var}(p_i) = \frac{\beta_i(\beta_0 - \beta_i)}{\beta_0^2(\beta_0 + 1)}. \tag{2.58}$$

    The posterior density is Dirichlet with parameters $X_i + \beta_i$, $i = 1, \ldots, k$ so that the posterior mean is

$$E(p_i | X_1, \ldots, X_k) = \frac{X_i + \beta_i}{n + \beta_0}. \tag{2.59}$$

    The Bayes estimator, for the loss function $L(p, a) = \sum_{i=1}^{k} (p_i - a_i)^2$, is the posterior mean given by (2.59). Note that the Bayes estimator given by (2.59) can

be rewritten as

$$E(p_i|X_1,\ldots,X_k) = \left(\frac{n}{n+\beta_0}\right)\frac{X_i}{n} + \left(\frac{\beta_0}{n+\beta_0}\right)\frac{\beta_i}{\beta_0}, \tag{2.60}$$

which is a weighted average of the observed sample proportion X_i/n and the prior guess at p_i, β_i/β_0. Note that as n gets large, the Bayes estimator approaches the frequentist estimator given by (2.39).

Properties

1. *Bayes Optimality of* $E(p|B=b)$. For the $Beta(r,s)$ prior and squared-error loss, $E(p|B=b)$ minimizes the Bayes risk.
2. *Bayes Optimality of* $E(p_i|X_1,\ldots,X_k)$. For the Dirichlet distribution prior and sum of squared-error loss, $E(p_i|X_1,\ldots,X_k)$ minimizes the Bayes risk.

Problems

28. Consider the canopy gap closure data of Example 2.1. Determine a Bayes estimate for p. Explain how you obtained your prior distribution.

29. Consider the cardiac preservation data of Problem 16. Determine a Bayes estimate for p. Explain how you obtained your prior distribution.

30. Consider the silicone elastomer data of Problem 17. Determine a Bayes estimate for p. Explain how you obtained your prior distribution.

31. Consider the panic attack data of Problem 18. Determine a Bayes estimate for p. Explain how you obtained your prior distribution.

32. Consider the tempting fate data of Example 2.4. Determine a Bayes estimate for p. Explain how you obtained your prior distribution.

33. Consider the data on smokers in Example 2.5. Suppose the data from STA 5507 is not available so that you only have the data from the STA 4321 class. Determine a Bayes estimate for p. Explain how you obtained your prior distribution.

34. Consider the tomato data of Problem 23. Determine a Bayes estimate for $\mathbf{p} = (p_1, p_2, p_3, p_4)$. Explain how you obtained your prior distribution.

35. Consider the insurance data of Problem 24, Determine a Bayes estimate for $\mathbf{p} = (p_1, p_2, p_3, p_4)$. Explain how you obtained your prior distribution.

36. Describe three situations in which the costs associated with obtaining sample observations are exorbitant and thus in those situations the Bayesian approach is particularly appealing.

Chapter 3

The One-Sample Location Problem

INTRODUCTION

The procedures of this chapter are designed for statistical analyses in which primary interest is centered on the location (median) of a population. We encounter two types of data for which such analyses are important. The first of these, referred to as *paired replicates data*, represents pairs of "pretreatment" and "posttreatment" observations; here, we are concerned with a shift in location due to the application of the "treatment." The second type of data, referred to as *one-sample data*, consists of observations from a single population about whose location we wish to make inferences.

In Sections 3.1–3.3, procedures are considered for analyzing paired replicates data using signed ranks. In particular, Section 3.1 presents a distribution-free signed rank test; Section 3.2, a point estimator associated with the signed rank statistic; and Section 3.3, a related distribution-free confidence interval. In Section 3.7, these procedures are applied to some one-sample data. An asymptotically distribution-free test for symmetry of the underlying population (one of the assumptions in Sections 3.1–3.3 and 3.7) is considered in Section 3.9. A distribution-free test for exchangeability of the paired replicates data is discussed in Section 3.10.

Procedures for analyzing paired replicates data using signs are discussed in Sections 3.4–3.6. A distribution-free sign test is considered in Section 3.4, a point estimator associated with the sign statistic in Section 3.5, and a related distribution-free confidence interval in Section 3.6. These sign procedures are applied to some one-sample data in Section 3.8.

The asymptotic relative efficiencies for translation alternatives of the procedures based on the signed rank statistic and those based on the sign statistic with respect to their normal theory counterparts based on the sample mean are discussed in Section 3.11.

PAIRED REPLICATES ANALYSES BY WAY OF SIGNED RANKS

Data. We obtain $2n$ observations, two observations on each of n subjects (blocks, patients, etc.).

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

Subject i	X_i	Y_i
1	X_1	Y_1
2	X_2	Y_2
.	.	.
.	.	.
.	.	.
n	X_n	Y_n

Assumptions

A1. We let $Z_i = Y_i - X_i$, for $i = 1, \ldots, n$. The differences $Z_1, \ldots, Z_n$ are mutually independent.

A2. Each $Z, i = 1, \ldots, n$, comes from a continuous population (not necessarily the same one) that is symmetric about a common median θ. If F_i represents the distribution function for $Z_i, i = 1, \ldots n$, this assumption requires that

$$F_i(\theta + t) + F_i(\theta - t) = 1, \text{ for every } t \text{ and } i = 1, \ldots, n.$$

The parameter θ is referred to as the *treatment effect.*

3.1 A DISTRIBUTION-FREE SIGNED RANK TEST (WILCOXON)

Hypothesis

The null hypothesis of interest here is that of zero shift in location due to the treatment, namely,

$$H_0 : \theta = 0. \tag{3.1}$$

This null hypothesis asserts that each of the distributions (not necessarily the same) for the differences (posttreatment minus pretreatment observations) is symmetrically distributed about 0, corresponding to no shift in location due to the treatment.

Procedure

To compute the Wilcoxon signed rank statistic T^+, form the absolute values $|Z_1|, \ldots, |Z_n|$ of the differences and order them from least to greatest. Let R_i denote the rank of $|Z_i|, i = 1, \ldots, n$, in this ordering. Define indicator variables $\psi_i, i = 1, \ldots, n$, where

$$\psi_i = \begin{cases} 1, & \text{if } Z_i > 0, \\ 0, & \text{if } Z_i < 0, \end{cases} \tag{3.2}$$

and obtain the n products $R_1\psi_1, \ldots, R_n\psi_n$. The product $R_i\psi_i$ is known as the *positive signed rank of* Z_i. It takes on the value zero if Z_i is negative and is equal to the rank of $|Z_i|$ when Z_i is positive. The Wilcoxon signed rank statistic T^+ is then the sum of the positive signed ranks, namely,

$$T^+ = \sum_{i=1}^{n} R_i\psi_i. \tag{3.3}$$

a. *One-Sided Upper-Tail Test.* To test

$$H_0 : \theta = 0$$

versus

$$H_1 : \theta > 0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } T^+ \geq t_\alpha; \text{ otherwise do not reject,} \tag{3.4}$$

where the constant t_α is chosen to make the type I error probability equal to α.

b. *One-Sided Lower-Tail Test.* To test

$$H_0 : \theta = 0$$

versus

$$H_2 : \theta < 0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } T^+ \leq \frac{n(n+1)}{2} - t_\alpha; \text{ otherwise do not reject.} \tag{3.5}$$

c. *Two-Sided Test.* To test

$$H_0 : \theta = 0$$

versus

$$H_3 : \theta \neq 0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } T^+ \geq t_{\alpha/2} \text{ or } T^+ \leq \frac{n(n+1)}{2} - t_{\alpha/2}; \text{ otherwise do not reject.} \tag{3.6}$$

This two-sided procedure is the two-sided symmetric test with $\alpha/2$ probability in each tail of the null distribution of T^+.

The tests can be performed using the `R` command `wilcox.test` (see Example 3.1). The t_α critical values can be obtained from the `R` command `psignrank` (see Comment 5).

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of T^+, suitably standardized. We first need to know the expected value and variance of T^+ when the null hypothesis is true. When H_0 is true, the expected value and variance of T^+ are

$$E_0(T^+) = \frac{n(n+1)}{4} \tag{3.7}$$

and

$$\text{var}_0(T^+) = \frac{n(n+1)(2n+1)}{24}, \tag{3.8}$$

respectively. These expressions for $E_0(T^+)$ and $\text{var}_0(T^+)$ are verified by direct calculations in Comment 6 for the special case of $n = 3$. General derivations of both expressions are presented in Comment 7.

The standardized version of T^+ is

$$T^* = \frac{T^+ - E_0(T^+)}{\{\text{var}_0(T^+)\}^{1/2}} = \frac{T^+ - \left\{\dfrac{n(n+1)}{4}\right\}}{\{n(n+1)(2n+1)/24\}^{1/2}}. \tag{3.9}$$

When H_0 is true, T^* has, as n tends to infinity, an asymptotic $N(0,1)$ distribution (see Comment 7 for indications of the proof). The normal theory approximation for procedure (3.4) is

$$\text{Reject } H_0 \text{ if } T^* \geq z_\alpha; \text{ otherwise do not reject;} \tag{3.10}$$

the normal theory approximation for procedure (3.5) is

$$\text{Reject } H_0 \text{ if } T^* \leq -z_\alpha; \text{ otherwise do not reject;} \tag{3.11}$$

and the normal theory approximation for procedure (3.6) is

$$\text{Reject } H_0 \text{ if } |T^*| \geq z_{\alpha/2}; \text{ otherwise do not reject.} \tag{3.12}$$

Ties

If there are zero values among the Z's, discard the zero values and redefine n to be the number of nonzero Z's. If there are ties among the (nonzero) $|Z|$'s, assign each of the observations in a tied group the average of the integer ranks that are associated with the tied group. After computing T^+ with these average ranks for nonzero Z's, use procedure (3.4), (3.5), or (3.6). Note, however, that this test associated with tied $|Z|$'s is only approximately, and not exactly, of significance level α. (To get an exact level α test even in this tied setting, see Comment 11.)

When applying the large-sample approximation, an additional factor must be taken into account. Although ties in the nonzero $|Z|$'s do not affect the null expected value of T^+, its null variance is reduced to

$$\text{var}_0(T^+) = (24)^{-1}\left[n(n+1)(2n+1) - \frac{1}{2}\sum_{j=1}^{g} t_j(t_j-1)(t_j+1)\right], \tag{3.13}$$

where g denotes the number of tied groups of nonzero $|Z|$'s and t_j is the size of the tied group j. We note that an untied observation is considered to be a tied "group" of size 1. In particular, if there are no ties among the $|Z|$'s, then $g = n$ and $t_j = 1$ for $j = 1, \ldots, n$. In this case, each term in (3.13) of the form $t_j(t_j - 1)(t_j + 1)$ reduces to zero, and the variance expression in (3.13) reduces to the usual null variance of T^+ when there are no ties, as given in (3.8). Note that the term $(48)^{-1}\sum_{j=1}^{g} t_j(t_j-1)(t_j+1)$ represents the reduction in the null variance of T^+ due to the presence of tied nonzero Z's.

As a consequence of the effect that ties have on the null variance of T^+, the following modification is needed to apply the large-sample approximation when there are tied nonzero Z's. Compute T^+ using average ranks and set

$$T^* = \frac{T^+ - \left\{\dfrac{n(n+1)}{4}\right\}}{\{\text{var}_0(T^+)\}^{1/2}}, \tag{3.14}$$

where $\text{var}_0(T^+)$ is now given by display (3.13). With this modified value of T^*, approximations (3.10), (3.11), or (3.12) can be applied.

EXAMPLE 3.1 ***Hamilton Depression Scale Factor IV.***

The data in Table 3.1 are a portion of the data obtained by Salsburg (1970). These data, based on nine patients who received tranquilizer T, were taken from a double-blind clinical trial involving two tranquilizers. The measure used was the Hamilton (1960) depression scale factor IV (the "suicidal" factor). The X (pre) value was obtained at the first patient visit after initiation of therapy, whereas the Y (post) value was obtained at the second visit after initiation of therapy. The patients had been diagnosed as having mixed anxiety and depression.

In this example, an improvement due to tranquilizer T corresponds to a reduction in factor IV values. Hence, we apply test (3.5), which is designed to detect the alternative $\theta < 0$. One obtains the value of T^+ by first calculating the nine $Z_i = Y_i - X_i$ differences, then ranking from least to greatest the nine absolute values $|Z_1|, \ldots, |Z_9|$, and finally adding the ranks of the $|Z|$'s that emanated from positive Z difference.

To perform the test using *R*, set

```
pre<-c(1.83, .50, 1.62, 2.48, 1.68, 1.88, 1.55, 3.06, 1.30),

post<-c(.878, .647, .598, 2.05, 1.06, 1.29, 1.06, 3.14, 1.29).
```

Then apply `wilcox.test(pre,post, paired=TRUE, alterative = "less")` to obtain $T^+ = 5$ with a P-value of .02.

Table 3.1 The Hamilton Depression Scale Factor IV Values

Patient i	X_i	Y_i
1	1.83	0.878
2	0.50	0.647
3	1.62	0.598
4	2.48	2.05
5	1.68	1.06
6	1.88	1.29
7	1.55	1.06
8	3.06	3.14
9	1.30	1.29

Source: D. S. Salsburg (1970).

i	Z_i	$\lvert Z_i \rvert$	R_i	ψ_i	$R_i \psi_i$
1	−0.952	0.952	8	0	0
2	0.147	0.147	3	1	3
3	−1.022	1.022	9	0	0
4	−0.430	0.430	4	0	0
5	−0.620	0.620	7	0	0
6	−0.590	0.590	6	0	0
7	−0.490	0.490	5	0	0
8	0.080	0.080	2	1	2
9	−0.010	0.010	1	0	0
					$T^+ = 5$

For the large-sample approximation, we find (since there are no ties) from (3.9) that

$$T^* = \frac{5 - (9(10)/4)}{\{9(10)(19)/24\}^{1/2}} = -2.07.$$

From `pnorm(-2.07)=.0192`, the smallest significance level at which we can reject H_0 in favor of $\theta < 0$ using the normal approximation is .0192. Both the exact test and the large-sample approximation indicate that there is strong evidence that tranquilizer T does lead to patient improvement, as measured by a reduction in the Hamilton scale factor IV values.

EXAMPLE 3.2 ***Government versus Private Sector Salaries.***

In an annual survey to determine whether federal pay scales were commensurate with private sector salaries, government and private workers were matched as closely as possible (with respect to type of job, educational background, years experience, etc.) and the salaries of the matched pairs were obtained. The data in Table 3.2 are the annual salaries (in dollars) for 12 such matched pairs, as reported by McClave and Benson (1978).

Letting X correspond to the government worker's salary and Y to the matched private sector salary, the tabular presentation of the associated positive signed ranks (using average ranks to break ties) is as follows:

Table 3.2 Annual Salaries

Pair i	Private	Government
1	12,500	11,750
2	22,300	20,900
3	14,500	14,800
4	32,300	29,900
5	20,800	21,500
6	19,200	18,400
7	15,800	14,500
8	17,500	17,900
9	23,300	21,400
10	42,100	43,200
11	16,800	15,200
12	14,500	14,200

Source: J. T. McClave and G. Benson (1978).

i	z_i	$\lvert Z_i \rvert$	R_i	ψ_i	$R_i\psi_i$
1	750	750	5	1	5
2	1400	1400	9	1	9
3	−300	300	1.5	0	0
4	2400	2400	12	1	12
5	−700	700	4	0	0
6	800	800	6	1	6
7	1300	1300	8	1	8
8	−400	400	3	0	0
9	1900	1900	11	1	11
10	−1100	1100	7	0	0
11	1600	1600	10	1	10
12	300	300	1.5	1	1.5

To test H_0 versus the alternative that government workers are generally paid less than their counterparts in the private sector, we use the signed rank test of $H_0 : \theta = 0$ versus $H_0 : \theta > 0$. From the signed rank computational array, we see that

$$T^+ = 5 + 9 + 12 + 6 + 8 + 11 + 10 + 1.5 = 62.5.$$

Using the R command `psignrank(62,12,lower.tail=F)`, we find that the smallest significance level at which these data lead to rejection of $H_0 : \theta = 0$ in favor of $H_1 : \theta > 0$ (i.e., the one-sided P-value) is $\alpha = .0320$. Hence, there is moderate evidence to indicate that federal government workers (at least in the type of jobs considered in this survey) are, indeed, paid less than their private sector counterparts. (We point out that the P-value for these data is only approximate, due to the tied \$300 absolute differences. For a discussion of how to obtain the exact conditional P-value in this case, see Comment 11.)

For the normal approximation with the data in Table 3.2, we need to use the ties-corrected version of T^* given in (3.14). For the salary data, we have $g = 11$ and (arbitrarily labeling the tied groups in the order of increasing ranks) $t_1 = 2, t_2 = t_3 = \ldots = t_{10} = t_{11} = 1$. Using the ties-corrected formula (3.13) for $\text{var}_0(T^+)$, we obtain

$$T^* = \frac{62.5 - \dfrac{12(13)}{4}}{\left\{\dfrac{12(12+1)(2(12)+1) - \frac{1}{2}(2)(1)(3)}{24}\right\}^{1/2}} = \frac{62.5 - 39}{\left\{\dfrac{3897}{24}\right\}^{1/2}} = 1.84.$$

To find the P-value associated with this normal approximation, we obtain `1-pnorm(1.84)=.0329`, which is in good agreement with the value of .0320 obtained without using the normal approximation.

Comments

1. *Motivation for the Test.* When θ is greater than 0, there will tend to be a large proportion of positive Z differences and they will tend to have the larger absolute values. Hence, when θ is greater than 0, we would expect a higher proportion of positive signed ranks with relatively large sizes, leading to a big value of T^+. This suggests rejecting H_0 in favor of $\theta > 0$ for large values of T^+ and motivates procedures (3.4) and (3.10). Similar rationales lead to procedures (3.5), (3.6), (3.11), and (3.12).

2. *Assumptions.* There is no requirement that the individual X_i and Y_i be independent, only that the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, and therefore the resulting differences $Z_1, \ldots, Z_n$, be mutually independent. Indeed, in most applications, the individual X_i and Y_i are dependent. For paired replicates data, the symmetry part of Assumption A2 is often inherently satisfied. In particular, if each X_i and $Y_i, i = 1, \ldots, n$, arise from populations differing only in location (i.e., the only treatment "effect" is a change in location), then the $(Z_i - \theta)$'s come from populations that are symmetric about zero. (This is, in fact, true under more general conditions.)
3. *Testing θ Equal to Some Specified Nonzero Value.* Procedures (3.4), (3.5), and (3.6) and the corresponding normal approximations (3.10), (3.11), and (3.12) are for testing θ equal to zero. To test $\theta = \theta_0$, where θ_0 is some specified nonzero number, subtract θ_0 from each of the differences $Z_1, \ldots, Z_n$ to form a modified sample $Z_1' = Z_1 - \theta_0, \ldots, Z_n' = Z_n - \theta_0$. Then compute T^+ as the sum of the positive signed ranks for these Z_i's. Procedures (3.4), (3.5), and (3.6) and their corresponding large-sample approximations (3.10), (3.11), and (3.12) are then applied as previously described.
4. *Equivalent Form.* It may appear that some of the information in the ranking of the sample Z-differences is being lost by using only the positive signed ranks to compute T^+. Such is not the case. If we define T^- to be the sum of ranks (of the absolute values) corresponding to the negative Z observations, then $T^- = \sum_{i=1}^{n} (1 - \psi_i) R_i$. It follows that $T^+ + T^- = \sum_{i=1}^{n} R_i = n(n+1)/2$. Thus, the test procedures defined in procedures (3.4), (3.5), and (3.6) and the corresponding approximations (3.10), (3.11), and (3.12) could equivalently be based on $T^- = [n(n+1)/2] - T^+$.
5. *Derivation of the Distribution of T^+ under H_0 (No Ties Case).* Let B be the number of positive Z's and let $r_1 < \cdots < r_B$ denote the ordered ranks of the absolute values of these positive Z's. Then the null (H_0) distribution can be obtained directly from the representation $T^+ = \sum_{i=1}^{B} r_i$. Under the assumption that the underlying Z_i distributions are all continuous, the probabilities are zero that there are ties among the absolute values of the Z's or that any of the Z's are exactly zero. In addition, under H_0, these underlying Z_i distributions are all symmetric about $\theta = 0$. It follows that under H_0, each of the 2^n possible outcomes for the ordered configuration $(r_1, \ldots, r_B)$ occurs with equal probability $\left(\frac{1}{2}\right)^n$. For example, in the case of $n = 3$, the $2^3 = 8$ possible outcomes for $(r_1, \ldots, r_B)$ and associated values of T^+ are given in the following table.

B	$(r_1, r_2, \ldots, r_B)$	Probability under H_0	$T^+ = \sum_{i=1}^{B} r_i$
0		$\frac{1}{8}$	0
1	$r_1 = 1$	$\frac{1}{8}$	1
1	$r_1 = 2$	$\frac{1}{8}$	2
1	$r_1 = 3$	$\frac{1}{8}$	3
2	$r_1 = 1, r_2 = 2$	$\frac{1}{8}$	3
2	$r_1 = 1, r_2 = 3$	$\frac{1}{8}$	4
2	$r_1 = 2, r_2 = 3$	$\frac{1}{8}$	5
3	$r_1 = 1, r_2 = 2, r_3 = 3$	$\frac{1}{8}$	6

Thus, for example, the probability is $\frac{2}{8}$ under H_0 that T^+ is equal to 3, since $T^+ = 3$ when either of the exclusive outcomes $B = 1,\ r_1 = 3$ or $B = 2, (r_1 = 1, r_2 = 2)$ occurs and each of these outcomes has null probability $\frac{1}{8}$. Simplifying, we obtain the null distribution.

Possible value of T^+	Probability under H_0
0	$\frac{1}{8}$
1	$\frac{1}{8}$
2	$\frac{1}{8}$
3	$\frac{2}{8}$
4	$\frac{1}{8}$
5	$\frac{1}{8}$
6	$\frac{1}{8}$

The probability, under H_0, that T^+ is greater than or equal to 5, for example, is therefore

$$P_0(T^+ \geq 5) = P_0(T^+ = 5) + P_0(T^+ = 6)$$
$$= .125 + .125 = .25.$$

This agrees with what is obtained from `psignrank(4,3,lower.tail=F)` which gives, for $n = 3$, the probability under the null hypothesis that $T^+ > 4$.

Note that we have derived the null distribution of T^+ without specifying the forms of the underlying Z populations under H_0 beyond the point of requiring that they be continuous and symmetric about zero. This is why the test procedures based on T^+ are called *distribution-free procedures*. From the null distribution of T^+ we can determine the critical value t_α and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific forms of the underlying continuous and symmetric (about 0) Z distributions.

6. *Calculation of the Mean and Variance of T^+ under the Null Hypothesis H_0.* In displays (3.7) and (3.8) we presented formulas for the mean and variance of T^+ when the null hypothesis is true. In this comment, we illustrate a direct calculation of $E_0(T^+)$ and $\text{var}_0(T^+)$ in the particular case of $n = 3$, using the null distribution of T^+ obtained in Comment 5. (Later, in Comment 7, we present general derivations of $E_0(T^+)$ and $\text{var}_0\ (T^+)$.) The null mean, $E_0(T^+)$, is obtained by multiplying each possible value of T^+ with its probability under H_0. Thus,

$$E_0(T^+) = 0(.125) + 1(.125) + 2(.125) + 3(.25) + 4(.125) + 5(.125)$$
$$+ 6(.125) = 3.$$

This is in agreement with what we obtain using (3.7), namely,

$$E_0(T^+) = \frac{n(n+1)}{4} = \frac{3(3+1)}{4} = 3.$$

A check on the expression for $\text{var}_0(T^+)$ is also easily performed, using the well-known fact that

$$\text{var}_0(T^+) = E_0[(T^+)^2] - \{E_0(T^+)\}^2.$$

The value of $E_0[(T^+)^2]$, the second moment of the null distribution of T^+, is again obtained by multiplying possible values (in this case, values of $(T^+)^2$) by the corresponding probabilities under H_0. We find

$$E_0[(T^+)^2] = [(0+1+4)(.125) + 9(.25) + (16+25+36)(.125)] = 12.5.$$

Thus,

$$\text{var}_0(T^+) = 12.5 - (3)^2 = 3.5,$$

which agrees with what we obtain using (3.8) directly, namely,

$$\text{var}_0(T^+) = \frac{3(3+1)(2(3)+1)}{24} = 3.5.$$

7. *Large-Sample Approximation.* In view of the representation $T^+ = \sum_{i=1}^{B} r_i$, it follows from the discussion in Comment 5 that $T^+ \underline{\underline{d}} \sum_{i=1}^{n} V_i$, where the symbol $\underline{\underline{d}}$ means "has the same distribution as" and $V_1, \ldots, V_n$ are mutually independent dichotomous random variables with probability distributions

$$P(V_i = i) = P(V_i = 0) = \frac{1}{2},$$

for $i = 1, \ldots, n$. From this distributionally equivalent form, we can immediately use well-known expressions for the mean and variance of a sum of mutually independent random variables to obtain

$$E_0(T^+) = E\left[\sum_{i=1}^{n} V_i\right] = \sum_{i=1}^{n} E[V_i] \tag{3.15}$$

and

$$\text{var}_0(T^+) = \text{var}\left(\sum_{i=1}^{n} V_i\right) = \sum_{i=1}^{n} \text{var}(V_i). \tag{3.16}$$

Since V_i is a dichotomous variable, we have, for $i = 1, \ldots, n$, that

$$E_0(V_i) = i\left(\frac{1}{2}\right) + 0\left(\frac{1}{2}\right) = \frac{i}{2}$$

and

$$\text{var}_0(V_i) = E_0(V_i^2) - [E_0(V_i)]^2 = \left[i^2\left(\frac{1}{2}\right) + 0^2\left(\frac{1}{2}\right)\right] - \left[\frac{i}{2}\right]^2$$
$$= \frac{i^2}{2} - \frac{i^2}{4} = \frac{i^2}{4}.$$

Using these results, along with the closed-form expressions for the sum of the first n positive integers and the sum of the squares of the first n positive integers, in (3.15) and (3.16), we obtain

$$E_0(T^+) = \frac{1}{2}\sum_{i=1}^{n} i = \frac{1}{2}\left[\frac{n(n+1)}{2}\right] = \frac{n(n+1)}{4}$$

and

$$\text{var}_0(T^+) = \frac{1}{4}\sum_{i=1}^{n} i^2 = \frac{1}{4}\left[\frac{n(n+1)(2n+1)}{6}\right] = \frac{n(n+1)(2n+1)}{24},$$

which agree with the general expressions stated in (3.7) and (3.8), respectively.

Also using the distributional equality between T^+ and $\sum_{i=1}^{n} V_i$, the asymptotic normality of the standardized form

$$T^* = \frac{T^+ - E_0(T^+)}{\{\text{var}_0(T^+)\}^{1/2}} = \frac{T^+ - \dfrac{n(n+1)}{4}}{\left\{\dfrac{n(n+1)(2n+1)}{24}\right\}^{1/2}}$$

follows from standard theory for sums of mutually independent, but not identically distributed, random variables, such as the Liapounov central limit theorem (cf. Randles and Wolfe (1979, p. 423)). Asymptotic normality results are also obtainable under general alternatives to H_0. See, for example, the Hoeffding (1948a) U-statistic theorem as stated and applied to the Wilcoxon signed rank statistic on pages 82–85 of Randles and Wolfe (1979).

8. *Symmetry of the Distribution of T^+ under the Null Hypothesis.* When H_0 is true, the distribution of T^+ is symmetric about its mean $n(n+1)/4$. (See Comment 5 for verification of this when $n = 3$.) This implies that

$$P_0(T^+ \le x) = P_0\left(T^+ \ge \frac{n(n+1)}{2} - x\right), \tag{3.17}$$

for $x = 0, 1, \ldots, n(n+1)/2$.

9. *Zero Z Values.* We have recommended dealing with zero values among the Z's by discarding them and redefining n to be the number of nonzero Z's. This approach is satisfactory as long as the zero values are a very small percentage of the Z differences. If, however, there is a relatively large number of zero Z's, it would be advisable to consider an appropriate statistical procedure designed

for analyzing such discrete data. See, for example, Chapter 10 or a book on categorical data analysis, such as Agresti (2013).

We should also point out that there are methods other than elimination that have been proposed for dealing with zero Z values. One could use individual randomization (e.g., flipping a fair coin) to decide whether each of the zero Z values is to be counted as positive or negative in the construction of T^+. (Although this approach maintains many of the nice properties of T^+ that hold when there are no zeros, it introduces extraneous randomness that could quite easily have a direct effect on the outcome of any subsequent inferences based on such a modified T^+.) A second alternative approach in the case of the one-sided test procedures (3.4), (3.5), (3.10), and (3.11) is to be conservative about rejecting the null hypothesis H_0; that is, we could count all the zero Z values as if they were in favor of not rejecting H_0. Thus, for example, in applying either procedure (3.4) or (3.10) to test H_0 against the alternative $\theta > 0$, we would treat all of the zero Z's as if they were negative (in favor of not rejecting H_0) in the calculation of T^+. (In the case of procedures (3.5) and (3.11), zero Z's would be considered positive in the calculation of T^+.) Any rejection of H_0 with this conservative approach to deal with zero Z values could then be viewed as providing strong evidence in favor of the appropriate alternative. For a more detailed discussion of methods for handling zero observations, see Pratt (1959).

10. *Tied Nonzero Absolute Z Values.* Methods for dealing with tied nonzero absolute Z values other than using average ranks have been discussed in the literature. These include analogs to the randomization and conservative approaches mentioned in Comment 9 with regard to zero Z values. For further discussion of these alternative methods for dealing with tied nonzero absolute Z's, see Pratt (1959).

11. *Exact Conditional Distribution of T^+ with Ties Among the Nonzero Absolute Z Values.* To have a test with exact significance level even in the presence of tied absolute Z's (assuming there are no zero Z values or they have been discarded and n reduced accordingly), one considers all 2^n possible outcomes for the ordered configuration $(r_1, \ldots, r_B)$, where B represents the number of positive Z's as in Comment 5 but where $r_1 < \cdots < r_B$ now denote the ordered ranks of the absolute values of the positive Z's using average ranks to break the ties. As in Comment 5, it still follows that under H_0, each of the 2^n possible outcomes for the ordered configurations $(r_1, \ldots, r_B)$ based on using average ranks to break ties occurs with the probability $(\frac{1}{2})^n$. For each such configuration, the value of T^+ is computed and the results are tabulated. We illustrate this construction for $n = 4$ and the data $Z_1 = -12$, $Z_2 = -10$, $Z_3 = 10$, $Z_4 = 12$. Using average ranks to break ties, the associated absolute value ranks are $R_1 = 3.5, R_2 = 1.5, R_3 = 1.5$, and $R_4 = 3.5$. Thus, $B = 2$ and the ordered ties-broken-ranks for the positive Z's are $r_1 = 1.5$ and $r_2 = 3.5$, leading to an attained value of $T^+ = 5$. To assess the significance of T^+, we obtain its conditional distribution by considering the $2^4 = 16$ equally likely (under H_0) possible values of $(r_1, \ldots, r_B)$ for the given tied rank vector (1.5,1.5,3.5,3.5). These 16 values of $(r_1, \ldots, r_B)$ and associated values of T^+ are shown in the following table.

B	$(r_1, r_2, \ldots, r_B)$	Probability under H_0	Value of T^+
0		$\frac{1}{16}$	0
1	$r_1 = 1.5$	$\frac{1}{16}$	1.5
1	$r_1 = 1.5$	$\frac{1}{16}$	1.5
1	$r_1 = 3.5$	$\frac{1}{16}$	3.5
1	$r_1 = 3.5$	$\frac{1}{16}$	3.5
2	$r_1 = 1.5, r_2 = 1.5$	$\frac{1}{16}$	3
2	$r_1 = 1.5, r_2 = 3.5$	$\frac{1}{16}$	5
2	$r_1 = 1.5, r_2 = 3.5$	$\frac{1}{16}$	5
2	$r_1 = 1.5, r_2 = 3.5$	$\frac{1}{16}$	5
2	$r_1 = 1.5, r_2 = 3.5$	$\frac{1}{16}$	5
2	$r_1 = 3.5, r_2 = 3.5$	$\frac{1}{16}$	7
3	$r_1 = 1.5, r_2 = 1.5, r_3 = 3.5$	$\frac{1}{16}$	6.5
3	$r_1 = 1.5, r_2 = 1.5, r_3 = 3.5$	$\frac{1}{16}$	6.5
3	$r_1 = 1.5, r_2 = 3.5, r_3 = 3.5$	$\frac{1}{16}$	8.5
3	$r_1 = 1.5, r_2 = 3.5, r_3 = 3.5$	$\frac{1}{16}$	8.5
4	$r_1 = 1.5, r_2 = 1.5, r_3 = 3.5, r_4 = 3.5$	$\frac{1}{16}$	10

This yields the null tail probabilities

$$
\begin{aligned}
P_0(T^+ \geq 10) &= \frac{1}{16}, \\
P_0(T^+ \geq 8.5) &= \frac{3}{16}, \\
P_0(T^+ \geq 7) &= \frac{4}{16}, \\
P_0(T^+ \geq 6.5) &= \frac{6}{16}, \\
P_0(T^+ \geq 5) &= \frac{10}{16}, \\
P_0(T^+ \geq 3.5) &= \frac{12}{16}, \\
P_0(T^+ \geq 3) &= \frac{13}{16}, \\
P_0(T^+ \geq 1.5) &= \frac{15}{16}, \\
P_0(T^+ \geq 0) &= 1.
\end{aligned}
$$

This distribution is called the *conditional distribution* or the *permutation distribution of* T^+, given the set of tied ranks {1.5,1.5,3.5,3.5}. For the particular observed value $T^+ = 5$, we have $P_0(T^+ \geq 5) = \frac{10}{16}$, so that such a value does not indicate a deviation from H_0 in the direction of $\theta > 0$.

12. *Some Power Results for the Wilcoxon Signed Rank Test.* We consider the upper-tail α-level test of $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ given by procedure (3.4). Under the additive shift model (see Assumption A2) and common underlying distribution $F_1 \equiv F_2 \equiv \cdots \equiv F_n \equiv F$ for the Z differences, the power, or probability of correctly rejecting H_0, for median θ_0 values "near" the null hypothesis value of 0 can be approximated by

$$\text{Power} \doteq \Phi(A_F), \tag{3.18}$$

where $\Phi(A_F)$ is the area under a standard normal density to the left of the point

$$A_F = \left\{ \frac{n(n-1)f^*(0) + nf(0)}{[n(n+1)(2n+1)/24]^{1/2}} \right\} \theta - z_\alpha, \tag{3.19}$$

where $f(0)$ is the common density function, evaluated at 0, for the Z differences and $f^*(0)$ is the density function, also evaluated at 0, of the sum of two independent random variables drawn from the Z population having distribution F (cf. Lehmann (1975, 167 and 403)).

When F is normal with standard deviation σ, we have $f(0) = (\sigma\sqrt{2\pi})^{-1}$ and $f^*(0) = (2\sigma\sqrt{\pi})^{-1}$. Under this setting, A_F in (3.19) reduces to

$$A_{\text{normal}} = \left\{ \frac{(n(n-1)/2) + n/\sqrt{2}}{[n(n+1)(2n+1)/24]^{1/2}} \right\} \frac{\theta}{\sigma\sqrt{\pi}} - z_\alpha. \tag{3.20}$$

Thus, when F is normal, the approximate power for the additive shift model depends on θ and σ only through their ratio θ/σ. This implies, for example, that the approximate power for the pair $(\theta = .5, \sigma = 4)$ is the same as the approximate power for the pair $(\theta = 1, \sigma = 8)$.

For the purpose of illustration, suppose that the additive shift model holds, with the common underlying population F taken to be normal with variance $\sigma^2 = 4$ and treatment effect $\theta = 1.5$. For the case where $n = 10$ and $\alpha = .053$, the test rejects H_0 if and only if $T^+ \geq 44$. Substituting the appropriate values in (3.20), we obtain

$$A_{\text{normal}} = \left\{ \frac{10(9)/2 + 10/\sqrt{2}}{[10(11)(21)/24]^{1/2}} \right\} \frac{1.5}{2\sqrt{\pi}} - 1.62 = .61$$

Thus, the approximate power of this test at $\theta = 1.5$ (and $\sigma^2 = 4$) is

$$\text{Power} \doteq \texttt{pnorm}(.61) = .73.$$

This compares with the exact power of .70 as given in Table 1 of Klotz (1963). Additional exact power values for the one-sided Wilcoxon signed rank test and sample sizes 5(1)10 can be found in Klotz (1963) for normal shift alternatives and in Arnold (1965) for shifted t-distributions with $\nu = \frac{1}{2}$, 1, 2, and 4 degrees of freedom.

13. *Sample Size Determination.* The Wilcoxon signed rank test detects a more general class of alternatives than the location-shift alternatives associated with model Assumption A2. When $Z_1, \ldots, Z_n$ are a random sample from a single continuous, symmetric population F, the one-sided upper-tail test defined by procedure (3.4) is consistent (i.e., has power tending to 1 as n tends to infinity) against those F populations for which $\eta > \frac{1}{2}$, with

$$\eta = P(Z_1 + Z_2 > 0), \tag{3.21}$$

where Z_1 and Z_2 are independent and identically distributed as F. The parameter η is the probability that a Z_1 randomly selected from the continuous and symmetric F will be greater than the negative of a second independent Z_2 also randomly selected from the same distribution F.

Noether (1987) shows how to determine an approximate sample size n so that the α-level one-sided test given by procedure (3.4) will have approximate power $1 - \beta$ against an alternative value of η greater than $\frac{1}{2}$. This approximate value of n is

$$n \doteq \frac{(z_\alpha + z_\beta)^2}{3\left(\eta - \frac{1}{2}\right)^2}. \tag{3.22}$$

As an illustration of the use of (3.22), suppose we are testing H_0 and we desire to have an upper-tail level $\alpha = .025$ test with power $1 - \beta$ of at least .95 against an alternative for which $\eta = P(Z_1 + Z_2 > 0) = .8$ (recall that under $H_0, \eta = .5$). Since $z_\alpha = z_{.025} = \texttt{qnorm}(.975) = 1.96$ and $Z_\beta = Z_{.05} = \texttt{qnorm}(.95) = 1.65$, we find that the approximate required sample size for the alternative $\eta = .8$ is

$$n \doteq \frac{(1.96 + 1.65)^2}{3(.8 - .5)^2} = 48.3.$$

To be conservative, we would take $n = 49$.

14. *Consistency of the T^+ Test.* Under the assumption that $Z_1, \ldots, Z_n$ is a random sample from a single continuous population F, the consistency of the tests based on T^+ depends on the parameter

$$\eta^* = P(Z_1 + Z_2 > 0) - \frac{1}{2},$$

where Z_1 and Z_2 are independent and identically distributed as F. The test procedures defined by (3.4), (3.5), and (3.6) are consistent against the classes of alternatives corresponding to $\eta^\star >, <$, and $\neq 0$, respectively.

Properties

1. *Consistency.* For our consistency statement we strengthen Assumption A2 to require that each Z has the same continuous population that is symmetric about θ. Then the tests defined by (3.4), (3.5), and (3.6) are consistent against the alternatives $\theta >, <$, and $\neq 0$, respectively. (See also Comment 14.)
2. *Asymptotic Normality.* See Randles and Wolfe (1979, pp. 83–85).
3. *Efficiency.* See Section 3.11.

Problems

1. The data in Table 3.3 are a subset of the data obtained by Kaneto, Kosaka, and Nakao (1967). The experiment investigated the effect of vagal nerve stimulation on insulin secretion. The subjects were mongrel dogs with varying body weights. Table 3.3 gives the amount of immunoreactive insulin in pancreatic venous plasma just before stimulation of the left vagus nerve (X) and the amount measured 5 min after stimulation (Y) for seven dogs. Test the hypothesis of no effect against the alternative that stimulation of the vagus nerve increases the blood level of immunoreactive insulin.
2. Change the value of X_3, in Table 3.1, from 1.62 to 16.2. What effect does this outlying observation have on the calculations performed in Example 3.1? What does this suggest about the relative insensitivity of the signed rank tests to outliers? Construct an example in which changing one observation has a marked effect on the final decision regarding rejection or acceptance of H_0.
3. Let $T^- = \sum_{i=1}^{n} R_i(1 - \psi_i)$, where $\psi_i = 1$ if $Z_i > 0$, and 0 otherwise. Verify directly, or illustrate using the data of Table 3.1, the equation $T^+ + T^- = n(n+1)/2$.
4. August, Hung, and Houck (1974) studied collagen metabolism in children deficient in growth hormone before and after growth hormone therapy. The data in Table 3.4 are the values of heat-insoluble hydroxyproline in the skin of children before and 3 months after growth hormone

Table 3.3 Blood Levels of Immunoreactive Insulin (μU/ml)

Dog i	X_i	Y_i
1	350	480
2	200	130
3	240	250
4	290	310
5	90	280
6	370	1450
7	240	280

Source: A. Kaneto, K. Kosaka, and K. Nakao (1967).

Table 3.4 Heat-Insoluble Hydroxyproline Micromoles per Gram of Dry Weight

Child i	Before	After
1	349	425
2	400	533
3	520	362
4	490	628
5	574	463
6	427	427
7	435	449

Source: G. P. August, W. Hung, and J. C. Houck (1974).

therapy. Can we conclude on the basis of these data that growth hormone therapy increases heat-insoluble hydroxyproline in the skin?

5. Assume that the additive shift model (see Assumption A2) holds with the common underlying distribution $F_1 \equiv F_2 \equiv \cdots \equiv F_n \equiv F$. If we have 15 observations and F is normal with variance 16, what is the approximate power of the level $\alpha = .076$ test of $H_0 : \theta = 0$ versus the alternative $\theta > 0$ when the treatment effect is $\theta = 1.25$?

6. For arbitrary number of observations n, what are the smallest and largest possible values for T^+? Justify your answers.

7. Consider the case $n = 8$ and use the R command `psignrank(0:18,8,lower.tail=T)` to produce the lower-tail probabilities of the null distribution of T^+. What are the possible α values between .05 and .10? Compare the $\alpha = .055$ test of $H_0 : \theta = 0$ versus $H_2 : \theta < 0$ with the corresponding $\alpha = .055$ test based on the large-sample approximation.

8. Consider a level $\alpha = .05$ test of $H_0 : \theta = 0$ versus the alternative $\theta > 0$ based on T^+ and let η be as given in (3.21). If our data $Z_1, \ldots, Z_n$ are a random sample from a single continuous, symmetric distribution $F(\cdot)$, how many observations n will we need to collect in order to have an approximate power of at least .84 against an alternative for which $\eta = .7$?

9. Suppose $n = 5$ and we observe the data $Z_1 = -1.3, Z_2 = 2.4, Z_3 = 1.3, Z_4 = 1.3$, and $Z_5 = 2.4$. What is the conditional probability distribution of T^+ under $H_0 : \theta = 0$ when average ranks are used to break ties among the absolute values of the Z's? How extreme is the observed value of T^+ in this conditional null distribution?

10. Apply the large-sample approximation test of $H_0 : \theta = 5$ versus $H_1 : \theta > 5$ based on T^+ to the beak-clapping data in Table 3.5. What is the P-value?

11. Consider procedure (3.6) with n observations for testing H_0: $\theta = 0$ versus $H_1 : \theta \neq 0$. If your critical region consists of the four values $T^+ = 0, 1, [n(n+1)/2] - 1, n(n+1)/2$, what is the significance level for your test?

12. Apply the one-sided upper-tail test based on T^+ to the data on Stanford Profile Scales of hypnotic susceptibility in Table 3.6. What is the P-value obtained?

13. For the case $n = 5$ untied Z observations, use the representation for T^+ discussed in Comment 5 to obtain the form of the exact null (H_0) distribution of T^+.

14. Let Z_1 and Z_2 be independent, identically distributed continuous random variables with a common probability distribution that is symmetric about 0. What is the value of η^* in Comment 14 for this setting?

15. Consider the test of H_0: $\theta = 0$ versus $H_1 : \theta > 0$ based on T^+ for the following $n = 10$ Z observations: $Z_1 = 2.5, Z_2 = 3.7, Z_3 = 0, Z_4 = -0.6, Z_5 = 4.7, Z_6 = 0, Z_7 = 1.4, Z_8 = 0, Z_9 = 1.9$, $Z_{10} = 5.2$. Compute the P -values for the competing T^+ procedures based on either (i) discarding the zero Z values and reducing n accordingly, as recommended in the Ties portion of this section, or (ii) treating the zero Z values in a conservative manner, as presented in Comment 9. Discuss the results.

16. Suppose you desire an upper-tail test of $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ based on T^+ and you want the test to have $\alpha = .05$ and a power of at least .90 when the distribution of $Z = Y - X$ is N(.5,1). Find the approximate required sample size.

17. What are the possible values of T^+ when $n = 8$? Suppose you are testing $H_0 : \theta = 0$ versus $H_2 : \theta < 0$ and you want your α level to be between .05 and .10. What are the tests available?

3.2 AN ESTIMATOR ASSOCIATED WITH WILCOXON'S SIGNED RANK STATISTIC (HODGES–LEHMANN)

Procedure

To estimate the treatment effect θ, form the $M = n(n+1)/2$ averages $(Z_i + Z_j)/2$, for $i \le j = 1, \ldots, n$. The estimator of θ associated with the Wilcoxon signed rank statistic T^+ is (see Comment 15)

$$\hat{\theta} = \text{median}\left\{\frac{Z_i + Z_j}{2}, i \le j = 1, \ldots, n\right\}. \tag{3.23}$$

Let $W^{(1)} \le \cdots \le W^{(M)}$ denote the ordered values of $(Z_i + Z_j)/2$. Then if M is odd, say $M = 2k+1$, we have $k = (M-1)/2$ and

$$\hat{\theta} = W^{(k+1)}, \tag{3.24}$$

the value that occupies position $k+1$ in the list of the ordered $(Z_i + Z_j)/2$ averages. If M is even, say $M = 2k$, then $k = M/2$ and

$$\hat{\theta} = \frac{W^{(k)} + W^{(k+1)}}{2}. \tag{3.25}$$

That is, when M is even, $\hat{\theta}$ is the average of the two $(Z_i + Z_j)/2$ values that occupy positions k and $k+1$ in the ordered list of the M $(Z_i + Z_j)/2$ averages. The $(Z_i + Z_j)/2$ averages are known as the *Walsh averages* (see Walsh (1949)).

EXAMPLE 3.3 ***Continuation of Example 3.1.***

To estimate θ for the Hamilton depression scale factor IV data in Table 3.1, we first calculate the $M = 9(10)/2 = 45$ $(Z_i + Z_j)/2$ averages. To facilitate this computation and eventual ordering of these averages, we first obtain the individual ordered Z values, denoted by $Z^{(1)} \le \cdots \le Z^{(9)}$, and place them as the headings for a 9×9 upper triangular array. The (i, j)th entry in this array then corresponds to the value of the sum $(Z^{(i)} + Z^{j})$, for $i \le j = 1, 2, \ldots, 9$. This representation for the Hamilton depression scale factor IV data is as follows:

	−1.022	−0.952	−0.620	−0.590	−0.490	−0.430	−0.010	0.080	0.147
−1.022	−2.044	−1.974	−1.642	−1.612	−1.512	−1.452	−1.032	−0.942	−0.875
−0.952		−1.904	−1.572	−1.542	−1.442	−1.382	−0.962	−0.872	−0.805
−0.620			−1.240	−1.210	−1.110	−1.050	−0.630	−0.540	−0.473
−0.590				−1.180	−1.080	−1.020	−0.600	−0.510	−0.443
−0.490					−0.980	−0.920	−0.500	−0.410	−0.343
−0.430						−0.860	−0.440	−0.350	−0.283
−0.010							−0.020	0.070	0.137
0.080								0.160	0.227
0.147									0.294

Thus, for example, the entry in the fourth row and sixth column of the array $(i = 4, j = 6)$ is $Z^{(4)} + Z^{(6)} = -.590 - .430 = -1.020$. The remaining 44 entries are calculated similarly. The ordered $Z^{(i)} + Z^{(j)}$ sums are then obtained by observation in this array, moving carefully from the upper left (the (1, 1) entry) across and down the array to the lower right (the (9,9) entry). The ordered $Z^{(i)} + Z^{(j)}$ sums for these data are as follows: $-2.044, -1.974, -1.904, -1.642, -1.612, -1.572, -1.542, -1.512, -1.452, -1.442, -1.382, -1.240, -1.210, -1.180, -1.110, -1.080, -1.050, -1.032, -1.020, -.980, -.962, -.942, -.920, -.875, -.872, -.860, -.805, -.630, -.600, -.540, -.510, -.500, -.473, -.443, -.440, -.410, -.350, -.343, -.283, -.020, .070, .137, .160, .227, .294$. The ordered values of the $(Z_i + Z_j)/2$ averages, namely, $W^{(1)} \leq \cdots \leq W^{(45)}$, then correspond to these ordered $Z^{(i)} + Z^{(j)}$ sums divided by 2. As $M = 45$ is odd, we use (3.24) with $k = (45 - 1)/2 = 22$ to obtain the estimate $\hat{\theta} = W^{(23)} = -.920/2 = -.460$ for the treatment effect θ. Thus, we estimate that a typical patient of the type included in this study will have a drop in the Hamilton depression scale factor IV value of roughly .460 due to treatment with tranquilizer T.

We can use the `R` command `owa` to compute the ordered Walsh averages and the Hodges–Lehmann estimator. Use `owa(pre,post)`.

[1]	−1.0220	−0.9870	−0.9520	−0.8210	−0.8060	−0.7860	−0.7710	−0.7560	−0.7260
[10]	−0.7210	−0.6910	−0.6200	−0.6050	−0.5900	−0.5550	−0.5400	−0.5250	−0.5160
[19]	−0.5100	−0.4900	−0.4810	−0.4710	−0.4600	−0.4375	−0.4360	−0.4300	−0.4025
[28]	−0.3150	−0.3000	−0.2700	−0.2550	−0.2500	−0.2365	−0.2215	−0.2200	−0.2050
[37]	−0.1750	−0.1715	−0.1415	−0.0100	0.0350	0.0685	0.0800	0.1135	0.1470

Comments

15. *Motivation for the Hodges–Lehmann Estimator.* The Hodges–Lehmann estimator $\hat{\theta}$, defined by (3.23), is associated with the Wilcoxon signed rank test. When $\theta = 0$, the distribution of the statistic T^+ is symmetric about its mean, $n(n+1)/4$ (see Comment 8). A natural estimator of θ is the amount $\hat{\theta}$ (say) that should be subtracted from each Z_i so that the value of T^+, when applied to the shifted sample $Z_1 - \hat{\theta}, \ldots, Z_n - \hat{\theta}$, is as close to $n(n+1)/4$ as possible. Roughly speaking, we estimate θ by the amount $(\hat{\theta})$ that the Z sample should be shifted in order that $Z_1 - \hat{\theta}, \ldots, Z_n - \hat{\theta}$ appears (when "viewed" by the signed rank statistic T^+) as a sample from a population with median 0. (Under Assumptions A1 and A2, each of the $Z_1 - \theta, \ldots, Z_n - \theta$ variables is from a population with median 0.)

    The Hodges–Lehmann method can be applied to a large class of statistics containing T^+. However, the forms of the resulting estimators for other members of this class are not always as convenient for calculation as is $\hat{\theta}$. See Hodges and Lehmann (1983) for an expository article on their method.

16. *Sensitivity to Gross Errors.* The estimator $\hat{\theta}$ is relatively insensitive to outliers. This is not the case with the classical estimator $\bar{Z} = \sum_{i=1}^{n} Z_i/n$. Thus the use of $\hat{\theta}$ provides protection against gross errors.

17. *The Walsh Averages.* Each of the $n(n+1)/2$ averages $(Z_i + Z_j)/2, i \leq j = 1, \ldots, n$, is called a *Walsh average* (see Walsh (1949)). If we define W^+ to be the number of positive Walsh averages, then (when there are no ties among the

$|Z|$'s and none of the Z's is zero) the statistic W^+ is identical to T^+ (3.3). (See Problem 22.) This result is due to Tukey (1949).

18. *Zero and Tied Absolute Z's.* Note that in calculating the estimator $\hat{\theta}$ we use *all* of the Z differences in computing the $(Z_i + Z_j)/2$ averages. Although we recommend (see Ties in Section 3.1) discarding the zero Z values (and reducing n accordingly) prior to applying the signed rank test to the data, it is not necessary to do so when calculating $\hat{\theta}$. In fact, the zero Z values contain important information about the magnitude of the treatment effect. This is also the case when we consider (Section 3.3) confidence intervals and bounds for θ.
19. *Pseudomedian.* A pseudomedian (cf. Høyland (1965)) of a distribution F is defined to be a median of the distribution of $(Z_1 + Z_2)/2$, where Z_1 and Z_2 are independent, each with the same distribution F. We assume here that our F is such that both the median and the pseudomedian of F are unique. The estimator $\hat{\theta}$ (3.23) is a consistent estimator of the pseudomedian, which in general may differ from the median θ. However, when F is symmetric as assumed in this section, the median and the pseudomedian coincide.

Properties

1. *Standard Deviation of $\hat{\theta}$.* For the asymptotic standard deviation of $\hat{\theta}$ (3.23), see Hodges and Lehmann (1963), Lehmann (1963c), and Comment 24.
2. *Asymptotic Normality.* See Hodges and Lehmann (1963) and Ramachandramurty (1966a).
3. *Efficiency.* See Hodges and Lehmann (1963), Bickel (1965), Høyland (1968), Gastwirth and Rubin (1969), and Section 3.11.

Problems

18. Consider the data of Table 3.2. Using the X and Y associations from Example 3.2, estimate θ for the salary data of that example.

19. Estimate θ for the blood-level data of Table 3.3.

20. Change the value of X_3, as given in Table 3.1, from 1.62 to 16.2. How does this affect the value of $\bar{Z} = \sum_{i=1}^{9} Z_i/9$? How does it affect the estimate of θ given by $\hat{\theta}$? Interpret these calculations in light of Comment 16.

21. Estimate θ for the heat-insoluble hydroxyproline data of Table 3.4.

22. Verify directly, or illustrate using the data of Table 3.1, that (when there are no ties among the absolute values of the Z's and none of the Z's is zero) T^+ is equal to the number of positive Walsh averages W^+. (See Comment 17.)

23. (a) What happens to $\hat{\theta}$ when we add a number b to each of the sample values $Z_1, \ldots, Z_n$?
(b) What happens to $\hat{\theta}$ when we multiply each sample value $Z_i, \ldots, Z_n$ by a number d?
(c) Let k be a positive integer such that $n > 2k$. What happens to $\hat{\theta}$ when we discard the k largest and the k smallest Z values from the sample?

24. (a) Do we need to calculate all of the $n(n+1)/2$ Walsh averages in order to compute the value of $\hat{\theta}$? Explain.

25. Explain why the Hodges–Lehmann estimator is less influenced by outlying observations than is the sample mean of the Z's.

26. Use R to obtain the Hodges–Lehmann estimator for the salary data of Table 3.2.

3.3 A DISTRIBUTION-FREE CONFIDENCE INTERVAL BASED ON WILCOXON'S SIGNED RANK TEST (TUKEY)

Procedure

For a symmetric two-sided confidence interval for θ, with confidence coefficient $1-\alpha$, set

$$C_\alpha = \frac{n(n+1)}{2} + 1 - t_{\alpha/2}, \tag{3.26}$$

where $t_{\alpha/2}$ is the upper $(\alpha/2)$th percentile point of the null distribution of T^+. The percentile points can be found using the R function `psignrank`.

The $100(1-\alpha)\%$ confidence interval (θ_L, θ_U) for θ that is associated with the two-sided Wilcoxon signed rank test (see Comment 20) of $H_0 : \theta = 0$ is then given by

$$\theta_L = W^{(C_\alpha)}, \theta_U = W^{(M+1-C_\alpha)} = W^{(t_{\alpha/2})}, \tag{3.27}$$

where $M = n(n+1)/2$ and $W^{(1)} \le \cdots \le W^{(M)}$ are the ordered values of the $(Z_i + Z_j)/2$ averages, $1 \le i \le j \le n$, used in computing the point estimator $\hat{\theta}$ (3.23); that is, θ_L is the $(Z_i + Z_j)/2$ average (i.e., the Walsh average; see Comment 17) that occupies position C_α in the list of M ordered $(Z_i + Z_j)/2$ averages. The upper end point θ_U is the $(Z_i + Z_j)/2$ average that occupies the position $M + 1 - C_\alpha = t_{\alpha/2}$ in this ordered list. With θ_L and θ_U given by display (3.27), we have

$$P_\theta(\theta_L < \theta < \theta_U) = 1 - \alpha \text{ for all } \theta. \tag{3.28}$$

(For upper or lower confidence bounds for θ associated with the appropriate one-sided Wilcoxon signed rank tests of H_0: $\theta = 0$, see Comment 21.)

Large-Sample Approximation

For large n, the integer C_α may be approximated by

$$C_\alpha \approx \frac{n(n+1)}{4} - z_{\alpha/2}\left\{\frac{n(n+1)(2n+1)}{24}\right\}^{1/2}. \tag{3.29}$$

In general, the value of the right-hand side of (3.29) is not an integer. To be conservative, take C_α to be the largest integer that is less than or equal to the right-hand side of (3.29).

EXAMPLE 3.4 ***Continuation of Examples 3.1 and 3.3.***

Consider the Hamilton depression scale factor IV data of Table 3.1. We illustrate how to obtain the 96% confidence interval for θ. With $1-\alpha = .96, \alpha/2 = .02$. From `psignrank(0:22,9,lower.tail=T)`, we find $P_0(T^+ \le 6) = P_0(T^+ \ge 40) = .02$. Thus, $t_{.02} = 40$. From (3.26), it follows that

$$C_{.04} = \left[\frac{9(9+1)}{2} + 1 - 40\right] = 6.$$

Using these values of $C_{.04} = 6$ and $t_{.02} = 40$ in display (3.27), we see that

$$\theta_L = W^{(6)} = -.786 \text{ and } \theta_U = W^{(40)} = -.010.$$

The value $\theta_L = -.786$ is the sixth smallest Walsh average and can be found from the list of ordered Walsh averages at the end of Example 3.3. Similarly, $\theta_U = -.010$ is the sixth largest Walsh average (or 40th ordered).

If we choose to apply the large-sample approximation, we find from approximation (3.29) that

$$C_{.04} \approx \left[\frac{9(9+1)}{4}\right] - 2.05\left\{\frac{9(9+1)(2(9)+1)}{24}\right\}^{1/2} = 5.2.$$

Thus, with a conservative approach and the large-sample approximation, we set $C_{.04} = 5$ and find that

$$(\theta_L, \theta_U) = (W^{(5)}, W^{(41)}) = (-.806, .035)$$

is the approximate 96% confidence interval for θ.

The exact 96% confidence internal can be found from the R command `wilcox.test(post-pre,conf.int=T, conf.level=.96)` yielding (−.786, −.010).

Comments

20. *Relationship of Confidence Interval to Two-Sided Test.* The 100 $(1-\alpha)$% confidence interval for θ given by display (3.27) can be obtained from the two-sided signed rank test as follows. The confidence interval (θ_L, θ_U) consists of those θ_0 values for which the two-sided α-level test of $\theta = \theta_0$ (see Comment 3) does not reject the hypothesis $\theta = \theta_0$. The confidence interval given by display (3.27) was defined by way of a graphical procedure by Lincoln Moses (who attributed it to John Tukey) in Chapter 18 of Walker and Lev (1953). See Lehmann (1986, p. 90) for a general result relating confidence intervals and acceptance regions of tests, and see Lehmann (1963c) for the specific result involving the signed rank test.
21. *Confidence Bounds.* In many settings, we are interested only in making one-sided confidence statements about the parameter θ; that is, we wish to assert with specified confidence that θ is no larger (or, in other settings, no smaller) than some upper (lower) confidence bound based on the sample data. To obtain such one-sided confidence bounds for θ, we proceed as follows. For the specified confidence coefficient $1-\alpha$, set

$$C_\alpha^* = \frac{n(n+1)}{2} + 1 - t_\alpha, \tag{3.30}$$

where t_α is the upper αth percentile point of the null distribution of T^+.
The $100(1-\alpha)$% lower confidence bound θ_L^* for θ that is associated with the one-sided Wilcoxon signed rank test of $H_0 : \theta = 0$ against the alternative $H_1 : \theta > 0$ is then given by

$$(\theta_L^*, \infty) = (W^{(C_\alpha^*)}, \infty), \tag{3.31}$$

where, as before, $M = n(n+1)/2$ and $W^{(1)} \le \cdots \le W^{(M)}$ are the ordered values of the $(Z_i + Z_j)/2$ averages, $1 \le i \le j \le n$. With θ_{L}^* given by display (3.31), we have

$$P_\theta(\theta_{\mathrm{L}}^* < \theta < \infty) = 1 - \alpha \text{ for all } \theta. \tag{3.32}$$

The corresponding $100(1-\alpha)\%$ upper confidence bound θ_{U}^* for θ that is associated with the one-sided Wilcoxon signed rank test of $H_0 : \theta = 0$ against the alternative $H_1 : \theta < 0$ is given by

$$(-\infty, \theta_{\mathrm{U}}^*) = (-\infty, W^{(M+1-C_\alpha^*)}) = (-\infty, W^{(t_a)}), \tag{3.33}$$

where C_α^* is given in (3.30). It follows that

$$P_\theta(-\infty < \theta < \theta_{\mathrm{U}}^*) = 1 - \alpha \text{ for all } \theta. \tag{3.34}$$

For large n, the integer C_α^* may be approximated by

$$C_\alpha^* \approx \frac{n(n+1)}{4} - z_\alpha \left\{ \frac{n(n+1)(2n+1)}{24} \right\}^{1/2}. \tag{3.35}$$

As with C_α (3.29) and the confidence interval for θ, the value of the right-hand side of (3.35) is not an integer. To be conservative, take C_α^* to be the largest integer that is less than or equal to the right-hand side of (3.35).

The $100(1-\alpha)\%$ lower and upper confidence bounds θ_{L}^* (3.31) and θ_{U}^* (3.33) are related to the acceptance regions of the one-sided Wilcoxon signed rank tests of $H_0 : \theta = \theta_0$ against the alternatives $\theta > \theta_0$ and $\theta < \theta_0$, respectively, in the same way that the confidence interval $(\theta_{\mathrm{L}}, \theta_{\mathrm{U}})$ is related to the acceptance region of the two-sided Wilcoxon signed rank test of $H_0 : \theta = \theta_0$. (See Comment 20.)

22. *Zero and Tied Absolute Z's.* Note that in calculating the confidence interval $(\theta_{\mathrm{L}}, \theta_{\mathrm{u}})$ from display (3.27) or the confidence bounds θ_{L}^* (3.31) or θ_{U}^* (3.33) for θ, we use *all* the Z differences in computing the $(Z_i + Z_j)/2$ averages. This is in common with our recommendation (see Comment 18) for computing the point estimator $\hat{\theta}$ (3.23), but different from the recommended policy (see Ties in Section 3.1) of discarding the zero Z values (and reducing n accordingly) prior to applying the signed rank test to the data. However, if there are zero Z's in the data, the equivalence (discussed in Comments 20 and 21) between the acceptance regions of the one-sided and two-sided signed rank tests and the appropriate confidence bound and confidence interval, respectively, are no longer valid. In addition, in cases with tied absolute Z's, the nominal confidence coefficient $1-\alpha$ used in displays (3.27), (3.31), and (3.33) is no longer exact. (See Comment 11.)

23. *Midpoint of Confidence Interval as an Estimator.* The midpoint of the interval (3.27), namely, $[W^{(C_\alpha)} + W^{(M+1-C_\alpha)}]/2$, suggests itself as a reasonable estimator of θ. (Note that this actually yields a class of estimators depending on the value of α.) In general, this midpoint is not the same as $\hat{\theta}$ (3.23). Lehmann (1963c) has also derived an asymptotically distribution-free confidence interval centered at $\hat{\theta}$. This asymptotically distribution-free confidence interval is based on the assumption that each of the n Z_i's comes from the *same* continuous population that is symmetric about θ. This assumption is more restrictive than Assumption A2.

24. *Estimating the Asymptotic Standard Deviation of* $\hat{\theta}$. Replace Assumption A2 by the stronger Assumption $A2'$: each Z comes from the *same* continuous population that is symmetric about θ. Then, it follows from Lehmann (1963c) that the statistic $(\theta_U - \theta_L)/2z_{\alpha/2}$, where (θ_L, θ_U) is the $100(1-\alpha)\%$ confidence interval for θ defined by display (3.27), is a consistent estimator for the asymptotic standard deviation of the point estimator $\hat{\theta}$ (3.23).

Properties

1. *Distribution-Freeness.* For populations satisfying Assumptions Al and A2, (3.28) holds. Hence, we can control the coverage probability to be $1-\alpha$ without having more specific knowledge about the forms of the underlying Z distributions. Thus, (θ_L, θ_U) is a distribution-free confidence interval for θ over a very large class of populations.
2. *Efficiency.* See Lehmann (1963c) and Section 3.11.

Problems

27. For the blood-level data of Table 3.3, obtain a confidence interval for θ with the exact confidence coefficient .954.

28. For the heat-insoluble hydroxyproline data of Table 3.4, obtain a confidence interval for θ with the exact confidence coefficient .922.

29. For the blood-level data of Table 3.3 and $\alpha = .078$, calculate the point estimator of θ defined in Comment 23. Compare with the value of $\hat{\theta}$ obtained in Problem 19.

30. Use the results of Example 3.4 to obtain an estimate of the asymptotic standard deviation of $\hat{\theta}$ for the Hamilton depression scale factor IV data of Table 3.1 (see Comment 24).

31. For the Hamilton depression scale factor IV data of Table 3.1, find an upper confidence bound for θ with the exact confidence coefficient .973 (see Comment 21).

32. For the salary data of Table 3.2, use (3.31) in Comment 21 and find a lower confidence bound for θ with approximate confidence coefficient .936. Why is the confidence coefficient only approximate and not exact?

33. Consider the $1-\alpha$ confidence interval for θ defined by display (3.27). Let $Z_{(1)} \leq \cdots \leq Z_{(n)}$ be the ordered Z's. Show that when $\alpha = 2/2^n$,

$$\theta_L = Z_{(1)} \text{ and } \theta_U = Z_{(n)}.$$

34. Consider the $1-\alpha$ upper confidence bound for θ given in (3.33) in Comment 21. If $Z_{(1)} \leq \cdots \leq Z_{(n)}$ denote the ordered Z's and $\alpha = 2/2^n$, show that

$$\theta_U^* = \frac{Z_{(n-1)} + Z_{(n)}}{2}.$$

35. Consider the $1-\alpha$ confidence interval for θ defined by display (3.27). Let $Z_{(1)} \leq \cdots \leq Z_{(n)}$ be the ordered Z's. If $\alpha = 4/2^n$, express the length $(\theta_U - \theta_L)$ of the confidence interval in terms of $Z_{(1)}, \ldots, Z_{(n)}$.

36. How does varying α affect the length of the confidence interval defined by display (3.27)? How does it affect the point estimator defined in Comment 23?

37. Consider the blood-level data of Table 3.3. Obtain an approximate 95% confidence interval for θ using the large-sample approximation of this section. Compare this approximate confidence interval with the exact 95.4% confidence interval obtained in Problem 27.

38. Consider the salary data of Table 3.2. Use the large-sample approximation of this section to obtain an approximate 90% confidence interval for θ.

39. Consider the heat-insoluble hydroxyproline data of Table 3.4. Use the large-sample approximation to obtain an approximate 99% lower confidence bound for θ. (See Comment 21.)

40. Consider the case $n = 10$ and compare the length of the exact 95.2% confidence interval for θ given by display (3.27) with the length of the approximate 95.2% confidence interval for θ obtained using the large-sample approximation of this section.

41. Consider the case $n = 15$ and compare the exact 96.8% upper confidence bound for θ given by (3.33) with the approximate 96.8% upper confidence bound for θ obtained from the large-sample approximation in Comment 21.

42. Use (3.26) and (3.27) to show that, for a fixed value of n, as α decreases the width of the confidence interval increases. Explain this trade-off.

PAIRED REPLICATES ANALYSES BY WAY OF SIGNS

Data. We obtain $2n$ observations, two observations on each of the n subjects (blocks, patients, etc.)

Subject i	X_i	Y_i
1	X_1	Y_1
2	X_2	Y_2
$\vdots$	$\vdots$	$\vdots$
n	X_n	Y_n

Assumptions

B1. We let $Z_i = Y_i - X_i$, for $i = 1, \ldots, n$. The differences $Z_1, \ldots, Z_n$ are mutually independent.

B2. Each $Z_i, i = 1, \ldots, n$, comes from a continuous population (not necessarily the same) that has a common median θ. If F_i represents the distribution function for $Z_i, i = 1, \ldots, n$, this assumption requires that

$$F_i(\theta) = P(Z_i \leq \theta) = P(Z_i > \theta) = 1 - F_i(\theta), \text{ for } i = 1, \ldots, n. \tag{3.36}$$

The parameter θ is referred to as the *unknown treatment effect*.

3.4 A DISTRIBUTION-FREE SIGN TEST (FISHER)

Hypothesis

The null hypothesis of interest here is that of zero shift in location due to the treatment, namely,

$$H_0 : \theta = 0. \tag{3.37}$$

This null hypothesis asserts that each of the distributions (not necessarily the same) for the differences (posttreatment minus pretreatment observations) has median 0, corresponding to no shift in location due to the treatment.

Procedure

To compute the sign statistic B, define indicator variables $\psi_i, i = 1, \ldots, n$, where

$$\psi_i = \begin{cases} 1, & \text{if } Z_i > 0 \\ 0, & \text{if } Z_i < 0, \end{cases} \tag{3.38}$$

and set

$$B = \sum_{i=1}^{n} \psi_i. \tag{3.39}$$

The sign statistic B is the number of positive Z's.

a. *One-Sided Upper-Tail Test.* To test

$$H_0 : \theta = 0$$

versus

$$H_1 : \theta > 0,$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } B \geq b_{\alpha,1/2}; \text{ otherwise do not reject,} \tag{3.40}$$

where the constant $b_{\alpha,1/2}$ is chosen to make the type I error probability equal to α and is the upper αth percentile point for the binomial distribution with sample size n and $p = \frac{1}{2}$. Values of $b_{\alpha,1/2}$ are found with the `R` command `qbinom`.

b. *One-Sided Lower-Tail Test.* To test

$$H_0 : \theta = 0$$

versus

$$H_2 : \theta < 0,$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } B \leq n - b_{\alpha,1/2}; \text{ otherwise do not reject.} \tag{3.41}$$

c. *Two-Sided Test.* To test

$$H_0 : \theta = 0$$

versus

$$H_3 : \theta \neq 0,$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } B \geq b_{\alpha/2,1/2} \text{ or } B \leq n - b_{\alpha/2,1/2}; \text{ otherwise do not reject.} \tag{3.42}$$

This two-sided procedure is the two-sided symmetric test with $\alpha/2$ probability in each tail of the null distribution of B.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of B, suitably standardized. As the distribution of B under the null hypothesis H_0: $\theta = 0$ is binomial with parameters n and $p = \frac{1}{2}$, we know that

$$E_0(B) = \frac{n}{2} \tag{3.43}$$

and

$$\text{var}_0(B) = \frac{n}{4}. \tag{3.44}$$

The standardized version of B is then

$$B^* = \frac{B - E_0(B)}{\{\text{var}_0(B)\}^{1/2}} = \frac{B - (n/2)}{\{n/4\}^{1/2}}. \tag{3.45}$$

When H_0 is true, B^* has, as n tends to infinity, an asymptotic $N(0, 1)$ distribution. (See Comment 32 for indications of the proof.) The normal theory approximation for procedure (3.40) is

$$\text{Reject } H_0 \text{ if } B^* \geq z_\alpha; \text{ otherwise do not reject,} \tag{3.46}$$

the normal theory approximation for procedure (3.41) is

$$\text{Reject } H_0 \text{ if } B^* \leq -z_\alpha; \text{ otherwise do not reject,} \tag{3.47}$$

and the normal theory approximation for procedure (3.42) is

$$\text{Reject } H_0 \text{ if } |B^*| \geq z_{\alpha/2}; \text{ otherwise do not reject.} \tag{3.48}$$

Ties

If there are zero values among the Z's, discard the zero values and redefine n to be the number of nonzero Z's.

EXAMPLE 3.5 ***Beak-Clapping Counts.***

The data in Table 3.5 are a subset of the data obtained by Oppenheim (1968) in an experiment investigating light responsivity in chick embryos. The subjects were white leghorn chick embryos, and the behavioral response measured in the investigation was beak-clapping (i.e., the rapid opening and closing of the beak that occurs during the latter one-third of incubation in chick embryos). (Gottlieb (1965) had previously shown that changes in the rate of beak-clapping constituted a sensitive indicator of auditory responsiveness in chick embryos.) The embryos were placed in a dark chamber 30 min before the initiation of testing. Then ten 1-min readings were taken in the dark, and at the end of this 10-min period, a single reading was obtained for a 1-min period of illumination. Table 3.5 gives the average number of claps per minute during the dark period (X) and the corresponding rate during the period of illumination (Y) for 25 chick embryos.

Table 3.5 Beak-Clapping Counts per Minute

Embryo i	X_i (Dark period)	Y_i (Illumination)	$Z_i = Y_i - X_i$	ψ_i
1	5.8	5	−0.8	0
2	13.5	21	7.5	1
3	26.1	73	46.9	1
4	7.4	25	17.6	1
5	7.6	3	−4.6	0
6	23.0	77	54.0	1
7	10.7	59	48.3	1
8	9.1	13	3.9	1
9	19.3	36	16.7	1
10	26.3	46	19.7	1
11	17.5	9	−8.5	0
12	17.9	25	7.1	1
13	18.3	59	40.7	1
14	14.2	38	23.8	1
15	55.2	70	14.8	1
16	15.4	36	20.6	1
17	30.0	55	25.0	1
18	21.3	46	24.7	1
19	26.8	25	−1.8	0
20	8.1	30	21.9	1
21	24.3	29	4.7	1
22	21.3	46	24.7	1
23	18.2	71	52.8	1
24	22.5	31	8.5	1
25	31.1	33	1.9	1

Source: R. W. Oppenheim (1968).

As responsivity of a chick embryo to a light stimulus is expected to correspond to positive Z differences, we apply procedure (3.40), which is designed to detect the alternative $\theta > 0$. To implement the sign test, one may use `qbinom` directly in procedure (3.40). If setting $\alpha = .05$, the appropriate command is

```
qbinom(p=0.05, size=25, prob=1/2, lower.tail=F)
```

where the argument `p` is α and `size` and `prob` are n and p, respectively, in the binomial distribution. The resulting value is 17. Recall that the argument `lower.tail=F` provides probabilities that are strictly greater than a specified value. Therefore, to be consistent with (3.40), one must use the value $b_{\alpha,1/2} = 18$. Procedure (3.40) is then given by

$$\text{Reject } H_0 \text{ if } B \geq 18.$$

Note that the critical value 18 given by `R` results in a significance level of $\alpha = .022$, not .05. Now, the sample value of B can be obtained directly from the indicator variables $\psi_1, \ldots, \psi_{25}$ listed in Table 3.5.

We find that $B = \sum_{i=1}^{25} \psi_i$= (number of positive Z's) = 21. As this value of B is greater than the critical value 18, we reject H_0 in favor of $\theta > 0$ at the $\alpha = .05$ level.

(We note that the actual magnitudes of the Z differences are not needed to calculate B. We require only the information as to whether or not Y_i is larger than X_i, for $i = 1, \ldots, n$, and this information is contained entirely in the indicator variables $\psi_1, \ldots, \psi_n$. However, the actual magnitude of the Z_i's will be necessary in Sections 3.5 and 3.6

to obtain point and interval estimates, respectively, of θ associated with the sign test.) It is simpler to use the `R` command `SIGN.test` from package `BSDA` (Arnholt, 2012). Running

```
SIGN.test(y, x, alt=''greater''),
```

where `y` is the vector of illumination data and `x` is the vector of dark period data from Table 3.5, results in a test statistic of $B = 21$ (the output actually uses S as the name of the test statistic rather than B) and a P-value of .0005. The P-value may also be found using the `pbinom` command. Partial output from `SIGN.test` is shown below:

```
Dependent-samples Sign-Test

data: y and x
S = 21, p-value =.0004553
alternative hypothesis: true median difference is greater
   than 0
95 percent confidence interval:
7.4519 Inf
sample estimates:
median of x-y
17.6
```

For the large-sample approximation, we find from (3.45) that

$$B^* = \frac{21 - \left(\frac{25}{2}\right)}{\left(\frac{25}{4}\right)^{1/2}} = 3.40.$$

Thus, the smallest significance level at which we can reject H_0 in favor of $\theta > 0$ using the normal approximation (i.e., the approximate P-value) is .0003. Clearly, both the exact test and the large-sample approximation indicate that there is strong evidence that chick embryos are indeed responsive to a light stimulus, as measured by an increase in the frequency of beak-claps.

Comments

25. *Motivation for the Test.* When θ is greater than 0, there will tend to be a large number of positive Z differences, leading to a big value of B. This suggests rejecting H_0 in favor of $\theta > 0$ for large values of B and motivates the procedures (3.40) and (3.46). Similar rationales lead to procedures (3.41), (3.42), (3.47), and (3.48).
26. *Assumptions.* Assumption B2 is implied by Assumption A2, but the converse is not true. Thus, Assumption B2 is less stringent than Assumption A2—an advantage of the sign test over the signed rank test. We can, when testing $\theta = 0$, weaken Assumption B2 further to Assumption B2', namely, $P(Z_i < 0) = P(Z_i > 0) = \frac{1}{2}, i = 1, \ldots, n$, when θ is the hypothesized value 0. When testing $\theta = \theta_0$ (see Comment 28), for $\theta_0 \neq 0$, Assumption B2 can be replaced by the weaker Assumption B_2'', namely, $P(Z_i < \theta_0) = P(Z_i > \theta_0) = \frac{1}{2}, i = 1, \ldots, n$, when θ is the hypothesized value θ_0.

    We also note that there is no requirement that the individual X_i and Y_i be independent, only that the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, and therefore the

resulting differences $Z_1, \ldots, Z_n$, be mutually independent. Indeed, in most applications, the individual X_i and Y_i are dependent.

27. *Binomial Test.* The test procedures based on the sign statistic B are actually special cases of the general binomial test procedures considered in Chapter 2. The sign test procedures are simply binomial procedures, with "success" corresponding to a positive Z difference, "failure" corresponding to a negative Z difference, and $p = P(\text{"success"}) = P(Z_i > 0)$ assuming the value $p_0 = \frac{1}{2}$ when the null hypothesis $H_0 : \theta = 0$ is true.

28. *Testing θ Equal to Some Specified Nonzero Value.* Procedures (3.40), (3.41), and (3.42) and the corresponding normal approximations in (3.46), (3.47), and (3.48) are for testing θ equal to zero. To test H_0: $\theta = \theta_0$, where θ_0 is some specified nonzero number, subtract θ_0 from each of the differences $Z_1, \ldots, Z_n$ to form a modified sample $Z_1' = Z_1 - \theta_0, \ldots, Z_n' = Z_n - \theta_0$. Then compute B as the number of these Z_i's that are positive. Procedures (3.40), (3.41), and (3.42) and their corresponding large-sample approximations in (3.46), (3.47), and (3.48) are then applied as previously described.

29. *Equivalent Form.* The statistic B (3.39) is the number of positive Z differences. If we define B^- to be the number of negative Z differences, then $B^- = \sum_{i=1}^{n}(1 - \psi_i) = n - \sum_{i=1}^{n} \psi_i = n - B$. Thus, the test procedures (3.40), (3.41), and (3.42) and (3.46), (3.47), and (3.48) could equivalently be based on $B^- = (n - B)$. (We point out that B^- also (as does B) has a binomial distribution with sample size n and $p = \frac{1}{2}$ when $H_0 : \theta = 0$ is true.)

30. *Derivation of the Distribution of B under H_0 (When There Are No Zero Z Values).* The null (H_0) distribution of B can be obtained directly from the representation $B = \sum_{i=1}^{n} \psi_i$. Under the assumption that the underlying Z_i distributions are all continuous, the probabilities are zero that any of the Z_i's are zero. Hence, under H_0, each of the 2^n possible outcomes for the configuration $(\psi_1, \ldots, \psi_n)$ occurs with the probability $\left(\frac{1}{2}\right)^n$. For example, in the case of $n = 3$, the $2^3 = 8$ possible outcomes for (ψ_1, ψ_2, ψ_3) and the associated values of B are given in the following table.

(ψ_1, ψ_2, ψ_3)	Probability under H_0	$B = \sum_{i=1}^{3} \psi_i$
(0,0,0)	$\frac{1}{8}$	0
(0,0,1)	$\frac{1}{8}$	1
(0,1,0)	$\frac{1}{8}$	1
(1,0,0)	$\frac{1}{8}$	1
(0,1,1)	$\frac{1}{8}$	2
(1,0,1)	$\frac{1}{8}$	2
(1,1,0)	$\frac{1}{8}$	2
(1,1,1)	$\frac{1}{8}$	3

Thus, for example, the probability is $\frac{3}{8}$ under H_0 that B is equal to 2, as $B = 2$ when any of the three exclusive outcomes $(\psi_1, \psi_2, \psi_3) = (0, 1, 1)$, $(1, 0, 1)$, or $(1, 0, 1)$ occurs, and each of these outcomes has null probability $\frac{1}{8}$. Simplifying, we obtain the null distribution.

Possible value of B	Probability under H_0
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

The probability, under H_0, that B is greater than or equal to 2, for example, is therefore

$$P_0(B \geq 2) = P_0(B = 2) + P_0(B = 3)$$
$$= .375 + .125 = .50.$$

(We note that this null distribution of B could alternatively be obtained from the binomial probability distribution in (2.15) in Comment 2.7 by taking $P_0 = \frac{1}{2}$.)

Note that we have derived the null distribution of B without specifying the forms of the underlying Z populations under H_0 beyond the requirement that they be continuous and have common median 0. This is why the test procedures based on B are called *distribution-free procedures*. From the null distribution of B we can determine the critical value $b_{\alpha/2,1/2}$ and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific forms of the underlying continuous Z distributions with common median 0.

31. *Calculation of the Mean and Variance of B under the Null Hypothesis H_0.* In displays (3.43) and (3.44), we presented formulas for the mean and variance of B when the null hypothesis is true. In this comment, we illustrate a direct calculation of $E_0(B)$ and $\text{var}_0(B)$ in the particular case of $n = 3$, using the null distribution of B obtained in Comment 30. (Later, in Comment 32, we present the general derivations of $E_0(B)$ and $\text{var}_0(B)$.) The null mean, $E_0(B)$, is obtained by multiplying each possible value of B with its probability under H_0. Thus,

$$E_0(B) = 0(.125) + 1(.375) + 2(.375) + 3(.125) = 1.5.$$

This is in agreement with what we obtain using (3.43), namely,

$$E_0(B) = \frac{n}{2} = \frac{3}{2} = 1.5.$$

A check on the expression for $\text{var}_0(B)$ is also easily performed using the well-known fact that

$$\text{var}_0(B) = E_0(B^2) - \{E_0(B)\}^2.$$

The value of $E_0(B^2)$, the second moment of the null distribution of B, is again obtained by multiplying possible values (in this case, value of B^2) by the corresponding probabilities under H_0. We find

$$E_0(B^2) = 0(.125) + 1(.375) + 4(.375) + 9(.125) = 3.0.$$

Thus,

$$\text{var}_0(B) = 3.0 - (1.5)^2 = 0.75,$$

which agrees with what we obtain using (3.44) directly, namely,

$$\text{var}_0(B) = \frac{n}{4} = \frac{3}{4} = 0.75.$$

32. *Large-Sample Approximation.* Under Assumption B1, the variables $Z_1, \ldots, Z_n$ are mutually independent. The variable ψ_i is a function of Z_i only, for $i = 1, \ldots, n$, therefore $\psi_1, \ldots, \psi_n$ are also mutually independent variables. In view of the representation $B = \sum_{i=1}^{n} \psi_1$ in (3.39), we can immediately use well-known expressions for the mean and variance of a sum of mutually independent random variables to obtain

$$E_0(B) = E_0\left[\sum_{i=1}^{n} \psi_i\right] = \sum_{i=1}^{n} E_0(\psi_i) \tag{3.49}$$

and

$$\text{var}_0(B) = \text{var}_0\left(\sum_{i=1}^{n} \psi_i\right) = \sum_{i=1}^{n} \text{var}_0(\psi_i). \tag{3.50}$$

Now, under H_0, the ψ_i's are also identically distributed, each following the Bernoulli probability distribution with $p = \frac{1}{2}$. Thus, for $i = 1, \ldots, n$, we see that

$$E_0(\psi_i) = 0\left(\frac{1}{2}\right) + 1\left(\frac{1}{2}\right) = \frac{1}{2}$$

and

$$\begin{aligned}\text{var}_0(\psi_i) &= E_0(\psi_i^2) - [E_0(\psi_i)]^2 \\ &= 0^2\left(\frac{1}{2}\right) + 1^2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.\end{aligned}$$

Using these results in (3.49) and (3.50), we obtain

$$E_0(B) = \sum_{i=1}^{n}\left(\frac{1}{2}\right) = \frac{n}{2}$$

and

$$\text{var}_0(B) = \sum_{i=1}^{n}\left(\frac{1}{4}\right) = \frac{n}{4},$$

which agree with the general expressions stated in (3.43) and (3.44).

The asymptotic normality of the standardized form

$$B^* = \frac{B - E_0(B)}{\{\text{var}_0(B)\}^{1/2}} = \frac{B - \frac{n}{2}}{\left\{\frac{n}{4}\right\}^{1/2}}$$

follows from standard central limit theory for sums of mutually independent, identically distributed random variables (cf. Randles and Wolfe (1979, p. 421)). Asymptotic normality results are also obtainable under general alternatives to H_0. (See Comment 35.)

33. *Symmetry of the Distribution of B under the Null Hypothesis.* When H_0 is true, the distribution of B is symmetric about its mean $n/2$. (See Comment 30 for verification of this when $n = 3$.) This implies that

$$P_0(B \leq x) = P_0(B \geq n - x), \tag{3.51}$$

for $x = 0, 1, \ldots, n$.

Equation (3.51) is used directly to convert upper-tail probabilities to lower-tail probabilities.

34. *Zero Z Values.* We have recommended discarding zero Z values and redefining n to be the number of nonzero Z's. This approach is satisfactory as long as the zero values do not represent a sizable percentage of the total number of Z differences. If, however, there is a relatively large number of zero Z's, it would be advisable to consider an appropriate statistical procedure designed specifically for analyzing such discrete data. See, for example, Chapter 10 or a book on categorical data analysis such as Agresti (2013).

We should also point out that there are methods other than elimination that have been proposed for dealing with zero Z values. One could use individual randomization (e.g., flipping a fair coin) to decide whether each of the zero Z values is to be counted as positive or negative in the computation of B. (Although this approach maintains many of the nice theoretical properties of B that hold when there are no zeros, it introduces extraneous randomness that could quite easily have a direct effect on the outcome of any subsequent inferences based on such a modified B.) A second alternative approach in the case of the one-sided test procedures in (3.40), (3.41), (3.46), and (3.47) is to be conservative about rejecting the null hypothesis H_0; that is, we could count all the zero Z values as if they were in favor of not rejecting H_0. Thus, for example, in applying either procedure (3.41) or (3.47) to test H_0 against the alternative $\theta < 0$, we would treat all the zero Z's as if they were "positive" (in favor of not rejecting H_0) in the calculation of B. (In the case of procedures (3.40) and (3.46), zero Z values would be considered "negative" in the calculation of B.) Any rejection of H_0 with this conservative approach to dealing with zero Z values could be viewed as providing strong evidence in favor of the appropriate alternative. For a more detailed discussion of methods for handling zero observations, see Pratt (1959).

35. *Some Power Results for the Sign Test.* We consider the upper-tail α-level test of $H_0: \theta = 0$ versus $H_1 : \theta > 0$ given by procedure (3.40). When we have a common underlying distribution with median θ and distribution function $F_1 \equiv F_2 \equiv \cdots \equiv F_n \equiv F$ for the Z differences, the sign statistic B has a binomial distribution with parameters n and $p_\theta = P_\theta(Z_1 > 0) = 1 - F(0)$. It follows (see (2.15) in Comment 2.7) that the exact power of the sign test procedure (3.40) against the alternative $\theta > 0$ is given by the expression

$$\text{Power}_\theta = \sum_{t=b_{\alpha,1/2}}^{n} \binom{n}{t} p_\theta^t (1 - p_\theta)^{n-t} = \sum_{t=b_{\alpha,1/2}}^{n} \binom{n}{t} [1 - F(0)]^t [F(0)]^{n-t}. \tag{3.52}$$

(As $p_\theta = 1 - F(0) > 1 - F(\theta) = 1 - \frac{1}{2} = \frac{1}{2}$ for all $\theta > 0$, it follows that $\text{Power}_\theta > \text{Power}_0 = \alpha$ for all alternatives $\theta > 0$.) Evaluation of Power_θ for a moderate sample size n and particular value of $\theta > 0$ (and associated $p_\theta = 1 - F(0) > \frac{1}{2}$) can thus be accomplished by direct computation.

For large sample sizes, we can make use of the standard central limit theorem for sums of mutually independent and identically distributed random variables to conclude that

$$\frac{B - np_\theta}{[np_\theta(1-p_\theta)]^{1/2}} = \frac{B - n(1-F(0))}{[n(1-F(0))(F(0))]^{1/2}} \tag{3.53}$$

has an asymptotic ($n \longrightarrow \infty$) standard normal distribution. Thus, for large n, we can approximate the exact power in (3.52) by

$$\begin{aligned}\text{Power}_\theta &\approx 1 - \Phi\left(\frac{b_{\alpha,1/2} - np_\theta}{[np_\theta(1-p_\theta)]^{1/2}}\right) \\ &= 1 - \Phi\left(\frac{b_{\alpha,1/2} - n(1-F(0))}{[n(1-F(0))(F(0))]^{1/2}}\right),\end{aligned} \tag{3.54}$$

where $\Phi(t)$ is the area under a standard normal density to the left of t.

We note that both the exact power (3.52) and the approximate power (3.54) against an alternative $\theta > 0$ depend on the common distribution only through the value of its distribution function $F(z)$ at $z = 0$. Thus, if two distributions have a common median $\theta > 0$ and distribution functions F_1 and F_2 such that $F_1(0) = F_2(0)$, then the exact power (3.52) of the sign test against the alternative $\theta > 0$ will be the same for both distributions F_1 and F_2. (The same is, of course, true for the approximate power in (3.54).)

For the purpose of illustration, consider the case where $n = 10$ and $\alpha = .05$. Then using `qbinom`, we see that $b_{.05,1/2} = 9$ and the test (3.40) rejects H_0 if and only if $B \geq 9$. If F is the distribution function for a probability distribution with median $\theta = 2$ (i.e., $F(2) = \frac{1}{2}$) and $F(0) = \frac{1}{4}$, then from (3.52), the exact power of the sign test in this setting is

$$\begin{aligned}\text{Power}_{\theta=2} &= \sum_{t=9}^{10} \binom{10}{t}\left(\frac{3}{4}\right)^t\left(\frac{1}{4}\right)^{10-t} \\ &= \texttt{pbinom(8,size = 10, prob = 3/4, lower.tail = F)} \\ &= .2440.\end{aligned}$$

The approximate power for the same setting is seen from (3.54) to be

$$\begin{aligned}\text{Power}_{\theta=2} &\approx 1 - \Phi\left(\frac{9 - 10\left(\frac{3}{4}\right)}{\left[10\left(\frac{3}{4}\right)\left(\frac{1}{4}\right)\right]^{1/2}}\right) \\ &= 1 - \Phi\left(\frac{1.5}{\left[\frac{30}{16}\right]^{1/2}}\right) = .1367.\end{aligned}$$

Thus, for n as small as 10 and $p_\theta = \frac{3}{4}$, the agreement between the exact power (3.52) and the approximate power (3.54) is not too good. (We note

that for an underlying normal distribution F with mean $\theta = 2$ and variance σ^2, the condition $F(0) = \frac{1}{4}$ corresponds to $\Phi((0-2)/\sigma) = \frac{1}{4}$, which in turn corresponds to $(-2/\sigma) = z_{.75} = -.675$, or $\sigma = (2/.675) = 2.96$. More generally, the test (3.40) for $n = 10$ and $\alpha = .05$ has an exact power of .2440 against any normal distribution with mean θ and variance σ^2 for which $-(\theta/\sigma) = z_{.75} = -6.75$, or $\theta = .675\sigma$. Although we specified $\alpha = .05$ in this example, the upper tail probability for $B \geq 9$ is actually .01.

36. *Sample Size Determination.* When $Z_1, \ldots, Z_n$ are a random sample from a single continuous F, the one-sided upper-tail test defined by procedure (3.40) is consistent (i.e., has power tending to 1 as n tends to infinity) against those F populations for which $p > \frac{1}{2}$, with

$$p = P(Z > 0), \tag{3.55}$$

where Z is distributed as F.

Noether (1987), among others, shows how to determine an approximate sample size n so that the α-level one-sided test given by procedure (3.40) will have approximate power $1 - \beta$ against an alternative value of p (3.55) greater than $\frac{1}{2}$. This approximate value of n is

$$n \doteq \frac{(z_\alpha + z_\beta)^2}{4\left(p - \frac{1}{2}\right)^2}. \tag{3.56}$$

As an illustration of the use of (3.56), suppose we are testing H_0 and we wish to have an upper-tail level $\alpha = .04$ test with power $1 - \beta$ of at least .975 against an alternative for which $p = P(Z > 0) = .7$ (recall that $p = .5$ under H_0). The critical values are $z_\alpha = z_{.04} = 1.75$ and $z_\beta = Z_{.025} - 1.96$, and we find that the required sample size for the alternative $p = .7$ is

$$n \doteq \frac{(1.75 + 1.96)^2}{4(.7 - .5)^2} = 86.03.$$

To be conservative, we take $n = 87$.

37. *Consistency of the B Test.* Under the assumption that $Z_1, \ldots, Z_n$ are a random sample from a single continuous population F, the consistency of the tests based on B depends on the parameter

$$p^* = P(Z_1 > 0) - \frac{1}{2}. \tag{3.57}$$

The test procedures defined by (3.40), (3.41), and (3.42) are consistent against the classes of alternatives corresponding to $p^* >, <$, and $\neq 0$, respectively.

Properties

1. *Consistency.* For our consistency statement, we strengthen Assumption B2 to require that each Z has the same continuous population with median θ. Then the test procedures defined by (3.40), (3.41), and (3.42) are consistent against the alternatives $\theta >, <$, and $\neq 0$, respectively. (See also Comment 37.)

2. *Asymptotic Normality.* Under a strengthened Assumption B2 that requires that each Z has the same continuous population with median θ, the asymptotic normality of the standardized form of the B statistic follows from the standard central limit theorem for sums of mutually independent and identically distributed random variables. (See also Comment 35.)
3. *Efficiency.* See Section 3.11.

Problems

43. The data in Table 3.6 are a portion of the data obtained by Cooper et al. (1967). The purpose of their investigation was to determine whether hypnotic susceptibility as measured on objective scales can be changed with practice and training. The objective measures used were the Stanford Profile Scales of Hypnotic Susceptibility, forms I and II (Hilgard, Lauer, and Morgan (1963)). The subjects were administered these Profile Scales, both forms I and II, by a hypnotist other than the experimenter. Each subject was then seen by one of the authors for an extensive period of "hypnotic training." After these sessions were concluded, each subject was retested by a different hypnotist (again not the experimenter) using equivalent forms of the Profile Scales, forms I′ and II′. Table 3.6 gives the average score obtained on forms I and II prior to hypnotic training (X) and the corresponding average score obtained on forms I′ and II′ after the training (Y) for the six subjects. Note that a high (or low) score on the Profile Scales indicates a high (or low) degree of hypnotic susceptibility.

Test the hypothesis of no change in hypnotic susceptibility versus the alternative that hypnotic susceptibility (as measured by the Profile Scales) can be increased with practice and training.

44. Change the value of Y_3 in Table 3.5 from 73 to 173. What effect does this outlying observation have on the calculations performed in Example 3.5? What does this suggest about the relative insensitivity of the sign tests to outliers? Construct an example in which changing one observation has an effect on the final decision regarding rejection or acceptance of H_0.

45. Suppose $n = 25$. Compare the exact P-value of test of $H_0 : \theta = 0$ versus $H_1 : \theta < 0$ based on $B = 8$, with the P-value found using the large-sample approximation.

46. In an investigation to determine the effect of aspirin on bleeding time and platelet adhesion, Bick, Adams, and Schmalhorst (1976) studied the reactions of normal subjects to aspirin. A subset of their data is presented in Table 3.7, where the X observation for each subject is the bleeding time (in seconds) before ingestion of 600 mg of aspirin and the Y observation is the bleeding time (again in seconds) 2 h after administration of the aspirin.

Table 3.6 Average Scores on the Stanford Profile Scales of Hypnotic Susceptibility

Subject i	X_i	Y_i
1	10.5	18.5
2	19.5	24.5
3	7.5	11.0
4	4.0	2.5
5	4.5	5.5
6	2.0	3.5

Source: L. M. Cooper, E. Schubot, S. A. Banford, and C. T. Tart (1967).

Perform the appropriate test of the hypothesis that a 600-mg dose of aspirin has no effect on bleeding time versus the alternative that it typically leads to an increase in bleeding time.

47. Assume that we have a common underlying distribution $F_1 \equiv F_2 \equiv \cdots \equiv F_n \equiv F$ (in Assumption B2). If we have 20 observations, what is the exact power of the level $\alpha = .05$ test of $H_0 : \theta = 0$ versus the alternative $\theta > 0$ when $F(0) = .3$?

48. Assume that we have a common underlying distribution $F_1 \equiv F_2 \equiv \cdots \equiv F_n \equiv F$ (in Assumption B2). If we have 18 observations and F is normal with variance 4, what is the exact power of the level $\alpha = .01$ test of $H_0 : \theta = 0$ versus the alternative $\theta < 0$ when the treatment effect is $\theta = -2$?

49. Consider a level $\alpha = .025$ test of $H_0 : \theta = 0$ versus the alternative $\theta > 0$ based on B. If our data $Z_1, \ldots, Z_n$ are a random sample from a single, continuous distribution $F(\cdot)$, how many n observations will we need to collect in order to have an approximate power of at least .75 against an alternative for which $F(0) = .20$?

50. Apply the appropriate form of the test based on B to the Hamilton depression scale factor IV data in Table 3.1.

51. Assume that we have a common underlying distribution $F_1 \equiv F_2 \equiv \cdots \equiv F_n \equiv F$ (in Assumption B2). If we have 20 observations, what is the approximate power of the level $\alpha = .05$ test of $H_0 : \theta = 0$ versus the alternative $\theta > 0$ when $F(0) = .3$? Compare this approximate power with the exact power from Problem 47.

52. Apply the large-sample approximation test of $H_0 : \theta = 1000$ versus $H_1 : \theta > 1000$ based on B to the salary data in Table 3.2.

53. For the case of $n = 5$ nonzero Z values, use the approach discussed in Comment 30 to obtain the form of the exact null (H_0) distribution of B. Verify numerically that this null distribution is, indeed, the binomial distribution with parameters $n = 5$ and $p_0 = .5$.

54. Consider the test of $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ based on B for the following $n = 15$ Z observations: $Z_1 = 2.5, Z_2 = 0, Z_3 = 3.7, Z_4 = -0.6, Z_5 = 1.7, Z_6 = 0, Z_7 = 5.9, Z_8 = 4.6, Z_9 = 0, Z_{10} = -1.4, Z_{11} = 5.4, Z_{12} = 4.6, Z_{13} = 3.1, Z_{14} = -2.0,$ and $Z_{15} = 6.3$. Compute the P-values for the competing B procedures based on either (i) discarding the zero Z values and reducing n accordingly, as recommended in the Ties portion of this section, or (ii) treating

Table 3.7 Bleeding Time (in seconds)

Subject i	X_i	Y_i
1	270	525
2	150	570
3	270	190
4	420	395
5	202	370
6	255	210
7	165	490
8	220	250
9	305	360
10	210	285
11	240	630
12	300	385
13	300	195
14	70	295

Source: R. L. Bick, T. Adams, and W. R. Schmalhorst (1976).

the zero Z values in a conservative manner, as presented in Comment 34. Discuss the results.

55. Consider the same setting as in Problem 54. Suppose that you had decided to use randomization to deal with the three zero Z values in the data (see Comment 34). Consider the various possible outcomes for this randomization process and compute the associated P-value for each of these outcomes. Discuss the implication of these findings in conjunction with the results of Problem 54.

56. Obtain the exact P-value for the test based on B for the bleeding time data in Table 3.7. Compare this to the P-value obtained using the large sample approximation.

57. Obtain the exact P-value for the test of $H_0 : \theta = 1000$ versus $H_1 : \theta > 1000$ based on B for the salary data in Table 3.2.

3.5 AN ESTIMATOR ASSOCIATED WITH THE SIGN STATISTIC (HODGES–LEHMANN)

Procedure

To estimate the treatment effect θ, order the sample observations and let $Z^{(1)} \leq \cdots \leq Z^{(n)}$ denote these ordered items. The estimator of θ associated with the sign statistic (see Comment 38) is

$$\tilde{\theta} = \text{median}\{Z_i,\ 1 \leq i \leq n\}. \tag{3.58}$$

Thus, if n is odd, say $n = 2k + 1$, we have $k = (n - 1)/2$ and

$$\tilde{\theta} = Z^{(k+1)}, \tag{3.59}$$

the value that occupies position $k + 1$ in the list of the ordered Z_i values. If n is even, say $n = 2k$, then $k = n/2$ and

$$\tilde{\theta} = \frac{Z^{(k)} + Z^{(k+1)}}{2}; \tag{3.60}$$

that is, when n is even, $\tilde{\theta}$ is the average of the two Z_i values that occupy positions k and $k + 1$ in the ordered list of the n data values.

EXAMPLE 3.6 ***Continuation of Example 3.5.***

To estimate θ for the beak-clapping data in Table 3.5, we first form the $n = 25$ ordered Z values, namely, $Z^{(1)} \leq \cdots \leq Z^{(25)}$: $-8.5, -4.6, -1.8, -0.8$, 1.9, 3.9, 4.7, 7.1, 7.5, 8.5, 14.8, 16.7, 17.6, 19.7, 20.6, 21.9, 23.8, 24.7, 24.7, 25.0, 40.7, 46.9, 48.3, 52.8, and 54.0. The sample size $n = 25$ is odd, so we use (3.59) with $k = (25 - 1)/2 = 12$ to obtain the estimate $\tilde{\theta} = Z^{(13)} = 17.6$ for the treatment effect θ. Thus, we estimate that a typical chick embryo of the type included in this study will produce 17.6 more beak-claps per minute during periods of illumination than during periods of darkness.

The `SIGN.test` command will provide this value automatically as seen in the `R` output in Example 3.5. Alternatively, one may use the command `median(z)` directly on the difference data $Z_i = Y_i - X_i$.

Comments

38. *Motivation for the Hodges–Lehmann Estimator.* The estimator $\tilde{\theta}$ defined by (3.58) is associated with the sign test in the same way as the estimator $\tilde{\theta}$ (3.23) is associated with the signed rank test (see Comment 15). When $\theta = 0$, the distribution of the statistic B (3.39) is symmetric about its mean $n/2$ (see Comment 33). A natural estimator of θ is the amount $\tilde{\theta}$ (say) that should be subtracted from each Z_i so that the value of B, when applied to the shifted sample $Z_1 - \tilde{\theta}, \ldots, Z_n - \tilde{\theta}$, is as close to $n/2$ as possible. Intuitively, we estimate θ by the amount ($\tilde{\theta}$) that the Z sample should be shifted in order that $Z_1 - \tilde{\theta}, \ldots, Z_n - \tilde{\theta}$ appears (when "viewed" by the sign statistic B) as a sample from a population with median 0. (Under Assumptions B1 and B2, each of the $Z_1 - \theta, \ldots, Z_n - \theta$ variables is from a population with median 0.)

    The Hodges–Lehmann method can be applied to a large class of statistics containing both B and T^+ (3.3). However, the forms of the resulting estimator for other members of this class are not always as convenient for calculation as are $\tilde{\theta}$ (3.58) or $\tilde{\theta}$ (3.23). See Hodges and Lehmann (1983) for an expository article on their method of estimation.

39. *Simplicity.* One of the virtues of $\tilde{\theta}$ (3.58) is its simplicity. While many estimators associated with distribution-free test statistics are tedious to compute (e.g., $\tilde{\theta}$ (3.23) requires computing the median of $n(n+1)/2$ values), $\tilde{\theta}$ requires only that we find the median of the n Z observations. However, although the signs of the Z differences provide sufficient information to conduct a sign test, the magnitudes of these differences are needed to obtain the value of the estimator $\tilde{\theta}$.

40. *Sensitivity to Gross Errors.* The estimator $\tilde{\theta}$ (3.58) is even less sensitive to outliers than the estimator $\hat{\theta}$ (3.23) associated with the signed rank statistic T^+ (3.3). (See Comment 16 and Problems 20 and 60.) As a result, $\tilde{\theta}$ protects well against gross errors. However, all the information contained in the collected sample is not utilized in computing $\tilde{\theta}$. Consequently, $\tilde{\theta}$ is rather inefficient for many populations.

41. *Zero Z Values.* Note that in calculating the estimator $\tilde{\theta}$, we use *all* the Z differences. Although we recommend (see Ties in Section 3.4) discarding the zero Z values (and reducing n accordingly) prior to applying the sign test to the data, it is not necessary to do so when calculating $\tilde{\theta}$. In fact, the zero Z values contain important information about the magnitude of the treatment effect. This is also the case when we consider (Section 3.6) confidence intervals and bounds for θ.

42. *Historical Perspective.* The use of the estimator $\tilde{\theta}$ predates most of the recent unified developments in the field of nonparametric statistics. A. T. Craig (1932) first found the sampling distribution of $\tilde{\theta}$, and its asymptotic properties were developed shortly thereafter by Smirnov (1935).

43. *Quasimedians.* Let $Z^{(1)} \leq \cdots \leq Z^{(n)}$ be the ordered sample observations, as in step 1 of the Procedure. Hodges and Lehmann (1967) defined the sample

quasimedians by

$$\tilde{\theta}_i = \begin{cases} \dfrac{Z^{(k+1-i)} + Z^{(k+1+i)}}{2}, & \text{if } n = 2k+1 \\ \dfrac{Z^{(k-i)} + Z^{(k+1+i)}}{2}, & \text{if } n = 2k, \end{cases}$$

for $i = 0, 1, \ldots, k$ if $n = 2k + 1$, or $i = 0, 1, \ldots, k - 1$ if $n = 2k$; that is, each quasimedian $\tilde{\theta}_i$ is an average of two symmetrically situated, ordered Z observations. (Note that this definition of a quasimedian generalizes the concept of a sample median, as the sample median $\tilde{\theta}$ (3.58) is equal to $\tilde{\theta}_0$.) These quasimedians are natural estimators for the parameter θ (see Comment 52) and were considered by Hodges and Lehmann (1967), who investigated some of the asymptotic properties of this class of statistics.

44. *Linear Combinations of Order Statistics.* Let $Z^{(1)} \leq \cdots \leq Z^{(n)}$ be the ordered sample observations, as in step 1 of the Procedure. Under the additional assumption that we have a common underlying distribution $F_1 \equiv F_2 \equiv \cdots \equiv F_n \equiv F$ (in Assumption B2), the n variables $Z^{(1)}, \ldots, Z^{(n)}$ are called the *order statistics* for the random sample $Z_1, \ldots, Z_n$. The estimator $\tilde{\theta}$ (3.58) is a special case of a general class of estimators of θ based on linear combinations of these sample order statistics, corresponding to estimators of the form

$$\tilde{\theta}_{\mathbf{b}} = \sum_{i=1}^{n} b_i Z^{(i)}, \tag{3.61}$$

where $\mathbf{b} = (b_1, \ldots, b_n)$ is a vector of n nonnegative constants such that $\sum_{i=1}^{n} b_i = 1$. For a more detailed discussion about estimators of the form $\tilde{\theta}_{\mathbf{b}}$ (3.61), see, for example, David and Nagaraja (2003) or Arnold, Balakrishnan, and Nagaraja (1992).

45. *Variance Approximation.* Hodges and Lehmann (1967) obtained an approximation for the variance of the estimator $\tilde{\theta}$ (3.58) under the additional assumption that we have a common underlying distribution $F_1 \equiv F_2 \equiv \cdots \equiv F_n \equiv F$ (in Assumption B2). (See equation (1.4) of their paper.) They point out that, up to the accuracy of their approximation, it is not wise to compute the sample median $\tilde{\theta}$ using an odd number of observations, say $n = 2k + 1$. The next smaller even number, $n = 2k$, yields a sample median that is just as accurate. This conclusion does not depend on the shape of the underlying population except that it be symmetric, although the degree of accuracy of the approximation is affected by the shape.

46. *Estimating the Asymptotic Standard Deviation of* $\tilde{\theta}$. Assume that we have a common underlying distribution $F_1 \equiv F_2 \equiv \cdots \equiv F_n \equiv F$ (in Assumption B2) and set

$$D = \sum_{i=1}^{n} a_i,$$

where

$$a_i = \begin{cases} 1, & \text{if } \left[\tilde{\theta} - (n)^{-1/5}\right] \leq Z_i \leq \left[\tilde{\theta} + (n)^{-1/5}\right] \\ 0, & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, n$. Let $A =$ maximum $\{1, D\}$. Under the additional assumption on the common distribution F that the probability of obtaining a Z observation in any (sufficiently) small interval I centered at the median θ is greater than or equal to some fixed constant (not depending on I) times the length of I, the statistic $C = n^{3/10}A^{-1}$ is a consistent estimator of the asymptotic standard deviation of the point estimator $\tilde{\theta}$ (3.58). The statistic C is related to general classes of estimators of probability density functions considered by Rosenblatt (1956), Parzen (1962), and Gupta (1967). The consistency of C follows directly from the results in Korwar (1971).

47. *Relative Merits of $\hat{\theta}$ and $\tilde{\theta}$.* The point estimator $\tilde{\theta}$ (3.58) associated with the sign test statistic B is to be preferred to the point estimator $\tilde{\theta}$ (3.23) associated with the signed rank test statistic T^+ when ease of computation is a consideration (see Comment 39). Generally (but not always), $\hat{\theta}$ is more efficient than $\tilde{\theta}$. (See Comment 40 and Section 3.11.)

Properties

1. *Standard Deviation of $\tilde{\theta}$.* For the asymptotic standard deviation of $\tilde{\theta}$ (3.58), see Fisz (1963, p. 383) and Comment 46.
2. *Asymptotic Normality.* See Fisz (1963, p. 383).
3. *Efficiency.* See Hodges and Lehman (1963) and Section 3.11.

Problems

58. Using the designated X and Y associations, estimate θ for the average Profile Scales data of Table 3.6.

59. Estimate θ for the bleeding time data of Table 3.7.

60. Change the value of Y_3 in Table 3.5 from 73 to 173. What effect does this have on the value of $\bar{Z} = \sum_{i=1}^{25} Z_i/25$? What is the new value of $\tilde{\theta}$ (3.58)? Interpret these calculations. (See Comment 40.)

61. Calculate $\tilde{\theta}$ for the heat-insoluble hydroxyproline data of Table 3.4. Compare with the value of $\hat{\theta}$ obtained in Problem 21.

62. (a) What happens to $\tilde{\theta}$ when we add a number b to each of the sample values $Z_1, \ldots, Z_n$?

(b) What happens to $\tilde{\theta}$ when we multiply each sample value by the number textitd?

(c) What happens to $\tilde{\theta}$ when we discard the k largest and k smallest values from the sample (assume $n > 2k$)? Compare your answers with the corresponding answers to Problem 23.

63. Calculate $\tilde{\theta}$ for the blood-level data of Table 3.3. Compare with the value of $\hat{\theta}$ obtained in Problem 19.

64. Calculate $\tilde{\theta}$ for the salary data in Table 3.2. Compare with the value of $\hat{\theta}$ obtained in Problem 18.

65. Calculate $\tilde{\theta}$ for the Hamilton depression scale factor IV data in Table 3.1. Compare with the value of $\tilde{\theta}$ obtained in Example 3.3.

66. Find the vector $\mathbf{b} = (b_1, \ldots, b_n)$ to show that $\tilde{\theta}$ can be written as a linear combination of the sample order statistics $Z^{(1)} \leq \cdots \leq Z^{(n)}$, as discussed in Comment 44.

67. Show that the class of quasimedian estimators of θ (see Comment 43) is a subset of the class of estimators of θ based on linear combinations of the sample order statistics $Z^{(1)} \leq \cdots \leq Z^{(n)}$, as discussed in Comment 44.

68. Find the sample quasimedians (see Comment 43) for the data in Table 3.2. How do these values compare with $\tilde{\theta}$?

69. Find the sample quasimedians (see Comment 43) for the data in Table 3.7. How do these values compare with $\tilde{\theta}$?

3.6 A DISTRIBUTION-FREE CONFIDENCE INTERVAL BASED ON THE SIGN TEST (THOMPSON, SAVUR)

Procedure

For a symmetric two-sided confidence interval for θ, with confidence coefficient $1-\alpha$, first obtain the upper $(\alpha/2)$nd percentile point $b_{\alpha/2,1/2}$ of the null distribution of B from `qbinom`. Set

$$C_\alpha = n + 1 - b_{\alpha/2,1/2}. \tag{3.62}$$

The $100(1-\alpha)\%$ confidence interval $(\theta_{\mathrm{L}}, \theta_{\mathrm{U}})$ for θ that is associated with the two-sided sign test (see Comment 48) of $H_0 : \theta = 0$ is then given by

$$\theta_{\mathrm{L}} = Z^{(C_\alpha)}, \theta_{\mathrm{U}} = Z^{(n+1-C_\alpha)} = Z^{(b_{\alpha/2,1/2})}, \tag{3.63}$$

where $Z^{(1)} \leq \cdots \leq Z^{(n)}$ are the ordered sample observations; that is, θ_L is the sample observation that occupies position C_α in the list of ordered sample data. The upper end point θ_U is the sample observation that occupies position $n+1-C_\alpha = b_{\alpha/2,1/2}$ in this ordered list. With θ_L and θ_{U} given by display (3.63), we have

$$P_\theta(\theta_{\mathrm{L}} < \theta < \theta_{\mathrm{U}}) = 1 - \alpha \text{ for all } \theta. \tag{3.64}$$

(For upper or lower confidence bounds for θ associated with appropriate one-sided sign tests of H_0: $\theta = 0$, see Comment 49.)

Large-Sample Approximation

For large n, the integer C_α may be approximated by

$$C_\alpha \approx \frac{n}{2} - z_{\alpha/2}\left(\frac{n}{4}\right)^{1/2}. \tag{3.65}$$

In general, the value of the right-hand side of (3.65) is not an integer. To be conservative, take C_α to be the largest integer that is less than or equal to the right-hand side of (3.65).

EXAMPLE 3.7 ***Continuation of Examples 3.5 and 3.6.***

Consider the beak-clapping data in Table 3.5. We illustrate how to obtain the 95% confidence interval for θ. With $1-\alpha = .95$, $n = 25$, and $p = 1/2$ we see that $b_{\alpha,1/2} = b_{0.025,1/2} = 18$. From (3.62), it follows that

$$C_\alpha = 25 + 1 - 18 = 8.$$

Using C_α and $b_{\alpha,1/2}$ in (3.63), we see that

$$\theta_{\mathrm{L}} = Z^{(8)} = 7.1 \text{ and } \theta_{\mathrm{U}} = Z^{(18)} = 24.7$$

so that the 95% confidence interval for θ is

$$(\theta_{\mathrm{L}}, \theta_{\mathrm{U}}) = \left(Z^{(8)}, Z^{(18)}\right) = (7.1, 24.7).$$

The size of this confidence interval is the same as the probability a binomial random variable with parameters $n = 25$ and $p = 1/2$ is in the interval (8,18). Using `pbinom`, this is .9567. Thus, the actual confidence level is not α =.05, but $\alpha = .0433$. This is due to the discrete nature of the statistic. `SIGN.test` provides this confidence interval. The value of $1-\alpha$ is specified through the argument `conf.level`. In the output below, three confidence intervals are provided. As it is not possible to get an interval with exactly .05 in the tails, intervals bracketing this α are given. The lower achieved interval is the exact intervals with $\alpha = 1 - .8922$, the upper achieved interval is the exact interval with $\alpha = 1 - .9567$ found above. The interpolated interval is found by linearly interpolating the lower and upper end points on $1-\alpha$. Two-sided confidence intervals are provided when the alternative hypothesis is two-sided. This is done with the argument `alternative="two.sided"` in `SIGN.test`.

```
                    Conf.Level   L.E.pt     U.E.pt
Lower Achieved CI       .8922   7.5000    23.8000
Interpolated CI         .9500   7.1417    24.6063
Upper Achieved CI       .9567   7.1000    24.7000
```

Comments

48. *Relationship of Confidence Interval to Two-Sided Test.* The $100(1-\alpha)\%$ confidence interval for θ given by display (3.63) can be obtained from the two-sided sign test as follows. The confidence interval $(\theta_{\mathrm{L}}, \theta_{\mathrm{U}})$ consists of those θ_0 values for which the two-sided α-level test of $\theta = \theta_0$ (see Comment 28) does not reject the hypothesis $\theta = \theta_0$.
49. *Confidence Bounds.* Often we are interested only in making one-sided confidence statements about the parameter θ; that is, we wish to assert with specified confidence that θ is no larger (or, in other settings, no smaller) than some upper (lower) confidence bound based on the sample data. To obtain such one-sided confidence bounds for θ, we proceed as follows. For the specified confidence coefficient $1-\alpha$, find the upper αth (not $(\alpha/2)$nd, as for the confidence interval) percentile point $b_{\alpha,1/2}$ of the null distribution of B. Set

$$C_\alpha^* = n+1 - b_{\alpha,1/2}. \tag{3.66}$$

The $100(1-\alpha)\%$ lower confidence bound θ_{L}^* for θ that is associated with the one-sided sign test of $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta > \theta_0$ is then given by

$$(\theta_{\mathrm{L}}^*, \infty) = (Z^{(C_\alpha^*)}, \infty), \tag{3.67}$$

where, as before, $Z^{(1)} \le \cdots \le Z^{(n)}$ are the ordered sample observations. With θ_L^* given by display (3.67), we have

$$P_\theta(\theta_{\mathrm{L}}^* < \theta < \infty) = 1 - \alpha \text{ for all } \theta. \tag{3.68}$$

The corresponding $100(1-\alpha)\%$ *upper* confidence bound $\theta_{\mathrm{U}}^\star$ for θ that is associated with the one-sided sign test of $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta < \theta_0$ is given by

$$(-\infty, \theta_{\mathrm{U}}^*) = (-\infty, Z^{(n+1-C_\alpha^*)}) = \left(-\infty, Z^{(b_{\alpha,1/2})}\right), \tag{3.69}$$

where C_α^* is given in (3.66). It follows that

$$P_\theta(-\infty < \theta < \theta_{\mathrm{U}}^*) = 1 - \alpha \text{ for all } \theta. \tag{3.70}$$

For large n, the integer C_α^* may be approximated by

$$C_\alpha^* \approx \frac{n}{2} - z_\alpha \left(\frac{n}{4}\right)^{1/2}. \tag{3.71}$$

As with C_α (3.65) and the confidence interval for θ, the value of the right-hand side of (3.71) is not an integer. To be conservative, take C_α^* to be the largest integer that is less than or equal to the right-hand side of (3.71).

The $100(1-\alpha)\%$ lower and upper confidence bounds θ_L^* (3.67) and θ_{U}^* (3.69) are related to the acceptance regions of the one-sided sign tests of H_0: $\theta = \theta_0$ against the alternatives $\theta > \theta_0$ and $\theta < \theta_0$, respectively, in the same way that the confidence interval $(\theta_{\mathrm{L}}, \theta_{\mathrm{U}})$ is related to the acceptance region of the two-sided sign test of H_0: $\theta = \theta_0$ (see Comment 48). When using `SIGN.test`, one-sided confidence intervals are produced by specifying a one-sided alternative.

50. *Zero Z Values.* Note that in calculating the confidence interval $(\theta_{\mathrm{L}}, \theta_{\mathrm{U}})$ from display (3.63) or the confidence bounds θ_{L}^* (3.67) or $\theta_{\mathrm{U}}^\star$ (3.69) for θ, we use *all* the Z differences. This is in common with our recommendation (see Comment 41) for computing the point estimator $\tilde{\theta}$ (3.58), but differs from the recommended policy (see Ties in Section 3.4) of discarding the zero Z values (and reducing n accordingly) prior to applying the sign test to the data. However, if there are zero Z's in the data, the equivalence (discussed in Comments 48 and 49) between the acceptance regions of the one-sided and two-sided sign tests and the appropriate confidence bound and confidence interval, respectively, are no longer valid.

51. *Necessity of Magnitudes.* The confidence interval and bounds (see Comment 49) for θ based on the sign tests are simple to compute, as the end points depend only on the ordered sample Z observations. However, for such a computation, knowledge of the signs of the Z differences is no longer sufficient as it was for the computation of B (3.39) for the various sign tests. We need the observation magnitudes to obtain $\theta_{\mathrm{L}}, \theta_{\mathrm{U}}, \theta_{\mathrm{L}}^*$, or θ_{U}^*.

52. *Midpoint of the Confidence Interval as an Estimator.* The midpoint of the interval (3.63), namely, $[Z^{(C_\alpha)} + Z^{(n+1-C_\alpha)}]/2$, is also a natural estimator of θ. (Note that this actually yields a class of estimators, depending on the value of α.) In general, this midpoint is not the same as $\tilde{\theta}$ (3.58). (See Hodges and Lehmann (1967) and Comment 43 for additional discussion of this midpoint class of estimators.)
53. *Comparison of Sign and Signed Rank Confidence Intervals for θ.* The confidence interval (3.63) for θ associated with the sign test and based on the n ordered Z differences is easier to compute than the confidence interval (3.27) for θ associated with the signed rank test and based on the $n(n+1)/2$ ordered Walsh averages (see Comment 17). However, the signed rank confidence interval (3.27) is generally (but not always) more efficient than the sign confidence interval (3.63). (See Section 3.11.)
54. *Extension to Discrete Distributions.* Consider the closed version $[\theta_L, \theta_U] = [Z^{(C_\alpha)}, Z^{(n+1-C_\alpha)}]$ of the $100(1-\alpha)\%$ confidence interval for θ given in display (3.63) under the alternative (to Assumptions B1 and B2) assumption that $Z_1, \ldots, Z_n$ are a random sample from an underlying distribution $F(\cdot)$ with a *unique* median θ. Suppose that this common distribution $F(\cdot)$ is such that in any bounded interval of the real line there are at most a finite number (could be zero) of values having positive probability. (If $F(\cdot)$ is continuous, this is trivially satisfied because in that case no real number has positive probability. However, the large majority of discrete probability distributions also satisfy this mild assumption.) Under these weakened conditions on the common $F(\cdot)$, the closed interval $[\theta_L, \theta_U]$ remains a conservative $100(1-\alpha)\%$ confidence interval for θ in the sense that

$$P_\theta(\theta_L \le \theta \le \theta_U) \ge 1 - \alpha \quad \text{for all} \quad \theta$$

is guaranteed for every such $F(\cdot)$. The closed versions of the upper and lower confidence bounds (see Comment 49), namely, $(-\infty, \theta_U^\star] = (-\infty, Z^{(b_{\alpha,1/2})}]$ and $[\theta_L^\star, \infty) = [Z^{(C_\alpha^\star)}, \infty)$, respectively, also remain conservative $100(1-\alpha)\%$ bounds over this expanded class of common distributions $F(\cdot)$. (For more details on the extension of these confidence intervals and bounds to common discrete distributions, see Scheffé and Tukey (1945) and Noether (1967a).)

Properties

1. *Distribution-Freeness.* For populations satisfying Assumptions B1 and B2, (3.64) holds. Hence, we can control the coverage probability to be $1-\alpha$ without having more specific knowledge about the forms of the underlying Z distributions. Thus, (θ_L, θ_U) is a distribution-free confidence interval for θ over a very large class of populations. (See also Comment 54.)
2. *Efficiency.* See Section 3.11.

Problems

70. For the Profile Scales data of Table 3.6, obtain a confidence interval for θ with the exact confidence coefficient .9688.

71. For the bleeding time data in Table 3.7, obtain a confidence interval for θ with the exact confidence coefficient .9426.

72. For the beak-clapping data of Table 3.5, obtain an estimate for the asymptotic standard deviation of $\tilde{\theta}$. (See Comment 46.)

73. For the beak-clapping data of Table 3.5 and $\alpha = .1078$, calculate the point estimator of θ defined in Comment 52. Compare with the value of $\tilde{\theta}$ obtained in Example 3.6.

74. For the Hamilton depression scale factor IV data of Table 3.1, find a confidence interval for θ with the exact confidence coefficient .9610.

75. For the bleeding time data in Table 3.7, obtain an approximate 94.26% confidence interval for θ using the large-sample approximation of this section. Compare this approximate confidence interval with the exact 94.26% confidence interval obtained in Problem 71.

76. How does varying α affect the length of the confidence interval defined by display (3.63)? How does it affect the point estimator of θ defined in Comment 52?

77. For the beak-clapping data of Table 3.5, find a lower confidence bound for θ with the exact confidence coefficient .9461. (See Comment 49.)

78. Consider the Stanford Profile Scores data of Table 3.6. Obtain an upper confidence bound for θ with the exact confidence coefficient .8906. (See Comment 49.)

79. For the salary data in Table 3.2, find a lower confidence bound for θ with the exact confidence coefficient .9270. How does this compare with the approximate 93.6% lower confidence bound for θ obtained in Problem 32?

80. Consider the beak-clapping data of Table 3.5. Use the large-sample approximation to obtain an approximate 95% lower confidence bound for θ (see Comment 49). Compare this approximate bound with the exact 94.61% lower confidence bound obtained in Problem 77.

81. Consider the bleeding time data of Table 3.7. Use the large-sample approximation to find an approximate 92% upper confidence bound for θ. (See Comment 49.)

82. Consider the case $n = 15$ and compare the length of the exact 96.48% confidence interval for θ given by display (3.63), with the length of the approximate 96.48% confidence interval for θ obtained using the large-sample approximation of this section.

83. Consider the case $n = 25$ and compare the exact 94.61% lower confidence bound for θ given by (3.67), with the approximate 94.61% lower confidence bound for θ obtained from the large-sample approximation in Comment 49.

84. For the bleeding time data in Table 3.7 and $\alpha = 0.05$, find the estimate of θ as described in Comment 52. Compare this with the estimate found in Problem 59.

85. Consider the two-sided confidence interval found in Problem 74. What range of α values results in the same upper and lower bounds?

86. Consider the one-sided confidence interval found in Problem 77. What range of α values results in the same lower bound?

ONE-SAMPLE DATA*

3.7 PROCEDURES BASED ON THE SIGNED RANK STATISTIC

Data. We obtain n observations $Z_1, \ldots, Z_n$.

*Sections 3.7–3.10 are optional. The contents of these sections are not used in the sequel.

Table 3.8 Estimated Values of θ from the Mariner and the Pioneer Spacecraft

Spacecraft	θ
Mariner 2 (Venus)	81.3001
Mariner 4 (Mars)	81.3015
Mariner 5 (Venus)	81.3006
Mariner 6 (Mars)	81.3011
Mariner 7 (Mars)	81.2997
Pioneer 6	81.3005
Pioneer 7	81.3021

Source: J. D. Anderson, L. Efron, and S. K. Wong (1970).

Assumptions

C1. The Z's are mutually independent.

C2. Each Z comes from a population (not necessarily the same) that is continuous and symmetric about θ.

Procedures

To test $H_0 : \theta = \theta_0$, where θ_0 is some specified number, we create the modified observations $Z_i' = Z_i - \theta_0$, for $i = 1, \ldots, n$. Then we apply any of the test procedures of Section 3.1 to these modified Z' observations.

To obtain a point estimator of θ or a confidence interval for θ, we apply the procedures of Sections 3.2 and 3.3 directly to the Z observations without modification.

EXAMPLE 3.8 ***The Mariner and the Pioneer Spacecraft Data.***

The data in Table 3.8 were reported by Anderson, Efron, and Wong (1970). The seven observations represent average measurements of θ, the ratio of the mass of the Earth to that of the moon, obtained from seven different spacecraft.

On the basis of the previous (2–3 years earlier) Ranger spacecraft findings, scientists had considered the value of θ to be approximately 81.3035. Thus, with the data of Table 3.8, we are interested in testing $H_0 : \theta = 81.3035$ versus the alternative $\theta \neq 81.3035$, and we perform test procedure (3.6). With $\alpha = .078$, we see that $t_{.078/2} = 26$.

Now, we form the modified Z' observations as follows.

i	Z_i	$Z_i' = Z_i - 81.3035$
1	81.3001	−.0034
2	81.3015	−.0020
3	81.3006	−.0029
4	81.3011	−.0024
5	81.2997	−.0038
6	81.3005	−.0030
7	81.3021	−.0014

Using the computational setup of Section 3.1 on the Z' observations, we calculate $T^+ = 0$. Thus, we reject $H_0 : \theta = 81.3035$ at the $\alpha = .078$ level, since $T^+ = 0 \leq [28 - t_{.039}] = 2$. The P-value for this symmetric test based on T^+ is `2 * psignrank(0, n=7, lower.tail=T)=.0156`. This test is implemented with `wilcox.test`. For this example,

```
wilcox.test(z, mu=81.3035)
```

results in the output

```
Wilcoxon signed rank test

data: z
V = 0, p-value = .01563
alternative hypothesis: true location is not equal to
    81.3035
```

Note the use of the symbol V in place of T^+

For the large-sample approximation, we see from (3.9) that

$$T^* = \frac{0 - [7(8)/4]}{[7(8)(15)/24]^{1/2}} = -2.366.$$

Thus, the smallest significance level at which we could reject H_0 by using a symmetric test based on the normal approximation is .018. This means that both the exact test and the large-sample approximation indicate the existence of strong evidence to reject the findings of the earlier Ranger spacecraft that $\theta = 81.3035$.

The ordered values of $(Z_i + Z_j)/2$ are $W^{(1)} \leq \cdots \leq W^{(28)}$: 81.2997, 81.2999, 81.3001, 81.3001, 81.30015, 81.3003, 81.30035, 81.3004, 81.3005, 81.30055, 81.3006, 81.3006, 81.3006, 81.3008, 81.3008, 81.30085, 81.3009, 81.3010, 81.30105, 81.3011, 81.3011, 81.3013, 81.3013, 81.30135, 81.3015, 81.3016, 81.3018, and 81.3021. If $M = 7(8)/2 = 28$, we see that $M = 2k$ with $k = 14$. Thus, from (3.25), we have

$$\hat{\theta} = \frac{W^{(14)} + W^{(15)}}{2} = \frac{81.3008 + 81.3008}{2} = 81.3008.$$

With $n = 7$ and $\alpha = .046$, we find that $t_{\alpha/2} = t_{.023} = 27$. Thus, $C_{.046} = \{7(8)/2\} + 1 - t_{.023} = 28 + 1 - 27 = 2$.

From (3.27), it follows that

$$\theta_{\mathrm{L}} = W^{(2)} = 81.2999 \text{ and } \theta_{\mathrm{U}} = W^{(27)} = 81.3018$$

so that our 95.4% confidence interval for θ is

$$(\theta_{\mathrm{L}}, \theta_{\mathrm{U}}) = (81.2999, 81.3018).$$

The above results may be produced in R through the function call

```
wilcox,test(z, mu=81.3035, exact=T, conf.int=T,
    conf.level=l-.046)
```

where z is a vector containing the seven measurements from Table 3.8. The above function call extends the R output given in Example 3.8:

```
  Wilcoxon signed rank test

data: z
V = 0, p-value = .01563
alternative hypothesis: true location is not equal to
    81.3035
95.4 percent confidence interval:
81.2999 81.3018
sample estimates:
(pseudo)median
81.3008
```

Applying the large-sample approximation, we find from (3.29) that

$$C_{.046} \approx [7(8)/4] - 1.996[7(8)(15)/24]^{1/2} \approx 2,$$

resulting in the same interval.

It is important to comment that in applying the procedures based on the signed rank statistic T^+ (3.3), we made the assumption that the population of average θ measurements for each of the satellites was symmetric about θ. (For a test of this basic assumption, see Section 3.9.) We also note that this set of data provides an example in which the populations of the Z observations are probably not the same (see Assumption C2).

Comments

55. *Assumptions.* Note that Assumption Al for the paired replicates procedures based on the signed rank statistic is not necessary for the one-sample data because these data need not consist of differences for paired observations.

Properties

1. The properties of the one-sample procedures based on the signed rank statistic are essentially the same as those of the corresponding paired replicates procedures. An exception occurs in the efficiencies of the procedures and is due to the difference in the type of data for the two problems. See Section 3.11 for a discussion of the difference in efficiencies of the procedures of Sections 3.1–3.3 when they are applied to single-sample problems.

Problems

87. The data in Table 3.9 are a subset of the data reported by Ijzermans (1970) from an investigation on the susceptibility to corrosion of 18Cr_10Ni_2Mo stainless steel (i.e., stainless steel containing 18% chromium, 10% nickel, and 2% molybdenum by weight).

Twelve specimens of steel were selected for use in the corrosion study. Although Ijzermans' experiment was directed toward corrosion, we are concerned here with the quality of the steel from which the stainless steel samples were chosen. Table 3.9 gives the percentage of chromium in the 12 samples used by Ijzermans.

Test the hypothesis that the median percentage of chromium content (θ) of the steel is 18% against the alternative that it is not 18%. Obtain a point estimate of θ and find a confidence interval for θ with the confidence coefficient .936.

88. For the percentage of chromium data in Table 3.9, obtain a point estimate of θ from the midpoint of the confidence interval calculated in Problem 87 (see Comment 23). Compare with the point estimate obtained in Problem 87.

89. Compute $\hat{\theta}$ for the settling velocity data of Table 3.12 and compare with the value of $\tilde{\theta}$ obtained in Example 3.9.

90. Lamp (1976) studied the age distribution of a common mayfly species, *Stenacron interpunctatum*, among various habitats in Big Darby Creek, Ohio. One of the measurements considered was head width (in micrometer divisions, 1 division = .0345 mm); a subset of Lamp's data from the mayflies in habitat A is presented in Table 3.10.

Test the hypothesis that the median head width for mayflies from habitat A (θ) is 22 μm divisions against the alternative that it is greater than 22 μm. Obtain a point estimate of θ and find a lower confidence bound (see Comment 21) for θ with the confidence coefficient .976.

91. The data in Table 3.11 are a subset of the data obtained by Poland et al. (1970) in an experiment concerned with the effect of occupational exposure to DDT on human drug and steroid metabolism. The DDT-exposed subjects were employees of the Montrose Chemical Corporation, who had been working in the DDT plant at Torrance, California, for more than 5 years.

Table 3.9 Percentage of Chromium in the Stainless Steel Samples

Steel sample	% of Cr
1	17.4
2	17.9
3	17.6
4	18.1
5	17.6
6	18.9
7	16.9
8	17.5
9	17.8
10	17.4
11	24.6
12	26.0

Source: A. B. Ijzermans (1970).

Table 3.10 Mayfly Head Width, Habitat A (Micrometer Divisions)

Mayfly i	Z_i
1	36
2	31
3	30
4	27
5	20
6	33
7	27
8	18
9	19
10	28

Source: W. O. Lamp (1976).

Table 3.11 6β-Hydroxycortisol Excretion (μg/24 h)

Worker i	Z_i
1	254
2	171
3	345
4	134
5	190
6	447
7	106
8	173
9	449
10	198

Source: A. Poland, D. Smith, R. Kuntzman, M. Jacobson, and A. H. Conney (1970).

Table 3.12 Settling Velocities at 22 °C

Sample i	Z_i, cm/s
1	12.9
2	13.7
3	14.5
4	13.3
5	12.8
6	13.8
7	13.4

Source: J. D. Smith (1969).

Table 3.13 Oxidant Content of Dew Water, Port Burwell, 1960

Sample i	Z_i, ppm ozone
1	.32
2	.21
3	.28
4	.15
5	.08
6	.22
7	.17
8	.35
9	.20
10	.31
11	.17
12	.11

Source: A. F. W. Cole and M. Katz (1966).

All these individuals had received moderate to intense occupational exposure to DDT, and all were in good health. One of the measures used in the study was the 24-h urinary excretion of 6β-hydroxycortisol.

Test the hypothesis that the median 6β-hydroxycortisol excretion rate for subjects with occupational exposure to DDT similar to the workers in this study (θ) is 175 μg/24 h against

the alternative that it is greater than 175. Obtain a point estimate of θ and find a confidence interval for θ with the confidence coefficient .916.

92. Consider the oxidant content of dew water data in Table 3.13. Use the computer software R to test the hypothesis that the median oxidant content of dew water (θ) was .25 against the alternative that it was less than .25. Also use R to obtain a point estimate of θ and find an upper confidence bound (see Comment 21) for θ with the confidence coefficient .961. Compare with the answers to Problem 94.

93. Consider the settling velocity data of Table 3.12. Use the computer software R to test the hypothesis that the median settling velocity for the Middle Ground sand ridge (θ) was 14 cm/s against the alternative that it was not equal to 14 cm/s. Also use R to obtain a point estimate of θ and find a confidence interval for θ with the confidence coefficient .890. Compare with the results obtained in Example 3.9.

3.8 PROCEDURES BASED ON THE SIGN STATISTIC

Data. We obtain n observations $Z_1, \ldots, Z_n$.

Assumptions

D1. The Z's are mutually independent.

D2. Each Z comes from the same continuous population with median θ, so that $P(Z_i > \theta) = P(Z_i > \theta) = \frac{1}{2}, i = 1, \ldots, n$.

Procedures

To test $H_0 : \theta = \theta_0$, where θ_0 is some specified number, we form the modified observations $Z_i' = Z_i - \theta_0$, for $i = 1, \ldots, n$. Then we can apply any of the test procedures of Section 3.4 to these modified Z' observations. (In the test of H_0, we can weaken Assumption D2 to D2′, namely, that each Z comes from a population, not necessarily the same population, such that $P(Z_i < \theta_0) = P(Z_i > \theta_0) = \frac{1}{2}, i = 1, \ldots, n$, when θ is equal to the hypothesized value θ_0.)

To obtain a point estimator of θ or a confidence interval for θ, we apply the procedures of Sections 3.5 and 3.6 directly to the Z observations without modification.

EXAMPLE 3.9 ***Sediment Settling Velocities.***

The data in Table 3.12 are a subset of the data obtained by Smith (1969) in an experiment investigating the geomorphology of the Middle Ground sand ridge, which is located in Vineyard Sound, Massachusetts.

Seven samples were obtained from a particular portion of the ridge using a Van Veen grab. One of the objective measurements reported by Smith was the settling velocity of the sediment at 22 °C. For sediment from a sand-wave crest section of a sand ridge, the settling velocity has a typical value of 14 cm/s. Table 3.12 gives the settling velocities for the seven sediment samples collected from a particular portion of the Middle Ground sand ridge.

We would like to detect whether the seven sediment samples came from a sand-wave crest section of the Middle Ground sand ridge. Let θ denote the median settling velocity for the population of sediment samples from this portion of Middle Ground. Then we are interested in testing $H_0 : \theta = 14$ cm/s versus the alternative $\theta \neq 14$ cm/s, and we perform test procedure (3.42). With $\alpha = .02$, we see that $b_{.02/2,1/2} = 7$.

Now, we create the modified Z' observations using the following setup.

i	Z_i	$Z_i' = Z_i - 14$
1	12.9	−1.1
2	13.7	−0.3
3	14.5	0.5
4	13.3	−0.7
5	12.8	−1.2
6	13.8	−0.2
7	13.4	−0.6

Using the computational setup of Section 3.4 on the Z' observations, we calculate $B = 1$. Thus, we accept $H_0 : \theta = 14$ cm/s at the $\alpha = .02$ level, since $[7 - b_{.02/2,1/2}] = 0 < B < 7 = b_{.02/2,1/2}$. The above results may be reproduced in R through the function call

```
SIGN.test(z, md=14)
```

where z is a vector containing the seven measurements from Table 3.12. This command also provides the P-value and, optionally, confidence intervals. For the current data and test, the P-value is .125. The partial R output of the above command is

```
    One-sample Sign-Test

data: z
s = 1, p-value = .125
alternative hypothesis: true median is not equal to 14
sample estimates:
median of x
13.4
```

For the large-sample approximation, we see from (3.45) that

$$B^* = \frac{1 - \left(\frac{7}{2}\right)}{\left(\frac{7}{4}\right)^{1/2}} \approx -1.89.$$

Thus the smallest significance level at which we could reject H_0 using a symmetric test based on the normal approximation is .0588.

The ordered Z observations are $Z^{(1)} \leq \cdots \leq Z^{(7)}$:12.8, 12.9, 13.3, 13.4, 13.7, 13.8, and 14.5. The sample size n is $(2k + 1)$ with $k = 3$, therefore (3.59) implies that

$$\tilde{\theta} = Z^{(4)} = 13.4.$$

With $n = 7$ and $\alpha = .1250$, we find that $b_{\alpha/2,1/2} = b_{.0625,1/2} = 6$. Thus, $C_{.1250} = 7 + 1 - 6 = 2$. From (3.63), it follows that

$$\theta_{\mathrm{L}} = Z^{(2)} = 12.9 \text{ and } \theta_{\mathrm{U}} = Z^{(6)} = 13.8,$$

so that our 87.50% confidence interval for θ is

$$(\theta_{\mathrm{L}}, \theta_{\mathrm{U}}) = (12.9, 13.8).$$

The confidence interval for θ is found with

```
SIGN.test(z, md=14, conf.level=1-.125)
```

This results in the following output being appended to the output give above:

```
                    Conf.Level      L.E.pt      U.E.pt
Lower Achieved CI       .8750        12.9        13.8
Interpolated CI         .8750        12.9        13.8
Upper Achieved CI       .9844        12.8        14.5
```

Note that the lower achieved interval is the desired interval for this α.

Applying the large-sample approximation, we find from (3.65) that

$$C_{.1250} \approx \left(\frac{7}{2}\right) - 1.534\left(\frac{7}{4}\right)^{1/2} \approx 1,$$

and as $Z^{(1)} = 12.8$ and $Z^{(n+1-1)} = Z^{(7)} = 14.5$, the approximate 87.50% confidence interval for θ is $(12.8, 14.5)$.

Comments

56. *Assumptions.* Note that Assumption B1 for the paired replicates procedures based on the sign statistic is not necessary for the one-sample data because these data do not consist of differences for paired observations.

57. *Procedures for Population Quantiles Other than the Median.* For one-sample data, the theory underlying the sign statistic can also be used to construct distribution-free test procedures for population quantiles other than the median. Such test procedures are similar to procedures (3.40), (3.41), and (3.42), but they have different P-values in the null hypothesis binomial distribution. For example, let $Z_1, \ldots, Z_n$ be a random sample from a population II. Define μ_ξ to be the unknown ξ quantile of the population. (For convenience, let us assume that μ_ξ is unique.) Consider the problem of testing $H_0 : \mu_\xi = \mu_0$ (specified) versus the one-sided alternative $\mu_\xi > \mu_0$. Define B to be the number of Z's that are greater than μ_0. Under H_0, B has the binomial distribution with parameters n and $p = 1 - \xi$. Large values of B indicate that $\mu_\xi > \mu_0$, so an appropriate one-sided α-level test is to reject H_0 in favor of $\mu_\xi > \mu_0$ if $B \geq b_{\alpha,1-\xi}$ and accept H_0 if $B < b_{\alpha,1-\xi}$. One-sided tests against $\mu_\xi < \mu_0$ and two-sided tests for alternatives $\mu_\xi \neq \mu_0$ are constructed in a similar manner. The natural point estimator of the parameter $P(Z > \mu_0)$ is the statistic B/n. Approximate confidence intervals for μ_ξ can also be obtained (cf. Conover (1999)).

Let $Z_1, Z_2, \ldots, Z_n$ be a random sample of size n from an unknown distribution. Let z_p denote the pth quantile of the distribution. Hayter (2013) constructs simultaneous confidence intervals for $z_{p_i}, 1 \leq i \leq k, 0 < p_1 < p_2 < \cdots < p_k < 1$. The intervals are of the form

$$z_{p_i} \in [Z_{(l_i)}, Z_{(u_i+1)}], \quad 1 \leq i \leq k,$$

where $Z_{(1)}, Z_{(2)} \ldots, Z_{(n)}$ are the order statistics. The integers l_i and u_i are suitably chosen as described by Hayter to provide an overall simultaneous confidence level of at least $1 - \alpha$, with the lower limits being $-\infty$ if $l_i = 0$ and the upper limits being ∞ if $u_i = n$. See Hayter (2013) for his methodology and for specific examples with $\alpha = .05$ and $n = 20, 50$ and 80.

Properties

1. The properties of the one-sample procedures based on the sign statistic are essentially the same as those of the corresponding paired replicates procedures. An exception occurs in the efficiencies of the procedures and is due to the difference in the type of data for the two problems. See Section 3.11 for a discussion of the difference in efficiencies for the procedures of Sections 3.4–3.6 when they are applied to single-sample problems.

Problems

94. The data in Table 3.13 are a subset of the data obtained by Cole and Katz (1966). They were investigating the relation between ozone concentrations and weather fleck damage to tobacco crops in southern Ontario, Canada. One of the objective measurements reported was oxidant content of dew water in parts per million (ppm) ozone. Twelve samples of dew were collected during the period August 25–30, 1960, at Port Burwell, Ontario; the resulting oxidant contents are given in Table 3.13.

Test the hypothesis that the median oxidant content (θ) of dew water was .25 against the alternative that it was less than .25. Obtain a point estimate of θ and find a confidence interval for θ with the confidence coefficient .9614.

95. For the oxidant content data of Table 3.13, obtain a point estimate of θ from the midpoint of the confidence interval calculated in Problem 94 (see Comment 52). Compare with $\tilde{\theta}$ obtained in Problem 94.

96. Compute $\tilde{\theta}$ for the mass ratio data of Table 3.8 and compare with the value of $\hat{\theta}$ obtained in Example 3.8.

97. Maxson (1977) studied the activity patterns of female ruffed grouse with broods. Using surveillance techniques, he recorded the movements of seven female ruffed grouse with broods over a fixed period. The percentage of time that these grouse were in active movement is recorded in Table 3.14.

Test the hypothesis that the median percentage time active for female ruffed grouse with broods (θ) is 50% against the alternative that it is greater than 50%. Obtain a point estimate of θ and find a lower confidence bound (see Comment 49) for θ with the confidence coefficient .99.

98. The data in Table 3.15 are a subset of the data obtained by Flores and Zohman (1970) in an experiment investigating the effect of the method of bed-making on the oxygen consumption for patients assigned to complete or modified bed rest. The subjects were inpatients of the Rehabilitation Medicine Service, Montefiore Hospital and Medical Center, Bronx, New York.

Table 3.14 Ruffed Grouse, Percentage Time in Active Movement

Grouse i	Z_i (% time active)
1	52.7
2	51.5
3	58.4
4	56.9
5	58.5
6	54.4
7	47.1

Source: S. J. Maxson (1977).

Table 3.15 Net Oxygen Consumption (cc)

Patient i	Z_i
1	339
2	349
3	387
4	159
5	579
6	586
7	519
8	275

Source: A. M. Flores and L. R. Zohman (1970).

The measure used was net oxygen consumption for the patients during bed-making. The data in Table 3.15 are the net oxygen consumptions (in cc) for the eight patients in the study during a cardiac top-to-bottom bed-making procedure, consisting of moving the patient to a sitting position and changing the sheets from the top to the bottom of the bed.

Test the hypothesis that the median oxygen consumption rate during cardiac bed-making for patients assigned to complete or modified bed rest (θ) is 350 cc against the alternative that it is not 350 cc. Obtain a point estimate of θ and find a confidence interval for θ with the confidence coefficient .95.

99. Consider the 6β-hydroxycortisol excretion data in Table 3.11. Use the computer software R to test the hypothesis that the median 6β-hydroxycortisol excretion rate for subjects with occupational exposure to DDT similar to the workers in the Poland et al. (1970) study (θ) is 175 μg/24 h against the alternative that it is greater than 175 μg/24 h. Obtain a point estimate of θ and find a confidence interval for θ with the confidence coefficient .925. Compare with the answers to Problem 91.

100. Consider the mayfly head width data in Table 3.10. Let $\mu_{.75}$ be the 75th percentile for the distribution of mayfly head widths in habitat A studied by Lamp (1976). Test the hypothesis that $\mu_{.75} = 25$ against the alternative that $\mu_{.75}$ is greater than 25. (See Comment 57.)

3.9 AN ASYMPTOTICALLY DISTRIBUTION-FREE TEST OF SYMMETRY (RANDLES–FLIGNER– POLICELLO–WOLFE, DAVIS–QUADE)

Data. We obtain n observations $Z_1, \ldots, Z_n$.

Assumptions

E1. The Z's are mutually independent.

E2. Each Z comes from the same continuous population having distribution function F and unknown median θ. This assumption requires that $F(\theta) = \frac{1}{2}$.

Hypothesis

The null hypothesis of interest here is that the common underlying distribution for the Z observations is symmetric about θ. This hypothesis of symmetry can be written as

$$H_0 : [F(\theta + b) + F(\theta - b) = 1, \text{ for every } b], \tag{3.72}$$

and it is equivalent to the statement that $P(0 < Z - \theta < b) = P(-b < Z - \theta < 0)$ for all $b > 0$.

Procedure

For each triple of observations $(Z_i, Z_j, Z_k), 1 \le i < j < k \le n$, obtain the value of

$$f^*(Z_i, Z_j, Z_k) = [\text{sign } (Z_i + Z_j - 2Z_k)] + \text{sign } (Z_i + Z_k - 2Z_j) + \text{sign } (Z_j + Z_k - 2Z_i)], \tag{3.73}$$

where sign $(t) = -1, 0, 1$ as $t <, =, > 0$. (Note that there are $n(n-1)(n-2)/6$ distinct triples in the sample.) We say that (Z_i, Z_j, Z_k) forms a **right triple** (looks skewed to the right) if $f^*(Z_i, Z_j, Z_k) = 1$. (Note that being a right triple is equivalent to the middle *ordered* observation in (Z_i, Z_j, Z_k) being closer to the smallest of the three observations than it is to the largest of them.) Conversely, (Z_i, Z_j, Z_k) is said to be a **left triple** (looks skewed to the left) if $f^*(Z_i, Z_j, Z_k) = -1$ (i.e., the middle *ordered* observation in (Z_i, Z_j, Z_k) is closer to the largest than to the smallest of the three observations). Finally, when $f^*(Z_i, Z_j, Z_k) = 0$, the triple (Z_i, Z_j, Z_k) is neither right nor left.

For the data $Z_1, \ldots, Z_n$, set

$$T = \sum\sum\sum_{1 \le i < j < k \le n} f^*(Z_i, Z_j, Z_k) = \{[\text{number of right triples}] - [\text{number of left triples}]\}. \tag{3.74}$$

For each fixed $t = 1, \ldots, n$, let

$$B_t = \{[\text{number of right triples involving } Z_t] - [\text{number of left triples involving } Z_t]\} = \left[\sum_{j=t+1}^{n-1} \sum_{k=j+1}^{n} f^*(Z_t, Z_j, Z_k) + \sum_{j=1}^{t-1} \sum_{k=t+1}^{n} f^*(Z_j, Z_t, Z_k) + \sum_{j=1}^{t-2} \sum_{k=j+1}^{t-1} f^*(Z_j, Z_k, Z_t) \right]. \tag{3.75}$$

For each fixed integer pair (s, t) such that $1 \le s < t \le n$, define

$$B_{s,t} = \{[\text{number of right triples involving } Z_s \text{ and } Z_t] - [\text{number of left triples involving } Z_s \text{ and } Z_t]\} = \left[\sum_{j=1}^{s-1} f^*(Z_j, Z_s, Z_t) + \sum_{j=s+1}^{t-1} f^*(Z_s, Z_j, Z_t) + \sum_{j=t+1}^{n} f^*(Z_s, Z_t, Z_j) \right]. \tag{3.76}$$

Using the expressions for B_t (3.75), $B_{s,t}$ (3.76), and the triple statistic T (3.74), set

$$V = \frac{T}{\hat{\sigma}}, \tag{3.77}$$

where

$$\hat{\sigma}^2 = \left[\frac{(n-3)(n-4)}{(n-1)(n-2)} \sum_{t=1}^{n} B_t^2 + \frac{(n-3)}{(n-4)} \sum_{s=1}^{n-1} \sum_{t=s+1}^{n} B_{s,t}^2 \right.$$

$$\left. + \frac{n(n-1)(n-2)}{6} - \left\{ 1 - \frac{(n-3)(n-4)(n-5)}{n(n-1)(n-2)} \right\} T^2 \right]. \tag{3.78}$$

When H_0 is true and the underlying distribution is symmetric, V has, as n tends to infinity, an asymptotic $N(0,1)$ distribution. (In order for this normal approximation to be reasonably effective, the sample size n should be at least 10. For further discussion along these lines, see Comment 60.)

To test H_0 (3.72), corresponding to symmetry of the underlying distribution, versus the general alternative of asymmetry, corresponding to

$$H_1 : [P(Z \leq \theta + b) + P(Z \leq \theta - b) \neq 1 \text{ for at least one } b], \tag{3.79}$$

at the approximate (n large) α level of significance,

$$\text{Reject } H_0 \text{ if } |V| \geq z_{\alpha/2}; \text{ otherwise do not reject.} \tag{3.80}$$

Ties

The test procedure in (3.80) is well-defined when zeros occur in the $(Z_i + Z_j - 2Z_k)$ variables and further adjustments are not necessary.

EXAMPLE 3.10 ***Percentage Chromium in Stainless Steel.***

In order to clearly illustrate the details of the rather involved calculations necessary to obtain the value of the test statistic V (3.77), we consider the application of the test for symmetry to the first five (i.e., $n = 5$) percentage chromium data values in Table 3.9, namely, $Z_1 = 17.4, Z_2 = 17.9, Z_3 = 17.6, Z_4 = 18.1$, and $Z_5 = 17.6$. (We emphasize that this application is for illustrative purposes only. The test for symmetry is totally ineffective at detecting asymmetry for sample sizes as small as $n = 5$. See Comment 60 for related discussion.) We must calculate $n(n-1)(n-2)/6 = 5(4)(3)/6 = 10$ values of the triple indicator $f^*(Z_i, Z_j, Z_k)$ given by (3.73). We have that

$$f^*(Z_1, Z_2, Z_3) = [\text{sign } (17.4 + 17.9 - 2(17.6)) + \text{sign } (17.4 + 17.6$$

$$- 2(17.9)) + \text{sign } (17.9 + 17.6 - 2(17.4))]$$

$$= [\text{sign } (.1) + \text{sign } (-.8) + \text{sign } (.7)] = 1 - 1 + 1 = 1. \tag{3.81}$$

Similarly, we obtain

$$f^*(Z_1, Z_2, Z_5) = f^*(Z_1, Z_3, Z_4) = f^*(Z_1, Z_4, Z_5)$$

$$= f^*(Z_2, Z_3, Z_5) = f^*(Z_3, Z_4, Z_5) = 1 \tag{3.82}$$

and

$$f^*(Z_1, Z_2, Z_4) = f^*(Z_1, Z_3, Z_5) = f^*(Z_2, Z_3, Z_4)$$
$$= f^*(Z_2, Z_4, Z_5) = -1. \tag{3.83}$$

Hence, from (3.74) we have that

$$T = \sum_{1 \le i < j < k \le 5} \sum \sum f^*(Z_i, Z_j, Z_k) = 6 - 4 = 2. \tag{3.84}$$

For the calculation of $\hat{\sigma}^2$, we first need to obtain the values of $B_1, \ldots, B_5$ and $B_{s,t}$ for $1 \le s < t \le 5$. From (3.75), (3.81), (3.82), and (3.83), we have that

$$B_1 = \sum_{j=2}^{4} \sum_{k=j+1}^{5} f^*(Z_1, Z_j, Z_k) = [1 - 1 + 1 + 1 - 1 + 1] = 2,$$

$$B_2 = \left[\sum_{j=3}^{4} \sum_{k=j+1}^{5} f^*(Z_2, Z_j, Z_k) + \sum_{k=3}^{5} f^*(Z_1, Z_2, Z_k) \right]$$
$$= [(-1 + 1 - 1) + (1 - 1 + 1)] = 0,$$

$$B_3 = \left[f^*(Z_3, Z_4, Z_5) + \sum_{j=1}^{2} \sum_{k=4}^{5} f^*(Z_j, Z_3, Z_k) + f^*(Z_1, Z_2, Z_3) \right]$$
$$= [1 + (1 - 1 - 1 + 1) + 1] = 2,$$

$$B_4 = \left[\sum_{j=1}^{3} f^*(Z_j, Z_4, Z_5) + \sum_{j=1}^{2} \sum_{k=j+1}^{3} f^*(Z_j, Z_k, Z_4) \right]$$
$$= [(1 - 1 + 1) + (-1 + 1 - 1)] = 0,$$

and

$$B_5 = \sum_{j=1}^{3} \sum_{k=j+1}^{4} f^*(Z_j, Z_k, Z_5) = [1 - 1 + 1 + 1 - 1 + 1] = 2.$$

It follows that

$$\sum_{t=1}^{5} B_t^2 = [2^2 + 0^2 + 2^2 + 0^2 + 2^2] = 12. \tag{3.85}$$

Furthermore, using (3.76), (3.81), (3.82), and (3.83), we obtain

$$B_{1,2} = \sum_{j=3}^{5} f^*(Z_1, Z_2, Z_j) = [1 - 1 + 1] = 1,$$

$$B_{1,3} = f^*(Z_1, Z_2, Z_3) + \sum_{j=4}^{5} f^*(Z_1, Z_3, Z_j) = [1 + (1 - 1)] = 1,$$

$$B_{1,4} = \sum_{j=2}^{3} f^*(Z_1, Z_j, Z_4) + f^*(Z_1, Z_4, Z_5) = [(-1 + 1) + 1] = 1,$$

$$B_{1,5} = \sum_{j=2}^{4} f^*(Z_1, Z_j, Z_5) = [1 - 1 + 1] = 1,$$

$$B_{2,3} = f^*(Z_1, Z_2, Z_3) + \sum_{j=4}^{5} f^*(Z_2, Z_3, Z_j) = [1 + (-1 + 1)] = 1,$$

$$\begin{aligned} B_{2,4} &= f^*(Z_1, Z_2, Z_4) + f^*(Z_2, Z_3, Z_4) + f^*(Z_2, Z_4, Z_5) \\ &= [-1 - 1 - 1] = -3, \end{aligned}$$

$$B_{2,5} = f^*(Z_1, Z_2, Z_5) + \sum_{j=3}^{4} f^*(Z_2, Z_j, Z_5) = [1 + (1 - 1)] = 1,$$

$$B_{3,4} = \sum_{j=1}^{2} f^*(Z_j, Z_3, Z_4) + f^*(Z_3, Z_4, Z_5) = [(1 - 1) + 1] = 1,$$

$$B_{3,5} = \sum_{j=1}^{2} f^*(Z_j, Z_3, Z_5) + f^*(Z_3, Z_4, Z_5) = [(-1 + 1) + 1] = 1,$$

and

$$B_{4,5} = \sum_{j=1}^{3} f^*(Z_j, Z_4, Z_5) = [1 - 1 + 1] = 1.$$

These $B_{s,t}$ values yield

$$\begin{aligned} \sum_{s=1}^{4} \sum_{t=s+1}^{5} B_{s,t}^2 &= [1^2 + 1^2 + 1^2 + 1^2 + 1^2 + (-3)^2 + 1^2 + 1^2 + 1^2 + 1^2] \\ &= 18. \end{aligned} \tag{3.86}$$

Using the computational results from (3.84), (3.85), and (3.86) in the formula for $\hat{\sigma}^2$ (3.78), we obtain

$$\begin{aligned} \hat{\sigma}^2 &= \left[\frac{2(1)}{4(3)}(12) + \frac{2}{1}(18) + \frac{5(4)(3)}{6} - \left\{ 1 - \frac{2(1)(0)}{5(4)(3)} \right\} (2)^2 \right] \\ &= [2 + 36 + 10 - 4] = 44. \end{aligned}$$

Finally, from (3.77), we have

$$V = \frac{T}{\hat{\sigma}} = \frac{2}{(44)^{1/2}} = .30.$$

(We note that the R command RFPW(**z**) can also be used to obtain the value of the test statistic $V = \frac{T}{\hat{\sigma}}$ for the data **z**. For this example, we have $n = 5$, $\mathbf{z} = (17.4, 17.9, 17.6, 18.1, 17.6)$, and RFPW(**z**) =.30.)

With significance level $\alpha = .05$, we use the R command qnorm(.) to obtain the critical value $z_{.025} = 1.96$ from the fact that qnorm(1-.025) = qnorm(.975)=1.96. Since $|V| = .30$ is less than 1.96, we cannot reject the null hypothesis of symmetry for the underlying distribution. In fact, using the R command pnorm (.), we see that the smallest significance level at which we could reject this distributional symmetry (i.e., the two-sided P-value for these data) is

$$\begin{aligned} \underline{\alpha} &= 2P(\text{standard normal variable exceeds } .30) \\ &= 2(1 - \texttt{pnorm}(.30)) \\ &= 2(.3821) = .7642, \end{aligned}$$

clearly indicating that there is virtually no evidence in this subset of the percentage chromium data to indicate asymmetry in the underlying probability distribution. (Remember, however, that this subset was a sample of only five observations. These are simply not sufficient data to detect asymmetry even if it were present. See Comment 60.)

Comments

58. *Motivation.* A right triple is indicative of skewness to the right and a left triple is indicative of skewness to the left. The absolute value of the statistic T (3.74) is the difference between the numbers of right and left triples among the $n(n-1)(n-2)/6$ triples in the sample. When the null hypothesis H_0 (3.72) of symmetry is true, we would expect half of the sample triples to be right triples and the other half to be left triples. Thus, when H_0 is true, we would expect T to be near zero. A substantial deviation in either direction from zero for T is therefore indicative of asymmetry in the population and serves as a partial motivation for the procedure defined in (3.80).
59. *Asymptotic Distribution-Freeness.* Asymptotically (i.e., for infinitely large samples), the true level of the test defined by (3.80) will agree with the nominal level. Subject to Assumptions E1 and E2, this asymptotic result does not depend on the underlying population of the Z's. More precisely, subject to Assumptions E1 and E2, V has an asymptotic $N(0, 1)$ distribution when H_0 is true. Since this asymptotic distribution does not depend on the underlying population of the Z's, we say that the test based on V is asymptotically distribution-free. Of course, in practice, we do not have the luxury of infinite samples. Thus in any particular case, with n large, we hope the level of a test based on V is close to the nominal level α but it may not be exactly equal to a. The closeness of the approximation depends on n and α and, for fixed α, the closeness generally improves as n increases. In the case of the V test, the reader is warned that the question of how large n should be, in order for the approximation to be good, is unanswered. Exact null distribution critical values for V cannot be provided because, for a specified value of n, the exact null distribution of V depends on the underlying Z population; thus exact critical values would vary with the form of the Z population. The procedure in (3.80) based on V, therefore, is not (strictly) distribution-free.

60. *Sample Size Requirement.* As noted in Comment 59, the test of symmetry described in (3.80) is not an exact distribution-free procedure. The nominal significance level α is guaranteed only asymptotically, as the number of observations, n, becomes infinite. In addition, symmetry is a rather complex property of a probability distribution. It is, therefore, virtually impossible to deny its presence without at least a moderate sample size. It is simply difficult to "see" asymmetry in a small number of sample observations. Both Randles et al. (1980) and Davis and Quade (1978) found this to be the case. They concluded that the symmetry test (3.80) is not effective at detecting asymmetry in the underlying population unless the sample size (n) is at least 20.

61. *One-Sided Tests for Right-Skewness or Left-Skewness.* The test procedure in (3.80) is a two-sided test of symmetry against a very general class of asymmetric alternatives. However, one-sided tests of symmetry versus specific classes of right-skewed (or left-skewed) asymmetric alternatives can also be based on the statistic V (3.77). In particular, a one-sided (approximate) level α test of H_0 (3.72) (symmetry) versus the specific class of right-skewed alternatives satisfying

$$F(\theta + b) \leq [1 - F(\theta - b)], \text{ for every } b > 0,$$
$$\text{with strict inequality for at least one positive } b, \tag{3.87}$$

is given by

$$\text{Reject } H_0 \text{ if } V \geq z_\alpha; \text{ otherwise do not reject.} \tag{3.88}$$

Similarly, a one-sided (approximate) level α test of H_0 (3.72) versus the specific class of left-skewed alternatives satisfying

$$F(\theta + b) \geq [1 - F(\theta - b)], \text{ for every } b > 0,$$
$$\text{with strict inequality for at least one positive } b, \tag{3.89}$$

is given by

$$\text{Reject } H_0 \text{ if } V \leq -z_\alpha; \text{ otherwise do not reject.} \tag{3.90}$$

These one-sided hypothesis tests in (3.88) and (3.90) are asymptotically distribution-free in the same sense as the two-sided test given by (3.80). See Comment 59 for further discussion of this property.

62. *Signed Rank Procedures.* One of the critical assumptions permitting the application of signed rank procedures to one-sample data is that of underlying distributional symmetry (see Assumption C2 in Section 3.7). Under this symmetry *assumption,* procedures based on the signed rank statistic T^+ (3.3) for one-sample data are used to make inferences about the median of a population. Procedure (3.80), on the other hand, is used to test for the symmetry of a population and is not directly concerned with the numerical value of the median of the population. Therefore, in an appropriate one-sample location problem we might wish to apply procedure (3.80) (to check the symmetry assumption) prior to using the signed rank procedures of Section 3.7 for making inferences about the actual value of the unknown median of the population. Procedure (3.80) is appropriate, but the *known* median test mentioned in Comment 63 is inappropriate as a pretest in this situation. (For the paired-replicates data in Sections 3.1–3.3, we remind the reader that the symmetry assumption

is most often inherently satisfied through the nature of the pairing. See Comment 2.)

63. *Case of Known Median.* For the situation when the median of the underlying population is *known* to be a specified value θ_0 (say), Gupta (1967) proposed a procedure for testing the hypothesis of symmetry about θ_0. However, situations in which the median of the underlying population is known but the symmetry of the distribution is not known are encountered considerably less frequently than situations in which both the median and the symmetry are not known (see Comment 62). (Gupta (1967) also proposed a test for symmetry when the underlying median is not known. His procedure in this case is a competitor to the test given by (3.80). He investigated the loss of efficiency that results from using his test for symmetry with unknown median when his known median procedure is applicable.)

64. *Alternative Determination of Right and Left Triples.* The original definitions of right and left triples in this section involve the sign function $f^*(Z_i, Z_j, Z_k)$ in (3.73). A more intuitive interpretation is associated with the comparison of two common sample measures of location. For a triple (Z_i, Z_j, Z_k), let $\bar{Z} = (Z_i + Z_j + Z_k)/3$ and $\tilde{Z} =$ median $\{Z_i, Z_j, Z_k\}$ be the average and median, respectively, for the observations in the triple. Then the triple (Z_i, Z_j, Z_k) is a right triple if $\bar{Z} > \tilde{Z}$ and it is a left triple if $\bar{Z} < \tilde{Z}$. (It is neither right nor left if $\bar{Z} = \tilde{Z}$.) This formulation provides a very natural interpretation of what it means to be a right or left triple, as we know that the population mean is greater than or less than the population median according to whether the population is skewed to the right or left, respectively. If the population is symmetric, its mean and median are equal and it would be a toss up as to which of $\bar{Z}$ or $\tilde{Z}$ would be greater. This should lead to about an equal number of right and left triples in the sample.

65. *Consistent Estimator of the Asymptotic Variance of $\sqrt{n}T$.* In order to insure that V (3.77) is asymptotically distribution-free, $n\hat{\sigma}^2$ (3.78) is taken to be a consistent estimator of the asymptotic null variance of $n^{1/2}T$. The consistency of this estimator $n\hat{\sigma}^2$ follows from a standard body of theory about a class of statistics introduced by Hoeffding (1948a) and referred to as U-statistics. (For more details about U-statistics and their application in the triples test, see Randles and Wolfe (1979).) The asymptotic normality (and, thereby, the asymptotic distribution-freeness) for V (3.77) follows from standard U-statistics theory and Slutsky's theorem (see, for example, Theorem A.3.13 in Randles and Wolfe (1979)).

66. *Consistency of the V Test.* Under Assumptions El and E2, the consistency of the tests based on V depend on the parameter

$$p^* = P(Z_1 + Z_2 - 2Z_3 > 0) - \frac{1}{2}. \tag{3.91}$$

The two-sided test defined by (3.80) is consistent against the class of asymmetric alternatives corresponding to $p^* \neq 0$. We point out that while asymmetry of a probability distribution implies that $p^* \neq 0$ for that distribution, the converse is not necessarily true; that is, there are asymmetric probability distributions for which $p^* = 0$ and against which, therefore, the two-sided test (3.80) based on V will not be consistent. Randles et al. (1980) note, however, that the class of distributions with this property is quite small. (The one-sided tests discussed in Comment 61 and defined by (3.88) and (3.90) are consistent against the classes of asymmetric alternatives corresponding to $p^* > 0$ and < 0, respectively.)

Properties

1. *Consistency.* See Comment 66 and Randles et al. (1980).
2. *Asymptotic Normality.* See Randles and Wolfe (1979, pp. 99–101).

Problems

101. Consider the percentage chromium data in Table 3.9. Test the hypothesis of symmetry versus general asymmetry. (Note that some of the necessary calculations for this test have been completed in Example 3.10.)

102. Show that a triple (Z_1, Z_2, Z_3) is a right triple if and only if $\bar{Z} = (Z_1 + Z_2 + Z_3)/3$ is greater than $\tilde{Z} = \text{median } (Z_1, Z_2, Z_3)$. (See also Comment 64.)

103. Consider the oxidant content data of Table 3.13. Test the hypothesis of symmetry versus general asymmetry.

104. What effect does the addition of a number b to each of the Z observations have on the value of the V (3.77) statistic? Comment on this as a desirable property for a test of population symmetry.

105. What effect does the multiplication of each of the Z observations by a number b have on the absolute value of the V (3.77) statistic? Comment on this as a desirable property for a test of population symmetry versus general asymmetry.

106. Consider the settling velocity data in Table 3.12. Test the hypothesis of symmetry against the alternative that the population of settling velocities is skewed to the right. (See Comment 61.)

107. Consider the Z differences for the beak-clapping data in Table 3.5. Test the hypothesis of symmetry against the alternative that the population of beak-clapping differences is skewed to the left. (See Comment 61.)

108. For n observations $Z_1, \ldots, Z_n$, what is the maximum possible value for T (3.74)? What is the minimum possible value for T? For $n = 4$, construct examples where these extreme values for T are achieved.

109. Consider the four observations $Z_1 = 2, Z_2 = 2.4, Z_3 = 3$, and $Z_4 = 3.5$. Compute the value of T (3.74) for these data. Indicate how to change only one of the sample observations in such a way that T achieves its maximum value (see Problem 108) on the altered data. Similarly, indicate how to change only one of the sample observations in such a way that T achieves its minimum value (see Problem 108) on the altered data.

BIVARIATE DATA

3.10 A DISTRIBUTION-FREE TEST FOR BIVARIATE SYMMETRY (HOLLANDER)

Data. We obtain $2n$ observations, two observations on each of n subjects.

Subject i	X_i	Y_i
1	X_1	Y_1
2	X_2	Y_2
⋮	⋮	⋮
n	X_n	Y_n

Assumptions

F1. The n bivariate observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are mutually independent.

F2. Each $(X_i, Y_i), i = 1, \ldots, n$, comes from the same bivariate population with joint distribution function $F(x, y)$.

Hypothesis

The null hypothesis of interest here is that the X and Y variables are exchangeable or, equivalently, that there is no treatment effect (see Comment 67). This hypothesis of exchangeability can be written as

$$H_0 : [F(x, y) = F(y, x), \text{for all } (x, y)]. \tag{3.92}$$

(Another way to state this exchangeability property is that the pairs (X, Y) and (Y, X) have the same joint bivariate distribution.)

Procedure

For each observation pair $(X_i, Y_i), i = 1, \ldots, n$, let

$$a_i = \min(X_i, Y_i), \quad b_i = \max(X_i, Y_i), \tag{3.93}$$

where, without loss of generality, we take $a_1 \leq a_2 \leq \cdots \leq a_n$. (We may simply relabel the n (X, Y) pairs so that the a's defined by (3.93) are increasing.) Define the n (observed) r values $r_1, r_2, \ldots, r_n$ by

$$r_i = \begin{cases} 1, & \text{if } X_i = a_i < b_i = Y_i \\ 0, & \text{if } X_i = b_i \geq a_i = Y_i. \end{cases} \tag{3.94}$$

That is, r_i is defined to be 1 if $X_i < Y_i$ and 0 if $X_i \geq Y_i$. (Note that the designation $r_i = 0$ for those cases where $X_i = Y_i$ is purely arbitrary, because such a tied situation makes no contribution to the overall test statistic to be defined by (3.98).)

Define the n^2 values d_{ij}, for $i, j = 1, \ldots, n$, by

$$d_{ij} = \begin{cases} 1, & \text{if } a_j < b_i \leq b_j \text{ and } a_i \leq a_j \\ 0, & \text{otherwise.} \end{cases} \tag{3.95}$$

For $j = 1, \ldots, n$, set

$$T_j = \sum_{i=1}^{n} s_i d_{ij}, \tag{3.96}$$

where d_{ij} is given by (3.95) and

$$s_i = 2r_i - 1. \tag{3.97}$$

Let A_{obs}, to be read as "A observed," be defined as

$$A_{\text{obs}} = \sum_{j=1}^{n} \frac{T_j^2}{n^2}. \tag{3.98}$$

Now, in addition to our observed r configuration $(r_1, \ldots, r_n)$ defined by (3.94), there are $2^n - 1$ other possible r configurations, corresponding to the cases in which each r_i can be either 0 or 1 and excluding the observed configuration (see Comment 68). For each of these $2^n - 1$ additional r configurations, calculate the corresponding value of A using (3.96) to (3.98). It is important to note that the d's defined by (3.95) remain the same for each of these additional calculations of A_{obs}.

Let

$$A^{(1)} \leq A^{(2)} \leq \cdots \leq A^{(2^n)} \tag{3.99}$$

denote the 2^n ordered values of the A's. (Note that A_{obs} will be one of these ordered A's.) Set

$$m = 2^n - [[2^n \alpha]], \tag{3.100}$$

where $[[2^n \alpha]]$ is the greatest integer less than or equal to $2^n \alpha$. Define M_1 to be the number of ordered values $A^{(1)} \leq \cdots \leq A^{(2^n)}$ that are greater than $A^{(m)}$ and take M_2 to be the number of $A^{(1)} \leq \cdots \leq A^{(2^n)}$ values that are equal to $A^{(m)}$, where $A^{(m)}$ is determined by (3.99) and (3.100).

To test H_0 (3.92), corresponding to the exchangeability of the X and Y variables, versus the general (two-sided) alternative that they are not exchangeable, corresponding to

$$H_1 : [F(x, y) \neq F(y, x) \text{ for at least one}(x, y)], \tag{3.101}$$

at the exact α level of significance,

$$\text{Reject } H_0 \text{ if } A_{\text{obs}} > A^{(m)}; \text{ do not reject } H_0 \text{ if } A_{\text{obs}} < A^{(m)}; \tag{3.102}$$

and

$$\text{If } A_{\text{obs}} = A^{(m)}, \text{ make a randomized decision to reject } H_0$$
$$\text{with probability } q \text{ and to not reject } H_0 \text{ with probability } 1 - q,$$

where

$$q = \frac{2^n \alpha - M_1}{M_2}. \tag{3.103}$$

The R program `HollBivSym` computes the statistic A given by (3.98). The `R` program `pHollBivSym` returns A and the exact P-value for $n \leq 20$. By default, if $n > 20$, the program uses a Monte Carlo approximation with 100,000 samples. The user can change both the number of Monte Carlo samples and the largest number of pairs for which he or she is willing to wait for the exact calculation (see Example 3.11).

Koziol (1979) developed a large-sample approximation but Kepner and Randles (1984) and Hilton and Gee (1997a) found that the large-sample approximation does not perform well. Thus one can use `pHollBivSym` understanding that it is exact for $n \leq 20$ but only approximate when $n \geq 21$.

Ties

No adjustment for ties is necessary. The calculation of A_{obs} (3.98) is well defined when ties occur. As a result, the procedure (3.102) handles ties automatically.

EXAMPLE 3.11 ***Inulin Clearance in Kidney Transplants.***

The data in Table 3.16 are a subset of the data obtained by Shelp et al. (1970) in a study of renal transplants. Part of their study dealt with living related donor kidneys and pertained to a comparison of clearance capacity of the donor and recipient after the transplant was done. Table 3.16 gives inulin clearance values for seven recipients and their corresponding donors. (We note that Assumption F2 may not be satisfied because the subjects are not homogeneous. They differ in various factors that may be pertinent to clearance, such as the basic disease, age when the transplant was performed, age of the donor, and sex of the donor. In order to illustrate the bivariate symmetry test, we neglect this heterogeneity of subjects.)

From (3.93) and Table 3.16, we find

$$a_1 = 61.4, a_2 = 63.3, a_3 = 63.7, a_4 = 67.1, a_5 = 77.3,$$
$$a_6 = 84.0, a_7 = 88.1 \tag{3.104}$$

$$b_1 = 70.8, b_2 = 89.2, b_3 = 65.8, b_4 = 88.0, b_5 = 87.3,$$
$$b_6 = 85.1, b_7 = 105.0. \tag{3.105}$$

Our observed r configuration is, from (3.94),

$$r_1 = 1, r_2 = 1, r_3 = 1, r_4 = 0, r_5 = 1, r_6 = 1, r_7 = 0. \tag{3.106}$$

Table 3.16 Inulin Clearance of Living Donors and Recipients of Their Kidneys

Patient[a]	Insulin clearance, ml/min	
	Recipient, X_i	Donor, Y_i
$1'$	61.4	70.8
$2'$	63.3	89.2
$3'$	63.7	65.8
$4'$	80.0	67.1
$5'$	77.3	87.3
$6'$	84.0	85.1
$7'$	105.0	88.1

Source: W. D. Shelp, F. H. Bach, W. A. Kisken, M. Newton, R. E. Rieselbach, and A. B. Weinstein (1970).

[a]The primes on the patient numbers indicate that our numbering is different from that in the study. We have renumbered so that the a's defined by (3.93) are in the order $a_1 < a_2 < a_3 < a_4 < a_5 < a_6 < a_7$.

We next calculate the $7^2 = 49$ d values using (3.95). We have

$$\begin{aligned}
&d_{11} = 1, d_{12} = 1, d_{13} = 0, d_{14} = 1, d_{15} = 0, d_{16} = 0, d_{17} = 0,\\
&d_{21} = 0, d_{22} = 1, d_{23} = 0, d_{24} = 0, d_{25} = 0, d_{26} = 0, d_{27} = 1,\\
&d_{31} = 0, d_{32} = 0, d_{33} = 1, d_{34} = 0, d_{35} = 0, d_{36} = 0, d_{37} = 0,\\
&d_{41} = 0, d_{42} = 0, d_{43} = 0, d_{44} = 1, d_{45} = 1, d_{46} = 0, d_{47} = 0,\\
&d_{51} = 0, d_{52} = 0, d_{53} = 0, d_{54} = 0, d_{55} = 1, d_{56} = 0, d_{57} = 0,\\
&d_{61} = 0, d_{62} = 0, d_{63} = 0, d_{64} = 0, d_{65} = 0, d_{66} = 1, d_{67} = 0,\\
&d_{71} = 0, d_{72} = 0, d_{73} = 0, d_{74} = 0, d_{75} = 0, d_{76} = 0, d_{77} = 1.
\end{aligned} \tag{3.107}$$

From (3.97) and (3.106), we have

$$s_1 = 1, s_2 = 1, s_3 = 1, s_4 = -1, s_5 = 1, s_6 = 1, s_7 = -1. \tag{3.108}$$

From (3.96), (3.107), and (3.108) we obtain

$$\begin{aligned}
T_1 &= d_{11}s_1 + d_{21}s_2 + d_{31}s_3 + d_{41}s_4 + d_{51}s_5 + d_{61}s_6 + d_{71}s_7\\
&= 1(1) + 0(1) + 0(1) + 0(-1) + 0(1) + 0(1) + 0(-1) = 1,\\
T_2 &= d_{12}s_1 + d_{22}s_2 + d_{32}s_3 + d_{42}s_4 + d_{52}s_5 + d_{62}s_6 + d_{72}s_7\\
&= 1(1) + 1(1) + 0(1) + 0(-1) + 0(1) + 0(1) + 0(-1) = 2,\\
T_3 &= d_{13}s_1 + d_{23}s_2 + d_{33}s_3 + d_{43}s_4 + d_{53}s_5 + d_{63}s_6 + d_{73}s_7\\
&= 0(1) + 0(1) + 1(1) + 0(-1) + 0(1) + 0(1) + 0(-1) = 1,\\
T_4 &= d_{14}s_1 + d_{24}s_2 + d_{34}s_3 + d_{44}s_4 + d_{54}s_5 + d_{64}s_6 + d_{74}s_7\\
&= 1(1) + 0(1) + 0(1) + 1(-1) + 0(1) + 0(1) + 0(-1) = 0,\\
T_5 &= d_{15}s_1 + d_{25}s_2 + d_{35}s_3 + d_{45}s_4 + d_{55}s_5 + d_{65}s_6 + d_{75}s_7\\
&= 0(1) + 0(1) + 0(1) + 1(-1) + 1(1) + 0(1) + 0(-1) = 0,\\
T_6 &= d_{16}s_1 + d_{26}s_2 + d_{36}s_3 + d_{46}s_4 + d_{56}s_5 + d_{66}s_6 + d_{76}s_7\\
&= 0(1) + 0(1) + 0(1) + 0(-1) + 0(1) + 1(1) + 0(-1) = 1,\\
T_7 &= d_{17}s_1 + d_{27}s_2 + d_{37}s_3 + d_{47}s_4 + d_{57}s_5 + d_{67}s_6 + d_{77}s_7\\
&= 0(1) + 1(1) + 0(1) + 0(-1) + 0(1) + 0(1) + 1(-1) = 0.
\end{aligned} \tag{3.109}$$

Equation (3.98) then yields

$$\begin{aligned}
A_{\text{obs}} &= \frac{T_1^2 + T_2^2 + T_3^2 + T_4^2 + T_5^2 + T_6^2 + T_7^2}{49}\\
&= \frac{1 + 4 + 1 + 0 + 0 + 1 + 0}{49} = \frac{7}{49}.
\end{aligned} \tag{3.110}$$

Now, to apply the exact procedure given by (3.102), we need to obtain the additional $2^7 - 1 = 127$ A values, corresponding to the other 127 possible r configurations. The 128 possible r configurations, including r observed, which is given by (3.106), are displayed in Table 3.17.

The parenthetical values to the right of each r configuration in Table 3.17 are the corresponding values of $49A$. These values are calculated in the same way we calculated A_{obs} in (3.107) to (3.110). The s's corresponding to (3.108) must be recalculated for each r configuration for use in the T_j equations, but the d's remain the same for each calculation. The ordered A's defined by (3.99) are $A^{(1)} \leq \cdots \leq A^{(128)}$.

We now list the ordered values of $49A : 49A^{(1)} = \cdots = 49A^{(8)} = 3, 49A^{(9)} = \cdots = 49A^{(40)} = 7, 49A^{(41)} = \cdots = 49A^{(88)} = 11, 49A^{(89)} = \cdots = 49A^{(120)} = 15, 49A^{(121)} = \cdots = 49A^{(128)} = 19$.

Let us illustrate the $\alpha = \frac{8}{128} = .0625$ test. The value of m (3.100) is

$$m = 2^7 - \left[\left[2^7\left(\frac{8}{128}\right)\right]\right] = 128 - 8 = 120,$$

and thus $A^{(m)} = A^{(120)} = \left(\frac{15}{49}\right)$. We then have

$$M_1 = \text{number of } A \text{ values greater than } \frac{15}{49} = 8,$$

$$M_2 = \text{number of } A \text{ values equal to } \frac{15}{49} = 32,$$

and

$$q_1 = \left(128\left(\frac{8}{128}\right) - 8\right) = 0.$$

Hence, procedure (3.102) reduces to, at the $\alpha = .0625$ level,

$$\text{Reject } H_0 \text{ if } A_{\text{obs}} > \frac{15}{49}. \tag{3.111}$$

As $A_{\text{obs}} = \frac{7}{49}$, we do not reject the hypothesis of bivariate symmetry at the .0625 level. Furthermore, because there are 120 configurations (including the one corresponding to A_{obs}) that yield a value greater than or equal to A_{obs}, the lowest level at which we can reject using a nonrandomized test based on A is $\frac{120}{128} = .9375$.

To perform the A test using R let

$$x < \text{-}c(61.4, 63.3, 63.7, 80, 77.3, 84, 105)$$

$$y < \text{-}c(70.8, 89.2, 65.8, 67.1, 87.3, 85.1, 88.1)$$

Then `HollBivSym(x, y)` returns $A = .1429$ and `pHollBivSym(x, y)` returns $A = .1429$ and the P-value .9375.

The command `pHollBivSym(x, y, approx=7)` signifies that the user is willing to use the approximate P-value when $n \geq 7$. Furthermore, the command `pHollBivSym(x, y, approx=7, n.mc=200,000)` changes the default from 100,000 to 200,000 Monte Carlo samples.

Table 3.17 The 128 Possible r Configurations and Corresponding Values of 49A

r_1	r_2	r_3	r_4	r_5	r_6	r_7	(49A)
1	1	1	1	1	1	1	(19)
1	1	1	1	1	1	0	(15)
1	1	1	1	1	0	1	(19)
1	1	1	1	1	0	0	(15)
1	1	1	1	0	1	1	(15)
1	1	1	1	0	1	0	(11)
1	1	1	1	0	0	1	(15)
1	1	1	1	0	0	0	(11)
1	1	1	0	1	1	1	(11)
1	1	1	0	1	0	1	(11)
1	1	1	0	1	1	0[a]	(7)
1	1	1	0	1	0	0	(7)
1	1	1	0	0	1	1	(15)
1	1	1	0	0	0	1	(15)
1	1	1	0	0	1	0	(11)
1	1	1	0	0	0	0	(11)
1	1	0	1	1	1	1	(19)
1	1	0	1	1	1	0	(15)
1	1	0	1	1	0	1	(19)
1	1	0	1	1	0	0	(15)
1	1	0	1	0	1	1	(15)
1	1	0	1	0	1	0	(11)
1	1	0	1	0	0	1	(15)
1	1	0	1	0	0	0	(11)
1	1	0	0	1	1	1	(11)
1	1	0	0	1	0	1	(11)
1	1	0	0	1	1	0	(7)
1	1	0	0	1	0	0	(7)
1	1	0	0	0	1	1	(15)
1	1	0	0	0	0	1	(15)
1	1	0	0	0	1	0	(11)
1	1	0	0	0	0	0	(11)
1	0	1	1	1	1	1	(11)
1	0	1	1	1	1	0	(15)
1	0	1	1	1	0	1	(11)
1	0	1	1	1	0	0	(15)
1	0	1	1	0	1	1	(7)
1	0	1	1	0	1	0	(11)
1	0	1	1	0	0	1	(7)
1	0	1	1	0	0	0	(11)
1	0	1	0	1	1	1	(3)
1	0	1	0	1	0	1	(3)
1	0	1	0	1	1	0	(7)
1	0	1	0	1	0	0	(7)
1	0	1	0	0	1	1	(7)
1	0	1	0	0	0	1	(7)
1	0	1	0	0	1	0	(11)
1	0	1	0	0	0	0	(11)
1	0	0	1	1	1	1	(11)
1	0	0	1	1	1	0	(15)
1	0	0	1	1	0	1	(11)
1	0	0	1	1	0	0	(15)
1	0	0	1	0	1	1	(7)
1	0	0	1	0	1	0	(11)

Table 3.17 (*Continued*)

r_1	r_2	r_3	r_4	r_5	r_6	r_7	(49*A*)
1	0	0	1	0	0	1	(7)
1	0	0	1	0	0	0	(11)
1	0	0	0	1	1	1	(3)
1	0	0	0	1	0	1	(3)
1	0	0	0	1	1	0	(7)
1	0	0	0	1	0	0	(7)
1	0	0	0	0	1	1	(7)
1	0	0	0	0	0	1	(7)
1	0	0	0	0	1	0	(11)
1	0	0	0	0	0	0	(11)
0	1	1	1	1	1	1	(11)
0	1	1	1	1	1	0	(7)
0	1	1	1	1	0	1	(11)
0	1	1	1	1	0	0	(7)
0	1	1	1	0	1	1	(7)
0	1	1	1	0	1	0	(3)
0	1	1	1	0	0	1	(7)
0	1	1	1	0	0	0	(3)
0	1	1	0	1	1	1	(11)
0	1	1	0	1	0	1	(11)
0	1	1	0	1	1	0	(7)
0	1	1	0	1	0	0	(7)
0	1	1	0	0	1	1	(15)
0	1	1	0	0	0	1	(15)
0	1	1	0	0	1	0	(11)
0	1	1	0	0	0	0	(11)
0	1	0	1	1	1	1	(11)
0	1	0	1	1	1	0	(7)
0	1	0	1	1	0	1	(11)
0	1	0	1	1	0	0	(7)
0	1	0	1	0	1	1	(7)
0	1	0	1	0	1	0	(3)
0	1	0	1	0	0	1	(7)
0	1	0	1	0	0	0	(3)
0	1	0	0	1	1	1	(11)
0	1	0	0	1	0	1	(11)
0	1	0	0	1	1	0	(7)
0	1	0	0	1	0	0	(7)
0	1	0	0	0	1	1	(15)
0	1	0	0	0	0	1	(15)
0	1	0	0	0	1	0	(11)
0	1	0	0	0	0	0	(11)
0	0	1	1	1	1	1	(11)
0	0	1	1	1	1	0	(15)
0	0	1	1	1	0	1	(11)
0	0	1	1	1	0	0	(15)
0	0	1	1	0	1	1	(7)
0	0	1	1	0	1	0	(11)
0	0	1	1	0	0	1	(7)
0	0	1	1	0	0	0	(11)
0	0	1	0	1	1	1	(11)
0	0	1	0	1	0	1	(11)
0	0	1	0	1	1	0	(15)
0	0	1	0	1	0	0	(15)
0	0	1	0	0	1	1	(15)

(*continued*)

Table 3.17 (*Continued*)

r_1	r_2	r_3	r_4	r_5	r_6	r_7	(49*A*)
0	0	1	0	0	0	1	(15)
0	0	1	0	0	1	0	(19)
0	0	1	0	0	0	0	(19)
0	0	0	1	1	1	1	(11)
0	0	0	1	1	1	0	(15)
0	0	0	1	1	0	1	(11)
0	0	0	1	1	0	0	(15)
0	0	0	1	0	1	1	(7)
0	0	0	1	0	1	0	(11)
0	0	0	1	0	0	1	(7)
0	0	0	1	0	0	0	(11)
0	0	0	0	1	1	1	(11)
0	0	0	0	1	0	1	(11)
0	0	0	0	1	1	0	(15)
0	0	0	0	1	0	0	(15)
0	0	0	0	0	1	1	(15)
0	0	0	0	0	0	1	(15)
0	0	0	0	0	1	0	(19)
0	0	0	0	0	0	0	(19)

[a] Note that (1,1,1,0,1,1,0) was our observed configuration (see (3.106)).

Comments

67. *Motivation.* The hypothesis H_0 (3.92) is a natural one when an experimenter is testing for a treatment effect and finds it convenient (or necessary) to have the same subjects receive the treatment and also act as controls. Since (X_i, Y_i) then represent two observations on the same subject, it is unrealistic to assume that X_i and Y_i are independent. The hypothesis of no treatment effect is precisely H_0. Terms used by various workers to describe H_0 include exchangeability, interchangeability, and bivariate symmetry. (See Hollander (1971).)

68. *Conditional Nature of the Test.* The hypothesis H_0 implies that the r's defined by (3.94) are independent and identically distributed, each r_i assuming values 1 and 0 with probabilities $\frac{1}{2}$ and $\frac{1}{2}$, respectively. This leads to a conditional distribution P_c that assigns probability $\left(\frac{1}{2}\right)^n$ to each of the A-values associated with each of the possible 2^n r configurations. (In the foregoing statement, we implicitly distinguish between all A-values, although, as we see in Example 3.11, two different r's may yield the same value of A.) The test defined by (3.102) investigates how large A_{obs} is with respect to this conditional distribution. For further information on conditional tests of this nature (which are known as *permutation tests*), see Hoeffding (1952), Box and Andersen (1955), Lehmann (1959), and Scheffé (1959).

69. *Alternative Computation of the d's.* In computing the d's defined by (3.95), life can be made easier by observing:

    (i) $d_{ij} = 0$ for all j, if $a_i = b_i$.
    (ii) $d_{ii} = 1$ if $a_i \neq b_i$, and $d_{ii} = 0$ if $a_i = b_i$.
    (iii) When $i > j$, if $a_i \neq a_j$, then $d_{ij} = 0$.

70. *Parametric Representation of the Null Hypothesis* (H_0). Consider (3.92) and define

$$A^*(x,y) = P(X \le x \text{ and } Y \le y) - P(X \le y \text{ and } Y \le x). \tag{3.112}$$

The hypothesis H_0 (3.92) is true if and only if $A^*(x,y) = 0$ for all (x,y). The statistic (A/n) estimates the parameter

$$\Delta(F) = E_F\{A^*(X',Y')\}^2, \tag{3.113}$$

where (X',Y') is a random member from the underlying bivariate population with distribution F. We may view $A^*(x,y)$ as a measure of the deviation from H_0 at the point (x,y) and $\Delta(F)$ (3.113) as the average value of the square of this deviation.

71. *Consistency: Comparison of A Test and Signed Rank Test.* The A test was designed by Hollander (1971) to detect a broad class of alternatives to the hypothesis of no treatment effect. Thus, although the A test will detect alternatives of the form associated with nonzero ($\theta \neq 0$) treatment effects as discussed for paired replicates data in Section 3.1, it will also be sensitive to differences in dispersion in the (marginal) X and Y populations, as well as to more general deviations from H_0. Of course, a price must be paid for this more general type of protection; namely, we cannot expect the A test to have power as good as that of, say, the Wilcoxon signed rank test (3.6) when the location model of Section 3.1 is true, because the signed rank test is directed to location changes. On the other hand, there are many alternatives to H_0 for which the signed rank test will have power remaining at α (for any sample size), whereas the A test will have power tending to 1 (as n tends to infinity). In fact, under mild conditions on the nature of the underlying bivariate population F, the A test is consistent when H_0 is false.

72. *Other Nonparametric Tests.* Other nonparametric tests for bivariate symmetry are proposed in Kepner and Randles (1984). See Randles and Kepner (1984) and Hilton and Gee (1997a) for power comparisons of A versus competitors when F is bivariate normal and when F is bivariate exponential. See Hilton and Gee (1997b) for an efficient algorithm for conducting the exact test based on A.

Properties

1. *Consistency.* The test defined by (3.102) is consistent against populations for which the parameter $\Delta(F)$ defined by (3.113) is positive. For conditions on F insuring that $\Delta(F)$ will be positive, see Hollander (1971).
2. *Asymptotic Distribution.* See Koziol (1979).

Problems

110. Cain, Mayer, and Jones (1970) have studied albumin and fibrinogen metabolism using the carbonate-^{14}C method to measure the synthetic rate of liver-produced plasma proteins before and after a 13-day course of prednisolone. The eight subjects were patients with hepatocellular

Table 3.18 Intravascular Albumin Pool Before and After Prednisolone

Patient	Intravascular albumin pool (g)	
	Before, X_i	After, Y_i
1	74.4	83.8
2	100.0	97.5
3	82.5	77.4
4	84.3	87.2
5	91.4	116.2
6	92.8	88.2
7	104.2	115.1
8	58.3	50.5

Source: G. D. Cain, G. Mayer, and E. A. Jones (1970).

disease as established by needle biopsy. Part of the study was related to changes in the intravascular albumin pool. Table 3.18 is based on a subset of the Cain–Mayer–Jones data.

Use R to find the exact conditional P-value for these data achieved by the test based on A.

111. Consider the intravascular albumin data in Table 3.18. Use R to determine an approximate P-value for these data based on the A test. Compare with the exact conditional P-value for these data as found in Problem 110.

112. Verify directly, or illustrate with a numerical example, remarks (i)–(iii) of Comment 69.

113. Consider the immunoreactive insulin blood-level data of Table 3.3. Use R to find the exact conditional P-value obtained by the A test for those data.

114. Consider the immunoreactive insulin blood-level data of Table 3.3. Use R to find an approximate P-value and compare it with the exact conditional P-value for these data as found in Problem 113.

115. Calculate $\Delta(F)$ for the bivariate population having joint distribution function $F(x, y) = 0$ for $x < 0, y < 0$; $= xy^2$ for $0 \le x \le 1, 0 \le y \le 1$; $= 1$ for $x > 1, y > 1$.

3.11 EFFICIENCIES OF PAIRED REPLICATES AND ONE-SAMPLE LOCATION PROCEDURES

Recall the normal theory one-sample t-test based on the statistic

$$V = \frac{\sqrt{n}\bar{Z}}{S_z}, \tag{3.114}$$

where $\bar{Z} = \sum_{i=1}^{n} Z_i/n$ and $S_z^2 = \sum_{i=1}^{n} (Z_i - \bar{Z})^2/(n-1)$. The Pitman asymptotic relative efficiency of the one-sample test procedure (one- or two-sided) based on the signed rank statistic T^+ (3.3) with respect to the corresponding normal theory test based on V is

$$e(T^+, V) = 12\sigma_F^2 \left\{ \int_{-\infty}^{\infty} f^2(u)du \right\}^2, \tag{3.115}$$

where σ_F^2 is the variance of the common (continuous and symmetric) distribution $F(\cdot)$ of $Z_1, \ldots, Z_n$ and $f(\cdot)$ is the probability density function corresponding to $F(\cdot)$. The

parameter $\int_{-\infty}^{\infty} f^2(u)du$ is the area under the curve associated with $f^2(\cdot)$, the square of the common probability density function.

The expression in (3.115) was first obtained by Pitman (1948) in the context of hypothesis testing. Hodges and Lehmann (1963) showed that the same expression, $e(T^+, V)$, also pertains to the asymptotic relative efficiency of the point estimator $\hat{\theta}$ (see (3.23)) with respect to $\bar{\theta} = \bar{Z}$. Finally, Lehmann (1963c) established that (3.115) also provides the asymptotic relative efficiency of the confidence interval (or bound) for θ derived from T^+ (see Section 3.3) relative to the corresponding confidence interval (or bound) for θ associated with the one-sample t-test based on V (3.114).

Hodges and Lehmann (1956) demonstrated that within the class of continuous and symmetric $F(\cdot), e(T^+, V)$ is always at least .864. Thus, in this class of distributions, the most efficiency that can be lost when employing a procedure (test, point estimator, or confidence interval/bound) based on T^+ instead of the corresponding normal theory procedure associated with V (3.114) is about 14%. Even when $F(\cdot)$ is normal (the proper setting for procedures based on V), $e(T^+, V) = .955$ and there is only a minor loss (4.5%) in efficiency from using a T^+-based procedure rather than the optimal procedure based on V. On the other hand, $e(T^+, V)$ exceeds 1 for many populations and it can be infinite (e.g., when $F(\cdot)$ is Cauchy). Some values of $e(T^+, V)$ for selected $F(\cdot)$ are

F :	Normal	Uniform	Logistic	Double Exponential	Cauchy
$e(T^+, V)$:	.955	1.000	1.097	1.500	∞

(3.116)

The Pitman asymptotic relative efficiency of the one-sample test procedure (one- or two-sided) based on the sign statistic B (3.39) with respect to the corresponding normal theory test based on V (3.114) is

$$e(B, V) = 4\sigma_F^2 f^2(0), \tag{3.117}$$

where σ_F^2 is the variance and $f(\cdot)$ is the probability density function for the common (continuous and symmetric) distribution $F(\cdot)$ of the Z observations.

Pitman (1948) established the general efficiency expression in (3.117) for the hypothesis tests based on B and V, although Cochran (1937) had previously obtained the particular efficiency value of .637 for the case of an underlying (F) normal distribution. Hodges and Lehmann (1963) showed that the expression $e(B, V)$ also holds for the asymptotic relative efficiency of the point estimator $\tilde{\theta}$ (see (3.58)) with respect to $\bar{\theta} = \bar{Z}$, and the results in Lehmann (1963c) lead to the same conclusion for the confidence interval (or bound) for θ based on B (see Section 3.6) relative to the corresponding confidence interval (or bound) for θ associated with the one-sample t-test based on V (3.114).

Hodges and Lehmann (1956) found that within a certain class of populations, $e(B, V)$ is always at least $\frac{1}{3}$ and it can be infinite. Some values of $e(B, V)$ for selected $F(\cdot)$ are

F :	Normal	Uniform	Logistic	Double Exponential	Cauchy
$e(B, V)$:	.637	.333	.822	2.000	∞

(3.118)

We note that for the paired replicates problem, each Z is actually a difference of two observations. For the efficiency calculation, the common F in the parameters $e(T^+, V)$

and $e(B, V)$ is a distribution for a difference of two independent and identically distributed random variables. Since neither all continuous distributions nor all continuous and unimodal distributions can be distributions for such a difference, the lower bounds for $e(T^+, V)$ and $e(B, V)$ for paired replicates data are obtained over smaller classes of distributions than for the one-sample data. In particular, in the paired case, Hollander (1967a) proved that the lower bound of .864 for $e(T^+, V)$ is no longer attainable. Similarly, Puri and Sen (1968) demonstrated that the lower bound of $\frac{1}{3}$ for $e(B, V)$ is not attainable in the paired case.

For the paired replicates data, the values of $e(T^+, V)$ and $e(B, V)$ remain the same as given in expressions (3.116) and (3.118), respectively, for an underlying (F) normal, logistic, double exponential, or Cauchy distribution. However, the uniform distribution cannot be a distribution for a difference of two independent and identically distributed random variables (see Puri and Sen (1968)).

We do not know of any results for the asymptotic efficiencies of the Randles et al. test for distributional symmetry (Section 3.9) or Hollander's bivariate symmetry test (Section 3.10).

Chapter 4

The Two-Sample Location Problem

INTRODUCTION

In this chapter the data consist of two random samples, a sample from the control population and an independent sample from the treatment population. On the basis of these samples, we wish to investigate the presence of a treatment effect that results in a shift of location. The basic hypothesis is that of no treatment effect; that is, the samples can be thought of as a single sample from one population.

Section 4.1 presents a distribution-free rank sum test for the hypothesis of no treatment effect; Section 4.2, a point estimator associated with the rank sum statistic; and Section 4.3, a related distribution-free confidence interval that emanates from the rank sum test. The basic model for Sections 4.1, 4.2 and 4.3 assumes the populations differ only by a location shift. In Section 4.4 we present a test for location differences that allows the population dispersions to differ. Section 4.5 considers the asymptotic relative efficiencies for translation alternatives of the procedures based on the rank sum statistic with respect to their normal theory counterparts based on sample means.

Data. We obtain $N = m + n$ observations $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$.

Assumptions

- **A1.** The observations $X_1, \ldots, X_m$ are a random sample from population 1; that is, the X's are independent and identically distributed. The observations $Y_1, \ldots, Y_n$ are a random sample from population 2; that is, the Y's are independent and identically distributed.
- **A2.** The X's and Y's are mutually independent. Thus, in addition to assumptions of independence within each sample, we also assume independence between the two samples.
- **A3.** Populations 1 and 2 are continuous populations.

4.1 A DISTRIBUTION-FREE RANK SUM TEST (WILCOXON, MANN AND WHITNEY)

Hypothesis

Let F be the distribution function corresponding to population 1 and let G be the distribution function corresponding to population 2.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

The null hypothesis is

$$H_0 : F(t) = G(t), \quad \text{for every } t. \tag{4.1}$$

The null hypothesis asserts that the X variable and the Y variable have the same probability distribution, but the common distribution is not specified.

The alternative hypothesis in a two-sample location problem typically specifies that Y tends to be larger (or smaller) than X. One model that is useful to describe such alternatives is the translation model—also called the *location-shift model*. The location-shift model is

$$G(t) = F(t - \Delta), \quad \text{for every } t. \tag{4.2}$$

Model (4.2) says that population 2 is the same as population 1 except it is shifted by the amount Δ. Another way of writing this is

$$Y \stackrel{d}{=} X + \Delta$$

where the symbol $\stackrel{d}{=}$ means "has the same distribution as." The parameter Δ is called the *location shift*. It is also known as the *treatment effect*. If X is a randomly selected value from population 1, the control population, and Y is a randomly selected value from population 2, the treatment population, then Δ is the expected effect due to the treatment. If Δ is positive, it is the expected increase due to the treatment, and if Δ is negative, it is the expected decrease due to the treatment. If the mean $E(X)$ of population 1 exists, then letting $E(Y)$ denote the mean of population 2,

$$\Delta = E(Y) - E(X),$$

the difference in population means. In terms of the location-shift model, the null hypothesis H_0 reduces to

$$H_0 : \Delta = 0,$$

the hypothesis that asserts the population means are equal or, equivalently, that the treatment has no effect.

We note that although we find it convenient to use the "treatment" and "control" terminology, many situations will arise in which we want to compare two random samples, neither one of which can be described as a sample from a control population. The procedures of this chapter are applicable even when there are no natural control or treatment designations.

Procedure

To compute the Wilcoxon two-sample rank sum statistic W, order the combined sample of $N = m + n$ X-values and Y-values from least to greatest. Let S_1 denote the rank of $Y_1, \ldots, S_n$ denote the rank of Y_n in this joint ordering. W is the sum of the ranks assigned to the Y-values. That is,

$$W = \sum_{j=1}^{n} S_j. \tag{4.3}$$

a. *One-Sided Upper-Tail Test*. To test

$$H_0 : \Delta = 0$$

versus

$$H_1 : \Delta > 0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } W \geq w_\alpha; \quad \text{otherwise do not reject,} \tag{4.4}$$

where the constant w_α is chosen to make the type I error probability equal to α. Values of w_α can be obtained from the R functions `pwilcox` and `qwilcox` as illustrated in Example 4.1 and Comment 3.

b. *One-Sided Lower-Tail Test*. To test

$$H_0 : \Delta = 0$$

versus

$$H_2 : \Delta < 0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } W \leq n(m+n+1) - w_\alpha; \quad \text{otherwise do not reject.} \tag{4.5}$$

c. *Two-Sided Test*. To test

$$H_0 : \Delta = 0$$

versus

$$H_3 : \Delta \neq 0$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } W \geq w_{\alpha/2} \text{ or if } W \leq n(m+n+1) - w_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{4.6}$$

The two-sided procedure given by (4.6) is the two-sided symmetric test with the $\alpha/2$ probability in each tail of the distribution.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of W, suitably standardized. We first need to know the mean and variance of W when the null hypothesis is true. When H_0 is true, the mean and variance of W are, respectively,

$$E_0(W) = \frac{n(m+n+1)}{2} \tag{4.7}$$

$$\text{var}_0(W) = \frac{mn(m+n+1)}{12}. \tag{4.8}$$

Comment 4 gives direct calculations of $E_0(W)$ and $\text{var}_0(W)$ in the special case where $m = 3$, $n = 2$. Comment 6 gives general derivations.

The standardized version of W is

$$W^* = \frac{W - E_0(W)}{\{\text{var}_0(W)\}^{1/2}} = \frac{W - \{n(m+n+1)/2\}}{\{mn(m+n+1)/12\}^{1/2}}. \tag{4.9}$$

When H_0 is true, W^* has, as $\min(m,n)$ tends to infinity, an asymptotic $N(0,1)$ distribution.

The normal theory approximation to procedure (4.4) is

$$\text{Reject } H_0 \text{ if } W^* \geq z_\alpha; \quad \text{otherwise do not reject.} \tag{4.10}$$

The normal theory approximation to procedure (4.5) is

$$\text{Reject } H_0 \text{ if } W^* \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{4.11}$$

The normal theory approximation to procedure (4.6) is

$$\text{Reject } H_0 \text{ if } |W^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{4.12}$$

Ties

If there are ties, give tied observations the average of the ranks for which those observations are competing. After computing W using average ranks, use procedure (4.4), (4.5), or (4.6). Now, however, the test is approximate rather than exact. (To get an exact test, even in the tied case, see Comment 5.)

When applying the large-sample approximation, the following modification should be made. What there are ties, the null mean of W is unaffected, but the null variance is reduced to

$$\text{var}_0(W) = \frac{mn}{12}\left[m+n+1 - \frac{\sum_{j=1}^{g}(t_j-1)t_j(t_j+1)}{(m+n)(m+n-1)}\right], \tag{4.13}$$

or, equivalently,

$$\text{var}_0(W) = \frac{mn(N+1)}{12} - \left\{\frac{mn}{12N(N-1)} \cdot \sum_{j=1}^{g}(t_j-1)t_j(t_j+1)\right\}. \tag{4.14}$$

In displays (4.13) and (4.14) g denotes the number of tied groups and t_j is the size of tied group j. Furthermore, an untied observation is considered to be a tied "group" of size 1. In particular, if there are no tied observations, $g = N$, $t_j = 1$ for $j = 1, \ldots, N$, and thus each term of the form $(t_j - 1)(t_j)(t_j + 1)$ reduces to 0 and $\text{var}_0(W)$ reduces to $mn(m+n+1)/12$, the null variance of W when there are no ties. Note also that the term in curly braces on the right-hand side of display (4.14) measures the reductions in the null variance due to the presence of ties.

To apply the large-sample approximation when ties are present, compute W using average ranks and compute

$$W^* = \frac{W - [n(m+n+1)/2]}{\{\text{var}_0(W)\}^{1/2}},$$

where $\text{var}_0(W)$ is given by display (4.13). With this modified value of W^*, approximations (4.10), (4.11), and (4.12) can be applied.

The Mann–Whitney Statistic

Procedures (4.4), (4.5), and (4.6) based on the rank sum statistic can also be performed using the Mann–Whitney statistic. Let

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i, Y_j), \tag{4.15}$$

where

$$\phi(X_i, Y_j) = \begin{cases} 1 & \text{if } X_i < Y_j, \\ 0 & \text{otherwise.} \end{cases}$$

The statistic U counts the number of "X before Y" predecessors. It is easy to show (see Comment 7) that

$$W = U + \frac{n(n+1)}{2}. \tag{4.16}$$

Thus tests based on W and U are equivalent. For example, the one-sided test given by (4.4) that rejects if $W \geq w_\alpha$ is equivalent to the one-sided test that rejects if $U \geq u_\alpha$ where u_α is the upper α percentile point of the null distribution of U. From (4.16) it follows that $w_\alpha = u_\alpha + (n(n+1)/2)$. Some textbooks and some software find it more convenient to use U rather than W. For example, the R functions `wilcox.test`, `pwilcox`, and `qwilcox`, illustrated in Comment 3 and Example 4.1, utilize

$$U' = U - mn, \tag{4.17}$$

the number of Y before X predecessors. The possible values of U and U' are 0, 1, . . . , mn. Furthermore, when H_0 is true, the mean and variance of U and U' are, respectively,

$$E_0(U) = E_0(U') = mn/2 \tag{4.18}$$

$$\text{Var}_0(U) = \text{Var}_0(U') = mn(m+n+1)/12. \tag{4.19}$$

The null distributions of U and U' are symmetric about the mean $mn/2$.

EXAMPLE 4.1 ***Water Transfer in Placental Membrane.***

The data in Table 4.1 are a portion of the data obtained by Lloyd et al. (1969). Among other things, these authors investigated whether there is a difference in the transfer of tritiated water (water containing tritium, a radioactive isotope of hydrogen) across the tissue layers in the term human chorioamnion (a placental membrane) and in the human chorioamnion between 3- and 6-months' gestational age. The objective measure used was the permeability constant Pd of the human chorioamnion to water. The tissues used for the study were obtained within 5 min of delivery from the placentas of healthy, uncomplicated pregnancies in the following two gestational age categories: (a) between 12 and 26 weeks following termination of pregnancy via abdominal hysterotomy (surgical incision of the uterus) for psychiatric indications and (b) term, uncomplicated vaginal deliveries. Tissues from 10 term pregnancies and five terminated pregnancies were used in the experiment. Table 4.1 gives the average permeability constant (in units of 10^{-4} cm/s) for six measurements on each of the 15 tissues in the study.

Table 4.1 Tritiated Water Diffusion Across Human Chorioamnion

$Pd(10^{-4}$ cm/s)	
At term	12–26 Weeks gestational age
0.80	1.15
0.83	0.88
1.89	0.90
1.04	0.74
1.45	1.21
1.38	
1.91	
1.64	
0.73	
1.46	

Source: S.J. Lloyd, K.D. Garlid, R.C. Reba and A.E. Seeds (1969).

In this example, the alternative of interest is greater permeability of the human chorioamnion for the term pregnancy. Thus, if we let X correspond to the Pd values of tissues from term pregnancies and Y to the Pd values of tissues from terminated pregnancies, we perform a one-sided test designed to detect the alternative $\Delta < 0$.

We list the combined sample in increasing order to facilitate the joint ranking. The ranks are given in parentheses

X	Y	X	X	Y	Y	X	Y
0.73	0.74	0.80	0.83	0.88	0.90	1.04	1.15
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Y	X	X	X	X	X	X	
1.21	1.38	1.45	1.46	1.64	1.89	1.91	
(9)	(10)	(11)	(12)	(13)	(14)	(15)	

We see that the Y-ranks are 2, 5, 6, 8, and 9 and thus

$$W = 2 + 5 + 6 + 8 + 9 = 30.$$

From (4.16) we find

$$U = W - n(n+1)/2 = 30 - 15 = 15.$$

The R function `wilcox.test` computes the value of $U' = U - mn$ and gives the P-value corresponding to U'. In the R output, U' is denoted by W, which is not to be confused with our use of W for the sum of the Y-ranks. Since $U = 15$, $U' = 50 - 15 = 35$, and that is the value (labeled W) provided by `wilcox.test`. If you let

```
at.term<−c (.80, .83, 1.89, 1.04, 1.45, 1.38, 1.91, 1.64, .73, 1.46),
gest.age<−c (1.15, .88, .90, .74, 1.21)
```

and perform `wilcox.test (at.term, gest.age, alternative="t", conf. int=T)` you get the two-sided P-value .2544 and the one-sided P-value for the test of $\Delta < 0$ is .127. This one-sided P-value is also obtained using `wilcox.test (x, y, alt="g")`.

The function `wilcox.test` not only performs the test but also provides the Hodges–Lehmann estimator of Section 4.2 and the confidence interval of Section 4.3.

There is no need to perform the large-sample approximation because we have the result for the exact test. Nevertheless, it is informative to see how close the P-value given by the large-sample approximation is to the exact P-value. From (4.9) we find $W^* = -1.225$ and the R function `pnorm` gives `pnorm` $(-1.225) = .110$. Thus, the one-sided P-value based on the large-sample approximation is .110 compared to the exact one-sided P-value of .127.

Both the exact test and the large-sample approximation indicate that there is no sufficiently strong evidence to support the hypothesis that human chorioamnion is more permeable to water transfer at term than at 12–26 weeks' gestational age.

EXAMPLE 4.2 ***Alcohol Intakes.***

Eriksen, Björnstad, and Götestam (1986) studied a social skills training program for alcoholics. Twenty-four "alcohol-dependent" male inpatients at an alcohol treatment center were randomly assigned to two groups. The control group patients were given a traditional treatment program. The treatment group patients were given the traditional treatment program plus a class in social skills training (SST). After being discharged from the program, each patient reported—in 2-week intervals—the quantity of alcohol consumed, the number of days prior to his first drink, the number of sober days, the days worked, the times admitted to an institution, and the nights slept at home. Reports were verified by other sources (wives or family members). (Such data can be unreliable!) One patient in the SST group, discovered to be an opiate addict, disappeared after discharge and submitted no reports. The remaining 23 patients reported faithfully for a year. The results for alcohol intake are given in Table 4.2. The ranks in the joint ranking of the 23 observations are given in parentheses in Table 4.2 and we find that the sum of the SST ranks is $W = 81$.

To test H_0 versus the alternative that the SST group tends to have lower alcohol intakes, we need to test $H_0 : \Delta = 0$ versus $H_2 : \Delta < 0$. We will use the R function `wilcox.test`. Let

```
x<−c(1042, 1617, 1180, 973, 1552, 1251, 1151, 1511, 728, 1079, 951, 1319)
y<−c(874, 389, 612, 798, 1152, 893, 541, 741, 1064, 862, 213)
```

Table 4.2 Alcohol Intake for 1 Year (Centiliter of Pure Alcohol)

Control		SST	
1042	(13)	874	(9)
1617	(23)	389	(2)
1180	(18)	612	(4)
973	(12)	798	(7)
1552	(22)	1152	(17)
1251	(19)	893	(10)
1151	(16)	541	(3)
1511	(21)	741	(6)
728	(5)	1064	(14)
1079	(15)	862	(8)
951	(11)	213	(1)
1319	(20)		

Source: L. Eriksen, S. Björnstad, and K. G. Götestam (1986).

The alternative that the SST groups tend to have lower alcohol intakes would be reflected in small W values or, in terms of the function `wilcox.test`, which uses $U' = U\text{-}mn$, large values of U. Thus we use `wilcox.test (x, y, alt="g")`. This yields a one-sided P-value of .00049. Thus there is strong evidence that the SST class in combination with the traditional treatment program tends to lower alcohol intake in alcoholics.

Comments

1. *Motivation for the Test*. When Δ is greater than 0, the Y-values will tend to be larger than the X-values, and thus the Y-ranks will tend to be larger than the X-ranks. Hence the value of W will tend to be large. This suggests rejecting H_0 in favor of $\Delta > 0$ for large values of W and motivates procedure (4.4). An analogous motivation leads to procedure (4.5).

   The test based on W was introduced by Wilcoxon in 1945. An equivalent test based on the number of X before Y occurrences in the jointly ordered sample (see Comment 7) was proposed by Mann and Whitney (1947). Kruskal (1957) gives a detailed history of the Wilcoxon statistic dating back to 1914.

2. *Testing Δ is Equal to Some Specified Nonzero Value*. Procedures (4.4), (4.5), and (4.6) and the corresponding large-sample approximations given by procedures (4.10), (4.11), and (4.12) are for testing if Δ is equal to zero. To test $\Delta = \Delta_0$, where Δ_0 is some specified nonzero number, subtract Δ_0 from each Y-value to form a pseudosample, namely, $Y_1' = Y_1 - \Delta_0$, $Y_2' = Y_2 - \Delta_0, \ldots, Y_n' = Y_n - \Delta_0$. Then compute W as the sum of the Y'-ranks in the joint ranking of the m X-values and the n Y'-values. Then procedures (4.4), (4.5), and (4.6), and their corresponding large-sample approximations given by displays (4.10), (4.11), and (4.12), can be applied as described earlier.

3. *Derivation of the Distribution of W under H_0 (No-Ties Case)*. Assume that the underlying distribution under H_0 is continuous so that ties have probability zero of occurring. Then under H_0, all $\binom{N}{n}$ possible assignments for the Y-ranks are equally likely, each having probability $1/\binom{N}{n}$. For example, in the case of $m = 3$, $n = 2$, the $\binom{5}{2} = 10$ possible outcomes for the ranks attained by the two Y observations and the corresponding values of W are given in the following table.

Y-ranks	Probability	W
1, 2	$\frac{1}{10}$	3
1, 3	$\frac{1}{10}$	4
1, 4	$\frac{1}{10}$	5
1, 5	$\frac{1}{10}$	6
2, 3	$\frac{1}{10}$	5
2, 4	$\frac{1}{10}$	6
2, 5	$\frac{1}{10}$	7
3, 4	$\frac{1}{10}$	7
3, 5	$\frac{1}{10}$	8
4, 5	$\frac{1}{10}$	9

Thus, for example, under H_0, the probability is $\frac{2}{10}$ that W is equal to 5, because $W = 5$ when either Y-rank configuration {1, 4} or Y-rank configuration {2, 3} occurs, each has a $\frac{1}{10}$ chance of occurring (and, of course, they cannot both occur simultaneously). Simplifying, we obtain the null distribution.

Possible value of W	Probability of value
3	.1
4	.1
5	.2
6	.2
7	.2
8	.1
9	.1

Thus, for example, under H_0, the probability that W is greater than or equal to 7 is

$$\begin{aligned} P_0(W \geq 7) &= P_0(W = 7) + P_0(W = 8) + P_0(W = 9) \\ &= .2 + .1 + .1 = .4. \end{aligned}$$

The R command `pwilcox` enumerates the null distribution of U', which is the same as the null distribution of U. The command `pwilcox(0:6,2,3, lower.tail=T)` gives the lower tail probabilities, that is, the cumulative distribution, corresponding to the six possible values of U. The output is .1, .2, .4, .6, .8, .9, and 1.0; that is,

$$P(U < 0) = .1,\ P(U < 1) = .2,\ P(U < 2) = .4,$$
$$P(U < 3) = .6,\ P(U < 4) = .8,\ P(U < 5) = .9,\ P(U < 6) = 1.0.$$

Recall that $W = U + n(n+1)/2 = U + 3$ to verify that this output agrees with results for W.

Observe that we have derived the null distribution of W (and equivalently U) without specifying the common underlying continuous distribution of the two populations. This is why the procedures based on W are called *distribution-free procedures*. From the null distribution of W, we can determine the critical values w_α and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the common underlying distribution.

4. *Calculation of the Mean and Variance of W under the Null Hypothesis*. In displays (4.7) and (4.8), we presented formulas for the mean and variance of W when the null hypothesis is true. In this comment, we illustrate a direct calculation of $E_0(W)$ and $\text{var}_0(W)$ in a particular case. We use the null distribution of W obtained in Comment 3. (Later, in Comment 6, we present general derivations of $E_0(W)$ and $\text{var}_0(W)$.) Comment 3 treated the case where $m = 3$, $n = 2$. The null mean of W, $E_0(W)$, is obtained by multiplying each possible value of W with its probability under H_0. Thus

$$E_0(W) = 3(.1) + 4(.1) + 5(.2) + 6(.2) + 7(.2) + 8(.1) + 9(.1) = 6.$$

This is in agreement with what we obtain using (4.7), namely,

$$E_0(W) = \frac{n(m+n+1)}{2} = \frac{2(3+2+1)}{2} = 6.$$

A check on the expression for $\text{var}_0(W)$ is also easily performed. Recall

$$\text{var}_0(W) = E_0(W^2) - \{E_0(W)\}^2,$$

where $E_0(W^2)$, the second moment of the distribution of W, is again obtained by multiplying possible values (in this case, values of W^2) by the corresponding probabilities under H_0. We find

$$E_0(W^2) = 9(.1) + 16(.1) + 25(.2) + 36(.2) + 49(.2) + 64(.1) + 81(.1) = 39.$$

Thus

$$\text{var}_0(W) = 39 - (6)^2 = 3.$$

This agrees with what we obtain using (4.8) directly, namely,

$$\text{var}_0(W) = \frac{3(2)(3+2+1)}{12} = 3.$$

5. *Exact Conditional Distribution of W with Ties*. To get an exact test in the presence of ties, we consider all $\binom{N}{n}$ possible assignments of the N observations with n observations serving as Y's and m observations serving as X's. For each such assignment, we compute a value of W. Then we see how extreme our observed value of W is in this "built-up" conditional distribution. To keep computations simple, we illustrate for the $n = 2$, $m = 3$ data

Y	X	X	Y	X
.7	1.2	1.7	1.7	2.8
(1)	(2)	(3.5)	(3.5)	(5)

Note that the two tied 1.7 values get the average ranks 3.5. We then find that W, the sum of the Y-ranks, is

$$W = 1 + 3.5 = 4.5.$$

To assess the significance of W, we obtain a conditional distribution by considering the $\binom{5}{2} = 10$ possible assignments of the observations

$$.7, \quad 1.2, \quad 1.7, \quad 1.7, \quad 2.8,$$

to serve as three X-values and two Y-values, or, equivalently, the 10 possible assignments of the ranks

$$1, \quad 2, \quad 3.5, \quad 3.5, \quad 5,$$

to serve as three X-ranks and two Y-ranks. These 10 assignments and the corresponding values of W are shown in the following table.

Y-ranks	Probability	W
1, 2	$\frac{1}{10}$	3
1, 3.5	$\frac{1}{10}$	4.5
1, 3.5	$\frac{1}{10}$	4.5
1, 5	$\frac{1}{10}$	6
2, 3.5	$\frac{1}{10}$	5.5
2, 3.5	$\frac{1}{10}$	5.5
2, 5	$\frac{1}{10}$	7
3.5, 3.5	$\frac{1}{10}$	7
3.5, 5	$\frac{1}{10}$	8.5
3.5, 5	$\frac{1}{10}$	8.5

Then, for the tail probabilities, we obtain

$$P_0(W \geq 8.5) = \tfrac{2}{10},$$
$$P_0(W \geq 7) = \tfrac{4}{10},$$
$$P_0(W \geq 6) = \tfrac{5}{10},$$
$$P_0(W \geq 5.5) = \tfrac{7}{10},$$
$$P_0(W \geq 4.5) = \tfrac{9}{10},$$
$$P_0(W \geq 3) = 1.$$

This distribution is called the *conditional distribution* or the permutation distribution of W. For the particular observed value $W = 4.5$, we see $P_0(W \leq 4.5) = 1 - P_0(W \geq 5.5) = \frac{3}{10}$, and such a value would not indicate a deviation from H_0.

The R package coin contains the program `wilcox_test` that computes the P-value attained by referring W to its conditional distribution. For our $n = 2$, $m = 3$ data, let x<-c(1.2, 1.7, 2.8) and y<-c(.7, 1.7). Then the command `wilcox_test(c(x,y)~factor(c(0, 0, 0, 1, 1)), distribution = "exact", alt = "g")` yields the P-value .3 agreeing with what we obtained by enumeration.

6. *Large-Sample Approximation*. The statistic W/n is the average of the Y-ranks. All $\binom{N}{n}$ possible outcomes of the Y-ranks are equally likely under H_0. It follows that the null distribution of W/n is the same as the distribution of the sample mean of a random sample of size n drawn without replacement from the finite population $\{1, 2, \ldots, N\}$ of the first N integers. Next, we use results (i) and (ii), which are basic results from finite population theory concerning the mean and variance of the distribution of the sample mean of a sample of size n drawn without replacement from a finite population of N elements:

   (i) The mean is equal to the mean μ_{pop} of the finite population.
   (ii) The variance is equal to

$$\frac{\sigma^2_{\text{pop}}}{n} \times \frac{N-n}{N-1},$$

where σ^2_{pop} denotes the variance of the finite population and the factor $(N - n)/(N - 1)$ is the finite population correction factor.

For the finite population $\{1, 2, \ldots, N\}$, direct calculations establish

(iii) $\mu_{\text{pop}} = \dfrac{1 + 2 + \cdots + N}{N} = \dfrac{N + 1}{2}$,

(iv) $\sigma^2_{\text{pop}} = \dfrac{1}{N}\{1^2 + 2^2 + \cdots + N^2\} - \left(\dfrac{N + 1}{2}\right)^2 = \dfrac{(N - 1)(N + 1)}{12}$.

From (i), (ii), (iii), and (iv), we then obtain

$$E_0\left(\frac{W}{n}\right) = \frac{N + 1}{2},$$

$$\text{var}_0\left(\frac{W}{n}\right) = \frac{(N - 1)(N + 1)}{12n} \times \frac{N - n}{N - 1} = \frac{m(N + 1)}{12n},$$

and it follows that

$$\text{var}_0(W) = \frac{mn(N + 1)}{12}.$$

Asymptotic normality of

$$W^* = \frac{W - \frac{n(N+1)}{2}}{\sqrt{\frac{mn(N+1)}{12}}} = \frac{W - E_0(W)}{\sigma_0(W)}$$

follows from standard theory for the mean of a sample from a finite population (cf. Wilks, 1962, p. 268).

Asymptotic normality results are also obtainable under general alternatives. See, for example, Lehmann's (1951) extension of Hoeffding's (1948a) U-statistic theorem as stated and applied to the Wilcoxon statistic on pages 92–94 of Randles and Wolfe (1979).

7. *The Mann–Whitney U Statistic*. For testing the hypothesis $H_0 : \Delta = 0$, Mann and Whitney (1947) proposed the statistic U given by (4.15),

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(X_i, Y_j),$$

where

$$\phi(X_i, Y_j) = \begin{cases} 1, & \text{if } X_i < Y_j, \\ 0, & \text{otherwise.} \end{cases}$$

The statistic U can be computed as follows. For each pair of values X_i and Y_j, observe which is smaller. If the X_i value is smaller, score 1 for that pair; if the Y_j value is smaller, score 0 for that pair. Add up the 0's and 1's and call the sum U. Mann and Whitney showed that, in the case of no ties,

$$W = U + \frac{n(n + 1)}{2}. \tag{4.20}$$

This implies that tests based on U are equivalent to tests based on W.

To establish (4.20), write

$$W = \sum_{j=1}^{n} R(Y_j), \tag{4.21}$$

where $R(Y_j)$ denotes the rank of Y_j in the joint ranking of the $m + n$ X's and Y's. Since the rank of Y_j is equal to the number of X's less than Y_j plus the number of Y's less than Y_j plus 1, write

$$R(Y_j) = \sum_{i=1}^{m} \phi(X_i, Y_j) + \sum_{j'=1}^{n} \phi(Y_{j'}, Y_j) + 1. \tag{4.22}$$

Substituting (4.22) into (4.21) yields

$$W = \sum_{j=1}^{n} \sum_{i=1}^{m} \phi(X_i, Y_j) + \sum_{j=1}^{n} \sum_{j'=1}^{n} \phi(Y_{j'}, Y_j) + n. \tag{4.23}$$

In (4.23), the first term on the right is U. The second term on the right is equal to the number of Y's less than the smallest Y plus the number of Y's less than the second smallest Y plus ... plus the number of Y's less than the largest Y, that is $0 + 1 + \cdots + n - 1$. Thus

$$W = U + \{1 + 2 + \cdots + n - 1\} + n = U + \frac{n(n+1)}{2},$$

recalling that the sum of the first n integers is equal to $n(n+1)/2$.

We illustrate the computation of U for the diffusion data of Table 4.1. The first row below counts the number of X-values less than 1.15, the second row counts the number of X-values less than .88, the third row the number of X-values less than .90, the fourth row the number of X-values less than .74, and the fifth row the number of X-values less than 1.21. The counts are given in brackets; each count is simply the number of 1s in the particular row. U is then obtained by summing the counts, which is equivalent to summing all the 1s.

$$\begin{aligned}
U = {} & 1+1+0+1+0+0+0+0+1+0 && [4]\\
& +1+1+0+0+0+0+0+0+1+0 && [3]\\
& +1+1+0+0+0+0+0+0+1+0 && [3]\\
& +0+0+0+0+0+0+0+0+1+0 && [1]\\
& +1+1+0+1+0+0+0+0+1+0 && [4]\\
= {} & 15.
\end{aligned}$$

Recall that in Example 4.1 we found $W = 30$. We could alternatively obtain W by computing U as before and then using (4.16) to find

$$W = 15 + \frac{5(6)}{2} = 30.$$

There is a generalization of (4.16) that holds when there are ties. If W is computed using average ranks and U is computed via

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \phi^*(X_i, Y_j),$$

where

$$\phi^*(X_i, Y_j) = \begin{cases} 1, & \text{if } X_i < Y_j \\ \frac{1}{2}, & \text{if } X_i = Y_j \\ 0, & \text{if } X_i > Y_j, \end{cases}$$

then we still have $W = U + n(n+1)/2$. In other words, when there are ties, instead of scoring 1 if X is less than Y and 0 otherwise, compute U by scoring 1 if X is less than Y, $\frac{1}{2}$ if X equals Y, and 0 if X is greater than Y.

Bohn and Wolfe (1992, 1994) developed statistical procedures based on an analog of the Mann–Whitney statistic U for data obtained under the structure of ranked-set sampling. This form of data collection is a preferable alternative to simple random sampling when the actual sample measurements are costly and/or difficult to obtain, but ranking a small set of items is relatively easy and inexpensive.

Bohn (1996) provides a nice review of the general concept of ranked-set sampling, as well as an overview of the related nonparametric literature in this area of research. See Chapter 15 for more on ranked-set sampling.

8. *Symmetry of the Distribution of W under the Null Hypothesis*. When H_0 is true, the distribution of W is symmetric about its mean. This implies that when H_0 is true,

$$P(W \leq x) = P(W \geq n(m+n+1) - x), \tag{4.24}$$

for $x = n(n+1)/2, \ldots, n(2m+n+1)/2$.

Equation (4.24) is useful for converting upper-tail probabilities to lower-tail probabilities.

9. *Some Power Results for the Wilcoxon Test*. We consider the upper-tail α-level test of $H_0 : \Delta = 0$ versus $H_1 : \Delta > 0$ given by procedure (4.4). Suppose that the Y-population is the X-population shifted by an amount Δ, so that model (4.2) holds. Recall that

$$\text{Power} = \text{probability of rejecting } H_0, \text{ given that } H_0 \text{ is false.}$$

Then for Δ values "near" the null hypothesis value of 0, the power can be approximated as

$$\text{Power} \doteq \Phi(A_F), \tag{4.25}$$

where $\Phi(A_F)$ is the area under a standard normal density to the left of the point

$$A_F = \left[\left(\frac{12mn}{N+1} \right)^{1/2} \cdot f^*(0) \cdot \Delta \right] - z_\alpha, \tag{4.26}$$

where $f^*(0)$ is the density function, evaluated at 0, of the difference between two independent values drawn from the X-population having distribution F (cf. Lehmann (1975, p. 72, 403)).

When F is normal with standard deviation σ, $f^*(0) = 1/\{2\sigma(\pi)^{1/2}\}$ and A_F reduces to

$$A_{\text{normal}} = \left(\sqrt{\frac{3mn}{(N+1)\pi}} \cdot \frac{\Delta}{\sigma}\right) - z_\alpha. \tag{4.27}$$

Equation (4.27) shows that when F is normal, the approximate power depends on Δ and σ only through their ratio Δ/σ. (This is also true of the exact power.) Thus, for example, the power for the pair ($\Delta = 1$, $\sigma = 2$) is the same as the power for the pair ($\Delta = .5$, $\sigma = 1$).

Exact power values for the one-sided Wilcoxon test for model (4.2) when F is normal are given in Table B-1 of Milton (1970). Exact power values for the two-sided Wilcoxon test when F is normal are given in Table B-2 of Milton (1970). Milton's tables give power values for all sample sizes $2 \le n \le m \le 7$ that yield nontrivial results. If the sample of size m (or n) is from a normal population with mean μ_1 (or μ_2), $\mu_2 > \mu_1$, and variance σ^2, the location-shift alternative is defined in terms of $d = \{(\mu_2 - \mu_1)/\sigma\} = \Delta/\sigma$. Values are given for $d = .2(.2)1.0, 1.5, 2.0, 3.0$. Entries in the tables are ordered according to increasing values of $m + n$, from $2 \le m + n \le 14$. In Tables B-1 and B-2, the nominal levels of α are $\alpha = .25, .10, .05, .025, .01, .005$. The α's appearing in the tables are the attainable levels of significance nearest to but less than the nominal α's.

We suppose, for purposes of illustration, that model (4.2) holds with the underlying population F taken to be normal with variance $\sigma^2 = 16$ and the treatment effect $\Delta = 4$. Suppose further that we wish to determine, in a case where $m = 7$ and $n = 7$, the power of the $\alpha = .082$ test that rejects H_0 if $W \ge 64$ and accepts H_0 if $W < 64$. Substituting into (4.26) yields

$$A_{\text{normal}} = \left\{\left(\sqrt{\frac{3(7)(7)}{(15)\pi}} \cdot \left(\frac{4}{4}\right) - 1.39\right)\right\} = .376$$

and thus the power is approximately

$$\text{Power} \doteq \Phi(.376) = 1 - .35 = .65.$$

The exact power in this case is found from Table B-1 of Milton (1970) to be .635.

10. *Sample-Size Determination*. The Wilcoxon rank sum test detects a more general class of alternatives than the location-shift alternatives described by model (4.2). The one-sided upper-tail test defined by procedure (4.4) is consistent (i.e., has power tending to 1 as m, n tend to infinity) against those (F, G) populations for which $\delta > \frac{1}{2}$, where

$$\delta = P(X < Y). \tag{4.28}$$

The parameter δ defined by (4.28) is the probability that an X randomly selected from the distribution F will be less than an independent Y randomly selected from the distribution G. We say more about δ in Comment 18.

Noether (1987) shows how to determine an approximate total sample size N so that the α-level one-sided test given by procedure (4.4) will have an approximate power $1-\beta$ against an alternative value δ, where δ is greater than $\frac{1}{2}$. With $m = cN$, the approximate value of N is

$$N \doteq \frac{(z_\alpha + z_\beta)^2}{12c(1-c)(\delta - \frac{1}{2})^2}. \tag{4.29}$$

We illustrate the use of (4.29). Suppose we are testing H_0 and we desire to use an upper-tail $\alpha = .05$ test with power $= 1-\beta$ at least .90 against an alternative where $\delta = P(X < Y) = .7$ (recall that under H_0, $\delta = .5$). For simplicity, we take $m = n$ so that $c = .5$. From (4.29) with $z_\alpha = z_{.05} = 1.65$, $z_\beta = z_{.10} = 1.28$, and $\delta = .7$, we find

$$N \doteq \frac{(1.65 + 1.28)^2}{12(.5)(.5)(.7 - .5)^2} = 71.54, \qquad m = n = \frac{N}{2} = 35.8.$$

To be conservative take $m = n = 36$ rather than 35.

11. *Robustness of Level*. The significance level of the rank sum test is not preserved if the two populations differ in dispersion or shape. This is also the case for the normal theory two-sample t-test. For the effect of shape differences between the populations on the level of the rank sum test and other two-sample location procedures, see Pratt (1964). For a test of location differences that does not assume equal dispersions, see Fligner and Policello (1981) and Section 4.4.

    The level of the rank sum test is not preserved if dependencies exist among the X's or among the Y's, or if the X's are not independent of the Y's. Recall we have assumed that the NX's and Y's are mutually independent. For the effect on the level when this assumption is relaxed so that dependencies are allowed, see Serfling (1968); Hollander, Pledger, and Lin (1974) and Pettitt and Siskind (1981).

    There are other situations and designs in which the exact conditional randomization distribution of the Wilcoxon statistic is different than the usual Wilcoxon null distribution, and different approaches need to be used to obtain a P-value for comparing two treatments. See, for example, Efron's (1971) biased coin design and other restricted randomization designs considered by Hollander and Peña (1988) and Mehta, Patel, and Wei (1988).

12. *van der Waerden's Test*. The van der Waerden's rank statistic is

$$c = \sum_{j=1}^{n} \Phi^{-1}\left(\frac{S_j}{N+1}\right), \tag{4.30}$$

where, as before, $S_1, \ldots, S_n$ are the Y-ranks and $\Phi^{-1}(t)$ is the tth percentile of the $N(0,1)$ distribution. That is, $\Phi^{-1}(t)$ is the point such that the area under a $N(0,1)$ curve to the left of $\Phi^{-1}(t)$ is equal to t. The test of H_0 based on c has competitive efficiency properties versus the test based on W (see Section 4.5) and therefore is a popular competitor of W. To test $H_0 : \Delta = 0$ versus $H_1 : \Delta > 0$, reject H_0 for significantly large values of c. To test $H_0 : \Delta = 0$ versus $H_1 : \Delta < 0$, reject H_0 for significantly small values of c. To test $H_0 : \Delta = 0$ versus $H_3 : \Delta \neq 0$, reject

H_0 for significantly large values of $|c|$. Under H_0, the distribution of c is symmetric about 0. Tables of critical values are given by van der Waerden and Nievergelt (1956). The R package agricolae contains the program `waerden.test`, which gives P-values for van der Waerden's test. We illustrate using the water transfer data of Table 4.1. Let `at.term <-c(.80, .83, 1.89, 1.04, 1.45, 1.38, 1.91, 1.64, .73, 1.46)` and `gest.age <-c(1.15, .88, .90, .74, 1.21)`. Then, the R command `waerden.test (c(at.term, gest.age), factor (c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1)), group = T)` yields the two-sided P-value .26, which is in agreement with the one-sided P-value .13, which we illustrate in the following text.

The large-sample approximation is easy to perform. Under H_0, c has mean 0 and variance

$$\text{var}_0(c) = \frac{mn\left[\sum_{i=1}^{N}\{\Phi^{-1}(i/(N+1))\}^2\right]}{N(N-1)}. \tag{4.31}$$

The normal theory approximation to the distribution of

$$c^* = \frac{c}{\sqrt{\text{var}_0(c)}}, \tag{4.32}$$

treats c^* as an approximate $N(0,1)$ random variable for large m, n.

We illustrate the large-sample test based on c^* using the chorioamnion permeability data of Table 4.1 for which $m = 10$, $n = 5$, and $N = 15$. From the symmetry of the normal distribution, we note $\Phi^{-1}(i/16) = -\Phi^{-1}((16-i)/16)$ for $i = 1, \ldots, 7$, and $\Phi^{-1}\left(\frac{8}{16}\right) = \Phi^{-1}\left(\frac{1}{2}\right) = 0$. The values of $\Phi^{-1}(i/16)$ are found by using the R command `qnorm(x, 0, 1)`. For example, `qnorm(1/16, 0, 1) = -1.534` and so forth.

i:	1	2	3	4	5	6	7	8
$\Phi^{-1}(i/16)$:	-1.534	-1.150	$-.887$	$-.674$	$-.489$	$-.319$	$-.157$	0
i:	9	10	11	12	13	14	15	
$\Phi^{-1}(i/16)$:	.157	.319	.489	.674	.887	1.150	1.534.	

Recall that the Y-ranks for the data of Table 4.1 are 2, 5, 6, 8, and 9. From (4.30) we obtain

$$\begin{aligned} c &= \Phi^{-1}\left(\tfrac{2}{16}\right) + \Phi^{-1}\left(\tfrac{5}{16}\right) + \Phi^{-1}\left(\tfrac{6}{16}\right) + \Phi^{-1}(\tfrac{8}{16}) + \Phi^{-1}\left(\tfrac{9}{16}\right) \\ &= -1.150 - .489 - .319 + 0 + .157 = -1.80. \end{aligned}$$

From (4.31) we obtain

$$\begin{aligned} \text{var}_0(c) = \frac{10(5)}{15(14)}\{&(-1.534)^2 + (-1.150)^2 + (-.887)^2 + (-.674)^2 \\ &+ (-.489)^2 + (-.319)^2 + (-.157)^2 + (.157)^2 + (.319)^2 + (.489)^2 \\ &+ (.674)^2 + (.887)^2 + (1.150)^2 + (1.534)^2\} = 2.52. \end{aligned}$$

Then from (4.28) we find

$$c^* = \frac{-1.80}{\sqrt{2.52}} = -1.14,$$

with a one-sided P-value of .13.

Note that the results based on c are very close to those based on W that we found in Example 4.1. The large-sample approximation based on W gave a one-sided P-value of .11 and the exact P-value from W is $P = P_0(W \leq 30) = .127$.

A test that is asymptotically equivalent to the test based on c is the Fisher–Yates–Terry–Hoeffding (cf. Terry (1952), Hoeffding (1951)) test based on

$$c_1 = \sum_{j=1}^{n} E(V^{(S_j)}),$$

where $V^{(1)} < V^{(2)} < \cdots < V^{(N)}$ are the order statistics of a sample of size N from a $N(0,1)$ distribution and $S_1, \ldots, S_n$ are the Y-ranks. Values of $E(V^{(i)})$, $i = 1, \ldots, N$ for $N \leq 100$ and some larger sizes are given in Harter (1961). Exact tables can be found in Terry (1952) and Klotz (1964). Under H_0, the distribution of c_1 is symmetric about 0. The large-sample normal theory approximation treats

$$c_1^* = \frac{c_1}{\sqrt{\text{var}_0(c_1)}},$$

as a $N(0,1)$ random variable under H_0, where

$$\text{var}_0(c_1) = \frac{mn \sum_{i=1}^{N} \{E(V^{(i)})\}^2}{N(N-1)}.$$

Because $E(V^{(i)}) \doteq \Phi^{-1}(i/(N+1))$, it can be shown that tests based on c and c_1 are asymptotically equivalent. Both tests are often referred to as the *normal scores test*. Exact power values for the one-sided and two-sided tests based on c_1 for model (4.2) when F is normal are given in Tables B-3 and B-4 of Milton (1970).

13. *The Location-Shift Function*. Model (4.2) implies that the treatment effect is the same constant value Δ for each possible value of X. In some instances, it will be more appropriate to use a model that allows the treatment effects to be a function $\Delta(X)$ that is allowed to vary with X. For example, the treatment effect may be the expected increase (decrease) in systolic blood pressure due to taking a tranquilizer. In such a case, $\Delta(X)$ would depend on the patient's pretranquilizer blood pressure level X. This suggests the model

$$Y \stackrel{d}{=} X + \Delta(X), \tag{4.33}$$

where Y is systolic blood pressure after taking the tranquilizer. Model (4.33) was introduced by Lehmann (1975, p. 68). The function $\Delta(X)$ is called the *location-shift function*. Properties of $\Delta(X)$ were developed by Doksum (1974) and Switzer (1976). Doksum and Sievers (1976) derived simultaneous confidence bands for $\Delta(X)$. Hollander and Korwar (1982) and Wells and Tiwari (1989) extended the results of Doksum (1974) and Switzer (1976) to a nonparametric Bayesian framework. Lu, Wells, and Tiwari (1994) studied the location-shift function when the two samples are censored.

14. *Consistency of the W Test*. Under Assumptions A1–A3, the consistency of the tests based on W depends on the parameter

$$\delta^* = P(X < Y) - \tfrac{1}{2}.$$

The test procedures defined by (4.4), (4.5), and (4.6) are consistent against the alternatives for which $\delta^* >$, $<$, and $\neq 0$, respectively.

Properties

1. *Consistency*. For the location-shift model defined by (4.2), the tests defined by (4.4), (4.5), and (4.6) are consistent against the alternatives $\Delta >$, $<$, and $\neq 0$, respectively. Also, see Comment 14.
2. *Asymptotic Normality*. See Lehmann (1975, pp. 365–366).
3. *Efficiency*. See Section 4.5.

Problems

1. The data in Table 4.3 are a subset of the data obtained by Thomas and Simmons (1969), who investigated the relation of sputum histamine levels to inhaled irritants or allergens. The histamine content was reported in micrograms per gram of dry weight of sputum. The subjects for this portion of the study consisted of 22 smokers; 9 of them were allergics and the remaining 13 were asymptomatic (nonallergic) individuals. Care was taken to avoid people who carried out part of their daily work in an atmosphere of noxious gases or other respiratory toxicants. Table 4.3 gives the ordered sputum histamine levels for the 22 individuals in the study.

Test the hypothesis of equal levels versus the alternative that allergic smokers have higher sputum histamine levels than nonallergic smokers. Use the large-sample approximation.

2. Let W' be the sum of the ranks of the X observations. Verify directly, or illustrate using the chorioamnion permeability data of Table 4.1, the equation $W + W' = (m + n)(m + n + 1)/2$.

Table 4.3 Sputum Histamine Levels ($\mu g/g$ Dry Weight Sputum)

Allergics	Nonallergics
1651.0	48.1
1112.0	48.0
102.4	45.5
100.0	41.7
67.6	35.4
65.9	34.3
64.7	32.4
39.6	29.1
31.0	27.3
	18.9
	6.6
	5.2
	4.7

Source: H. V. Thomas and E. Simmons (1969).

3. Suppose a sixth Y observation is added to the five Y's of Table 4.1, and assume that the value $W = 30$ based on the original 10 X's and five Y's has already been calculated. How would you calculate the new value of W? Compare the method of reranking (to obtain new Y ranks) with a method based on using the Mann–Whitney statistic U in conjunction with the equation relating U and W (see Comment 7). Generalize the problem to different m and n values and make the same comparison.

4. Let U' denote the number of (X_i, Y_j) pairs for which $X_i > Y_j$. Assume that there are no $X = Y$ ties, and either establish directly or illustrate with the chorioamnion permeability data of Table 4.1, the relation $U' + U = mn$.

5. Molitor (1989) conducted a study to see if children who watched TV or film violence were significantly more tolerant of "real-life" violent behavior than children who instead watched a nonviolent TV show or film. Half of the 42 children in the study were shown violent TV (an edited version of *The Karate Kid*), whereas the other half watched exciting but nonviolent sports (highlights from the 1984 Summer Olympic Games). Each child was asked to "watch over" two younger children, supposedly in the next room, via a television monitor. Each child was instructed to go and get the research assistant (who stated she had to leave for an emergency) if the younger children "got into trouble." What each child witnessed, while alone, was actually a videotaped sequence depicting two small children first play with blocks and then progressively get more violent. That is, they called each other names, then pushed each other, chased each other, fought, and then supposedly broke a video camera while fighting.

Toleration of violence was measured by the time (in seconds) each child stayed in the room after he or she witnessed the two younger children's first act of violence. As soon as the subject child left the room, the timing clock was stopped. Each child was subsequently assured that an adult had entered the room where the two children were and that they were not hurt and the video camera was not damaged.

Do the data of Table 4.4 indicate that the children who viewed the violent TV tend to take longer to seek help (were more tolerant) than the children who viewed the nonviolent sports-action TV? Use Wilcoxon's W.

6. Assume that model (4.2) holds and that F is normal with variance 13. We have eight X observations and eight Y observations. If we use the $\alpha = .065$ test of $H_0 : \Delta = 0$ versus the alternative $\Delta > 0$, what is the approximate power of this test when the treatment effect is $\Delta = 2$?

7. For testing $H_0 : \Delta = 0$ versus the alternative $\Delta > 0$, you choose to use a type I error probability $\alpha = .10$. Using equal sample sizes, what should the common value of m, n be to have power at least .88 against an alternative where $\delta = .8$?

8. We observe $X_1 = 2.1, X_2 = 1.9, X_3 = 2.6, X_4 = 3.3, Y_1 = 1.9, Y_2 = 2.6$, and $Y_3 = 3.7$. What is the conditional distribution of W obtained by considering all $\binom{7}{3}$ possible choices of three data points to serve as the Y-values? How extreme is the observed value of W in this conditional distribution?

9. Apply van der Waerden's test based on c to the data of Table 4.4. Compare your result with that obtained in Problem 5 using Wilcoxon's W.

10. Apply the test based on W to the plasma glucose data of Table 4.6.

11. Apply the test based on c to the plasmas glucose data of Table 4.6. Compare with the results obtained in Problem 10.

12. Show directly, or illustrate via an example, that the maximum value of W is $n(2m + n + 1)/2$. What is the minimum value of W?

13. Suppose you reject H_0 if $W = n(2m + n + 1)/2$ or if $W = n(n + 1)/2$, and you accept H_0 otherwise. What is α for this test?

14. Suppose $m = n = 7$. Compare the exact $\alpha = .049$ test of $H_0 : \Delta = 0$ versus $H_1 : \Delta > 0$ based on W with its corresponding test based on large-sample approximation. What is the

Table 4.4 Seconds Spent in Room after Witnessing Violence

Olympics watchers	*Karate Kid* watchers
12	37
44	39
34	30
14	7
9	13
19	139
156	45
23	25
13	16
11	146
47	94
26	16
14	23
33	1
15	290
62	169
5	62
8	145
0	36
154	20
146	13

Source: F. T. Molitor (1989).

exact α value of the test based on the large-sample approximation whose nominal α value is .049?

15. Phadke et al. (2006) conducted a study to evaluate the soleus Hoffman reflex (H-reflex) for two different leg loading conditions on people who have not experienced spinal cord injuries (non-injured subjects) and people with incomplete spinal cord injuries (i-SCI subjects). The Phadke et al. (2006) paper was selected by Erin Easton (2006) for her term project in M. Hollander's 2006 Applied Nonparametric Statistics class. This problem is based on a portion of her analysis. Decreasing the load of weight on the leg is one way that patients with SCI undergo rehabilitation in order to relearn how to stand and walk. Leg loading is controlled through a body weight support (BWS) system that consists of a harness and a suspension system. The typical setting for BWS during rehabilitation for post-SCI patients is 60% leg loading (or 40% BWS). In the Phadke et al. study, 40% BWS was compared to 0% BWS (or 100% leg loading) for both i-SCI and noninjured subjects in order to determine whether a change in percent BWS changed the soleus H-reflex response for subjects in a standing position. Here, we focus on a portion of their data comparing noninjured to i-SCI subjects for 40% BWS.

The soleus muscle is one of the muscles that run from the just below the back of the knee down to the heel, and contraction of this muscle results in plantar flexion of the foot (pointing of the toes) and in maintenance of the body in a stable standing position. The H-reflex is an involuntary response (or flexion) in a muscle on electrical stimulation of the nerves that controls contraction and relaxation of the muscle. The tibialis anterior muscle is a muscle that runs along the front side of the tibia from below the knee to the top of the foot, and contraction of this muscle results in the dorsal flexion of the foot (rise of the foot toward the front of the leg). The tibial nerve runs along the entire back side of the leg, and it supplies electrical impulses to the muscles of the back of the leg, including the soleus. An electromyogram (EMG) is used to measure the electrical current in a muscle. The current is generally proportional to the activity level of the muscle, where an inactive muscle has no current. The H/M ratio is the ratio of the maximum soleus H-reflex to the maximum soleus muscle potential (or to a preset percentage

Table 4.5 H/M Ratios of Noninjured Subjects and i-SCI Subjects For 40% BWS

Noninjured H/M ratios	Ranks	i-SCI H/M ratios	Ranks
.19	4	.89	13
.14	3	.76	10
.02	1.5	.63	8
.44	6	.69	9
.37	5	.58	7
		.79	11.5
		.02	1.5
		.79	11.5

Source: C.P. Phadke, S.S. Wu, F.J. Thompson, and A.L. Behrman (2006).

of the maximum potential). Table 4.5 gives the H/M ratios for five noninjured subjects and eight i-SCI subjects for 40% BWS.

Is there evidence, for this 40% BWS situation, that the H/M ratios of the i-SCI subjects are significantly larger than the H/M ratios of the noninjured subjects? What is the approximate P-value achieved by your test.

16. Apply the exact conditional test based on W (see Comment 5) to the H/M ratios data of Table 4.5. Compare your result with that obtained in Problem 15.

17. Apply van der Waerden's test to the H/M ratios data of Table 4.5. Compare your result with the results of Problems 15 and 16.

4.2 AN ESTIMATOR ASSOCIATED WITH WILCOXON'S RANK SUM STATISTIC (HODGES–LEHMANN)

Procedure

To estimate Δ of model (4.2), form the mn differences $Y_j - X_i$, for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. The estimator of Δ associated with the Wilcoxon rank sum statistic (see Comment 15) is

$$\widehat{\Delta} = \text{median}\{(Y_j - X_i), i = 1, \ldots, m; j = 1, \ldots, n\}. \tag{4.34}$$

Let $U^{(1)} \leq \cdots \leq U^{(mn)}$ denote the ordered values of $Y_j - X_i$. Then if mn is odd, say $mn = 2k + 1$, we have $k = (mn - 1)/2$ and

$$\widehat{\Delta} = U^{(k+1)}, \tag{4.35}$$

the value that occupies the position $k + 1$ in the list of the ordered $Y - X$ differences. If mn is even, say $mn = 2k$, then $k = mn/2$ and

$$\widehat{\Delta} = \frac{U^{(k)} + U^{(k+1)}}{2}. \tag{4.36}$$

That is, $\widehat{\Delta}$ is the average of the two $Y - X$ differences that occupy the positions k and $k + 1$ in the ordered list of the mn differences.

EXAMPLE 4.3 ***Continuation of Example 4.1.***

To estimate Δ for the chorioamnion permeability data of Table 4.1, we obtain, using R, the ordered values $U^{(1)} \leq \cdots \leq U^{(50)}$.

```
diff<-numeric(0); m<-10; n<-5
for (i in 1:m) for (j in 1:n) diff<-c(diff,
gest.age[j]-at.term[i])
diff<-sort(diff)
```

Then we obtain the 50 ordered $Y_j - X_i$ differences displayed in Table 4.6.
The value of $mn = 50$ is even and thus we use (4.36) with $k = \frac{50}{2} = 25$ to obtain

$$\widehat{\Delta} = \frac{U^{(25)} + U^{(26)}}{2} = \frac{-.31 - .30}{2} = -.305.$$

The estimate $\widehat{\Delta} = -.305$ is directly available from the `R` command `wilcox.test`. If you perform `wilcox.test(at.term, gest.age, alternative="t", conf.int=T)`, the output .305 is the estimate for the difference in location because `wilcox.test` computes the median of the X-Y differences. To get the median of the Y-X differences perform `wilcox.test(gest.age, at.term, alternative="t", conf.int=T)` and obtain $\widehat{\Delta} = -.305$.

Comments

15. *Motivation for the Hodges–Lehmann Estimator*. The Hodges–Lehmann (1963) estimator $\widehat{\Delta}$ defined by (4.34) is associated with the Wilcoxon rank sum test. When $\Delta = 0$, the distribution of the statistic W is symmetric about its mean $n(m + n + 1)/2$ (see Comment 8). A reasonable estimator of Δ is the amount $\widehat{\Delta}$ (say) that should be subtracted from each Y_j so that the value of W, when applied to the aligned samples $X_1, \ldots, X_m, Y_1 - \widehat{\Delta}, \ldots, Y_n - \widehat{\Delta}$, is $n(m + n + 1)/2$. Roughly speaking, we estimate Δ by the amount $(\widehat{\Delta})$ that the Y sample should be shifted in order that $X_1, \ldots, X_m$ and $Y_1 - \widehat{\Delta}, \ldots, Y_n - \widehat{\Delta}$ appear (when "viewed" by the rank sum statistic W) as two samples from the same population. (Under Assumptions A1–A3, the variables $X_1, \ldots, X_m$ and $Y_1 - \Delta, \ldots, Y_n - \Delta$ can be taken as a single sample of size $N = m + n$ from the underlying population.)

The Hodges–Lehmann method can be applied to large classes of statistics, which, for example, include van der Waerden's V. The forms of the resulting

Table 4.6 Ordered Y-X Differences for the Chorioamnion Permeability Data

$U^{(1)} \leq U^{(2)} \leq \cdots \leq U^{(50)}$											
−1.17	−1.15	−1.03	−1.01	−1.01	−0.99	−0.90	−0.76	−0.76	−0.74	−0.74	−0.72
−0.71	−0.70	−0.68	−0.64	−0.58	−0.57	−0.56	−0.55	−0.50	−0.49	−0.48	−0.43
−0.31	−0.30	−0.30	−0.25	−0.24	−0.23	−0.17	−0.16	−0.14	−0.09	−0.06	0.01
0.05	0.07	0.08	0.10	0.11	0.15	0.17	0.17	0.32	0.35	0.38	0.41
0.42	0.48										

estimators are not always as convenient for calculation as in the case of $\widehat{\Delta}$. See Hodges and Lehmann (1983) for an expository article on their method. See McKean and Ryan (1977) for an algorithm for computing $\widehat{\Delta}$.

16. *Sensitivity to Gross Errors*. The estimator $\widehat{\Delta}$ is less sensitive to gross errors than its normal theory analog $\overline{Y} - \overline{X}$, the difference of the sample averages.

17. *Competing Estimators*. Observe that the estimator $\widehat{\Delta}$ cannot be written as a difference of a statistic based on the Y observations only and a second statistic based on the X observations only. The classical estimator $\overline{\Delta} = \overline{Y} - \overline{X}$ can be written as such a difference. When the underlying population is symmetric, Lehmann (1963a) proposed to estimate Δ by

$$\widehat{\widehat{\Delta}} = \widehat{\theta}_2 - \widehat{\theta}_1,$$

where $\widehat{\theta}_1(\widehat{\theta}_2)$ is the estimator (3.10) associated with the signed rank statistic T^+ for estimating the location of the population corresponding to the $X(Y)$ observations. That is,

$$\widehat{\widehat{\Delta}} = \text{median}\left\{\frac{Y_i + Y_j}{2}, 1 \le i \le j \le n\right\} - \text{median}\left\{\frac{X_i + X_j}{2}, 1 \le i \le j \le m\right\}. \tag{4.37}$$

The standard deviation of $\widehat{\widehat{\Delta}}$ can be estimated by

$$\widehat{\sigma}_{\widehat{\widehat{\Delta}}} = \left\{\left(\frac{\theta_{2U} - \theta_{2L}}{2z_{\alpha_2/2}}\right)^2 + \left(\frac{\theta_{1U} - \theta_{1L}}{2z_{\alpha_1/2}}\right)^2\right\}^{1/2}, \tag{4.38}$$

where θ_{2U} and θ_{2L} are the upper and lower endpoints, respectively, of the $100(1-\alpha_2)\%$ confidence interval obtained from the method of Section 3.3 by replacing the Z's of Section 3.3 by the n Y's of sample 2. Similarly, θ_{1U} and θ_{1L} are the end points of the $100(1-\alpha_1)\%$ confidence interval obtained by the method of Section 3.3 by replacing the Z's of Section 3.3 by the m X's of sample 1.

An approximate confidence interval for Δ, with the confidence coefficient $1-\alpha$, is

$$\Delta_\ell = \widehat{\widehat{\Delta}} - z_{\alpha/2}\widehat{\sigma}_{\widehat{\widehat{\Delta}}}, \qquad \Delta_u = \widehat{\widehat{\Delta}} + z_{\alpha/2}\widehat{\sigma}_{\widehat{\widehat{\Delta}}}. \tag{4.39}$$

Lehmann (1963a), Høyland (1965), and Ramachandramurty (1966a) investigated the properties of $\overline{\Delta}$, $\widehat{\Delta}$ and $\widehat{\widehat{\Delta}}$ for various deviations from the assumptions, including asymmetry and non-location-differences between the populations.

Other competing estimators of $\widehat{\Delta}$ include those in classes initiated by Serfling (1984), Akritas (1986), and Serfling (1992).

18. *The Probability That X Is Less Than Y*. A quantity of interest in the two-sample location problem is the parameter $\delta = P(X_1 < Y_1)$, where X_1 is a random member from the X population, Y_1 is a random member from the Y population, and X_1 and Y_1 are independent; that is, δ is the probability that a single Y observation will be larger than a single X observation. Pitman (1948) and Birnbaum (1956) discussed a point estimator for δ given by $\widehat{\delta} = U/mn$, where U is the

Mann–Whitney form of the rank sum statistic (see Comment 7). Upper bounds for the variance of U, which are useful when using $\widehat{\delta}$ as a point estimator for δ, were obtained in terms of δ by van Dantzig (1951). See Birnbaum and Klose (1957) for lower bounds. Lehmann (1951) showed that $\widehat{\delta}$ is the uniform minimum variance unbiased estimator of δ over the class of continuous populations (also see Blyth (1950)). For the use of the sign statistic in obtaining a point estimator for δ, see Saxena (1969).

Many statisticians, including Wolfe and Hogg (1971), have emphasized the importance of natural parameters such as δ. Consider a medical application in which X represents the response to treatment A and Y is the response to treatment B. Let μ_1, μ_2 be the respective means of the X and Y populations and let σ denote the (assumed) common standard deviation. Then, $P(X < Y) = .76$ will usually make more sense to a doctor than the statement $\{(\mu_2 - \mu_1)/\sigma\} = 1$. (If X and Y are normal, and independent, with means μ_1, μ_2, and common standard deviation σ, then $\{(\mu_2 - \mu_1)/\sigma\} = 1$ implies $P(X < Y) = .76$.) Furthermore, we are often more interested in the probability that X is less than Y than, say, in the difference between the Y and X means. This is true in a good deal of biological research, where, for example, a large liver is a large liver, but how large it is makes little difference except possibly in comparison with other livers (rather than in comparison with scale measurements on a weighing machine). In situations such as these, the estimator $\widehat{\delta}$ may be more useful than the estimator $\widehat{\Delta}$.

Birnbaum (1956) and Birnbaum and McCarty (1958) considered a distribution-free upper confidence bound for $\delta = P(X_1 < Y_1)$ based on the Mann–Whitney U when the underlying populations are continuous. Owen, Craswell, and Hanson (1964) extended this to discrete populations, and Govindarajulu (1968) sharpened the Birnbaum–McCarty upper bound and provided corresponding two-sided distribution-free confidence intervals for δ. Sen (1967) and Govindarajulu (1968) considered asymptotically distribution-free confidence bounds for δ based on consistent estimators of the variance of the Mann–Whitney U. Saxena (1969) discussed distribution-free confidence bounds for δ based on the sign statistic.

The parameter δ also arises naturally in reliability. Let X be the stress on a component and let Y be the strength of the component. Then, $\delta = P(X < Y)$ is the probability that the component functions properly. Johnson (1988) surveys many of the methods referenced in this comment in the context of reliability. His focus is on getting estimators and confidence bounds on system reliability in reliability systems such as "k out of n" systems.

Sen's (1967) asymptotic nonparametric interval for δ is relatively easy to obtain. Sen's interval is based on the asymptotic normality of $\sqrt{n_0}(\widehat{\delta} - \delta)/s$, where $n_0 = mn/(m+n)$. Here, s is a consistent estimator of the standard deviation of $\sqrt{n_0}\widehat{\delta}$. Many estimators are available. A particularly convenient one defined by Sen is

$$s^2 = \frac{nS_{10}^2 + mS_{01}^2}{m+n},$$

where

$$S_{10}^2 = \frac{\sum_{i=1}^{m}(R_i - i)^2 - m(\overline{R} - (m+1)/2)^2}{(m-1)n^2},$$

and

$$S_{01}^2 = \frac{\sum_{j=1}^{n}(S_j - j)^2 - n(\overline{S} - (n+1)/2)^2}{(n-1)m^2}.$$

Here, R_i is the rank of $X_{(i)}$ in the joint ranking of the X's and Y's, S_j is the rank of $Y_{(j)}$ in the joint ranking of the X's and Y's, $\overline{R} = \sum_{i=1}^{m} R_i/m$, and $\overline{S} = \sum_{j=1}^{n} S_j/n$. Recall that $X_{(1)} \le \cdots \le X_{(m)}$ are the ordered X-values and $Y_{(1)} \le \cdots \le Y_{(n)}$ are the ordered Y-values. The lower and upper end points, δ_{L} and δ_{U}, respectively, of the asymptotic $1-\alpha$ confidence interval are

$$\delta_{\mathrm{L}}^{S} = \widehat{\delta} - z_{\alpha/2}\sqrt{\frac{nS_{10}^2 + mS_{01}^2}{mn}},$$

$$\delta_{\mathrm{U}}^{S} = \widehat{\delta} + z_{\alpha/2}\sqrt{\frac{nS_{10}^2 + mS_{01}^2}{mn}}. \tag{4.40}$$

A competing interval has been proposed by Halperin, Gilbert, and Lachin (1987). Their $1-\alpha$ confidence interval is

$$\delta_{\mathrm{L}}^{H} = \frac{A - B}{C}, \qquad \delta_{\mathrm{U}}^{H} = \frac{A + B}{C}, \tag{4.41}$$

where

$$A = \widehat{\delta} + \frac{\gamma z_{\alpha/2}^2}{2mn},$$

$$B = \left(\frac{(\widehat{\delta}(1-\widehat{\delta})\gamma z_{\alpha/2}^2 + \gamma^2 z_{\alpha/2}^4/4mn)}{mn}\right)^{1/2},$$

$$C = 1 + \frac{\gamma z_{\alpha/2}^2}{mn},$$

$$\gamma = \widehat{\theta}(m+n-2) + 1,$$

$$\widehat{\theta} = \frac{\dfrac{\widehat{K} + 2(n-1)\widehat{\delta}}{m+n-2} - \widehat{\delta}^2}{\widehat{\delta}(1-\widehat{\delta})},$$

$$\widehat{K} = \left\{\frac{\sum_{j=1}^{n} r_{1j}(r_{1j}-1)}{mn}\right\} + \left\{\frac{\sum_{i=1}^{m} s_{1i}(s_{1i}-1)}{mn}\right\} - (n-1),$$

where r_{1j} is the number of X-observations that are less than $Y_{(j)}$ and s_{1i} is the number of Y-observations that are less than $X_{(i)}$.

Halperin, Gilbert, and Lachin point out that δ_{U}^{H} is less than 1 and δ_{L}^{U} is greater than 0. Also, $\widehat{\theta} \le 1$ if $\widehat{\delta}$ is neither 0 nor 1, but for some samples, $\widehat{\theta}$ may be less than 0. If that happens, take $\widehat{\theta} = 0$ in the definition of γ. If $\widehat{\delta} = 0$ or 1, take $\widehat{\theta} = 1$.

Halperin, Gilbert, and Lachin did simulations that indicated their method generally yields coverage probabilities closer to the nominal $1-\alpha$ than does the Sen method.

For the chorioamnion permeability data of Table 4.1, the approximate 95% Sen confidence interval for δ and the approximate 95% Halperin–Gilbert–Lachin confidence interval for δ are as follows. Recall that for these data, we have found (see Comment 7) $U = 15$ and thus

$$\widehat{\delta} = \frac{15}{10(5)} = .3.$$

For the Sen interval,

$$S_{10}^2 = .171, \qquad S_{01}^2 = .015.$$

From display (4.40), we obtain, with $\alpha = .05$,

$$\delta_{\mathrm{L}}^S = .02, \qquad \delta_{\mathrm{U}}^S = .58.$$

For the Halperin–Gilbert–Lachin interval, with $\alpha = .05$, we find

$$\widehat{K} = -.76, \qquad \widehat{\theta} = .172, \qquad \gamma = 3.24$$
$$A = .424, \qquad B = .260, \qquad C = 1.25.$$

From display (4.41), we obtain

$$\delta_{\mathrm{L}}^H = .13, \qquad \delta_{\mathrm{U}}^H = .55.$$

Properties

1. *Standard Deviation of* $\widehat{\Delta}$. For the asymptotic standard deviation of $\widehat{\Delta}$, see Hodges and Lehmann (1963), Lehmann (1963c), and Comment 21.
2. *Asymptotic Normality*. See Hodges and Lehmann (1963) and Ramachandramurty (1966a).
3. *Efficiency*. See Hodges and Lehmann (1963), Høyland (1965), Ramachandramurty (1966a), and Section 4.5.

Problems

18. Consider the data of Table 4.3. Associate the Y's (X's) with the allergies (nonallergies) and estimate Δ of model (4.2) using $\widehat{\Delta}$.

19. Again consider the data of Table 4.3. Estimate Δ using $\widehat{\widehat{\Delta}}$ and compare your estimate with $\widehat{\Delta}$ obtained in Problem 15.

20. Consider the data of Table 4.3. Use display (4.35) to obtain an approximate 95% confidence interval for Δ.

21. Consider the data of Table 4.3. Estimate $\delta = P(X < Y)$ and determine an approximate 90% confidence interval for δ.

22. Consider the data of Table 4.4. Estimate Δ of model (4.2) using $\widehat{\Delta}$.

23. Consider the data of Table 4.4. Estimate Δ using $\widehat{\widehat{\Delta}}$ and compare your estimate with $\widehat{\Delta}$ obtained in Problem 22.

24. Consider the data of Table 4.4. Use Comment 17 to obtain an approximate 93% confidence interval for Δ.

25. Consider the data of Table 4.4. Estimate $\delta = P(X < Y)$ and determine (a) an approximate 93% confidence interval for δ using the Sen's interval and (b) an approximate 93% confidence interval for δ using the Halperin–Gilbert–Lachin interval.

26. Change the value 102.4, appearing in Table 4.3, to 1024. How does this affect the estimate of Δ given by $\widehat{\Delta}$? How does this affect the estimate of Δ given by $\overline{\Delta} = \overline{Y} - \overline{X}$?

27. (a) What happens to $\widehat{\Delta}$ when we add a number b to each of the m X values and a number c to each of the n Y values? In particular, what happens when $b = c$?

(b) What happens to $\widehat{\Delta}$ when we multiply each of the X and Y values by the same number d?

28. Answer parts (a) and (b) of Problem 27 with $\widehat{\Delta}$ replaced by $\widehat{\widehat{\Delta}}$.

29. Do you need to calculate the values of all mn $Y - X$ differences in order to compute the value of $\widehat{\Delta}$? Explain.

4.3 A DISTRIBUTION-FREE CONFIDENCE INTERVAL BASED ON WILCOXON'S RANK SUM TEST (MOSES)

Procedure

For a symmetric two-sided confidence interval for Δ, with the confidence coefficient $1 - \alpha$, let $w_{\alpha/2}$ denote the upper $\alpha/2$ percentile point of the null distribution of W.

Then with

$$C_\alpha = \frac{n(2m + n + 1)}{2} + 1 - w_{\alpha/2}, \tag{4.42}$$

the $1 - \alpha$ confidence interval (Δ_L, Δ_U) is given by

$$\Delta_L = U^{(C_\alpha)}, \qquad \Delta_U = U^{(mn+1-C_\alpha)}. \tag{4.43}$$

That is, Δ_L is the $Y - X$ difference that occupies the position C_α in the list of the mn ordered $Y - X$ differences. The upper endpoint Δ_U is the $Y - X$ difference that occupies the position $mn + 1 - C_\alpha$ in the ordered list. With Δ_L and Δ_U given by display (4.43), we have, for all Δ,

$$P_\Delta(\Delta_L < \Delta < \Delta_U) = 1 - \alpha. \tag{4.44}$$

The confidence interval is found directly from the `R` command `wilcox.test`. We illustrate this in Example 4.4.

Large-Sample Approximation

For large m and n, the integer C_α may be approximated by

$$C_\alpha \approx \frac{mn}{2} - z_{\alpha/2}\left\{\frac{mn(m + n + 1)}{12}\right\}^{1/2}. \tag{4.45}$$

In general, the value of the right-hand side of (4.45) is not an integer. To be conservative, take C_α to be the largest integer that is less than or equal to the right-hand side of (4.45).

EXAMPLE 4.4 ***Continuation of Example 4.1.***

Consider the chorioamnion permeability data of Table 4.1. We will illustrate how to obtain the 96% confidence interval for Δ. Use the `R` command `wilcox.test (gest.age, at.term, conf.int = T, conf.level = .96)` to obtain the interval $\Delta_L = -.76$, $\Delta_U = .15$. Note $\Delta_L = U^{(9)}$ and $\Delta_U = U^{(42)}$ (see Table 4.6).

Applying the large-sample approximation, we find from approximation (4.45)

$$C_{.04} \approx \frac{10(5)}{2} - 2.05\left\{\frac{10(5)(10+5+1)}{12}\right\}^{1/2} = 8.3.$$

Thus, with the large-sample approximation, we set $C_{.04}$ equal to 8 and

$$\Delta_L = U^{(8)} = -.76, \qquad \Delta_U = U^{(43)} = .17.$$

Comments

19. *Relationship of Confidence Interval to Test*. The $1-\alpha$ confidence interval given by display (4.43) can be obtained from the two-sided rank sum test as follows. The confidence interval (Δ_L, Δ_U) consists of those Δ_0 values for which the two-sided α-level test of $\Delta = \Delta_0$ (see Comment 2) accepts the hypothesis $\Delta = \Delta_0$. The confidence interval given by display (4.43) was defined by way of a graphical procedure by Lincoln Moses in Chapter 18 of Walker and Lev (1953). See Lehmann (1986, p. 90) for a general result relating confidence intervals and acceptance regions of tests, and see Lehmann (1963c) for the specific result involving the rank sum test.
20. *Midpoint of Confidence Interval as an Estimator*. The midpoint of the interval (4.43), namely, $\{U^{(C_\alpha)} + U^{(mn+1-C_\alpha)}\}/2$, suggests itself as a reasonable estimator of Δ. (Note that this actually yields a class of estimators depending on the value of α.) In general this midpoint is not the same as $\widehat{\Delta}$. Lehmann (1963b) has also dealt with an asymptotically distribution-free confidence interval for Δ centered at $\widehat{\Delta}$, and Lehmann (1963c) has shown that the asymptotically distribution-free confidence interval has the same asymptotic behavior as the distribution-free confidence interval given by display (4.43).
21. *Estimating the Asymptotic Standard Deviation of* $\widehat{\Delta}$. The quantity $(\Delta_U - \Delta_L)/(2z_{\alpha/2})$, where (Δ_L, Δ_U) is the $1-\alpha$ confidence interval defined by display (4.43), provides us with a consistent estimator for the asymptotic standard deviation of the point estimator $\widehat{\Delta}$ (see Lehmann, (1963c)).
22. *Confidence Bounds*. To obtain a lower confidence bound for Δ, with the confidence coefficient $1-\alpha$, set

$$C_\alpha^* = \frac{n(2m+n+1)}{2} + 1 - w_\alpha, \tag{4.46}$$

where w_α, the upper α percentile point of the null distribution of W. The $100(1-\alpha)\%$ lower confidence bound Δ_L^* for Δ that is associated with the one-sided Wilcoxon rank sum test of $H_0 : \Delta = 0$ against the alternative $H_1 : \Delta > 0$ is given by

$$(\Delta_L^*, \infty) = (U^{(C_\alpha^*)}, \infty), \tag{4.47}$$

where $U^{(1)} \leq \cdots \leq U^{(mn)}$ are the ordered values of $Y_j - X_i$. With Δ_{L}^* defined by (4.47), we have, for all Δ,

$$P_\Delta(\Delta_{\mathrm{L}}^* < \Delta < \infty) = 1 - \alpha. \tag{4.48}$$

The $100(1-\alpha)\%$ upper confidence bound Δ_{U}^* for Δ that is associated with the one-sided Wilcoxon rank sum test of $H_0 : \Delta = 0$ against the alternative $H_1 : \Delta < 0$ is given by

$$(-\infty, \Delta_{\mathrm{U}}^*) = (-\infty, U^{(mn+1-C_\alpha^*)}), \tag{4.49}$$

where C_α^* is given by (4.46). With Δ_{U}^* defined by (4.49), we have, for all Δ,

$$P_\Delta(-\infty < \Delta < \Delta_{\mathrm{U}}^*) = 1 - \alpha. \tag{4.50}$$

For large m, n, the integer C_α^* can be approximated by

$$C_\alpha^* \cong \frac{mn}{2} - z_\alpha \left\{ \frac{mn(m+n+1)}{12} \right\}^{1/2}. \tag{4.51}$$

Properties

1. Under Assumptions A1–A3 and model (4.2), (4.44) holds. Hence, we can control the coverage probability to be $1 - \alpha$ without having more specific knowledge about the form of the underlying distribution. Thus $(\Delta_{\mathrm{L}}, \Delta_{\mathrm{U}})$ is a distribution-free confidence interval for Δ over a very large class of populations.
2. *Efficiency*. See Lehmann (1963c) and Section 4.5.

Problems

30. Refer to Problem 18 and obtain a confidence interval for Δ with approximate confidence coefficient .95.

31. For the chorioamnion permeability data of Table 4.1, compute an estimate of Δ utilizing the estimator defined in Comment 20. Compare with the value of $\widehat{\Delta}$ obtained in Example 4.3.

32. Use the results of Example 4.4 to obtain an estimate for the asymptotic standard deviation of $\widehat{\Delta}$ (see Comment 21).

33. Consider the $1 - \alpha$ confidence interval defined by display (4.43). Show that when $\alpha = 2/\binom{N}{n}$,

$$\Delta_{\mathrm{L}} = Y_{(1)} - X_{(m)}, \qquad \Delta_{\mathrm{U}} = Y_{(n)} - X_{(1)},$$

where $X_{(1)} \leq \ldots \leq X_{(m)}$ are the ordered X's and $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the ordered Y's.

34. Consider the $1 - \alpha$ confidence interval defined by display (4.43). Show that when $\alpha = 4/\binom{N}{n}$,

$$\Delta_{\mathrm{L}} = \text{minimum}\{Y_{(2)} - X_{(m)}, Y_{(1)} - X_{(m-1)}\},$$
$$\Delta_{\mathrm{U}} = \text{maximum}\{Y_{(n)} - X_{(2)}, Y_{(n-1)} - X_{(1)}\}.$$

35. Consider the data of Table 4.3. Obtain an approximate 95% confidence interval for Δ using the large-sample approximation of this section. Compare your result with the approximate 95% confidence interval obtained in Problem 20.

36. Consider the data of Table 4.2 and obtain an approximate 90% confidence interval for Δ using the large-sample approximation of this section.

37. Consider the data of Table 4.4 and obtain an approximation 99% confidence interval for Δ using the large-sample approximation of this section.

38. Consider the case $m = n = 8$ and compare the exact 91.8% confidence interval given by display (4.43) with that obtained by the large-sample approximation.

39. Consider the case $m = n = 10$ and compare the exact 91% confidence interval given by display (4.43) with that obtained by the large-sample approximation.

40. Consider the data of Table 4.5 and obtain a 95% confidence interval for Δ.

4.4 A ROBUST RANK TEST FOR THE BEHRENS–FISHER PROBLEM (FLIGNER–POLICELLO)

Hypothesis

In this section we introduce new assumptions. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from continuous distributions that are symmetric about the population medians θ_x and θ_y, respectively. Note that we do not require the X and Y populations to have the same distributional form nor do we assume that the variances of the two populations are equal. We are interested in testing $H_0' : \theta_x = \theta_y$ versus $\theta_x < \theta_y$ [or $\theta_x > \theta_y$ or $\theta_x \neq \theta_y$]. This problem of testing $H_0' : \theta_x = \theta_y$ without assuming equal variances is often referred to as the *Behrens–Fisher problem*.

Procedure

Let

$$P_i = [\text{number of sample } Y \text{ observations less than } X_i], \tag{4.52}$$

for $i = 1, \ldots, m$. Similarly, set

$$Q_j = [\text{number of sample } X \text{ observations less than } Y_j], \tag{4.53}$$

for $j = 1, \ldots, n$. We call P_i and Q_j the *placements* of X_i and Y_j, respectively. Compute

$$\overline{P} = \frac{1}{m}\sum_{i=1}^{m} P_i = \text{average } X \text{ sample placement}, \tag{4.54}$$

and

$$\overline{Q} = \frac{1}{n}\sum_{j=1}^{n} Q_j = \text{average } Y \text{ sample placement}. \tag{4.55}$$

Let

$$V_1 = \sum_{i=1}^{m}(P_i - \overline{P})^2 \quad \text{and} \quad V_2 = \sum_{j=1}^{n}(Q_j - \overline{Q})^2, \tag{4.56}$$

and set

$$\widehat{U} = \frac{\sum_{j=1}^{n} Q_j - \sum_{i=1}^{m} P_i}{2(V_1 + V_2 + \overline{P}\,\overline{Q})^{1/2}}. \tag{4.57}$$

a. *One-Sided Upper-Tail Test*. For a one-sided test of $H_0' : \theta_x = \theta_y$ versus the one-sided alternative $H_1' : \theta_y > \theta_x$ at the approximate α level of significance,

$$\text{Reject } H_0' \text{ if } \widehat{U} \geq u_\alpha; \qquad \text{otherwise do not reject,} \tag{4.58}$$

where u_α is a constant satisfying $P_0(\widehat{U} \geq u_\alpha) \approx \alpha$.

b. *One-Sided Lower-Tail Test*. For a one-sided test of $H_0' : \theta_x = \theta_y$ versus the alternative $H_2' : \theta_y < \theta_x$ at the approximate α level of significance, we

$$\text{Reject } H_0' \text{ if } \widehat{U} \leq -u_\alpha; \qquad \text{otherwise do not reject.} \tag{4.59}$$

c. *Two-Sided Test*. For a two-sided test of $H_0' : \theta_x = \theta_y$ versus the alternative $H_3' : \theta_y \neq \theta_x$ at the approximate α level of significance, we

$$\text{Reject } H_0' \text{ if } |\widehat{U}| \geq u_{\alpha/2}; \qquad \text{otherwise do not reject.} \tag{4.60}$$

Any u_α can be computed exactly or estimated using Monte Carlo simulation for large m and n using the `R` command `cFligPoli`.

Large-Sample Approximation

When $H_0' : \theta_x = \theta_y$ is true, the statistic $\widehat{U}$ has an asymptotic ($\min(m, n)$ tending to infinity) $N(0, 1)$ distribution. Thus the normal theory approximations to procedures (4.58), (4.59), and (4.60) are obtained by replacing u_α and $u_{\alpha/2}$ by z_α and $z_{\alpha/2}$, respectively.

The `R` function `pFligPoli` (with `method="Monte carlo"`) performs a Monte Carlo approximation to the P-value of the statistic $\hat{U}$ and the `R` function `pFligPoli` (with `method="Asymptotic"`) performs the large-sample approximation.

Ties

If there are ties among the N sample observations, replace the placement formulas (4.52) and (4.53) by

$$P_i = \Big\{[\text{number of sample } Y \text{ observations less than } X_i] + \tfrac{1}{2}[\text{number of sample } Y \text{ observations equal to } X_i]\Big\} \tag{4.61}$$

and

$$Q_j = \Big\{[\text{number of sample } X \text{ observations less than } Y_j] + \tfrac{1}{2}[\text{number of sample } X \text{ observations equal to } Y_j]\Big\}, \tag{4.62}$$

respectively.

Table 4.7 Plasma Glucose Values

Healthy geese	Lead-poisoned geese
297	293
340	291
325	289
227	430
277	510
337	353
250	318
290	

Source: G. L. March, T. M. John, B. A. McKeown, L. Sileo and J. C. George (1976).

EXAMPLE 4.5 ***Plasma Glucose in Geese.***

March et al. (1976) were interested in, among other things, examining the differences between healthy (normal) and lead-poisoned Canadian geese. In particular, one of the measures examined was plasma glucose (mg/100 ml plasma). The data they obtained for eight healthy and seven lead-poisoned geese are given in Table 4.7.

Labeling the lead-poisoned geese as the Y-sample (because there are fewer lead-poisoned observations), the authors were interested in testing H_0': $\theta_x = \theta_y$ versus H_1': $\theta_y > \theta_x$; that is, do lead-poisoned Canadian geese tend to have larger plasma glucose values than healthy geese? Computing the placements for the X and Y observations, we obtain

$$P_1 = 3, \quad P_2 = 4, \quad P_3 = 4, \quad P_4 = 0, \quad P_5 = 0, \quad P_6 = 4, \quad P_7 = 0, \quad P_8 = 1$$

and

$$Q_1 = 4, \quad Q_2 = 4, \quad Q_3 = 3, \quad Q_4 = 8, \quad Q_5 = 8, \quad Q_6 = 8, \quad Q_7 = 5.$$

Thus,

$$\overline{P} = \frac{3+4+4+0+0+4+0+1}{8} = 2$$

and

$$\overline{Q} = \frac{4+4+3+8+8+8+5}{7} = \frac{40}{7}.$$

Using the values in (4.56), we have

$$\begin{aligned} V_1 = [&(3-2)^2 + (4-2)^2 + (4-2)^2 + (0-2)^2 + (0-2)^2 \\ &+ (4-2)^2 + (0-2)^2 + (1-2)^2] = 26 \end{aligned}$$

and

$$V_2 = \left[\left(4-\frac{40}{7}\right)^2+\left(4-\frac{40}{7}\right)^2+\left(3-\frac{40}{7}\right)^2+\left(8-\frac{40}{7}\right)^2 + \left(8-\frac{40}{7}\right)^2+\left(8-\frac{40}{7}\right)^2+\left(5-\frac{40}{7}\right)^2\right] = \frac{206}{7}.$$

Combining these quantities, we obtain

$$\widehat{U} = \frac{(40-16)}{2\left[26+\frac{206}{7}+2\left(\frac{40}{7}\right)\right]^{1/2}} = 1.468.$$

From the R commands `cFligPoli(alpha=0.05035),m=8,n=7` and `cFligPolio(alpha= 0.1001, m=8,n=7)`, we find $u_{.05035} = 1.807$ and $u_{.1001} = 1.310$. Thus for these data, the P-value obtained for testing $H_0' : \theta_x = \theta_y$ versus $H_1' : \theta_y = \theta_x$ is between .05035 and .1001. Also see Comment 27.

Comments

23. *Relationship of $\widehat{U}$ to U.* The statistic $\widehat{U}$ defined by (4.57) is of the form

$$\widehat{U} = \frac{n^{1/2}\left\{(U/mn)-\frac{1}{2}\right\}}{\widehat{\sigma}}, \tag{4.63}$$

where U is the Mann–Whitney statistic defined by (4.15) and

$$\widehat{\sigma}^2 = \frac{\sum_{j=1}^{n}(Q_j-\overline{Q})^2+\sum_{i=1}^{m}(P_i-\overline{P})^2+\overline{PQ}}{m^2n}.$$

Fligner and Policello (1981) point out that when written in the form (4.57), namely,

$$\widehat{U} = \frac{\sum_{j=1}^{n}Q_j-\sum_{i=1}^{m}P_i}{2\{V_1+V_2+\overline{P}\,\overline{Q}\}^{1/2}},$$

$\widehat{U}$ resembles Welch's t statistic (Welch 1937, 1947) for the normal theory Behrens–Fisher problem.

24. *Symmetry of the Distribution of $\widehat{U}$.* When H_0: [Identical X and Y distributions] is true, the distribution of $\widehat{U}$ is symmetric about its mean 0, which implies that

$$P_0(\widehat{U} \geq x) = P_0(\widehat{U} \leq -x)$$

for every x. From this, it follows that the lower αth percentile for the null H_0 distribution of $\widehat{U}$ is $-u_\alpha$; hence, its use in the test of H_0' versus H_2' defined by (4.59).

25. *Maintaining Levels.* The test procedures in (4.58), (4.59), and (4.60) have *exact* significance levels equal to α for testing H_0: [Identical X and Y distributions]. However, they also maintain *approximate* level α for the more general null hypothesis $H_0' : \theta_x = \theta_y$, without requiring equal variances or identical distributional forms for the two underlying population.

26. *Consistency of the Test Based on* $\widehat{U}$. Fligner and Policello (1981) consider the consistency of their test based on $\widehat{U}$. To test $H_0 : \theta_x = \theta_y$ versus $\theta_x < \theta_y$, it is necessary to impose conditions on F and G to ensure that whenever $\theta_x = \theta_y$, we have $P(X < Y) = \frac{1}{2}$ and whenever $\theta_x < \theta_y$, we have $P(X < Y) > \frac{1}{2}$. Fligner and Policello point out that a sufficient condition is that F and G be symmetric.
27. Exact Fligner-Policello test. The exact P-value for the Fligner-Policello test can be obtained from the `R` function `pFligPoli`. The `R` function `pFligPoli` computes the P-value of the statistic $\hat{U}$ based on the exact calculations, a Monte Carlo simulation, or the large-sample approximation. By default, the exact calculations are used when $\binom{m+n}{n} \leq 10{,}000$ and a Monte Carlo simulation otherwise. The user may specify which method to use by the `method=option` and, if applicable, the number of Monte Carlo samples to use by the `n.mc=option`. Applying `pFligPoli` to the plasma glucose data of Table 4.7 yields an exact P-value of .0808.

Properties

1. *Consistency*. Assuming F, G are symmetric, the tests defined by (4.58), (4.59), and (4.60) are consistent against the alternatives for which $\theta_x < \theta_y$, $\theta_x > \theta_y$, and $\theta_x \neq \theta_y$, respectively.
2. *Asymptotic Normality*. See Fligner and Policello (1981).
3. *Efficiency*. See Fligner and Policello (1981) and Section 4.5.

Problems

41. Apply the test based on $\widehat{U}$ to the data of Table 4.1. Compare your results with those of Example 4.1.

42. Apply the test based on $\widehat{U}$ to the data of Table 4.2. Compare your results with those of Example 4.2.

43. Apply the test based on $\widehat{U}$ to the data of Table 4.3. Compare your results with those of Problem 1.

44. Apply the test based on $\widehat{U}$ to the data of Table 4.4. Compare your results with those of Problem 5.

45. Apply the test based on $\widehat{U}$ to the data of Table 4.5. Compare your results with those of Problems 15, 16, and 17.

46. Establish (4.63) directly or illustrate it using an example.

47. Show that $\widehat{U}$ is a rank statistic. That is, show that you can compute $\widehat{U}$ from knowledge of $S_1, \ldots, S_n$, where S_j = rank of Y_j in the joint ranking of the N X's and Y's.

4.5 EFFICIENCIES OF TWO-SAMPLE LOCATION PROCEDURES

Recall the normal theory t-test based on

$$t = \frac{\overline{Y} - \overline{X}}{s_p \sqrt{\frac{m+n}{mn}}},$$

where

$$s_p^2 = \frac{\sum_{i=1}^{m}(X_i - \overline{X})^2 + \sum_{j=1}^{n}(Y_j - \overline{Y})^2}{m+n-2}$$

is the pooled variance. The Pitman asymptotic relative efficiency of the test based on W versus the test based on t is

$$E(W,t) = 12\sigma_F^2 \left\{ \int f^2 \right\}^2. \tag{4.64}$$

In (4.64), σ_F^2 is the variance of the population with distribution F and f is the probability density corresponding to F. The parameter $\int f^2$ is the area under the curve of f^2.

Equation (4.64) was derived by Pitman (1948) in the testing context and shown by Hodges and Lehmann (1963) to hold also for the asymptotic relative efficiency of the point estimator $\widehat{\Delta}$ with respect to $\overline{\Delta} = \overline{Y} - \overline{\Delta}$. Lehmann (1963c) showed that (4.64) also gives the asymptotic relative efficiency of the confidence interval derived from W to that of the confidence interval derived from $\overline{Y} - \overline{X}$.

Hodges and Lehmann (1956) showed that for all populations, $E(W,t)$ is at least .864. Thus the most efficiency one can lose when employing the Wilcoxon test instead of the t-test is about 14%. When F is the normal distribution (the home turf of the t-test), $E(W,t) = .955$. For many populations, $E(W,t)$ exceeds 1, and it can be infinite, as it is in the case when F is Cauchy. Some values of $E(W,t)$ are in the following table.

F	Normal	Uniform	Logistic	Double exponential	Cauchy	Exponential
$E(W,t)$	.955	1.000	1.097	1.500	∞	3.00

Some asymptotic relative efficiency values of van der Waerden's c test relative to the test based on W are in the following table.

F	Normal	Uniform	Logistic	Double exponential	Cauchy	Exponential
$E(c,W)$	1.047	∞	.955	.847	.708	∞

These values are also the values of the asymptotic relative efficiency $E(c_1, W)$, where c_1 is the Fisher–Yates–Terry–Hoeffding statistic.

Chernoff and Savage (1958) showed that for all populations, the asymptotic relative efficiency of the Fisher–Yates–Terry–Hoeffding test with respect to the t-test is always greater than or equal to 1. It equals 1 when F is normal.

For model (4.2), the asymptotic relative efficiency of the Fligner–Policello test based on $\widehat{U}$ with respect to W is 1 for all F.

Chapter 5

The Two-Sample Dispersion Problem and Other Two-Sample Problems

INTRODUCTION

In this chapter the data once again consist of two independent random samples, one sample from each of two underlying populations. This is the same as the data setting considered in Chapter 4, where we discussed procedures designed for statistical analyses in which the primary interest was on possible differences in the locations (medians) of the populations. In this chapter we deal with statistical procedures designed to make inferences about possible differences other than location between two populations.

In Section 5.1 we present a distribution-free rank test for the hypothesis of equal scale parameters when the two underlying populations have a common median. Section 5.2 is devoted to an asymptotically distribution-free test for equality of scale parameters when the assumption of common medians is not justified. In Section 5.3 we consider a distribution-free rank test for the dual hypothesis of equal location and equal scale parameters for the underlying populations. Section 5.4 contains a distribution-free test of the general hypothesis that two populations are identical in all respects. Some aspects of the asymptotic relative efficiencies of the procedures in this chapter with respect to their normal theory counterparts are discussed in Section 5.5.

Data. We obtain $N = m + n$ observations $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$.

Assumptions

A1. The observations $X_1, \ldots, X_m$ are a random sample from a continuous population 1; that is, the X's are mutually independent and identically distributed. The observations $Y_1, \ldots, Y_n$ are a random sample from a continuous population 2, so that the Y's are also mutually independent and identically distributed.

A2. The X's and Y's are mutually independent. Thus, in addition to assumptions of independence within each sample, we also assume independence between the two samples.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

5.1 A DISTRIBUTION-FREE RANK TEST FOR DISPERSION—MEDIANS EQUAL (ANSARI–BRADLEY)

Hypothesis

Let F and G be the distribution functions corresponding to populations 1 and 2, respectively. The null hypothesis of interest here is that the X and Y variables have the same probability distribution but that their common distribution is not specified. Formally stated, this null hypothesis is

$$H_0 : [F(t) = G(t), \text{ for every } t]. \tag{5.1}$$

The typical alternative hypothesis in a two-sample dispersion problem specifies that the Y population has greater (or less) variability associated with it than does the X population. One model that is often used to describe such alternatives is the location-scale parameter model. In our two-sample setting, this location-scale parameter model corresponds to taking

$$F(t) = H\left(\frac{t-\theta_1}{\eta_1}\right) \quad \text{and} \quad G(t) = H\left(\frac{t-\theta_2}{\eta_2}\right), \quad -\infty < t < \infty, \tag{5.2}$$

where $H(u)$ is the distribution function for a continuous distribution with median 0, so that $F(\theta_1) = G(\theta_2) = \frac{1}{2}$. Thus, θ_1 and θ_2 are the population medians for the X and Y distributions, respectively. Moreover, η_1 and η_2 are the scale parameters associated with the X and Y distributions, respectively. Model (5.2) states that the Y population has the same general form as the X population, but they could have different medians and scale parameters. Another way to express this is to write

$$\frac{X-\theta_1}{\eta_1} \stackrel{d}{=} \frac{Y-\theta_2}{\eta_2}, \tag{5.3}$$

where the symbol $\stackrel{d}{=}$ means "has the same distribution as."

This two-sample location-scale problem will be further discussed in this most general context in Sections 5.2 and 5.3. In this section, however, we impose the further restriction that $\theta_1 = \theta_2$; that is, we also assume

A3. The median (θ_1) of the X population is equal to the median (θ_2) of the Y population.

Under this additional assumption, A3, the equal-in-distribution statement in (5.3) simplifies to

$$\frac{X-\theta}{\eta_1} \stackrel{d}{=} \frac{Y-\theta}{\eta_2}, \tag{5.4}$$

where θ is the common median and the only possible difference between the X and Y populations is in their respective scale parameters, as illustrated in Figure 5.1. (If the medians θ_1 and θ_2 of the X and Y populations are not necessarily equal but are known, the shifted variables $X_1 - \theta_1, \ldots, X_m - \theta_1$ and $Y_1 - \theta_2, \ldots, Y_n - \theta_2$ will satisfy Assumptions A1, A2, and A3. In such a situation, the procedures of this section can be applied to the shifted $(X - \theta_1)$ and $(Y - \theta_2)$ sample observations. For more about this known medians setting, see Comment 1.)

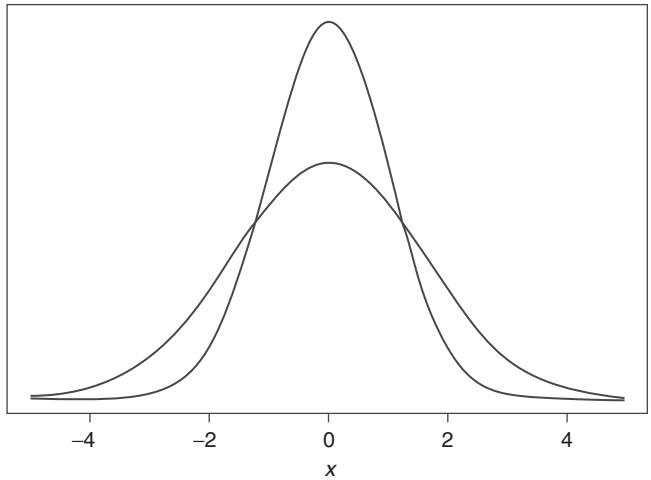

Figure 5.1 Probability distributions with the same general form and equal medians but different scale parameters.

Under Assumptions A1–A3, the parameter of interest in this section is the ratio of the scale parameters, $\gamma = (\eta_1/\eta_2)$ (see Comment 3). If the variance of population 1, var(X), exists and (5.4) is satisfied, then the variance of population 2, var(Y), also exists and

$$\gamma^2 = \left[\frac{\text{var}(X)}{\text{var}(Y)}\right], \tag{5.5}$$

the ratio of population variances (also see Comment 7). In terms of this location-scale parameter model with equal location parameters, as given in (5.4), the null hypothesis H_0 (5.1) reduces to $H_0 : \gamma^2 = 1$, corresponding to the assertion that the population scale parameters are equal.

Procedure

To compute the Ansari–Bradley two-sample scale statistic C, order the combined sample of $N = (m + n)$ X-values and Y-values from least to greatest. Assign the score 1 to both the smallest and largest observations in this combined sample, assign the score 2 to the second smallest and second largest, and continue in the manner. If N is an even integer, the array of assigned scores is $1, 2, 3, \ldots, N/2, N/2, \ldots, 3, 2, 1$. If N is an odd integer, the array of assigned scores is $1, 2, 3, \ldots, (N-1)/2, (N+1)/2, (N-1)/2, \ldots, 3, 2, 1$. Let R_j denote the score assigned in this manner to Y_j, for $j = 1, \ldots, n$, and set

$$C = \sum_{j=1}^{n} R_j. \tag{5.6}$$

Thus the statistic C is the sum of the scores assigned via this scheme to the Y observations.

a. *One-Sided Upper-Tail Test.* To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_1 : \gamma^2 > 1,$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } C \geq c_\alpha; \quad \text{otherwise do not reject,} \tag{5.7}$$

where the constant c_α is chosen to make the type I error probability equal to α. The constant c_α is the upper α percentile for the null ($\gamma^2 = 1$) distribution of C. Comment 4 explains how to obtain the critical value c_α for sample sizes m and n and available levels of α.

b. *One-Sided Lower-Tail Test*. To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_2 : \gamma^2 < 1,$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } C \leq [c_{1-\alpha} - 1]; \quad \text{otherwise do not reject,} \tag{5.8}$$

where, as with the upper-tail test in (5.7), the appropriate value of $c_{1-\alpha}$ is obtained as stipulated in Comment 4.

c. *Two-Sided Test*. To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_3 : \gamma^2 \neq 1,$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } C \geq c_{\alpha_1} \text{ or } C \leq [c_{1-\alpha_2} - 1]; \text{ otherwise do not reject,} \tag{5.9}$$

where $\alpha_1 + \alpha_2 = \alpha$ and the appropriate values of c_{α_1} and $c_{1-\alpha_2}$ are obtained as directed in Comment 4. We note that the null distribution of C is symmetric when $N = (m + n)$ is an even number (see Comment 5). In such a case, it is most natural to place an equal amount of probability in each tail of the null distribution of C, corresponding to setting $\alpha_1 = \alpha_2 = \alpha/2$. Thus, when N is even, the two-sided symmetric version of procedure (5.9) uses the critical values $c_{\alpha/2}$ and $[c_{1-(\alpha/2)} - 1]$.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of C, suitably standardized. For this purpose we first need to know the expected value and variance of C when the null hypothesis is true. Since the set of scores being assigned to the jointly ranked sample X and Y observations (see Procedure) depends on whether N is an even or odd integer, it is not surprising that the form of the mean and variance for C also depends on whether N is even or odd. When H_0 is true and $N = m + n$ is an even number, the expected value and variance of C are

$$E_0(C) = \frac{n(N+2)}{4} \tag{5.10}$$

and

$$\text{var}_0(C) = \frac{mn(N+2)(N-2)}{48(N-1)}, \tag{5.11}$$

respectively. When N is an odd integer, the null expected value and variance of C are

$$E_0(C) = \frac{n(N+1)^2}{4N} \tag{5.12}$$

and

$$\text{var}_0(C) = \frac{mn(N+1)(3+N^2)}{48N^2}, \tag{5.13}$$

respectively. These expressions for $E_0(C)$ and $\text{var}_0(C)$ are verified by direct calculations in Comment 8 for the special cases of $m = n = 2$ (where $N = 4$ is even) and $m = 3$, $n = 2$ (where $N = 5$ is odd). General derivations of the null expected value and variance expressions in (5.10), (5.11), (5.12), and (5.13) are presented in Comment 9.

For general N (even or odd), the standardized version of C is given by

$$C^* = \frac{C - E_0(C)}{\{\text{var}_0(C)\}^{1/2}}, \tag{5.14}$$

where $E_0(C)$ and $\text{var}_0(C)$ correspond to expressions (5.10) and (5.11), respectively, if N is even or to expressions (5.12) and (5.13), respectively, if N is odd. In either case, when H_0 is true, C^* has, as $\min(m, n)$ tends to infinity, an asymptotic $N(0, 1)$ distribution (see Comment 9 for indications of the proof). The normal theory approximation for procedure (5.7) is

$$\text{Reject } H_0 \text{ if } C^* \geq z_\alpha; \quad \text{otherwise do not reject,} \tag{5.15}$$

the normal theory approximation for procedure (5.8) is

$$\text{Reject } H_0 \text{ if } C^* \leq -z_\alpha; \quad \text{otherwise do not reject,} \tag{5.16}$$

and the normal theory approximation for procedure (5.9) is

$$\text{Reject } H_0 \text{ if } |C^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{5.17}$$

Ties

If there are ties among the X and/or Y observations, assign each of the observations in a tied group the average of the integer scores that are associated with the tied group. After computing C with these average scores for tied observations, use the appropriate procedure (5.7), (5.8), or (5.9) with this tie-averaged value of C. Note, however, that this test associated with tied X and/or Y observations is only approximately, and not exactly, of significance level α. (To get an exact level α test even in this tied setting, see Comment 11.)

When applying the large-sample approximation, an additional factor must be taken into account. Although ties among the X and/or Y observations do not affect the null expected value of C, its null variance is reduced to

$$\text{var}_0(C) = \frac{mn\left[16\sum_{j=1}^{g} t_j r_j^2 - (N)(N+2)^2\right]}{16N(N-1)} \tag{5.18}$$

in the presence of ties when N is even and to

$$\text{var}_0(C) = \frac{mn\left[16N\sum_{j=1}^{g} t_j r_j^2 - (N+1)^4\right]}{16N^2(N-1)} \tag{5.19}$$

when N is odd, where in (5.18) and (5.19) g denotes the number of tied groups among the N sample observations, t_j is the size of tied group j, and r_j is the average score associated with the observations in tied group j. We note that an untied observation is considered to be a tied "group" of size 1. In particular, if there are no ties among the X's and/or Y's, then $g = N$ and $t_j = 1$ for $j = 1, \ldots, N$. In this case of no tied sample observations, we have

$$\sum_{j=1}^{g} t_j r_j^2 = 2\sum_{j=1}^{N/2} j^2 = \frac{2\left(\frac{N}{2}\right)\left(\frac{N}{2}+1\right)\left(2\left(\frac{N}{2}\right)+1\right)}{6} = \frac{N(N+1)(N+2)}{12},$$

when N is an even integer, and

$$\begin{aligned}\sum_{j=1}^{g} t_j r_j^2 &= 2\sum_{j=1}^{(N-1)/2} j^2 + \left(\frac{N+1}{2}\right)^2 \\ &= \frac{2}{6}\left(\frac{N-1}{2}\right)\left(\frac{N-1}{2}+1\right)\left(2\left(\frac{N-1}{2}\right)+1\right) + \left(\frac{N+1}{2}\right)^2 \\ &= \left(\frac{N+1}{12}\right)(N^2+2N+3),\end{aligned}$$

when N is an odd integer. Using these expressions for $\sum_{j=1}^{g} t_j r_j^2$, the associated ties-adjusted expressions for $\text{var}_0(C)$ given in (5.18) and (5.19) reduce to the corresponding untied null variances in (5.11) and (5.13), respectively, in the case of no tied observations.

As a consequence of the effect that ties have on the null variance of C, the following modification is needed to apply the large-sample approximation when there are ties among the X and/or Y observations. Compute C using average scores and set

$$C^* = \frac{C - E_0(C)}{\{\text{var}_0(C)\}^{1/2}}, \tag{5.20}$$

where $E_0(C)$ and $\text{var}_0(C)$ are now given by displays (5.10) and (5.18), respectively, if N is even or by displays (5.12) and (5.19), respectively, if N is odd. With this modified form of C^*, approximations (5.15), (5.16), or (5.17) can be applied.

EXAMPLE 5.1 ***Serum Iron Determination.***

The data in Table 5.1 are a portion of the data obtained by Jung and Parekh (1970) in a study concerned with techniques for direct determination of serum iron. In particular, they attempted to eliminate some of the problems associated with other commonly used methods, which often result in turbidity of the analyzed serum, as well as requiring large samples and slow, tedious analyses. To accomplish this, the authors proposed an improved method for serum iron determination based on a different detergent. One of the purposes of their investigation was to study the accuracy of their method for serum iron determination in comparison to a method due to Ramsay (1957). Twenty duplicate analyses were made, each by the proposed method and by the method of Ramsay, using Hyland control sera containing 105 μg of serum iron per 100 ml. Table 5.1 gives the serum iron detected (in μg/100 ml) for the 40 analyses in the study.

From the point of view of procedural technique, the Jung–Parekh method competes favorably with the Ramsay method for serum iron determination. An additional concern,

Table 5.1 Serum Iron (μg/100 ml) Determination Using Hyland Control Sera

Ramsay method	Jung–Parekh method
111	107
107	108
100	106
99	98
102	105
106	103
109	110
108	105
104	104
99	100
101	96
96	108
97	103
102	104
107	114
113	114
116	113
113	108
110	106
98	99

Source: D.H. Jung and A.C. Parekh (1970).

however, is whether there is a loss of accuracy when the Jung–Parekh procedure is used instead of the Ramsay procedure. As a result, the alternative of interest in this example is greater dispersion or variation for the Jung–Parekh method of serum iron determination than for the method of Ramsay. Hence, letting Y correspond to the Ramsay determinations and X to the Jung–Parekh determinations, we are interested in a one-sided test designed to detect the alternative $H_1 : \gamma^2 > 1$. Since there are ties among the X and Y sample observations and $N = m + n = 20 + 20 = 40$ is an even integer, we will apply the large-sample approximation (with ties), as detailed in (5.15) and (5.20), to procedure (5.7).

For the purpose of illustration, we consider the approximate level $\alpha = .05$. Hence, using the R command `pnorm(·)`, we set `1-pnorm(`$z_{.05}$`)` $= .05$ and obtain $z_{.05} = 1.645$, and the large-sample approximation to procedure (5.7) is given by

$$\text{Reject } H_0 \text{ if} \frac{C - E_0(C)}{\{\text{var}_0(C)\}^{1/2}} \geq 1.645,$$

where $E_0(C)$ and $\text{var}_0(C)$ are given by expressions (5.10) and (5.18), respectively.

To calculate C (5.6) we need the Ansari–Bradley ranks of the 20 Y (Ramsay) observations. In the following display we list in order (from least to greatest) the combined sample of 40 X (Jung–Parekh) and Y (Ramsay) values and assign the ranks according to the Ansari–Bradley scheme.

Ansari–Bradley Ranking Scheme for the Data of Table 5.1

Y	*X*	*Y*	*Y*	*X*	*Y*	*Y*	*X*	*X*	*Y*
96	96	97	98	98	99	99	99	100	100
1.5	1.5	3	4.5	4.5	7	7	7	9.5	9.5
Y	*Y*	*Y*	*X*	*X*	*X*	*Y*	*X*	*X*	*X*
101	102	102	103	103	104	104	104	105	105
11	12.5	12.5	14.5	14.5	17	17	17	19.5	19.5
X	*X*	*Y*	*Y*	*X*	*Y*	*Y*	*X*	*X*	*X*
106	106	106	107	107	107	108	108	108	108
19	19	19	16	16	16	12.5	12.5	12.5	12.5
Y	*X*	*Y*	*Y*	*X*	*Y*	*Y*	*X*	*X*	*Y*
109	110	110	111	113	113	113	114	114	116
10	8.5	8.5	7	5	5	5	2.5	2.5	1

Thus $C = \sum_{i=1}^{20} R_i = 185.5$. In order to calculate C^*, we need to evaluate expressions (5.10) and (5.18). We illustrate the calculation of $\sum_{j=1}^{g} t_j r_j^2$ in the following table, where for our data there are $g = 19$ tied groups.

Thus, we have $\sum_{j=1}^{19} t_j r_j^2 = 5721$, and from (5.10) and (5.18), we obtain

$$C^* = \frac{185.5 - [20(42)/4]}{\{(20)(20)[16(5721) - 40(42)^2]/[16(40)(39)]\}^{1/2}} = -1.34,$$

which tells us not to reject H_0 at the approximate $\alpha = .05$ level, since $C^* = -1.34 < 1.645 = z_{.05}$. Hence, there is not sufficient evidence to indicate loss of accuracy when the Jung–Parekh method is used instead of the Ramsay method.

Tied group	t_j	r_j^2	$t_j r_j^2$
1	2	2.25	4.5
2	1	9	9
3	2	20.25	40.5
4	3	49	147
5	2	90.25	180.5
6	1	121	121
7	2	156.25	312.5
8	2	210.25	420.5
9	3	289	867
10	2	380.25	760.5
11	3	361	1083
12	3	256	768
13	4	156.25	625
14	1	100	100
15	2	72.25	144.5
16	1	49	49
17	3	25	75
18	2	6.25	12.5
19	1	1	1

Since the one-sided P-value for these data is the lowest significance level at which we can reject H_0 in favor of $\gamma^2 > 1$ with the observed value of the test statistic C^*, we see, using the R command `pnorm(·)`, that the P-value for these data is approximately $P_0(C^* \geq -1.34) \approx 1 -$ `pnorm(-1.34)` $= (1 - .0901) = .9099$. Thus, there is absolutely no evidence in the sample data to indicate any loss of accuracy with the Jung–Parekh method. In fact, the C^* value of -1.34 actually provides evidence pointing in the other direction, namely, $\gamma^2 < 1$, corresponding to improved accuracy with the Jung–Parekh method.

Comments

1. *Known Population Medians*. When the population median, θ_2, for the Y observations is known to be equal to $\theta_1 + \xi$, where θ_1 is the population median for the X observations and ξ is a known constant, we can create modified observations $X_i' = X_i + \xi, i = 1, \ldots, m$ and apply the Ansari–Bradley procedures of this section to the modified X' observations and the unchanged Y observations.
2. *Testing γ^2 Equal to Some Specified Value Other Than One*. To test the hypothesis $\gamma^2 = \gamma_0^2$, where γ_0^2 is some specified positive number different from 1, when the common median for the underlying X and Y populations has known value θ_0, we obtain the modified observations $X_i' = (X_i - \theta_0)/\gamma_0$, for $i = 1, \ldots, m$, and $Y_j' = (Y_j - \theta_0)$, for $j = 1, \ldots, n$, and compute C (5.6) using the X''s and Y''s (instead of the X's and Y's). Procedures (5.7), (5.8), or (5.9) or the corresponding large-sample approximations (5.15), (5.16), or (5.17) may then be applied as described.

3. *Motivation for the Test*. Under Assumptions A1– A3, the X and Y populations have the same median. Suppose, for example, that γ^2 is greater than 1. Then the X values would tend to be more spread out than the Y values. Thus, the Y's would tend to get larger scores than the X's from the scheme described in the Procedure and C (5.6) would tend to be larger. (Visualize an extreme sample where the sample values, when ordered, fall in the pattern *XXYYYXX*.) This serves as partial motivation for the one-sided upper-tail test procedure given in (5.7).
4. *Derivation of the Distribution of C under H_0 (No-Ties Case)*. Under H_0 (5.1), each of the $\binom{N}{n}$ possible "meshings" of the X's and Y's has probability $1/\binom{N}{n}$. This fact can be used to obtain the null distribution of C (5.6). We illustrate the steps involved in constructing this null distribution for the two cases $m = 3$, $n = 2$ (where $N = 5$ is odd) and $m = 2$, $n = 2$ (where $N = 4$ is even). First, for $m = 3$ and $n = 2$, we use the set of scores $\{1, 1, 2, 2, 3\}$. Let $R^{(1)} < R^{(2)}$ denote the ordered Y scores so that $C = R_1 + R_2 = R^{(1)} + R^{(2)}$. The $\binom{5}{2} = 10$ possible meshings and associated values of $(R^{(1)}, R^{(2)})$ and C are given in the following table.

Meshing	Probability	$(R^{(1)}, R^{(2)})$	$C = R^{(1)} + R^{(2)}$
YYXXX	$\frac{1}{10}$	(1, 2)	3
YXYXX	$\frac{1}{10}$	(1, 3)	4
YXXYX	$\frac{1}{10}$	(1, 2)	3
YXXXY	$\frac{1}{10}$	(1, 1)	2
XYYXX	$\frac{1}{10}$	(2, 3)	5
XYXYX	$\frac{1}{10}$	(2, 2)	4
XYXXY	$\frac{1}{10}$	(1, 2)	3
XXYYX	$\frac{1}{10}$	(2, 3)	5
XXYXY	$\frac{1}{10}$	(1, 3)	4
XXXYY	$\frac{1}{10}$	(1, 2)	3

Thus, for example, the probability is $\frac{3}{10}$ under H_0 that C is equal to 4, because $C = 4$ when either of the exclusive outcomes $(R^{(1)}, R^{(2)}) = (1, 3)$ or $(R^{(1)}, R^{(2)}) = (2, 2)$ occurs. These two outcomes for $(R^{(1)}, R^{(2)})$ are associated with three mutually exclusive meshings, each with null probability $\frac{1}{10}$. Hence, it follows that $P_0(C = 4) = 3(\frac{1}{10})$. Proceeding in the same manner for all possible values for C and simplifying, we obtain the null distribution.

Possible value of C	Probability under H_0
2	$\frac{1}{10}$
3	$\frac{4}{10}$
4	$\frac{3}{10}$
5	$\frac{2}{10}$

The probability, under H_0, that C is greater than or equal to 4, for example, is therefore

$$P_0(C \geq 4) = P_0(C = 4) + P_0(C = 5) = .3 + .2 = .5,$$

so that $c_{.5} = 4$. Note also that $c_{.2} = 5$.

For the case of $m = n = 2$ (where $N = 4$ is even), we use the set of scores $\{1, 1, 2, 2\}$. The $\binom{4}{2} = 6$ possible meshings, as well as the associated ordered Y-scores $(R^{(1)}, R^{(2)})$ and values of C are given in the following table.

Meshing	Probability	$(R^{(1)}, R^{(2)})$	$C = R^{(1)} + R^{(2)}$
XXYY	$\frac{1}{6}$	(1, 2)	3
XYXY	$\frac{1}{6}$	(1, 2)	3
YXXY	$\frac{1}{6}$	(1, 1)	2
XYYX	$\frac{1}{6}$	(2, 2)	4
YXYX	$\frac{1}{6}$	(1, 2)	3
YYXX	$\frac{1}{6}$	(1, 2)	3

Proceeding as for the previous case of $m = 3$, $n = 2$, we obtain the null distribution for C.

Possible value of C	Probability under H_0
2	$\frac{1}{6}$
3	$\frac{4}{6}$
4	$\frac{1}{6}$

Note that we have derived the null distribution of C without specifying the form of the common (under H_0) underlying X and Y populations beyond the point of requiring that they be continuous. This is why the test procedures based on C are called *distribution-free procedures*. From the null distribution of C, we can determine the critical value c_α and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying continuous distribution for the X and Y observations.

For given sample sizes m and n, the R command `cAnsBrad`(α, m, n) can be used to find the available upper-tail critical values c_α for possible values of C. For a given available significance level α, the critical value c_α then corresponds to $P_0(C \geq c_\alpha) = \alpha$ and is given by `cAnsBrad`$(\alpha, m, n) = c_\alpha$. Thus, for example, for $m = 8$ and $n = 4$, we have $P_0(C \geq 20) = .0283$ so that $c_{.0283} = 20$ for $m = 8$ and $n = 4$.

5. *Symmetry of the Distribution of C under the Null Hypothesis When $N = m + n$ Is Even*. When H_0 is true and $N = m + n$ is an even integer, the distribution of

C is symmetric about its mean $n(N+2)/4$. (See Comment 4 for verification of this when $m = n = 2$.) This implies that when N is even

$$P_0(C \le x) = P_0\left(C \ge \frac{n(N+2)}{2} - x\right), \tag{5.21}$$

for every possible value of x.

Equation (5.21) is directly used to convert upper-tail probabilities, as obtained from `cAnsBrad(`$\cdot$, *m, n*), to lower-tail probabilities when N is even. Thus, the lower-tail critical value $[c_{1-\alpha} - 1]$ used in test procedures (5.8) or (5.9) can be expressed in terms of the upper-tail critical value c_α by

$$[c_{1-\alpha} - 1] = \left[\frac{n(N+2)}{2} - c_\alpha\right], \tag{5.22}$$

when N is even.

6. *Equivalent Form*. The statistic C (5.6) is the sum of the scores assigned to the Y observations by the Ansari–Bradley scoring scheme described in the Procedure. Test procedures (5.7), (5.8), and (5.9) could equivalently be based on the statistic $C' = [\text{sum of the scores assigned by this scheme to the } X \text{ observations}]$, because $C' = [N(N+2)/4] - C$ when $N = m + n$ is even and $C' = [(N-1)^2/4] - C$ when N is odd (see Problem 6).
7. *Assumptions*. We can use the Ansari–Bradley test procedures in (5.7), (5.8), or (5.9) without even requiring that the variances for the X and Y populations exist. Indeed, our Assumptions A1–A3 for this section do not specify anything about the existence of even the first moments of the X and Y populations. However, when the first two moments (and, therefore, the variance) for the underlying distributional model $H(u)$ in (5.2) exist, we see from the equal-in-distribution statement in (5.3) that

$$\text{var}\left(\frac{X}{\eta_1}\right) = \text{var}\left(\frac{Y}{\eta_2}\right),$$

which, in turn, implies that

$$[\text{var}(X)]/\eta_1^2 = [\text{var}(Y)]/\eta_2^2.$$

Thus, when the variances exist, we see that $\gamma^2 = [\eta_1^2/\eta_2^2] = [\text{var}(X)/\text{var}(Y)]$.

Assumptions A1–A3 do imply that the only possible difference between the X and Y populations is the difference in scale parameters. In particular, these assumptions imply that the two populations do not differ in location, as they have a common median θ (see Comment 1 for a slight relaxation of this condition). While the requirement of equal medians is not necessary for the classical $\mathcal{F}$-test based on the ratio of the X and Y sample variances, this requirement is essential for the Ansari–Bradley test. For example, suppose that $m = 5$, $n = 4$, and the X and Y probability distributions are such that $P(X < Y) = 0$. Then, for *all* possible X and Y samples, the joint ordering of the five X observations and four Y observations would *always* result in a value of $C = 10$, regardless of the scale parameters for the two populations. That is, in such a setting, *no* information

about γ^2 can be obtained from the joint ranking and the Ansari–Bradley scoring scheme.

Moses (1963) has emphasized this bizarre behavior of tests for dispersion based on joint rankings of the sample X and Y observations and has shown that such tests are inadequate unless strong assumptions (such as equal or known medians) are made concerning the locations of the X and Y populations. For an asymptotically distribution-free test that does not require equal or known medians, see Section 5.2.

8. *Calculation of the Mean and Variance of C under the Null Hypothesis*, H_0. In (5.10) and (5.11), we presented formulas for the mean and variance of C when the null hypothesis is true and $N = (m + n)$ is an even number. The corresponding expressions for the null mean and variance of C when N is an odd number are given in (5.12) and (5.13). In this comment, we illustrate a direct calculation of $E_0(C)$ and $\text{var}_0(C)$ in the particular cases of $m = 3$, $n = 2$ (where $N = 5$ is odd) and $m = n = 2$ (where $N = 4$ is even), using the null distributions of C obtained in Comment 4. (Later, in Comment 9, we present general derivations of $E_0(C)$ and $\text{var}_0(C)$.) The null mean, $E_0(C)$, is obtained by multiplying each possible value of C by its probability under H_0. Thus, for $m = 3$, $n = 2$, we have

$$E_0(C) = 2(.1) + 3(.4) + 4(.3) + 5(.2) = 3.6.$$

This is in agreement with what we obtain using (5.12), namely.

$$E_0(C) = \frac{n(N+1)^2}{4N} = \frac{2(5+1)^2}{4(5)} = 3.6.$$

Similarly, for $m = n = 2$, we have by direct computation from the null distribution of C in Comment 4 that

$$E_0(C) = 2\left(\frac{1}{6}\right) + 3\left(\frac{4}{6}\right) + 4\left(\frac{1}{6}\right) = 3,$$

in agreement with the value obtained from (5.10), namely,

$$E_0(C) = \frac{n(N+2)}{4} = \frac{2(4+2)}{4} = 3.$$

Checks on the expressions for $\text{var}_0(C)$ are also easily performed, using the well-known fact that

$$\text{var}_0(C) = E_0(C^2) - \{E_0(C)\}^2.$$

The required values of $E_0(C^2)$, the second moment of the null distribution of C, are again obtained by multiplying the possible values of C^2 by the corresponding probabilities under H_0. For the case of $m = 3$, $n = 2$, we find that

$$E_0(C^2) = 2^2(.1) + 3^2(.4) + 4^2(.3) + 5^2(.2) = 13.8,$$

yielding

$$\text{var}_0(C) = 13.8 - (3.6)^2 = 13.8 - 12.96 = .84,$$

which is in agreement with the value obtained from (5.13), namely,

$$\begin{aligned}\text{var}_0(C) &= \frac{mn(N+1)(3+N^2)}{48N^2} \\ &= \frac{3(2)(5+1)(3+5^2)}{48(5)^2} = .84.\end{aligned}$$

Similarly, for $m = n = 2$, we have by direct computation from the null distribution in Comment 4 that

$$E_0(C^2) = 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{4}{6}\right) + 4^2\left(\frac{1}{6}\right) = \frac{56}{6},$$

yielding

$$\text{var}_0(C) = \tfrac{56}{6} - (3)^2 = \tfrac{1}{3},$$

which is in agreement with the value obtained from (5.11), namely,

$$\begin{aligned}\text{var}_0(C) &= \frac{mn(N+2)(N-2)}{48(N-1)} \\ &= \frac{2(2)(4+2)(4-2)}{48(4-1)} = \frac{1}{3}.\end{aligned}$$

9. *Large-Sample Approximation.* The statistic C/n is the average of the scores assigned to the Y observations. Since all $\binom{N}{n}$ possible distributions of the appropriate scores (depending on whether N is even or odd) to the X and Y observations are equally likely under H_0, the null distribution of C/n is the same as the distribution of the sample mean for a random sample of size n drawn without replacement from the finite population of scores S_N, where $S_N = \{1, 2, 3, \ldots, N/2, N/2, \ldots, 3, 2, 1\}$ if N is an even number and $S_N = \{1, 2, 3, \ldots, (N-1)/2, (N+1)/2, (N-1)/2, \ldots, 3, 2, 1\}$ if N is odd.

From basic results for a random sample of size n drawn without replacement from a finite population of N elements, we know that

(i) the expected value of the sample average is equal to the average, μ_{pop}, of the finite population,

(ii) the variance of the sample average is equal to

$$\frac{\sigma^2_{\text{pop}}}{n}\left(\frac{N-n}{N-1}\right),$$

where σ^2_{pop} is the variance of the finite population and the factor $(N-n)/(N-1)$ is known as the finite population correction factor.

For the case of N even and the finite population $S_N = \{1, 2, 3, \ldots, N/2, N/2, \ldots, 3, 2, 1\}$, we see that

(iii)
$$\mu_{\text{pop}} = \frac{2}{N}\sum_{i=1}^{N/2} i = \frac{(N/2)[(N/2)+1]}{2(N/2)} = \frac{N+2}{4}$$

and

(iv)
$$\begin{aligned}\sigma^2_{\text{pop}} &= \left[\frac{2}{N}\left(\sum_{i=1}^{N/2} i^2\right) - \left(\frac{N+2}{4}\right)^2\right] \\ &= \left[\frac{(N/2)[(N/2)+1][2(N/2)+1]}{6(N/2)} - \left(\frac{N+2}{4}\right)^2\right] \\ &= \left[\frac{(N+2)(N+1)}{12} - \frac{(N+2)(N+2)}{16}\right] \\ &= \frac{(N+2)(N-2)}{48}.\end{aligned}$$

From (i), (ii), (iii), and (iv), it follows that

$$E_0\left(\frac{C}{n}\right) = \frac{N+2}{4}$$

and

$$\text{var}_0\left(\frac{C}{n}\right) = \left[\frac{(N+2)(N-2)}{48n}\right]\left[\frac{N-n}{N-1}\right] = \frac{m(N+2)(N-2)}{48n(N-1)}.$$

Thus,

$$E_0(C) = nE_0\left(\frac{C}{n}\right) = \frac{n(N+2)}{4}$$

and

$$\text{var}_0(C) = n^2\text{var}_0\left(\frac{C}{n}\right) = \frac{mn(N+2)(N-2)}{48(N-1)},$$

in agreement with the formulas in (5.10) and (5.11). The corresponding expressions for $E_0(C)$ and $\text{var}_0(C)$ when N is an odd integer, as given in (5.12) and (5.13), respectively, can be similarly obtained using the expressions in (i) and (ii) and the finite population

$$S_N = \left\{1, 2, 3, \ldots, \frac{N-1}{2}, \frac{N+1}{2}, \frac{N-1}{2}, \ldots, 3, 2, 1\right\}.$$

For any N (even or odd), the asymptotic normality under H_0 of the standardized

$$C^* = \frac{C - E_0(C)}{\sqrt{\text{var}_0(C)}}$$

follows from standard theory for the mean of a sample from a finite population (cf. Wilks, 1962, p. 268). Asymptotic normality results for C^* are also available under general alternatives to H_0 (see, for example, Ansari and Bradley (1960), Randles and Wolfe (1979), or Hájek and Šidák (1967)).

10. *Lower-Tail Critical Values*. In the expression for the one-sided lower-tail test in (5.9), the critical value is given to be $c_{1-\alpha} - 1$, where $c_{1-\alpha}$ is the upper $(1-\alpha)$th

percentile of the null distribution of C. This means that

$$P_0(C \leq c_{1-\alpha} - 1) = 1 - P_0(C > c_{1-\alpha} - 1) = 1 - P_0(C \geq c_{1-\alpha}),$$

where the last equality follows from the fact that C is a discrete random variable assuming only positive integer values. Since $c_{1-\alpha}$ is the upper $(1-\alpha)$th percentile for the null distribution of C, it follows that

$$P_0(C \leq c_{1-\alpha} - 1) = 1 - (1 - \alpha) = \alpha.$$

Hence, $c_{1-\alpha} - 1$ is, indeed, the *lower* αth percentile for the null distribution of C, as required for the level α one-sided lower-tail test procedure in expression (5.8).

When N is an even integer, we have already noted in Comment 5 that the null distribution of C is symmetric about its mean, $n(N+2)/4$. It follows that $[c_{1-\alpha} - 1] = [\{n(N+2)/2\} - c_\alpha]$ when N is even.

11. *Exact Conditional Distribution of C with Ties*. To have a test with exact significance level even in the presence of ties among the X's and/or Y's, we need to consider all $\binom{N}{n}$ possible assignments of the N observations with n observations serving as Y's and m observations serving as X's. As in Comment 4, it still follows that, under H_0 (5.1), each of the $\binom{N}{n}$ possible "meshings" of the X's and Y's has probability $1/\binom{N}{n}$. The only difference in the case of ties is that we now use average scores in the computation of C for each of these $\binom{N}{n}$ "meshings" leading to the tabulation of the null distribution. We illustrate this construction for N odd (a similar approach will work for N even) and the following $m = 3$, $n = 2$ data: $X_1 = 3.2$, $X_2 = 5.7$, $X_3 = 6.3$, $Y_1 = 1.9$, $Y_2 = 6.3$. The associated average scores assignments (taking into account the tie between X_3 and Y_2) are 2, 3, 1.5, 1, and 1.5, respectively, and the corresponding value of C, the sum of the scores for the Y observations, is $C = 1.5 + 1 = 2.5$. To assess the significance of this value of C, we obtain its conditional distribution by considering the $\binom{5}{2} = 10$ possible assignments of the observations 1.9, 3.2, 5.7, 6.3, and 6.3 to serve as three X observations and two Y observations, or, equivalently the 10 possible assignments of the average scores 1, 1.5, 1.5, 2, and 3 to serve as three X scores and two Y scores. These 10 assignments and the corresponding values of C are as follows.

Y scores	Probability under H_0	Value of C
1, 1.5	$\frac{1}{10}$	2.5
1, 1.5	$\frac{1}{10}$	2.5
1, 2	$\frac{1}{10}$	3
1, 3	$\frac{1}{10}$	4
1.5, 1.5	$\frac{1}{10}$	3
1.5, 2	$\frac{1}{10}$	3.5
1.5, 3	$\frac{1}{10}$	4.5
1.5, 2	$\frac{1}{10}$	3.5
1.5, 3	$\frac{1}{10}$	4.5
2, 3	$\frac{1}{10}$	5

This yields the null tail probabilities

$$P_0(C \geq 5) = \tfrac{1}{10},$$

$$P_0(C \geq 4.5) = \tfrac{3}{10},$$

$$P_0(C \geq 4) = \tfrac{4}{10},$$

$$P_0(C \geq 3.5) = \tfrac{6}{10},$$

$$P_0(C \geq 3) = \tfrac{8}{10},$$

$$P_0(C \geq 2.5) = 1.$$

This distribution is called the *conditional null distribution* or the *permutation null distribution* of C, given the set of tied scores $\{1, 1.5, 1.5, 2, 3\}$. For the particular observed value $C = 2.5$, we have $P_0(C \geq 2.5) = 1$, so that such a value does not indicate a deviation from H_0 in the direction of $\gamma^2 > 1$ (although it would provide marginal support for the alternative $\gamma^2 < 1$).

12. *Confidence Intervals, Confidence Bounds, and Point Estimators for* γ^2. The Ansari–Bradley statistic, C (5.6), is a member of a large class of rank statistics (referred to as *linear rank statistics* in the literature; see, for example, Section 9.3 of Randles and Wolfe (1979)) that can be used to test for equality of scale parameters under the strict assumption of equal or known medians for the X and Y populations. Bauer (1972) has shown how to invert some of these linear rank tests of $\gamma^2 = 1$, including the Ansari–Bradley procedure, to obtain point estimators and confidence intervals or bounds for γ^2 in such a setting.

13. *Unequal and Unknown Medians*. If the medians of the X and Y populations are not known and it is questionable whether or not they are equal, Ansari and Bradley (1960) suggested the following modification to their test procedures. Define the adjusted observations $X_i' = X_i - \tilde{X}, i = 1, \ldots, m$, and $Y_j' = Y_j - \tilde{Y}, j = 1, \ldots, n$, where $\tilde{X}$ and $\tilde{Y}$ are the sample medians for the X and Y observations, respectively. Let C' be C (5.6) calculated for these adjusted X' and Y' observations. Depending on the alternative to $H_0 : \gamma^2 = 1$ that is of interest, the appropriate procedure (5.7), (5.8), or (5.9), or the corresponding large-sample approximation, can then be applied directly to the modified statistic C' instead of C. Such tests based on C' are no longer strictly distribution-free. However, Gross (1966) has given sufficient conditions under which such procedures are asymptotically distribution-free. Under such conditions, the various tests based on C' maintain an approximate (both m and n large) significance level α over a large class of continuous underlying distributions.

14. *Consistency of the C-Tests*. Under Assumptions A1–A3, the consistency of the tests based on C depends on the parameter

$$\Delta^* = \left[P(X > Y > \theta) + P(X < Y < \theta) - \tfrac{1}{4}\right].$$

The test procedures defined by (5.7), (5.8), and (5.9) are consistent against the alternatives corresponding to $\Delta^* >, <$, and $\neq 0$, respectively.

15. *More General Alternatives*. In many two-sample situations, we are interested in simultaneously detecting either location or scale differences between the X and Y populations. One solution to this broader problem is to use a test procedure designed to detect quite general alternatives. One such test procedure based on the two-sample Kolmogorov (1933)–Smirnov (1939) statistic is discussed in Section 5.4. A second approach is to conduct simultaneously a test such as the Wilcoxon rank sum procedure based on W (4.3) for detecting differences in location and a second test such as the Ansari–Bradley procedure based on C (5.6) for detecting differences in scale. One such simultaneous testing approach, due to Lepage (1971, 1973), for dealing with general alternatives is the topic of Section 5.3. Randles and Hogg (1971) have shown that in such a situation, W and C are uncorrelated and, in fact, asymptotically independent when H_0 (5.1) is true. This implies, among other things, that if we conduct the Wilcoxon rank sum test at a significance level α_1, and the Ansari–Bradley test at a significance level α_2, then the probability of incorrectly rejecting with at least one of the two tests, given that H_0 (5.1) is true, is approximately $\alpha_1 + \alpha_2 - \alpha_1\alpha_2$.

Properties

1. *Consistency*. For our statement, we consider the more stringent location-scale parameter model described in (5.4). Then the tests defined by (5.7), (5.8), and (5.9) are consistent against the alternative $\gamma^2 >, <$, and $\neq 1$, respectively. See also Comment 14.
2. *Asymptotic Normality*. See Randles and Wolfe (1979, pp. 315–320).
3. *Efficiency*. See Section 5.5.

Problems

1. Consider the chorioamnion permeability data in Table 4.1. In Section 4.1 we saw that a test procedure based on the Wilcoxon rank sum statistic did not reject the null hypothesis that the human chorioamnion is as permeable to water transfer at 12–26 weeks gestational age as it is at term. With this in mind and using the same data, test the hypothesis of equal dispersions versus the alternative that the variation in tritiated water diffusion across human chorioamnion is different at term than at 12–26 weeks gestational age.

2. Find or construct an example in which there exists a level α and a constant d such that when C (5.6) is computed for the original data $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$, the level α procedure in (5.7) does not lead to rejection of $H_0 : \gamma^2 = 1$, but when C is computed for the values $X_1, \ldots, X_m, Y_1 + d, \ldots, Y_n + d$, the level α procedure in (5.7) does lead to rejection of H_0. Note that such an example exposes an undesirable aspect of the test procedures based on C. Let Y be a random member from a population II. Then, the population II* formed by adding the constant d to each member of II must, by any reasonable definition of dispersion, have the same dispersion as the II population. Thus, the difference in dispersions between the X and Y populations must be the same as the difference in dispersions between the X and $Y + d$ populations. Yet the tests based on C, applied to the data $X_1, \ldots, X_m, Y_1 + d, \ldots, Y_n + d$, can yield a decision that differs from the one that results from applying the same C-test to $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. (For a related discussion, see Comment 7.)

3. Consider the television-viewing behavior data in Table 4.4. For these data, use the R command `pAnsBrad(x,y)` to find the P-value for an appropriate test of the hypothesis of equal dispersions versus the alternative that there is more variability in the time spent in the room after

witnessing the violent behavior for those children who had previously watched the *Karate Kid* than for those children who had previously watched parts of the 1984 Summer Olympic Games. Comment on the importance of the results of Problem 4.5 in relationship to this dispersion test.

4. Verify the expressions for $E_0(C)$ and $\text{var}_0(C)$ in (5.12) and (5.13), respectively, when $N = (m + n)$ is an odd integer. (See Comment 9 for guidance.)
5. Consider the following two-sample data for $m = 3$, $n = 3 : X_1 = -3.7$, $X_2 = 4.6$, $X_3 = 1.5$, $Y_1 = 1.5$, $Y_2 = 4.6$, $Y_3 = 1.5$. Here, $N = 3 + 3 = 6$ is an even integer. Using the approach discussed in Comment 11, find the exact conditional null distribution of the Ansari–Bradley statistic, C (5.6). Compare and contrast this conditional null distribution with the null distribution of C for $m = n = 3$ and no tied observations.
6. Let C' be the sum of the scores assigned to the X observations by the Ansari–Bradley scoring scheme described in the Procedure. Verify directly, or illustrate using the serum iron determination data in Table 5.1, that $C' = [N(N + 2)/4] - C$, when $N = (m + n)$ is even.
7. For an arbitrary total number of observations $N = (m + n)$, find an expression for the smallest and largest possible values of C. Consider the two cases of N even and N odd.
8. Let X and Y be independent, identically distributed continuous random variables with a common probability distribution with median θ. What is the value of Δ^* in Comment 14 for this setting?
9. Consider the alcoholic intake data in Table 4.2. In Example 4.2 the Wilcoxon rank sum test procedure led to the rejection of $H_0 : \Delta = 0$ in favor of $H_1 : \Delta < 0$. What does this result imply about the appropriateness of the Ansari–Bradley procedure in (5.9) or its approximate large-sample counterpart in (5.17) as a test of $H_0 : \gamma^2 = 1$ versus $H_1 : \gamma^2 \neq 1$ for these data? In view of this fact, find the approximate P-value for an appropriate modification of the large-sample procedure in (5.17) to test for possible differences in dispersions between the control and SST data. (See Comment 13.)
10. Consider the television-viewing behavior data in Table 4.4. *Without* the assumption of equal medians, find the approximate P-value for an appropriate test (see Comment 13) of the hypothesis of equal dispersions versus the alternative that there is more variability in the time spent in the room after witnessing the violent behavior for those children who had previously watched the *Karate Kid* than for those children who had previously watched parts of the 1984 Summer Olympic Games. Compare the P-value obtained here with the one found in Problem 3. Interpret the similarity or lack thereof between the two P-values.
11. Suppose $m = n = 10$. Compare the critical region for the exact level $\alpha = .056$ test of $H_0 : \gamma^2 = 1$ versus $H_1 : \gamma^2 > 1$ based on C with the critical region for the corresponding nominal level $\alpha = .056$ test based on the large-sample approximation. What is the exact significance level of this .056 nominal level test based on the large-sample approximation?
12. Generate the conditional permutation distribution of C (see Comments 4 and 11), given the set of tied values for the serum iron data in Example 5.1. From this conditional permutation distribution of C, obtain the exact conditional P-value, $P_0(C \geq 185.5)$, for the corresponding test of $H_0 : \gamma^2 = 1$ versus $H_1 : \gamma^2 > 1$. Compare this exact conditional P-value with the approximate P-value for the large-sample test procedure applied to these data in Example 5.1.

5.2 AN ASYMPTOTICALLY DISTRIBUTION-FREE TEST FOR DISPERSION BASED ON THE JACKKNIFE–MEDIANS NOT NECESSARILY EQUAL (MILLER)

Hypothesis

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples satisfying Assumptions A1 and A2 from continuous populations with distribution functions F and G, respectively,

satisfying the location-scale parameter model relationship in (5.2) and (5.3). In addition, we assume that the continuous distribution associated with the distribution function $H(\cdot)$ in (5.2) has finite fourth moment; that is, we assume

A4. If V is a continuous random variable with distribution function H, then $E(V^4) < \infty$.

Under Assumptions A1, A2, and A4, but without the equal median Assumption A3, we are once again interested in the ratio of scale parameters $\gamma = (\eta_1/\eta_2)$. In view of Assumption A4 (see Comment 7), we note that $\gamma^2 = [\text{var}(X)/\text{var}(Y)]$, the ratio of population variances. We are interested in testing the null hypothesis H_0 (5.1), which reduces to $H_0 : \gamma^2 = 1$, corresponding to the assertion that the population variances are equal, under the location-scale parameter model (5.2).

Procedure

Consider the X sample data with the first observation deleted and set

$$\bar{X}_1 = \sum_{s=2}^{m} \frac{X_s}{m-1} \quad \text{and} \quad D_1^2 = \sum_{s=2}^{m} \frac{(X_s - \bar{X}_1)^2}{m-2}. \tag{5.23}$$

Thus, $\bar{X}_1$ and D_1^2 are the sample average and sample variance for the data $X_2, \ldots, X_m$, corresponding to the X sample less X_1. Similarly, let

$$\bar{X}_i = \sum_{s \neq i}^{m} \frac{X_s}{m-1} \quad \text{and} \quad D_i^2 = \sum_{s \neq i}^{m} \frac{(X_s - \bar{X}_i)^2}{m-2} \tag{5.24}$$

be the sample average and sample variance, respectively, for the data $X_1, \ldots, X_{i-1}$, $X_{i+1}, \ldots, X_m$, corresponding to the X sample less X_i, for $i = 1, \ldots, m$. In the same fashion, let

$$\bar{Y}_j = \sum_{t \neq j}^{n} \frac{Y_t}{n-1} \quad \text{and} \quad E_j^2 = \sum_{t \neq j}^{n} \frac{(Y_t - \bar{Y}_j)^2}{n-2} \tag{5.25}$$

be the sample average and sample variance, respectively, for the data $Y_1, \ldots, Y_{j-1}$, $Y_{j+1}, \ldots, Y_n$, corresponding to the Y sample less Y_j, for $j = 1, \ldots, n$. Define $S_1, \ldots, S_m$ and $T_1, \ldots, T_n$ by

$$S_i = \ln D_i^2, \quad i = 1, \ldots, m, \tag{5.26}$$

and

$$T_j = \ln E_j^2, \quad j = 1, \ldots, n. \tag{5.27}$$

In addition, let

$$S_0 = \ln \left[\sum_{s=1}^{m} \frac{(X_s - \bar{X}_0)^2}{m-1} \right] \tag{5.28}$$

and

$$T_0 = \ln \left[\sum_{t=1}^{n} \frac{(Y_t - \bar{Y}_0)^2}{n-1} \right], \tag{5.29}$$

where $\bar{X}_0 = \sum_{s=1}^{m} X_s/m$ and $\bar{Y}_0 = \sum_{t=1}^{n} Y_t/n$, be the corresponding statistics for the complete samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, respectively. Compute

$$A_i = mS_0 - (m-1)S_i, \quad \text{for } i = 1, \ldots, m, \tag{5.30}$$

and

$$B_j = nT_0 - (n-1)T_j, \quad \text{for } j = 1, \ldots, n. \tag{5.31}$$

(This is what is referred to as the *jackknifing process*, as applied to the sample variance.) Set

$$\bar{A} = \sum_{i=1}^{m} \frac{A_i}{m} \quad \text{and} \quad \bar{B} = \sum_{j=1}^{n} \frac{B_j}{n}, \tag{5.32}$$

and compute

$$V_1 = \sum_{i=1}^{m} \frac{(A_i - \bar{A})^2}{m(m-1)} \tag{5.33}$$

and

$$V_2 = \sum_{j=1}^{n} \frac{(B_j - \bar{B})^2}{n(n-1)}. \tag{5.34}$$

Finally, set

$$Q = \frac{\bar{A} - \bar{B}}{\sqrt{V_1 + V_2}}. \tag{5.35}$$

a. *One-Sided Upper-Tail Test*. To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_1 : \gamma^2 > 1,$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } Q \geq z_\alpha; \quad \text{otherwise do not reject,} \tag{5.36}$$

where, as previously, z_α is the upper αth percentile for the standard normal distribution.

b. *One-Sided Lower-Tail Test*. To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_1 : \gamma^2 < 1,$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } Q \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{5.37}$$

c. *Two-Sided Test*. To test

$$H_0 : \gamma^2 = 1$$

versus

$$H_1 : \gamma^2 \neq 1,$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } |Q| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{5.38}$$

This two-sided procedure is the two-sided symmetric test with $\alpha/2$ probability in each tail of the approximating standard normal distribution.

When m and n are small and equal, the approximate level α test procedures given by (5.36), (5.37), and (5.38) can be improved slightly by replacing z_α and $z_{\alpha/2}$ by $t_{m+n-2,\alpha}$ and $t_{m+n-2,\alpha/2}$, respectively, where $t_{m+n-2,\alpha}$ is the upper α percentile point of the t distribution with $m + n - 2$ degrees of freedom. To find $t_{m+n-2,\alpha}$ for given sample sizes m and n, we use the `R` command `qt(1-α, m+n-2)`. For example, to find $t_{14,.05}$, we apply `qt(.95,14)` and obtain $t_{14,.05} = 1.761$.

Ties

The jackknife procedures are well defined when ties within or between the X's and Y's occur and further adjustments are not necessary.

EXAMPLE 5.2 ***Southern Armyworm and Pokeweed.***

Burnett and Jones (1973) investigated the idea of coevolution between the southern armyworm and pokeweed. They suspected that armyworms might have developed a greater resistance to the toxins from pokeweed populations that lie within their geographic range than to the toxins of pokeweeds found in other areas of the country. Pokeweed plants from Florida populations (within the range of southern armyworms) and Kentucky populations (well north of the range of southern armyworms) were raised under similar conditions in greenhouses for the study. Larval southern armyworms were then used in feeding experiments to determine whether they would eat less of the Kentucky pokeweed possessing toxins to which they are not resistant. Five samples of Kentucky pokeweed and five samples of Florida pokeweed were used, with each such sample being exposed to 10 separate southern armyworm larvae. (There were 100 *different* larvae used in the experiment.) Following an individual larva's 24-h feeding period (in darkness at $25 \pm 1^\circ$C) on a moist filter paper in a disposable petri dish, the fecal material of the larva was dried overnight in an oven and weighed the following day. This was then used as a measure of quantity of the plant material ingested by the armyworm larva during its feeding. The data in Table 5.2 are the average (over the 10 armyworm larvae replications) dry feces weights (in milligrams) for the five Kentucky pokeweed and five Florida pokeweed plant samples.

Table 5.2 Average Dry Feces Weight (mg)

Kentucky pokeweed	Florida pokeweed
6.2	9.5
5.9	9.8
8.9	9.5
6.5	9.6
8.6	10.3

Source: W. C. Burnett, Jr., and S. B. Jones, Jr. (1973).

It is clear from the data in Table 5.2 that the southern armyworm larvae had a tendency to eat more (on the average) of the Florida pokeweed than the Kentucky pokeweed. As a result, if we are interested in assessing whether there is any difference in the variability or dispersion of the southern armyworm's consumption of the two pokeweed varieties, it would not be appropriate to directly apply one of the Ansari–Bradley procedures discussed in Section 5.1, because they require equality of the respective population medians. However, the jackknifed variances procedure of this section makes no such assumption and can be applied directly to the sample data.

Letting X correspond to the Kentucky pokeweed observations and Y to the Florida pokeweed data, we consider testing the null hypothesis of no difference in dispersion against the alternative that the variability is greater for Kentucky pokeweed; that is, we want to use procedure (5.36) to test $H_0 : \gamma^2 = 1$ against the alternative $H_1 : \gamma^2 > 1$.

The five Kentucky pokeweed subgroups of four observations each, corresponding to the five different ways to delete a single measurement, are given by

$$G_1 = \{6.2, 5.9, 8.9, 6.5\}, \quad G_2 = \{6.2, 5.9, 8.9, 8.6\}, \quad G_3 = \{6.2, 5.9, 6.5, 8.6\},$$

$$G_4 = \{6.2, 8.9, 6.5, 8.6\} \quad \text{and} \quad G_5 = \{5.9, 8.9, 6.5, 8.6\}.$$

Following (5.23), the sample average and sample variance associated with subgroup G_1 are

$$\bar{X}_1 = \frac{6.2 + 5.9 + 8.9 + 6.5}{4} = 6.875 \tag{5.39}$$

and

$$D_1^2 = \frac{(6.2 - 6.875)^2 + (5.9 - 6.875)^2 + (8.9 - 6.875)^2 + (6.5 - 6.875)^2}{3} = 1.8825. \tag{5.40}$$

In a similar manner, it follows from (5.24) that the sample averages and sample variances for the other four Kentucky pokeweed subgroups are

Subgroup G_i	$\bar{X}_i$	D_i^2
G_2	7.4	2.46
G_3	6.8	1.50
G_4	7.55	1.95
G_5	7.475	2.2425

(5.41)

Proceeding in the same fashion with the Florida pokeweed data, we obtain the five deleted-observation subgroups

$$H_1 = \{9.5, 9.8, 9.5, 9.6\}, \quad H_2 = \{9.5, 9.8, 9.5, 10.3\}, \quad H_3 = \{9.5, 9.8, 9.6, 10.3\},$$

$$H_4 = \{9.5, 9.5, 9.6, 10.3\}, \quad \text{and} \quad H_5 = \{9.8, 9.5, 9.6, 10.3\}.$$

Using (5.25), we obtain the associated subgroup sample means and sample variances to be

Subgroup H_j	$\bar{Y}_j$	E_j^2
H_1	9.6	.02
H_2	9.775	.1425
H_3	9.8	.1267
H_4	9.725	.1492
H_5	9.8	.1267

(5.42)

Taking natural logarithms of the D_i^2's (in (5.40) and (5.41)) and the E_j^2's (in (5.42)), it follows from (5.26) and (5.27) that

$$S_1 = .6326, \quad S_2 = .9002, \quad S_3 = .4055, \quad S_4 = .6678, \quad S_5 = .8076 \tag{5.43}$$

and

$$T_1 = -3.9120, \quad T_2 = -1.9484, \quad T_3 = -2.0662, \quad T_4 = -1.9027, \quad T_5 = -2.0662. \tag{5.44}$$

Finally, using all five of the X sample observations, we see from (5.28) that

$$\bar{X}_0 = 7.22 \quad \text{and} \quad S_0 = \ln\left[\sum_{s=1}^{5} \frac{(X_s - 7.22)^2}{4}\right] = \ln 2.007 = .6966.$$

Similarly, using all five of the Y sample observations, it follows from (5.29) that

$$\bar{Y}_0 = 9.74 \quad \text{and} \quad T_0 = \ln\left[\sum_{t=1}^{5} \frac{(Y_t - 9.74)^2}{4}\right] = \ln .113 = -2.1804.$$

Combining these complete sample values with the subgroup calculations in (5.43) and (5.44) via the jackknifing process in (5.30) and (5.31), we obtain

$$A_1 = 5(.6966) - 4(.6326) = .9526, \quad A_2 = 5(.6966) - 4(.9002) = -.1178,$$

$$A_3 = 5(.6966) - 4(.4055) = 1.861, \tag{5.45}$$

$$A_4 = 5(.6966) - 4(.6678) = .8118, \quad A_5 = 5(.6966) - 4(.8076) = .2526$$

and

$$\begin{aligned} B_1 &= 5(-2.1804) - 4(-3.9120) = 4.746, \\ B_2 &= 5(-2.1804) - 4(-1.9484) = -3.1084, \\ B_3 &= 5(-2.1804) - 4(-2.0662) = -2.6372, \\ B_4 &= 5(-2.1804) - 4(-1.9027) = -3.2912, \\ B_5 &= 5(-2.1804) - 4(-2.0662) = -2.6372. \end{aligned} \tag{5.46}$$

From (5.32), (5.33), and (5.45), we see that

$$\bar{A} = .7520 \quad \text{and} \quad V_1 = \sum_{i=1}^{5} \frac{(A_i - \bar{A})^2}{5(4)} = .1140.$$

Similarly, from (5.32), (5.34), and (5.46), we have

$$\bar{B} = -1.3856 \quad \text{and} \quad V_2 = \sum_{j=1}^{5} \frac{(B_j - \bar{B})^2}{5(4)} = 2.3664.$$

It then follows from (5.35) that

$$Q = \frac{.7520 - (-1.3856)}{(.1140 + 2.3664)^{1/2}} = 1.36.$$

(We note that the R command `MillerJack(pokeweed$x, pokeweed$y)` can also be used to obtain the value $Q = 1.36$ for the (x, y) data in Table 5.2.)

Hence, from (5.36) and the R command `pnorm(·)`, we see that the lowest significance level at which we can reject H_0 in favor of $\gamma^2 > 1$ with the observed value of the test statistic Q (i.e., the one-sided P-value) is approximately $1 - \texttt{pnorm(1.36)} = .0869$. (Since the sample sizes $m = n = 5$ are small and equal, we can also use the t-distribution with $m + n - 2 = 5 + 5 - 2 = 8$ degrees of freedom to approximate the P-value. Using the R command `pt(·)`, we have P-value $\approx 1 - \texttt{pt(1.36, 8)} = \texttt{1 - .8945} = .1055$, in general agreement with the normal approximation.) Thus, there is only mild evidence in the sample data to indicate greater variability for the Kentucky pokeweed population.

Comments

16. *Assumptions*. Note that Assumptions A1, A2, and A4 do not impose the severe condition that the two underlying populations have equal medians. Although these assumptions do require that the two underlying populations have finite fourth moments, this is not a serious restriction for most common data collection settings. This means that the Miller procedures are applicable in more general settings than the Ansari–Bradley procedures based on C (5.6). (See Comment 7.)

17. *Testing γ^2 Equal to Some Specified Value Other Than One*. To test the hypothesis $\gamma^2 = \gamma_0^2$, where γ_0^2 is some specific positive number different from 1, we obtain the modified observations $Y_j' = \gamma_0 Y_j, j = 1, \ldots, n$, and compute Q(5.35) using the X's and the Y''s (instead of the X's and the Y's). The appropriate procedure (5.36), (5.37), or (5.38) may then he applied as described.

18. *Asymptotic Distribution-Freeness*. Asymptotically (i.e., for infinitely large samples) the true level of the tests defined by (5.36), (5.37), and (5.38) will agree with the nominal level α. Subject to Assumptions A1, A2, and A4, this asymptotic result does not depend on the underlying populations of the X's and Y's. More precisely, subject to Assumptions A1, A2, and A4, Q(5.35) has an asymptotic $N(0,1)$ distribution when H_0 is true. Since this asymptotic distribution does not depend on the underlying populations of the X's and Y's, we say that the tests based on Q are asymptotically distribution-free. Of course, in practice, we do not have the luxury of infinite samples. Thus, in any particular setting with m and n large, although the level of any of the tests based on Q is not necessarily exactly equal to the nominal level α, we hope it is close to it. The closeness of this approximation depends on m, n, α, and the underlying populations, but, for fixed α, the closeness generally improves as m and n increase, regardless of the underlying populations. In the case of the Q tests, the reader is cautioned that the question of how large m and n should be, in order for the normal approximation to be good, is relatively unanswered. Exact null distribution tables for Q cannot be provided for specified values of m and n, because the exact null distribution of Q depends on the underlying X and Y populations; thus, exact critical points would vary with the forms of the X and Y populations. The procedures (5.36), (5.37), and (5.38) based on Q, therefore, are not (strictly) distribution-free.

19. *Alternative Method of Calculation*. For $i = 1, \ldots, m$, S_i is the natural log of the sample variance for the $(m-1)$ X observations $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m$. Similarly, for $j = 1, \ldots, n$, T_j is the natural log of the sample variance for the $(n-1)$ Y observations $Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_n$. The following equivalent formulas for D_i^2 (5.24) and E_j^2 (5.25), namely,

$$D_i^2 = \frac{\sum_{s \neq i}^{m} X_s^2 - \dfrac{\left(\sum_{s \neq i}^{m} X_s\right)^2}{m-1}}{m-2}$$

and

$$E_j^2 = \frac{\sum_{t \neq j}^{n} Y_t^2 - \dfrac{\left(\sum_{t \neq j}^{n} Y_t\right)^2}{n-1}}{n-2},$$

are computationally more convenient than the definitions given in (5.24) and (5.25), respectively.

20. *General Jackknife Technique*. The jackknife technique applied in this section to the problem of testing two-sample dispersion hypotheses is a tool that can be used successfully in certain statistical problems to accomplish two

goals: (a) reducing the bias of point estimators and (b) generating broadly applicable and reasonably powerful test procedures for problems where classical test procedures are sensitive to nonnormality of the underlying populations. Although the jackknife technique is not always effective in achieving these goals (see Miller (1964)), it performs well in the two-sample dispersion problem, providing us with asymptotically distribution-free test procedures for a problem where deviations from normality of the underlying populations can be disastrous for the classical $\mathcal{F}$-test for equal variances (cf. Box (1953), Shorack (1969), and Comment 26) and where distribution-free rank tests for the problem are limited in their applicability (see Comment 7 and Moses (1963)). Miller (1968) discussed in detail the advantages gained by applying the jackknife technique to this two-sample dispersion problem. (See also Comment 21.)

21. *Motivation*. The jackknife is an extension of an idea due to Quenouille (1949) and is designed to reduce the bias of an estimator. Suppose we have a sample of N independent observations, each from the same distribution that depends on an unknown parameter θ. Assume that we have a general method for estimating θ and let $\hat{\theta}$ denote this estimator based on all N observations. Divide the data into n groups of size k. Let $\hat{\theta}_{-i}, i = 1, \ldots, n$, denote the estimator of θ obtained by deleting the ith group and estimating θ from the remaining $(n-1)k$ observations. Define $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$. The jackknife estimator of θ is $\tilde{\theta} = \sum_{i=1}^{n} \tilde{\theta}_i / n$. In certain situations, the jackknife can be shown to be less biased than the estimator $\hat{\theta}$. Tukey (1958, 1962) extended the jackknife to construct approximate significance tests and confidence intervals for θ.

The traditional estimator of the variance of $\tilde{\theta}$, in the case of $k = 1$, is

$$\widehat{V}^2 = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\widehat{\theta}_{-i} - \widehat{\theta})^2.$$

Asymptotic $100(1-\alpha)\%$ confidence intervals for θ are then

$$(\tilde{\theta} - z_{\alpha/2}\widehat{V}, \tilde{\theta} + z_{\alpha/2}\widehat{V}).$$

Under certain conditions (i.e., when $\widehat{\theta}$ is not "sufficiently smooth"), $\widehat{V}^2$ may be inconsistent. One such situation is when $\widehat{\theta}$ is a sample quantile (see, e.g., Miller (1974)). To overcome this difficulty, Shao (1988), Shao and Wu (1989), and Wu (1990) have studied a "delete-d" jackknife variance estimator. See also Maesono (1996) for further details.

In the dispersion problem, Miller jackknifed the natural logs of the sample variances rather than the sample variances themselves because the natural log transformation tends to stabilize the variance and create a distribution that is "closer" to the normal distribution. The statistic $\bar{B}$ (5.32) is an estimator of $\ln\{\mathrm{var}(Y)\}$, the statistic $\bar{A}$ (5.32) is an estimator of $\ln\{\mathrm{var}(X)\}$, and $\bar{A} - \bar{B}$ estimates $\ln\{\mathrm{var}(X)/\mathrm{var}(Y)\} = \ln \gamma^2$. The quantity $(V_1 + V_2)^{1/2}$ in the denominator of Q (5.35) is an estimator of the standard deviation of $\bar{A} - \bar{B}$. If, for example, the X's are more disperse than the Y's, $\bar{A} - \bar{B}$ would tend to be large, and this is partial motivation for procedure (5.36).

22. *Generalization*. In its most general formulation (see Comment 21), the jackknife process can be applied to any randomly selected partition of the data set into subsets of size k each, where k can be any positive integer that is a factor of the number of observations in the data set. In fact, Miller (1968) discussed the test for dispersion based on jackknifing the natural logs of the sample variances in the context of this most general formulation. However, for any integer $k > 1$, the associated Miller jackknifed variances procedures have the rather severe deficiency that it is possible for two different people to arrive at different conclusions when analyzing the same data set with the same test and at the same significance level. This possibility arises because of the variety of ways that the data set could be randomly partitioned into subsets of size k each. To avoid this undesirable feature, we have chosen to discuss the jackknifed variances procedure only for $k = 1$. In this case, there is no flexibility in partitioning a data set and the associated Miller test procedures are unambiguous in their conclusions.
23. *t Distribution Approximation*. The standard normal percentiles used in (5.36) to (5.38) should be replaced by the corresponding percentile points for a t distribution with $m + n - 2$ degrees of freedom only when m and n are small and equal. For other situations, the matter of which t distribution (i.e., what degrees of freedom) should be used to find the approximating percentile is somewhat ambiguous.
24. *Point Estimators and Confidence Intervals and Bounds for* γ^2. Point estimators of and approximate confidence intervals and bounds for γ^2 can be readily obtained from the jackknife procedures. In particular, the estimator for γ^2 associated with the jackknifed variances procedures is

$$\tilde{\gamma}^2 = e^{\bar{A}-\bar{B}}. \tag{5.47}$$

Moreover, an asymptotically distribution-free confidence interval for γ^2, with approximate confidence coefficient $1 - \alpha$, based on the jackknifed variances procedures is given by

$$(\gamma_L^2, \gamma_U^2), \tag{5.48}$$

where

$$\gamma_L^2 = e^{[(\bar{A}-\bar{B})-z_{\alpha/2}(V_1+V_2)^{1/2}]} \tag{5.49}$$

and

$$\gamma_U^2 = e^{[(\bar{A}-\bar{B})+z_{\alpha/2}(V_1+V_2)^{1/2}]}. \tag{5.50}$$

With γ_L^2 and γ_U^2 given by (5.49) and (5.50), we have

$$P_{\gamma^2}\{\gamma_L^2 < \gamma^2 < \gamma_U^2\} \approx 1 - \alpha. \tag{5.51}$$

The corresponding asymptotically distribution-free approximate 100 $(1 - \alpha)\%$ *lower* and *upper* confidence bounds for γ^2 based on the jackknifed variances procedures are

$$\gamma_L^{*2} = e^{[(\bar{A}-\bar{B})-z_{\alpha}(V_1+V_2)^{1/2}]} \tag{5.52}$$

and

$$\gamma_U^{*2} = e^{[(\bar{A}-\bar{B})+z_\alpha(V_1+V_2)^{1/2}]}, \tag{5.53}$$

respectively, satisfying

$$P_{\gamma^2}\{\gamma_L^{*2} < \gamma^2\} \approx 1-\alpha \quad \text{and} \quad P_{\gamma^2}\{\gamma^2 < \gamma_U^{*2}\} \approx 1-\alpha. \tag{5.54}$$

For the armyworm/pokeweed data in Example 5.2, the point estimate of γ^2 is $\tilde{\gamma}^2 = e^{[.7520-(-1.3856)]} = e^{2.1376} = 8.479$ and, with $\alpha = .0548$, the approximate 94.52% lower confidence bound for γ^2 is

$$\begin{aligned}\gamma_L^{*2} &= e^{[(.7520-(-1.3856))-1.6(.1140+2.3664)^{1/2}]} \\ &= e^{-.3823} = .6823.\end{aligned}$$

25. *Asymptotic Coverage Probability*. Asymptotically (i.e., for infinitely large samples), the true coverage probabilities of the confidence interval defined by (5.48) and the confidence bounds defined by (5.52) and (5.53) will agree with the nominal confidence coefficient $1-\alpha$. Subject to Assumptions A1, A2, and A4, this asymptotic result does not depend on the form of the distribution function $H(\cdot)$ in (5.2). Hence, we say that the interval given by (5.48) and the bounds given by (5.52) and (5.53) are an asymptotically distribution-free confidence interval and asymptotically distribution-free confidence bounds, respectively, for γ^2.

    The interval (5.48) has also been defined so that it is "asymptotically symmetric." Here, the word symmetric refers to the equal-tail probabilities of $\alpha/2$. The $1-\alpha$ confidence interval for γ^2 defined by (5.48) can be called asymptotically symmetric, because it is constructed so that $P_{\gamma^2}(\gamma_U^2 \le \gamma^2) \approx P_{\gamma^2}(\gamma_L^2 \ge \gamma^2) \approx \alpha/2$. The approximation is a result of approximating the true distribution of the statistic $\bar{A}-\bar{B}$ by its asymptotic normal distribution.

26. *Lack of Robustness of the Classical $\mathcal{F}$-Test for Equal Variances*. The classical normal theory $\mathcal{F}$-test for equality of variances is not robust with respect to the assumption of normality in the sense that when the underlying populations are not normal, the true level of an $\mathcal{F}$-test that is supposed to be of size α may be quite far from α. Box (1953) gave examples in which the level of the $\mathcal{F}$-test is specified to be .05, although the actual level is as large as .166 or as small as .0056. Furthermore, there exist nonnormal populations in which, even with large samples, the level of the $\mathcal{F}$-test will not be what it is supposed to be. This nonrobustness, which was pointed out as early as 1931 by Pearson (1931), has been emphasized by Box (1953) and more recently by Miller (1968) and Shorack (1969).

Properties

1. *Consistency*. The tests given by (5.36), (5.37), and (5.38) are consistent against the alternatives $\gamma^2 >$, $<$, and $\neq 1$, respectively.
2. *Asymptotic Normality.* See Miller (1968).
3. *Efficiency.* See Miller (1968) and Section 5.5.

Problems

13. The data in Table 5.3 are a portion of those collected by Bugyi et al. (1969) in a study concerned with ascertaining sodium ion content in erythrocytes (red blood cells). Such determinations are helpful in the diagnoses of certain diseases, where merely knowing the sodium ion content in plasma does not provide sufficient information. However, erythrocyte sodium ion determination is extremely variable and subject to error. This prompted the authors to propose using the flame photometric method to determine sodium ion content in erythrocytes, with the hope of providing better accuracy than can be obtained with the inefficient procedures commonly used at that time.

One of the ways to assess the accuracy of the proposed method is to compare the variation in erythrocyte sodium ion measurements with the variation in plasma sodium ion determinations, where it is known that the measurement variation for the flame photometric method is acceptably low. Sodium ion determinations were obtained by the flame photometric method on each of 10 plasma and 10 erythrocyte samples. Table 5.3 gives the sodium ion content in mequiv/l for the 20 samples.

Use a Miller jackknife procedure to test the hypothesis of equal dispersions for the plasma and erythrocyte sodium ion measurements against the alternative of interest in the study. Find the approximate P-value for the test.

14. Consider the chorioamnion permeability data given in Table 4.1. Find the approximate P-value for the Miller jackknife test of the hypothesis of equal dispersions versus the alternative that the variation in tritiated water diffusion across human chorioamnion is different at term than at 12–26 weeks gestational age. Compare your findings with those obtained in Problem 1 using an Ansari–Bradley procedure to analyze the data.

15. Consider the television-viewing behavior data in Table 4.4. Find the approximate P-value for the Miller jackknife test of equal dispersions versus the alternative that there is more variability in the time spent in the room after witnessing the violent behavior for those children who had previously watched the *Karate Kid* than for those children who had previously watched parts of the 1984 Summer Olympic Games. Comment on your analysis in conjunction with the findings of Problems 3 and 4.5.

16. Consider the alcoholic intake data in Table 4.2. Find the approximate P-value for the Miller jackknife test of whether there is any difference in dispersions for the control and SST data. Comment on your analysis relative to the findings in Problem 9 and Example 4.2.

17. For the sodium ion determination data of Table 5.3, compute the value of the estimator $\tilde{\gamma}^2$ defined in expression (5.47).

18. For the chorioamnion permeability data given in Table 4.1, obtain the value of the estimator $\tilde{\gamma}^2$ defined in expression (5.47).

Table 5.3 Sodium Ion Content (mequiv/l)

Plasma	Erythrocytes
147.0	10.3
147.0	12.2
146.0	16.5
145.0	19.3
146.5	8.3
161.0	15.2
141.0	27.0
146.5	26.3
145.0	17.5
153.5	21.7

Source: H. I. Bugyi, E. Magnier, W. Joseph, and G. Frank (1969).

19. Obtain the value of the estimator $\tilde{\gamma}^2$ defined in expression (5.47) for the television-viewing behavior data in Table 4.4.

20. Compute the value of the estimator $\tilde{\gamma}^2$ defined in expression (5.47) for the alcoholic intake data in Table 4.2.

21. With respect to the chorioamnion permeability data given in Table 4.1, find an approximate 96.6% confidence interval for γ^2 utilizing the procedure discussed in Comment 24.

22. Consider the alcoholic intake data in Table 4.2. Using the procedure discussed in Comment 24, find an approximate 93.72% confidence interval for γ^2.

23. Consider the television-viewing behavior data in Table 4.4. Labeling the Olympic watchers data as the X sample, use the procedure discussed in Comment 24 to find an approximate 98.96% upper confidence bound for γ^2.

24. Consider the sodium ion determination data of Table 5.3. Labeling the erythrocyte sodium ion measurements as the X sample, use the procedure discussed in Comment 24 to find an approximate 91.92% lower confidence bound for γ^2.

5.3 A DISTRIBUTION-FREE RANK TEST FOR EITHER LOCATION OR DISPERSION (LEPAGE)

Hypothesis

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples satisfying Assumptions Al and A2 from continuous populations with distribution functions F and G, respectively, satisfying the location-scale parameter model relationships in (5.2) and (5.3).

Under these assumptions, we are interested in assessing whether there are differences in *either* the location parameters (i.e., medians) θ_1 and θ_2 *or* the scale parameters η_1 and η_2 for the X and Y populations. Thus, we are interested in testing the null hypothesis H_0 (5.1) versus the general alternative H_1: [$\theta_1 \neq \theta_2$ and/or $\eta_1 \neq \eta_2$]. Note that under the location-scale parameter model, as stated in (5.2) and (5.3), the null hypothesis H_0 (5.1) reduces to H_0: [$\theta_1 = \theta_2$ and $\eta_1 = \eta_2$], corresponding to the assertion that both the population location parameters and the population scale parameters are equal.

Procedure

To compute the Lepage two-sample location-scale statistic D, order the combined sample of $N = (m + n)$ X-values and Y-values from least to greatest. Let S_j denote the combined samples rank of Y_j, for $j = 1, \ldots, n$, and let $W = \sum_{j=1}^{n} S_j$ be the Wilcoxon rank sum statistic defined in (4.3). In addition, for $j = 1, \ldots, n$, let R_j be the score assigned to Y_j by the Ansari–Bradley scoring scheme discussed in the Procedure of Section 5.2 and let $C = \sum_{j=1}^{n} R_j$ be the Ansari–Bradley scale statistic defined in (5.6). The Lepage rank statistic is then defined by

$$D = \frac{[W - E_0(W)]^2}{\text{var}_0(W)} + \frac{[C - E_0(C)]^2}{\text{var}_0(C)}, \tag{5.55}$$

where $E_0(W)$ and $\text{var}_0(W)$ are the expected value and variance of W under H_0 (5.1), as given in (4.7) and (4.8), respectively, and $E_0(C)$ and $\text{var}_0(C)$ are the corresponding expected value and variance of C under H_0 (5.1), as stated in (5.10) and

(5.11), respectively, when $N = (m + n)$ is an even number, or in (5.12) and (5.13), respectively, when N is odd. Thus, if we let W^* (4.9) and C^* (5.20) represent the standardized forms for the Wilcoxon rank sum statistic and Ansari–Bradley scale statistic, respectively, then the Lepage statistic D can be written as

$$D = (W^*)^2 + (C^*)^2. \tag{5.56}$$

To test H_0 (5.1), corresponding to the equality of both the location and the scale parameters for the X and Y populations, versus the general alternatives that the location parameters are different or the scale parameters are different or both, corresponding to

$$H_1 : [\theta_1 \neq \theta_2 \text{ and/or } \eta_1 \neq \eta_2], \tag{5.57}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } D \geq d_\alpha; \text{ otherwise do not reject,} \tag{5.58}$$

where the constant d_α is chosen to make the type I error probability equal to α. The constant d_α is the upper α percentile for the null H_0(5.1) distribution of D. Comment 28 explains how to obtain the critical value d_α for sample sizes m and n and available values of α.

Large-Sample Approximation

The large-sample approximation is based on the fact that when H_0 (5.1) is true, the statistic D has, as $\min(m, n)$ tends to infinity, a chi-square distribution with 2 degrees of freedom (see Comment 31 for indications of the proof). The large-sample approximation for the exact level α procedure in (5.58) is

$$\text{Reject } H_0 \text{ if } D \geq \chi^2_{2,\alpha}; \text{ otherwise do not reject,} \tag{5.59}$$

where $\chi^2_{2,\alpha}$ is the upper α percentile point of the chi-square distribution with 2 degrees of freedom. To find $\chi^2_{2,\alpha}$, we use the R command `qchisq(1 - α, 2)`. For example, to find $\chi^2_{2,.05}$ we apply `qchisq(.95, 2)` and obtain $\chi^2_{2,.05} = 5.991$.

Ties

If there are ties among the X and/or Y observations, we modify the standardized Wilcoxon rank sum statistic W^* and the standardized Ansari–Bradley scale statistic C^* in the manners prescribed for the large-sample approximations in the Ties portions of Sections 4.1 and 5.1, respectively. When applying either the small-sample procedure in (5.58) or the large-sample approximation in (5.59), the Lepage statistic D should be computed using these ties-modified versions of W^* and C^*. The corresponding modified version of procedure (5.58) in the case of ties among the X and/or Y observations is only approximately, and not exactly, of significance level α. (To get an exact level α test even in this tied setting, see Comment 32.)

EXAMPLE 5.3 ***Effect of Maternal Steroid Therapy on Platelet Counts of Newborn Infants.***

Autoimmune thrombocytopenic purpura (ATP) is a disease in which the patient produces antibodies to his/her own platelets. Due to transplacental passage of antiplatelet antibodies during pregnancy, children of women with ATP are often born with low platelet counts. For this reason, there is medical concern that a vaginal delivery for a mother with ATP could result in intracranial hemorrhage for the infant. However, the proper obstetrical management of pregnant women with ATP is controversial. Most doctors have advocated cesarean section as the preferable method of delivery for mothers with ATP. Others suggest that cesarean section, with its obvious complications for both mother and infant, be avoided unless there is some additional obstetrical reason for it. Karpatkin, Porges, and Karpatkin (1981) studied the effect of administering the corticosteroid prednisone to pregnant women with ATP with the intent of raising the infants' platelet counts to safe levels during their deliveries. The rationale for this treatment is the fact that steroids, in general, increase the platelet counts in patients with ATP by blocking splenic destruction of antibody-coated platelets. In theory, then, the corticosteroid prednisone should cross the placenta, enter the infant's circulation, and prevent splenic removal of those infant's platelets that are coated by the mother's antibodies.

The data in Table 5.4 are a subset of the data obtained by Karpatkin et al. in their study of the effect that administration of prednisone to pregnant women with ATP had on their infants' platelet counts. All the infants included in this example were delivered vaginally. Table 5.4 gives the platelet counts (per cubic millimeter) of 10 infants whose mothers received the steroid prednisone prior to delivery and 6 infants whose mothers were not treated with prednisone prior to delivery. All 16 mothers in the study were diagnosed with ATP.

The primary interest in the study is in whether or not the predelivery administration of prednisone typically leads to an increased newborn platelet count. Thus, the principal statistical issue in the study is that of a possible difference in locations for the prednisone and nonprednisone populations. However, there is some concern that the administration of predelivery prednisone could also lead to a rather large increase in variability in the newborn platelet counts. (Such a finding would certainly affect our interpretation of any possible increase in typical platelet count resulting from the prednisone.) As a result, we will apply the Lepage test procedure to test H_0 (5.1) versus the general alternative H_1 5.57). For the purposes of illustration, we consider the exact procedure (5.58) with level of

Table 5.4 Platelet Counts of Newborn Infants (per Millimeter^3)

Mothers given prednisone	Mothers not given prednisone
120,000	12,000
124,000	20,000
215,000	112,000
90,000	32,000
67,000	60,000
95,000	40,000
190,000	
180,000	
135,000	
399,000	

Source: M. Karpatkin, R. F. Porges, and S. Karpatkin (1981).

significance $\alpha = .02$. For convenience, we take the infant platelet count data for mothers given prednisone to be the Y sample ($n = 10$) and the corresponding control (nonprednisone) data to be the X sample ($m = 6$). Using the R command `cLepage(.02, 6, 10)`, we obtain $d_{.02} = 6.903$ so that procedure (5.58) is

$$\text{Reject } H_0 \text{ if } D \geq 6.903. \tag{5.60}$$

To calculate D, we need first to calculate the standardized versions of the Wilcoxon rank sum and Ansari–Bradley statistics. Proceeding as in (4.3), we note that the combined samples ranks for the 10 Y observations are 10, 11, 16, 7, 6, 8, 14, 13, 12, and 15, yielding a value of

$$W = 10 + 11 + 16 + 7 + 6 + 8 + 14 + 13 + 12 + 15 = 112$$

for the Wilcoxon rank sum statistic. Since there are no ties among the 16 X and Y observations, it follows from (4.7), (4.8), and (4.9) that the standardized form of W for the data in Table 5.4 is

$$W^* = \frac{112 - \{10(6 + 10 + 1)/2\}}{\{6(10)(6 + 10 + 1)/12\}^{1/2}} = 2.929. \tag{5.61}$$

For the calculation of the standardized Ansari–Bradley statistic C^*, we observe that the Ansari–Bradley scores (as defined in the Procedure of Section 5.1) for the 10 Y observations are 7, 6, 2, 7, 6, 8, 3, 4, 5, and 1. From (5.6), this produces a value of

$$C = 7 + 6 + 2 + 7 + 6 + 8 + 3 + 4 + 5 + 1 = 49.$$

Since $N = 6 + 10 = 16$ is an even number and there are no ties among the 16 X and Y observations, it follows from (5.10), (5.11), and (5.14) that the standardized form of C for the data in Table 5.4 is

$$C^* = \frac{49 - \{10(16 + 2)/4\}}{\left\{\dfrac{10(6)(16 + 2)(16 - 2)}{48(16 - 1)}\right\}^{1/2}} = .873. \tag{5.62}$$

Using these values of W^* (5.61) and C^* (5.62) in (5.56) yields

$$D = (2.929)^2 + (.873)^2 = 9.34, \tag{5.63}$$

which, in view of expression (5.60), tells us to reject H_0 at the $\alpha = .02$ level, because $D = 9.34 > d_{.02} = 6.903$. Hence, there is rather strong evidence that there are differences in locations or scales (or both) between the prednisone and control infant platelet count populations. In fact, using the R command `pLepage(`**`platelet.counts$x`**`, `**`platelet.counts$y`**`)`, the P-value for these data with observed value $D = 9.34$ is given by P-value = `pLepage(`**`platelet.counts$x`**`, `**`platelet.counts$y`**`)` = .0035, providing an even stronger statement in favor of the alternative H_1 (5.57).

For the large-sample approximation, we see from (5.59) that the approximate P-value for these data is

$$P\text{-value} \approx P(Q \geq 9.34),$$

where Q has a chi-square distribution with 2 degrees of freedom. This approximate P-value is then given by $1 - \texttt{pchisq(9.34, 2)} = 1 - .9906 = .0094$, in general agreement with the exact P-value of .0035 previously obtained.

We conclude this example by noting that the large value of D is due primarily to a large value of W^*. This would suggest intuitively that the rejection of H_0 is due primarily to a difference in locations between the infant platelet counts for the prednisone and control populations. However, we emphasize that such a conclusion is not statistically justified through the application of the general Lepage procedure (5.58). The only valid conclusion based on the Lepage procedure is that of the general alternative H_1 (5.57). (If you do, however, apply the Wilcoxon rank sum procedure of Section 4.1 to the data in Table 5.4, you would be able to conclude that there is, indeed, a difference in locations between the infant platelet counts for the prednisone and control populations. In view of this fact, would it be legitimate to then apply the Ansari–Bradley procedure of Section 5.1 directly to the data in Table 5.4 to test for possible scale differences in the two populations?)

Comments

27. *Motivation for the Test*. From Section 4.1 we know that a large value of $(W^*)^2$ is indicative of a possible difference in locations for the X and Y populations. We also know from Section 5.1 that a large value of $(C^*)^2$ is indicative of a possible difference in dispersions for the X and Y populations. Since D (5.56) will be large if and only if $(W^*)^2$ is large or $(C^*)^2$ is large or both, then such a large value of D is indicative of $\theta_1 \neq \theta_2$ or $\eta_1 \neq \eta_2$ or both. This serves as partial motivation for the test procedure given by (5.58).
28. *Derivation of the Distribution of D under H_0 (No-Ties Case)*. Under H_0 (5.1), each of the $\binom{N}{n}$ possible "meshings" of the X's and Y's has probability $1/\binom{N}{n}$. This fact can be used to obtain the null distribution of D (5.56). We illustrate the steps involved in constructing this null distribution for the simple case $m = 2$, $n = 2$. Since $N = 4$, we must consider $\binom{4}{2} = 6$ possible meshings of the X and Y observations. For this setting, it follows from (4.7) and (4.8) that $E_0(W) = 2(2 + 2 + 1)/2 = 5$ and $\text{var}_0(W) = 2(2)(2 + 2 + 1)/12 = \frac{5}{3}$. Similarly, from (5.10) and (5.11), we have $E_0(C) = 2(4 + 2)/4 = 3$ and $\text{var}_0(C) = [2(2)(4 + 2)(4 - 2)]/48(4 - 1) = \frac{1}{3}$. Thus, for $m = n = 2$, we have $(W^*)^2 = 3(W - 5)^2/5$ and $(C^*)^2 = 3(C - 3)^2$. Using these facts and the same approach taken in Comments 4.3 and 5.4 for the calculations of W and C, respectively, the values of D for these six meshings are given in the following table.

Meshing	Probability	$D = (W^*)^2 + (C^*)^2$
XXYY	$\frac{1}{6}$	2.4
XYXY	$\frac{1}{6}$	.6
YXXY	$\frac{1}{6}$	3.0
XYYX	$\frac{1}{6}$	3.0
YXYX	$\frac{1}{6}$	.6
YYXX	$\frac{1}{6}$	2.4

Thus, for example, the probability is $\frac{1}{3}$ under H_0 that D is equal to .6, since $D = .6$ when either of the exclusive meshings *XYXY* or *YXYX* occurs and each of these meshings has null probability $\frac{1}{6}$. Proceeding in the same manner for all possible values for D and simplifying, we obtain the null distribution.

Value of D	Probability under H_0
0.6	$\frac{1}{3}$
2.4	$\frac{1}{3}$
3.0	$\frac{1}{3}$

Thus, for example, the probability under H_0 that D is greater than or equal to 3 is, therefore, $P_0(D \geq 3) = \frac{1}{3}$, which implies that $d_{1/3} = 3$ for the setting $m = n = 2$.

Note that we have derived the null distribution of D without specifying the form of the common (under H_0) underlying X and Y populations beyond the point of requiring that they be continuous. This is why the test procedure based on D is called a *distribution-free procedure*. From the null distribution of D we can determine the critical value d_α and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying distribution for the X and Y populations.

For given sample sizes m and n, the R command `cLepage(`α`,m,n)` can be used to find the available upper-tail critical values d_α for possible values of D. For a given available significance level α, the critical value d_α then corresponds to $P_0(D \geq d_\alpha) = \alpha$ and is given by `cLepage(`α`,m,n)` $= d_\alpha$. Thus, for example, for $m = 5$ and $n = 8$, we have $P_0(D \geq 6.875) = .0194$ so that $d_{.0194} =$ `cLepage(.0194,5,8)` $= 6.875$ for $m = 5$ and $n = 8$.

29. *Equivalent Form.* In computing the Wilcoxon rank sum statistic W (4.3), we use the combined samples ranks of the Y observations. In computing the Ansari–Bradley statistic C (5.6), we use the scores assigned to the Y observations by the Ansari–Bradley outside-in scoring scheme. However, both W and C, and therefore D (5.55), can be computed solely from knowledge of the combined samples ranks of the Y observations. This follows directly from the fact that the Ansari–Bradley statistic can also be represented (in the case of no tied X and/or Y observations) as

$$C = \frac{n(N+1)}{2} - \sum_{j=1}^{n} \left| S_j - \frac{N+1}{2} \right|, \tag{5.64}$$

where, as in the calculation of W, S_j is the combined samples rank of Y_j, for $j = 1, \ldots, n$. In fact, both W and C are members of a very large class of statistics based solely on the combined samples ranks in a special way. This collection is referred to as the class of two-sample linear rank statistics, and they have been extensively studied in the literature (see, e.g., Randles and Wolfe (1979)).

Thus, although the Ansari–Bradley scoring scheme is useful in helping to motivate the statistic C as one appropriate for assessing possible scale differences in the X and Y populations, we could, in view of (5.64), just as easily have initially defined C in terms of the combined samples ranks as we

did for W (4.3). This means, of course, that D (5.55) is also a function of the X and Y observations only through their combined samples ranks.

30. *Assumptions.* We can use the Lepage test procedure in (5.58) without even requiring that the variances for the X and Y populations exist. Indeed, neither Assumptions A1 and A2 nor the location-scale parameter model in (5.2) and (5.3) specify anything about the existence of even the first moments of the X and Y populations. However, when the first two moments (and, therefore, the variance) for the underlying distributional model $H(u)$ in (5.2) exist, we see from the equal-in-distribution statement in (5.3) that

$$\text{var}\left(\frac{X}{\eta_1}\right) = \text{var}\left(\frac{Y}{\eta_2}\right),$$

which, in turn, implies that

$$\frac{\text{var}(X)}{\eta_1^2} = \frac{\text{var}(Y)}{\eta_2^2}.$$

Thus, when the variances exist, we see that $\gamma^2 = [\eta_1^2/\eta_2^2] = [\text{var}(X)/\text{var}(Y)]$.

It also follows from (5.3) that

$$E\left[\frac{X-\theta_1}{\eta_1}\right] = E\left[\frac{Y-\theta_2}{\eta_2}\right],$$

provided only that the first moment exists for $H(u)$ in (5.2). Thus, if this first moment exists, we have the relationship

$$E[Y] - \theta_2 = \frac{\eta_2}{\eta_1}\{E[X] - \theta_1\}.$$

As a result, if $\gamma^2 = \eta_1^2/\eta_2^2 = 1$, then $E[Y] - E[X] = \theta_2 - \theta_1$, corresponding to the standard interpretation of a location-only difference between two populations. This is the setting previously considered in Chapter 4 with the identification $\Delta = \theta_2 - \theta_1$. (We emphasize, however, that the existence of the first moment is not a necessary assumption for any of the statistical procedures developed in Chapter 4.)

31. *Large-Sample Approximation.* We have previously seen that both W^* (see Comment 4.6) and C^* (see Comment 9) have asymptotic standard normal distributions under H_0 (5.1) as $\min(m, n)$ becomes infinite. Moreover, it can be shown (see, e.g., Lepage (1971)) that W^* and C^* are asymptotically independent under H_0 (5.1) as $\min(m, n)$ becomes infinite. The conclusion that the statistic D (5.56) has, as $\min(m, n)$ tends to infinity, an asymptotic distribution under H_0 (5.1) that is chi-square with 2 degrees of freedom then follows from the properties (i) the square of a standard normal variable has a chi-square distribution with 1 degree of freedom and (ii) the sum of independent chi-square variables with degrees of freedom f_1 and f_2 has a chi-square distribution with $f_1 + f_2$ degrees of freedom.

32. *Exact Conditional Distribution of D with Ties.* To have a test with exact significance level even in the presence of ties among the X's and/or Y's, we need to consider all $\binom{N}{n}$ possible assignments of the N observations,

with n observations serving as Y's and m observations serving as X's. As in Comment 28, it still follows that under H_0 (5.1), each of the $\binom{N}{n}$ possible "meshings" of the X's and Y's has probability $1/\binom{N}{n}$. The only difference in the case of ties (see Ties in this section) is that we now use average scores and the appropriately modified $\text{var}_0(C)$ in the computation of C^* and average ranks and the appropriately modified $\text{var}_0(W)$ in the computation of W^* to calculate the value of D for each of these $\binom{N}{n}$ meshings leading to the tabulation of the exact conditional null distribution of D.

An example illustrating how to obtain such a conditional null distribution of D for a specific case of tied observations is not included here, because the details are much the same as those provided in Comments 4.5 and 11 for the conditional null distributions of W and C, respectively, in the presence of tied observations.

33. *More General Alternatives.* In his original discussion of the test procedure based on D, Lepage (1971) considered a slightly more general setting than that dictated by the location-scale parameter model in (5.2). In addition to Assumptions A1 and A2, he required that the X distribution function F and the Y distribution function G be related by the equation

$$G(t) = F(at + b), \quad \text{for every } t, \tag{5.65}$$

for some constants $a > 0$ and $-\infty < b < \infty$. He then considered tests of H_0^* : $[a = 1, b = 0]$ versus H_1^*: $[a \neq 1$ or $b \neq 0$ or both]. His null hypothesis H_0^* is, of course, identical to H_0 (5.1) considered in this section. However, his alternative H_1^* is more general than the location-scale parameter alternative H_1 (5.57) discussed here. The alternative H_1 (5.57) represents a slightly reduced subset of H_1^* corresponding to the identifications $a = \eta_1/\eta_2$ and $b = (\eta_2\theta_1 - \eta_1\theta_2)/\eta_2$.

34. *Consistency of the D Tests.* Let $\delta^* = [P(X < Y) - \frac{1}{2}]$ and $\Delta_\theta^* = [P(X > Y > \theta) + P(X < Y < \theta) - \frac{1}{4}]$. Under the minimal Assumptions A1 and A2 only, the test procedure (5.58) based on D is consistent if either $\delta^* \neq 0$ or $\theta_1 = \theta_2 = \theta$ and $\Delta_\theta^* \neq 0$. (See Comments 4.14 and 14.)

Under Assumptions A1, A2, and the additional general distributional relationship given by (5.65), the test procedure (5.58) based on D is consistent against any alternative for which either $a \neq 1$ or $b \neq 0$.

Properties

1. *Consistency.* For our statement we consider the more stringent location-scale parameter model described in (5.2). Then the test defined by (5.58) is consistent against alternatives for which either $\eta_1 \neq \eta_2$ or $\theta_1 \neq \theta_2$. (See also Comment 34.)
2. *Asymptotic Chi-Squareness.* See Lepage (1971) and Comment 31.
3. *Efficiency.* See Section 5.5.

Problems

25. It has long been generally accepted by medical doctors that exercise tends to stimulate the release of growth hormones in adolescents. However, little previous research had been directed toward assessment of possible effects that various medications might have on this phenomenon.

This fact led Falkner et al. (1981) to investigate whether the use of the drug clonidine to treat hypertension in adolescents has any effect on this exercise-induced release of growth hormones. Two groups of adolescents were involved in the study. The first was a control group consisting of 10 teenagers who had been diagnosed as hypertensive but were not being treated with clonidine. (Note that the "control" group considered by Falkner et al. included an additional seven nonhypertensive teenagers. In order not to possibly confound the effects of hypertension itself and the treatment clonidine on the release of the growth hormone during exercise, these seven subjects are not included in the control group presented in the problem. In addition, two subjects studied both as controls and again later after clonidine treatment are included here only in the control sample.) The second treatment group consisted of 13 hypertensive teenagers who were being treated with clonidine.

The experiment proceeded as follows. First, the basal level of growth hormone in the blood was measured for each of the subjects prior to exercising. Then each subject exercised on a treadmill until attaining a heart rate of 180–200 beats/min, at which time the blood level of growth hormone was once again obtained. The data in Table 5.5 represent these pre- and postexercise growth hormone blood levels (ng/ml) for the 23 subjects in the study.

Use an appropriate nonparametric test procedure to assess whether there are significant location or dispersion differences between the control hypertension population and the clonidine-treated population in their increases in growth hormone levels following exercise. Find the approximate P-value for the test.

26. In Example 5.3 we used a Lepage test procedure to assess whether or not the administration of the corticosteroid prednisone to pregnant women with ATP resulted in any location or dispersion changes in the platelet counts of their newborn infants. It would also be of interest to know whether there were any baseline (predelivery) differences in the platelet counts of those mothers in the study who were given the prednisone and those who served as the

Table 5.5 Growth Hormone Level (ng/ml)

	Preexercise	Postexercise
Control		
1	1.3	19.0
2	1.3	40.0
3	5.8	3.8
4	2.0	6.5
5	2.7	16.0
6	1.7	13.0
7	1.8	18.0
8	1.7	2.6
9	1.8	18.0
10	4.7	5.8
Clonidine-treated		
1	1.2	5.1
2	1.2	7.2
3	5.8	14.0
4	.3	4.0
5	3.3	25.0
6	2.2	15.0
7	4.1	10.0
8	1.2	7.6
9	6.4	10.0
10	1.8	10.0
11	1.8	8.0
12	5.2	40.0
13	1.3	21.0

Source: B. Falkner, G. Onesti, T. Moshang, Jr., and D. T. Lowenthal (1981).

Table 5.6 Maternal Platelet Counts (per mm^3)

Mothers given prednisone	Mothers not given prednisone
12,000	15,000
25,000	44,000
30,000	52,000
38,000	64,000
50,000	65,000
80,000	80,000
85,000	
126,000	
130,000	
180,000	

Source: M. Karpatkin, R. F. Porges, and S. Karpatkin (1981).

no-prednisone control group. The platelet count (per cubic millimeter) data for the mothers are given in Table 5.6.

Find the P-value for an appropriate nonparametric test procedure to assess whether are any significant location or dispersion differences in the predelivery maternal platelet counts for the control and prednisone-treated groups.

27. When there are no tied X and/or Y observations, show that the representation for C given in (5.64) in Comment 29 is indeed equivalent to the original definition of C in (5.6).

28. Generate the exact null distribution of D for the setting $m = 2$, $n = 3$. (See Comment 28.)

29. Consider the general relationship between the distribution functions for the X and Y populations prescribed in (5.65) of Comment 33. Verify that the location-scale parameter model relationship given in (5.2) corresponds to the special case of (5.65) with $a = \eta_1/\eta_2$ and $b = (\eta_2\theta_1 - \eta_1\theta_2)/\eta_2$.

30. Consider the television-viewing behavior data in Table 4.4. For these data, find the approximate P-value for an appropriate test of whether there are either location or dispersion differences in the time spent in the room after witnessing the violent behavior of those children who had previously watched the *Karate Kid* versus those children who had previously watched parts of the 1984 Summer Olympic Games. Comment on your finding in view of the results of Problems 3 and 4.5.

31. Consider the alcoholic intake data in Table 4.2. For these data, find the P-value for an appropriate test of whether there are either location or dispersion differences between the control and SST data. Discuss the result in conjunction with the previous findings in Example 4.2 and Problem 9.

32. Consider the following two-sample data for $m = 3$, $n = 3$: $X_1 = -3.7$, $X_2 = 4.6$, $X_3 = 1.5$, $Y_1 = 1.5$, $Y_2 = 4.6$, $Y_3 = 1.5$. Using the approach discussed in Comment 32, find the exact conditional null distribution of the Lepage statistic D (5.55). Compare and contrast the upper $\alpha = .10$ percentile for this exact conditional null distribution with the corresponding upper $\alpha = .10$ percentile for the null distribution of D for $m = n = 3$ and no tied observations.

5.4 A DISTRIBUTION-FREE TEST FOR GENERAL DIFFERENCES IN TWO POPULATIONS (KOLMOGOROV–SMIRNOV)

Hypothesis

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples satisfying Assumptions A1 and A2 from continuous populations with distribution functions F and G, respectively.

Under these assumptions we are interested in assessing whether there are *any* differences whatsoever between the X and Y probability distributions. Thus, we are interested in testing the null hypothesis H_0 (5.1) against the most general alternative possible, namely,

$$H_1 : [F(t) \neq G(t) \text{ for at least one } t]. \tag{5.66}$$

Procedure

To compute the two-sided two-sample Kolmogorov–Smirnov general alternative statistic J, we first need to obtain the empirical distribution functions for the X and Y samples. For every real number t, let

$$F_m(t) = \frac{\text{number of sample } X\text{'s} \leq t}{m} \tag{5.67}$$

and

$$G_n(t) = \frac{\text{number of sample } Y\text{'s} \leq t}{n}. \tag{5.68}$$

(The functions $F_m(t)$ and $G_n(t)$ are called the *empirical distribution functions* for the X and Y samples, respectively.) Let

$$d = \text{greatest common divisor of } m \text{ and } n \tag{5.69}$$

and set

$$J = \frac{mn}{d} \max_{(-\infty<t<\infty)} \{|F_m(t) - G_n(t)|\}. \tag{5.70}$$

The statistic J is the two-sided two-sample Kolmogorov–Smirnov statistic. To actually calculate J for the given X and Y samples, we use the fact that $F_m(t)$ and $G_n(t)$ are step functions changing functional values only at the observed X and Y sample observations, respectively. Thus, if we let $Z_{(1)} \leq \cdots \leq Z_{(N)}$ denote the $N = (m + n)$ ordered values for the combined sample of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, then we can rewrite J (5.70) in the computational form

$$J = \frac{mn}{d} \max_{i=1,\ldots,N} \{|F_m(Z_{(i)}) - G_n(Z_{(i)})|\}. \tag{5.71}$$

To test H_0 (5.1), corresponding to identical X and Y probability distributions, versus the general alternative H_1 (5.66), corresponding to *any* possible difference between the X and Y probability distributions, at the α level of significance,

$$\text{Reject } H_0 \text{ if } J \geq j_\alpha; \quad \text{otherwise do not reject,} \tag{5.72}$$

where the constant j_α is chosen to make the type I error probability equal to α. The constant j_α is the upper α percentile for the null H_0(5.1) distribution of J. Comment 38 explains how to obtain the critical value j_α for sample sizes m and n and available values of α.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic distribution of J, suitably normalized, as $\min(m,n)$ tends to infinity. Set

$$J^* = \left(\frac{mn}{N}\right)^{1/2} \max_{i=1,\ldots,N} \{|F_m(Z_{(i)}) - G_n(Z_{(i)})|\} = \frac{d}{(mnN)^{1/2}} J. \tag{5.73}$$

As $\min(m,n)$ tends to infinity,

$$P_0(J^* < s) \longrightarrow \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}, 0 \quad \text{for} \quad s >, \leq 0. \tag{5.74}$$

Defining the function Q(s) by

$$Q(s) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}, \quad s > 0, \tag{5.75}$$

the large-sample approximation to procedure (5.72) based on (5.74) and (5.75) is

$$\text{Reject } H_0 \text{ if } J^* \geq q_\alpha^*; \quad \text{otherwise do not reject,} \tag{5.76}$$

where q_α^* is defined by

$$Q(q_\alpha^*) = \alpha. \tag{5.77}$$

To find q_α^*, we use the `R` command `qKolSmirnLSA(`α`)`. For example, to find $q_{.05}^*$, we apply `qKolSmirnLSA(.05)` and obtain $q_{.05}^* = 1.358$.

Ties

The empirical distribution functions $F_m(t)$ and $G_n(t)$, given by (5.67) and (5.68), respectively, are well defined in the case of ties and no adjustments are necessary in the calculation of J (5.70). (See Comment 39.) The test is then conducted using the same critical point j_α (5.72) as specified for the untied case. This approach is conservative; it yields a test with a significance level that does not exceed the nominal level α (see Hájek and Šidák (1967, p. 123), Noether (1963), and Walsh (1963)). For different methods of treating ties when using the Kolmogorov–Smirnov statistic J, see Hájek (1969, p. 134, 145).

EXAMPLE 5.4 ***Effect of Feedback on Salivation Rate.***

The effect of enabling a subject to hear himself salivate while trying to increase or decrease his salivary rate has been studied by Delse and Feather (1968). Two groups of subjects were told to attempt to increase their salivary rates upon observing a light to the left and decrease their salivary rates upon observing a light to the right. The apparatus for collecting and recording the amounts of saliva was described by Delse and Feather (1968) and also Feather and Wells (1966). Members of the feedback group received a 0.2-s, 1000-cps tone for each drop collected, whereas members of the no-feedback group did not receive any indication of their salivary rates. Table 5.7 gives differences of the form mean number of drops over 13 increase signals minus mean number of drops

Table 5.7 Mean Drop Differences

Feedback group	No-Feedback group
−.15	2.55
8.60	12.07
5.00	.46
3.71	.35
4.29	2.69
7.74	−.94
2.48	1.73
3.25	.73
−1.15	−.35
8.38	−.37

Source: F. C. Delse and B. W. Feather (1968).

over 13 decrease signals for the feedback group and the no-feedback group, each group consisting of 10 subjects.

Since both sample sizes are equal to 10, we arbitrarily choose to label the feedback group data as the X sample and the no-feedback group data as the Y sample. Thus, we have $m = n = 10$, $N = (10 + 10) = 20$, and $d = 10$. We simultaneously illustrate the calculation of the values of the empirical distribution functions $F_{10}(t)$ and $G_{10}(t)$ at the ordered combined sample values $Z_{(1)} \leq \cdots \leq Z_{(20)}$ from Table 5.7, as well as the absolute differences $|F_{10}(Z_{(i)}) - G_{10}(Z_{(i)})|$, in the following display.

i	$Z_{(i)}$	$F_{10}(Z_{(i)})$	$G_{10}(Z_{(i)})$	$\lvert F_{10}(Z_{(i)}) - G_{10}(Z_{(i)})\rvert$
1	−1.15	$\frac{1}{10}$	$\frac{0}{10}$	$\frac{1}{10}$
2	−.94	$\frac{1}{10}$	$\frac{1}{10}$	0
3	−.37	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$
4	−.35	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{2}{10}$
5	−.15	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{1}{10}$
6	.35	$\frac{2}{10}$	$\frac{4}{10}$	$\frac{2}{10}$
7	.46	$\frac{2}{10}$	$\frac{5}{10}$	$\frac{3}{10}$
8	.73	$\frac{2}{10}$	$\frac{6}{10}$	$\frac{4}{10}$
9	1.73	$\frac{2}{10}$	$\frac{7}{10}$	$\frac{5}{10}$
10	2.48	$\frac{3}{10}$	$\frac{7}{10}$	$\frac{4}{10}$
11	2.55	$\frac{3}{10}$	$\frac{8}{10}$	$\frac{5}{10}$
12	2.69	$\frac{3}{10}$	$\frac{9}{10}$	$\frac{6}{10}$
13	3.25	$\frac{4}{10}$	$\frac{9}{10}$	$\frac{5}{10}$
14	3.71	$\frac{5}{10}$	$\frac{9}{10}$	$\frac{4}{10}$
15	4.29	$\frac{6}{10}$	$\frac{9}{10}$	$\frac{3}{10}$
16	5.00	$\frac{7}{10}$	$\frac{9}{10}$	$\frac{2}{10}$
17	7.74	$\frac{8}{10}$	$\frac{9}{10}$	$\frac{1}{10}$
18	8.38	$\frac{9}{10}$	$\frac{9}{10}$	0
19	8.60	$\frac{10}{10}$	$\frac{9}{10}$	$\frac{1}{10}$
20	12.07	$\frac{10}{10}$	$\frac{10}{10}$	0

For example, consider the evaluation of $F_{10}(Z_{(4)})$. We must count the number of X's less than or equal to $Z_{(4)} = -.35$, and divide this count by 10. From Table 5.7, we find that only one of the X values (-1.15) is less than $-.35$, none is equal to $-.35$, and thus $F_{10}(Z_{(4)}) = \frac{1}{10}$. Similarly, $G_{10}(Z_{(4)})$ is equal to {the number of Y's that are less than or equal to $-.35$}/10. From Table 5.7, we find two Y-values ($-.94$ and $-.37$) that are less than $-.35$ and one Y-value that is equal to $-.35$; thus, $G_{10}(Z_{(4)}) = \frac{3}{10}$. From this computational Table for the $|F_{10}(Z_{(i)}) - G_{10}(Z_{(i)})|$ values, we find

$$\max_{i=1,\ldots,20} \{|F_{10}(Z_{(i)}) - G_{10}(Z_{(i)})|\} = \tfrac{6}{10},$$

corresponding to $Z_{(12)}$. It follows from (5.71) that $J = [(10)(10)/10](6/10) = 6$.

Applying the R command `pKolSmirn(mean.drop$x,mean.drop$y)`, we find that `pKolSmirn(mean.drop$x,mean.drop$y)` $= P_0(J \geq 6) = .0524$. That is, in the notation of (5.72) with $m = n = 10$, we have $j_{.0524} = 6$. Thus, the lowest level at which we can reject H_0 (5.1) with our observed value of $J = 6$ (i.e., the P-value for the data) using procedure (5.72) is .0524, indicating some marginal evidence in the samples that feedback might have an effect on salivation rate.

To perform the large-sample approximation, we compute J^* (5.73). We find that $J^* = \{10/[10(10)(20)]^{1/2}\}(6) = 1.34$. Since `qKolSmirnLSA(.0551)` $= J^* = 1.34$, the smallest significance level at which we reject H_0, using the large-sample approximation to the Kolmogorov–Smirnov test, is approximately .0551.

Comments

35. *Motivation for the Test.* The empirical distribution functions $F_m(t)$ (5.67) and $G_n(t)$ (5.68) are estimators of the underlying distribution functions $F(t) = P\{X \leq t\}$ and $G(t) = P\{Y \leq t\}$, respectively. Thus, dJ/mn may be viewed as an estimator of $\max_{-\infty<t<\infty} |F(t) - G(t)| = \max_{-\infty<t<\infty} |P\{X \leq t\} - P\{Y \leq t\}|$, and this parameter is zero when H_0 (5.1) is true. Hence, large J values indicate a deviation from H_0 in the direction of the general alternative specified by (5.66).

36. *Equivalent Form.* In the case of no ties among the N combined $Z_{(i)}$ values, there is an alternative counting formulation for the test statistic J (5.70). Define the variables $\delta_i, i = 1, \ldots, N$, by

$$\delta_i = \begin{cases} 1, & \text{if } Z_{(i)} \text{ is an } X \text{ observation,} \\ 0, & \text{if } Z_{(i)} \text{ is a } Y \text{ observation.} \end{cases} \tag{5.78}$$

Set

$$s_j = \left[\frac{jm}{N} - \delta_1 - \cdots - \delta_j\right], \quad j = 1, \ldots, N. \tag{5.79}$$

Then the Kolmogorov–Smirnov statistic J (5.70) can also be expressed as

$$J = (N/d) \max\{|s_1|, \ldots, |s_N|\}. \tag{5.80}$$

(We note that, unlike expression (5.70), the formulation in (5.80) is not well defined in the case of ties among the $Z_{(i)}$'s.)

37. *Equal Sample Sizes.* In settings where $m = n$, the computational expression for J (5.71) can be simplified to

$$J = \max_{i=1,\ldots,N} |Q(Z_{(i)}) - S(Z_{(i)})|, \tag{5.81}$$

where, for every real number t,

$$Q(t) = mF_m(t) = [\text{number of sample } X\text{'s} \leq t] \tag{5.82}$$

and

$$S(t) = nG_n(t) = [\text{number of sample } Y\text{'s} \leq t]. \tag{5.83}$$

Thus, for the salivation data in Example 5.4, we have

$$J = |Q(Z_{(12)}) - S(Z_{(12)})| = |3 - 9| = 6,$$

in agreement with the value obtained via (5.71).

38. *Derivation of the Distribution of J under H_0 (No-Ties Case).* The null (H_0) distribution of J in the case of no ties can be obtained by using the fact that under H_0 (5.1) all possible $\binom{N}{n}$ meshings of the X's and Y's are equally likely, each having probability $1/\binom{N}{n}$. In the ensuing illustration, we derive the null distribution of J (5.70) for the sample sizes $m = 1$, $n = 3$. Here, $N = 4, d = 1$, and thus $J = 3 \max_{i=1,\ldots,4} |F_1(Z_{(i)}) - G_3(Z_{(i)})|$. We now list the $\binom{4}{1} = 4$ possible meshings, and for each of these meshings we give the associated values of $(F_1(Z_{(1)}), \ldots, F_1(Z_{(4)}))$, the associated values of $(G_3(Z_{(1)}), \ldots, G_3(Z_{(4)}))$, and finally the values of J. Thus, $P_0\{J = 2\} = \binom{2}{4} = .5$ and $P_0\{J = 3\} = .5$.

Meshings	$(F_1(Z_{(1)}), \ldots, F_1(Z_{(4)}))$	$(G_3(Z_{(1)}), \ldots, G_3(Z_{(4)}))$	J
XYYY	(1,1,1,1)	$\left(0, \frac{1}{3}, \frac{2}{3}, 1\right)$	3
YXYY	(0,1,1,1)	$\left(\frac{1}{3}, \frac{1}{3}, \frac{2}{3}, 1\right)$	2
YYXY	(0,0,1,1)	$\left(\frac{1}{3}, \frac{2}{3}, \frac{2}{3}, 1\right)$	2
YYYX	(0,0,0,1)	$\left(\frac{1}{3}, \frac{2}{3}, 1, 1\right)$	3

For given sample sizes m and n, the R command `cKolSmirn`(α, m, n) can be used to find the available upper-tail critical values j_α for possible values of J. For a given available significance level α, the critical value j_α then corresponds to $P_0(J \geq j_\alpha) = \alpha$ and is given by `cKolSmirn`$(\alpha, m, n) = j_\alpha$. Thus, for example, for $m = 4$ and $n = 6$, we have `cKolSmirn(.04762,4,6)` $= 10$ so that $P_0(J \geq 10) = .04762$ and $j_{.04762} = 10$ for $m = 4$ and $n = 6$.

39. *Ties.* To illustrate how the computational formula for J given in expression (5.71) is well defined in the case of ties, we consider the following artificial set of tied data: $X_1 = 3$, $X_2 = 3$, $X_3 = 5$, $X_4 = 7$, $X_5 = 9$ and $Y_1 = 3$, $Y_2 = 4$, $Y_3 = 4$, $Y_4 = 6$, $Y_5 = 7$, $Y_6 = 8$, $Y_7 = 10$, $Y_8 = 10$, $Y_9 = 11$, $Y_{10} = 12$. Here we

have $m = 5$, $n = 10$, $N = 15$, and $d = 5$. Following the tabular approach of Example 5.4 for computation of J, we obtain

i	$Z_{(i)}$	$F_5(Z_{(i)})$	$G_{10}(Z_{(i)})$	$\lvert F_5(Z_{(i)}) - G_{10}(Z_{(i)})\rvert$
1	3	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{3}{10}$
2	3	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{3}{10}$
3	3	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{3}{10}$
4	4	$\frac{2}{5}$	$\frac{3}{10}$	$\frac{1}{10}$
5	4	$\frac{2}{5}$	$\frac{3}{10}$	$\frac{1}{10}$
6	5	$\frac{3}{5}$	$\frac{3}{10}$	$\frac{3}{10}$
7	6	$\frac{3}{5}$	$\frac{4}{10}$	$\frac{3}{10}$
8	7	$\frac{4}{5}$	$\frac{5}{10}$	$\frac{3}{10}$
9	7	$\frac{4}{5}$	$\frac{5}{10}$	$\frac{3}{10}$
10	8	$\frac{4}{5}$	$\frac{6}{10}$	$\frac{2}{10}$
11	9	$\frac{5}{5}$	$\frac{6}{10}$	$\frac{4}{10}$
12	10	$\frac{5}{5}$	$\frac{8}{10}$	$\frac{2}{10}$
13	10	$\frac{5}{5}$	$\frac{8}{10}$	$\frac{2}{10}$
14	11	$\frac{5}{5}$	$\frac{9}{10}$	$\frac{1}{10}$
15	12	$\frac{5}{5}$	$\frac{10}{10}$	0

Thus, the empirical distribution function $F_5(t)$ for the X sample jumps from 0 to $\frac{2}{5}$ at $Z_{(1)} = Z_{(2)} = Z_{(3)} = 3$, since two of the five X values are 3's. Similarly, the empirical distribution function $G_{10}(t)$ for the Y sample jumps from $\frac{1}{10}$ to $\frac{3}{10}$ and $\frac{6}{10}$ to $\frac{8}{10}$ at $Z_{(4)} = Z_{(5)} = 4$ and $Z_{(12)} = Z_{(13)} = 10$, respectively, since there are two 4's and two 10's among the Y observations. For these tied data, we find

$$\max_{i=1,\ldots,15} |F_5(Z_{(i)}) - G_{10}(Z_{(i)})| = |F_5(Z_{(11)}) - G_{10}(Z_{(11)})| = \tfrac{4}{10}.$$

From (5.71) it then follows (as in the case of no ties) that $J = \left[\frac{5(10)}{5}\right]\left(\frac{4}{10}\right) = 4$.

40. *Exact Conditional Distribution of J with Ties.* To have a test with exact significance level even in the presence of ties among the X's and/or Y's, we need to consider all $\binom{N}{n}$ possible assignments of the N observations with n observations serving as Y's and m observations serving as X's. As in Comment 38, it still follows that, under H_0 (5.1), each of the $\binom{N}{n}$ possible meshings of the X's and Y's has probability $1/\binom{N}{n}$. The only difference in the case of ties is that now in the computation of J for each of these $\binom{N}{n}$ meshings, the jumps in the X and Y empirical distribution functions can occur at common observations and the sizes of these jumps can be greater than $1/m$ or $1/n$, respectively. We illustrate this construction for the following $m = 2, n = 3$ data: $X_1 = 3.2, X_2 = 6.3, Y_1 = 1.9, Y_2 = 1.9, Y_3 = 6.3$. The associated

ordered $Z_{(i)}$ values are $Z_{(1)} = Z_{(2)} = 1.9 < Z_{(3)} = 3.2 < Z_{(4)} = Z_{(5)} = 6.3$ and the corresponding value of J (5.71) is $\frac{2(3)}{1}|F_2(Z_{(1)}) - G_3(Z_{(1)})| = 6|F_2(Z_{(2)}) - G_3(Z_{(2)})| = 6|F_2(1.9) - G_3(1.9)| = 6|0 - \frac{2}{3}| = 4$. To assess the significance of this value of J, we obtain its conditional distribution by considering the $\binom{5}{3} = 10$ possible assignments of the observations 1.9, 1.9, 3.2, 6.3, and 6.3 to serve as two X observations and three Y observations. These 10 assignments and the corresponding values of J are

X observations	Y observations	Probability under H_0	Value of J
1.9, 1.9	3.2, 6.3, 6.3	$\frac{1}{10}$	6
1.9, 3.2	1.9, 6.3, 6.3	$\frac{1}{10}$	4
1.9, 3.2	1.9, 6.3, 6.3	$\frac{1}{10}$	4
1.9, 6.3	1.9, 3.2, 6.3	$\frac{1}{10}$	1
1.9, 6.3	1.9, 3.2, 6.3	$\frac{1}{10}$	1
1.9, 6.3	1.9, 3.2, 6.3	$\frac{1}{10}$	1
1.9, 6.3	1.9, 3.2, 6.3	$\frac{1}{10}$	1
3.2, 6.3	1.9, 1.9, 6.3	$\frac{1}{10}$	4
3.2, 6.3	1.9, 1.9, 6.3	$\frac{1}{10}$	4
6.3, 6.3	1.9, 1.9, 3.2	$\frac{1}{10}$	6

This yields the null tail probabilities

$$P_0(J \geq 6) = \frac{2}{10}, \quad P_0(J \geq 4) = \frac{6}{10}, \quad P_0(J \geq 1) = 1.$$

This distribution is called the *conditional null distribution* or the *permutation null distribution* of J, given the set of tied observations {1.9, 1.9, 3.2, 6.3, 6.3}. For the particular observed value $J = 4$, we have that $P_0(J \geq 4) = \frac{6}{10}$. (Note that the particular observed X and Y sample values are not important to the calculation of this conditional null distribution of J. It is critical only that the two smallest observations are tied in value, the middle ordered value is untied, and the two largest observations are tied. Thus, for example, the two sets of sample observations $\{X_1 = 3.2,\ X_2 = 6.3,\ Y_1 = 1.9,\ Y_2 = 1.9,\ Y_3 = 6.3\}$ and $\{X_1 = -12.1, X_2 = 13.7, Y_1 = -12.1, Y_2 = 0, Y_3 = 13.7\}$ yield the same exact conditional null distribution of J.)

41. *Large-Sample Approximation.* Smirnov (1939) derived the asymptotic (min (m, n) tending to infinity) distribution of the standardized Kolmogorov–Smirnov statistic J^* (5.73) using the work of Kolmogorov (1933) on the asymptotic (m tending to infinity) distribution of the one-sample statistic

$$J_0 = \sqrt{m} \max_{-\infty < a < \infty} |F_m(a) - F_0(a)|, \tag{5.84}$$

where $F_m(\cdot)$ is the empirical distribution function for a random sample of size m from the (assumed) continuous distribution with distribution function $F(a) =$

$P(X \leq a)$ and $F_0(a)$ is a completely specified distribution function. The statistic J_0 can be used to test the goodness-of-fit hypothesis that the random sample $X_1, \ldots, X_m$ has been drawn from a population with distribution function F_0, namely,

$$H_0' : [P(X \leq a) = F_0(a) \text{ for all } -\infty < a < \infty], \tag{5.85}$$

versus the broad alternative that the population from which the sample was drawn does not have distribution function F_0.

42. *Test Based on the One-Sample Limit of the Wilcoxon Rank Sum Statistic.* It is of interest to note that the two-sample Wilcoxon test discussed in Section 4.1 can be reduced to a test of H_0' (5.85) by allowing one of the sample sizes, say n, to become infinite. Moses (1964) showed how this leads to a test based on $W_0 = \sum_{j=1}^{m} F_0(X_j)$. (The normal approximation to W_0 treats $[W_0 - (m/2)]/(m/12)^{1/2}$ as an approximate $N(0, 1)$ random variable under H_0'.) Moses pointed out that a test based on W_0 is particularly convenient when F_0 is known but is specified by tabular data, such as demographic data on age of death distributions, rather than being given by a mathematical expression.
43. *Consistency of the J Tests.* Define the class $\mathcal{C}$ of pairs of distribution functions F and G by

$$\mathcal{C} = \{(F, G) : F(x) \neq G(x) \text{ for at least one } x\}. \tag{5.86}$$

Under the minimal Assumptions A1 and A2 only, the test procedure (5.72) is consistent for any $(F, G) \in \mathcal{C}$; that is, the test is consistent against *any* differences between the F and G distributions (i.e., *whenever* H_0 (5.1) is false). In gaining this extra protection against all differences, we do, however, sacrifice power against specific subclasses of alternatives (such as location shifts or differences in dispersions).

Properties

1. *Consistency.* See Comment 43.
2. *Asymptotic Distribution*. See Smirnov (1939) and Comment 41.
3. *Efficiency.* See Capon (1965), Ramachandramurty (1966b), Yu (1971), and Section 5.5.

Problems

33. The data in Table 5.8 are a subset of the data obtained by Friedman et al. (1971) in an experiment comparing the average concentrations of human plasma growth hormone both resting and after arginine hydrochloride infusion in relatively coronary-prone subjects (persons with type A behavior patterns) with the corresponding concentrations of relatively coronary-resistant individuals (subjects with type B behavior patterns). Type A behavior is characterized by an excessive sense of time urgency, drive, and competitiveness; type B denotes a converse type of behavior. Earlier studies (cf. Friedman and Rosenman (1959)) indicated that type A individuals may be more prone to coronary heart disease than type B individuals.

Table 5.8 Peak Levels of Human Plasma Growth Hormone after Arginine Hydrochloride Infusion (Initial Test, ng/ml)

Type A subjects	Type B subjects
3.6	16.2
2.6	17.4
4.7	8.5
8.0	15.6
3.1	5.4
8.8	9.8
4.6	14.9
5.8	16.6
4.0	15.9
4.6	5.3
	10.5

Source: M. Friedman, S. O. Byers, R. H. Rosenman, and R. Neuman (1971).

Find the P-value for an appropriate test of whether there is any difference between the probability distribution of peak level human plasma growth hormone (after arginine hydrochloride infusion) for type A subjects and that for type B subjects.

34. Consider the alcoholic intake data in Table 4.2. For these data, find the P-value for an appropriate test of whether there are *any* differences between the control and SST probability distributions. Discuss this result in conjunction with the previous findings in Example 4.2, Problem 9, and Problem 31.

35. Verify directly, or illustrate with a numerical example, that representations (5.71) and (5.80) for J are indeed equivalent.

36. When $m = n$, show that both representations (5.71) and (5.80) for J are equivalent to the expression

$$J = \max\{|t_1|, |t_2|, \ldots, |t_N|\}, \tag{5.87}$$

where

$$t_j = (1 - 2\delta_1) + (1 - 2\delta_2) + \cdots + (1 - 2\delta_j), \tag{5.88}$$

and the δ's are given by (5.78).

37. Calculate the value of J for the salivation data in Table 5.7 using the equivalent (when $m = n$) expression in (5.87).

38. Apply the two-sided Wilcoxon rank sum test procedure from Section 4.1 to the salivation data in Table 5.7 by finding the appropriate P-value. Compare the conclusion indicated by this Wilcoxon rank sum procedure with that indicated by the Kolmogorov–Smirnov procedure in Example 5.4. Comment on your findings.

39. Generate the exact null distribution of J (5.70) for the setting $m = 3$, $n = 3$. (See Comment 38.)

40. Consider the growth hormone level data found in Table 5.5. Use the Kolmogorov–Smirnov test procedure to assess whether there are significant differences of *any* kind between the control hypertension population and the clonidine-treated population in their increases in growth hormone levels following exercise. Find the appropriate P-value for the test and compare it with the P-value obtained in Problem 25.

41. Consider the following two-sample data for $m = 3$, $n = 3$: $X_1 = -3.7$, $X_2 = 4.6$, $X_3 = 1.5$, $Y_1 = 1.5$, $Y_2 = 4.6$, $Y_3 = 1.5$. Using the approach discussed in Comment 40, find the exact conditional null distribution of the Kolmogorov–Smirnov statistic J (5.70). Compare and

contrast this exact conditional null distribution with the corresponding null distribution of J for $m = n = 3$ and no tied observations, as obtained in Problem 39.

42. Consider the serum iron data in Table 5.1. Use the Kolmogorov–Smirnov test procedure to assess whether there are significant differences of *any* kind between the distribution of serum iron values obtained by the Ramsay method and the distribution of serum iron values obtained by the Jung–Parekh method. Find the P-value for the test and compare it with the results discussed in Example 5.1.

5.5 EFFICIENCIES OF TWO-SAMPLE DISPERSION AND BROAD ALTERNATIVES PROCEDURES

Recall the classical normal theory $\mathcal{F}$-test for equality of variances based on the statistic

$$D = \frac{S_x^2}{S_y^2}, \tag{5.89}$$

where $S_x^2 = \sum_{i=1}^{m}(X_i - \bar{X})^2/(m-1), S_y^2 = \sum_{j=1}^{n}(Y_j - \bar{Y})^2/(n-1), \bar{X} = \sum_{i=1}^{m} X_i/m$, and $\bar{Y} = \sum_{j=1}^{n} Y_j/n$. The significance level of this $\mathcal{F}$-test is extremely sensitive to nonnormality. (See Comment 26.) This is also true of the coverage probability of the confidence intervals for σ_2^2/σ_1^2 that are based on the ratio of sample variances and derived from the $\mathcal{F}$-test. The Box–Andersen (1955) test "adjusts" the $\mathcal{F}$-test to remedy this difficulty. Since this Box–Andersen approach has desirable properties, we report asymptotic efficiencies of the test procedures of Sections 5.1 and 5.2, as well as the point estimators and confidence intervals/bounds associated with the jackknife approach (see Comment 24), with respect to the corresponding Box–Andersen procedures. (The specific Box–Andersen procedures that we refer to are (a) the APF test of Shorack (1969), which is a slight variation of the test used by Box and Andersen for the case where the parameters θ_1 and θ_2 of model (5.2) are known; (b) an associated estimator given by Shorack (1965); and (c) the associated confidence interval and bounds discussed in Shorack (1969).)

The Pitman asymptotic relative efficiency for scale alternatives of the Ansari–Bradley test based on C (5.6) relative to the Box–Andersen adjusted $\mathcal{F}$-test based on D (5.89) is

$$e(C,D) = 12(\beta_G - 1)\left[\int_{-\infty}^{0} x g^2(x)dx - \int_{0}^{\infty} x g^2(x)dx\right]^2, \tag{5.90}$$

where

$$\beta_G = \frac{\int_{-\infty}^{\infty}(x-\mu)^4 g(x)dx}{\left\{\int_{-\infty}^{\infty}(x-\mu)^2 g(x)dx\right\}^2}$$

is the kurtosis and $\mu = \int_{-\infty}^{\infty} x g(x)dx$ is the mean of the population with distribution function $G(\cdot)$ and probability density function $g(\cdot)$.

The expression in (5.90) was obtained by Ansari and Bradley (1960). Some values of $e(C,D)$ for selected $G(\cdot)$ are

G	Normal	Uniform	Double exponential
$e(C,D)$	.61	.60	.94

Miller (1968) pointed out that the asymptotic relative efficiency of the jackknife procedures (tests, point estimators, and confidence intervals/bounds) with respect to the Box–Andersen procedures has the value 1 for *any* underlying distribution $F(\cdot)$; that is, $e(Q, D) \equiv 1$, where Q is given by (5.35) and D represents the Box–Andersen adjusted $\mathcal{F}$-test procedures.

We do not know of any results for the asymptotic efficiencies of the Lepage test for location or scale differences (Section 5.3).

The determination of asymptotic relative efficiencies for the Kolmogorov–Smirnov test based on J (5.70) is difficult, owing to the complicated form of the asymptotic distribution of the Kolmogorov–Smirnov statistic. Capon (1965) obtained lower bounds for the asymptotic relative efficiency of the Kolmogorov–Smirnov test. In particular, for normal translation alternatives, Capon derived the lower bound of .637 for the asymptotic relative efficiency of the Kolmogorov–Smirnov test with respect to the normal theory two-sample t test (see Section 4.5 and also Ramachandramurty (1966b) and Yu (1971)). For related efficiency results using different notions of asymptotic efficiency, see Klotz (1967), Hájek and Šidák (1967, p. 272), and Anděl (1967).

Chapter 6

The One-Way Layout

INTRODUCTION

The procedures of this chapter are designed for statistical analyses in which primary interest is centered on the relative locations (medians) of three or more populations. This development represents a direct generalization of the two-sample location problem (discussed in Chapter 4) to situations in which the data consist of $k\,(\geq 3)$ random samples, one sample from each of k populations. The basic null hypothesis of interest in that of no differences in locations (medians), under which the k samples can be treated as a single (combined) sample from one population. The alternatives considered here correspond to a variety of restricted nonnull relationships between the locations (medians). We encounter two types of data for which such analyses are important. The first of these corresponds to a general setting of k populations (referred to as *treatments* for convenience) with no additional conditions. The second deals with the setting where one of the *treatments* represents a *control* (or placebo) population, and we are interested in detecting which, if any, of the other $(k-1)$ treatments are different from this control.

Section 6.1 presents a distribution-free test directed at general alternatives for the setting of k treatments. A distribution-free test designed for detecting ordered alternatives among k treatments is considered in Section 6.2 and generalized in Section 6.3 to the broader class of umbrella alternatives. In Section 6.4 a distribution-free test procedure is presented for the simultaneous comparison of $(k-1)$ treatments with a control. In Sections 6.5–6.7 we introduce multiple comparison procedures designed to detect which particular populations, if any, differ from one another. Sections 6.5 and 6.6 are devoted to procedures for making the total of $\binom{k}{2}$ pairwise comparisons between all k treatments in the general and ordered alternatives settings, respectively. Section 6.7 presents multiple comparison procedures based on simple random samples for deciding which, if any, of $(k-1)$ treatments are different from a control. Section 6.8 considers estimators of contrasts in the treatment effects, and Section 6.9 deals with simultaneous confidence intervals for simple contrasts. The asymptotic relative efficiencies for translation alternatives of the procedures discussed in this chapter with respect to their normal theory counterparts based on sample averages are discussed in Section 6.10.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

Data. The data consist of $N = \sum_{j=1}^{k} n_j$ observations, with n_j observations from the jth treatment, $j = 1, \ldots, k$.

Treatments			
1	2	...	k
X_{11}	X_{12}	...	X_{1k}
X_{21}	X_{22}	...	X_{2k}
$\vdots$	$\vdots$		$\vdots$
$X_{n_1 1}$	$X_{n_2 2}$	...	$X_{n_k k}$

Assumptions

A1. The N random variables $\{X_{1j}, X_{2j}, \ldots, X_{n_j j}\}$, $j = 1, \ldots, k$, are mutually independent.

A2. For each fixed $j \in \{1, \ldots, k\}$, the n_j random variables $\{X_{1j}, X_{2j}, \ldots, X_{n_j j}\}$ are a random sample from a continuous distribution with distribution function F_j.

A3. The distribution functions $F_1, \ldots, F_k$ are connected through the relationship

$$F_j(t) = F(t - \tau_j), \quad -\infty < t < \infty, \tag{6.1}$$

for $j = 1, \ldots, k$, where F is a distribution function for a continuous distribution with unknown median θ and τ_j is the unknown treatment effect for the jth population.

We note that Assumptions A1–A3 correspond directly to the usual one-way layout model commonly associated with normal theory assumptions; that is, Assumptions A1–A3 are equivalent to the representation

$$X_{ij} = \theta + \tau_j + e_{ij}, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, k,$$

where θ is the overall median, τ_j is the *treatment j effect*, and the N e's form a random sample from a continuous distribution with median 0. (Under the additional assumption of normality, the medians θ and 0 are, of course, also the respective means.)

Hypothesis

The null hypothesis of interest in Sections 6.1–6.4 of this chapter is that of no differences among the treatment effects $\tau_1, \ldots, \tau_k$, namely,

$$H_0 : [\tau_1 = \cdots = \tau_k]. \tag{6.2}$$

This null hypothesis asserts that each of the underlying distributions $F_1, \ldots, F_k$ is the same, corresponding to $F_1 \equiv F_2 \equiv \cdots \equiv F_k \equiv F$ in (6.1).

6.1 A DISTRIBUTION-FREE TEST FOR GENERAL ALTERNATIVES (KRUSKAL–WALLIS)

In this section, we present a procedure for testing H_0 (6.2) against the general alternative that at least two of the treatment effects are not equal, namely,

$$H_1 : [\tau_1, \ldots, \tau_k \text{ not all equal}]. \tag{6.3}$$

Procedure

To compute the Kruskal–Wallis statistic, H, we first combine all N observations from the k samples and order them from least to greatest. Let r_{ij} denote the rank of X_{ij} in this joint ranking and set

$$R_j = \sum_{i=1}^{n_j} r_{ij} \quad \text{and} \quad R_{\cdot j} = \frac{R_j}{n_j}, \quad j = 1, \ldots, k. \tag{6.4}$$

Thus, for example, R_1 is the sum of the joint ranks received by the treatment 1 observations and $R_{\cdot 1}$ is the average rank for these same observations. The Kruskal–Wallis statistic H is then given by

$$\begin{aligned} H &= \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left(R_{\cdot j} - \frac{N+1}{2} \right)^2 \\ &= \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} \right) - 3(N+1), \end{aligned} \tag{6.5}$$

where $(N+1)/2 = \left(\sum_{j=1}^{k} \sum_{i=1}^{n_j} r_{ij} / N \right)$ is the average rank assigned in the joint ranking.

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the general alternative

$$H_1 : [\tau_1, \ldots, \tau_k \text{ not all equal}],$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } H \geq h_\alpha; \text{ otherwise do not reject,} \tag{6.6}$$

where the constant h_α is chosen to make the type I error probability equal to α. The constant h_α is the upper α percentile for the null $(\tau_1 = \cdots = \tau_k)$ distribution of H. Comment 6 explains how to obtain the critical value h_α for k treatments and sample sizes $n_1, \ldots, n_k$ and available levels of α.

Large-Sample Approximation

When H_0 is true, the statistic H has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic chi-square (χ^2) distribution with $k-1$ degrees of freedom (see Comment 9 for indications of the proof). The chi-square approximation for procedure (6.6) is

$$\text{Reject } H_0 \text{ if } H \geq \chi^2_{k-1,\alpha}; \quad \text{otherwise do not reject,} \tag{6.7}$$

where $\chi^2_{k-1,\alpha}$ is the upper α percentile point of a chi-square distribution with $k-1$ degrees of freedom. To find $\chi^2_{k-1,\alpha}$, we use the R command `qchisq`$(1-\alpha, k-1)$. For example, to find $\chi^2_{4,.025}$, we apply `qchisq (.975, 4)` and obtain $\chi^2_{4,.025} = 11.143$.

Ties

If there are ties among the N X's, assign each of the observations in a tied group the average of the integer runks that are associated with the tied group and compute H with these average ranks. As a consequence of the effect that ties have on the null distribution of H, the following modification is needed to apply either procedure (6.6) or the large-sample approximation in procedure (6.7) when there are tied X's. In either of these procedures, we replace H by

$$H' = \frac{H}{1 - \left(\sum_{j=1}^{g} \left(t_j^3 - t_j\right) / [N^3 - N]\right)}, \tag{6.8}$$

where, in (6.8), H is computed using average ranks, g denotes the number of tied X groups, and t_j is the size of tied group j. We note that an united observation is considered to be a tied group of size 1. In particular, if there are no ties among the X's then $g = N$ and $t_j = 1$ for $j = 1, \ldots, N$. In this case, each term in (6.8) of he form $t_j^3 - t_j$ reduces to zero, the denominator of the right-hand side of expression (6.8)) reduces to 1, and H' (6.8) reduces to H, as given in (6.5).

We note that even the small-sample procedure (6.6) is only approximately, and not exactly, of the significance level α in the presence of tied X observations. To get an exact level α test in this tied setting, see Comment 8.

EXAMPLE 6.1 ***Half-Time of Mucociliary Clearance.***

Thomson and Short (1969) have assessed mucociliary efficiency from the rate of removal of dust in normal subjects, subjects with obstructive airway disease, and subjects with asbestosis. Table 6.1 is based on a subset of the Thomson–Short data. The joint ranks (r_{ij}'s) of the observations are given in Table 6.1 in parentheses after the data values and the treatment rank sums (R_1, R_2, and R_3) are provided at the bottom of the columns.

We are interested in using procedure (6.6) to test if there are any differences in median mucociliary clearance half-times for the three subject populations. For purpose of illustration, we take the significance level to be $\alpha = .0502$. Applying the R command `cKW`$(\alpha, \boldsymbol{n})$, we find `cKW(.0502, c(5, 4, 5), "Exact")` = 5.643; that is, $P_0(H \geq 5.643) = .0502$, and, in the notation of (6.6) with $k = 5, n_1 = 5, n_2 = 4$, and $n_3 = 5$, we have $h_{.0502} = 5.643$ and procedure (6.6) reduces to

$$\text{Reject } H_0 \text{ if } H \geq 5.643.$$

Table 6.1 Half-Time of Mucociliary Clearance (h)

	Subjects with	
Normal subjects	Obstructive airways disease	Asbestosis
2.9 (8)	3.8 (13)	2.8 (7)
3.0 (9)	2.7 (6)	3.4 (11)
2.5 (4)	4.0 (14)	3.7 (12)
2.6 (5)	2.4 (3)	2.2 (2)
3.2 (10)		2.0 (1)
$R_1 = 36$	$R_2 = 36$	$R_3 = 33$

Source: M. L. Thomson and M. D. Short (1969).

Now, we illustrate the computations leading to the sample value of H (6.5). For these data, we have $n_1 = n_3 = 5$, $n_2 = 4$, and $N = 14$. Combining these facts with the treatment rank sums in Table 6.1, we find from (6.5) that

$$H = \frac{12}{14(14+1)}\left(\frac{(36)^2}{5} + \frac{(36)^2}{4} + \frac{(33)^2}{5}\right) - 3(14+1) = .771.$$

As this value of H is less than the critical value 5.643, we do not reject H_0 at the $\alpha = .0502$ level. In fact, from the observed value $H = .771$, we see, using the R command `pKW(mucociliary,"Exact")`, that $P_0(H \geq .771) =$ `pKW(mucociliary,"Exact")` $= .7108$. Thus, the lowest significance level at which we can reject H_0 in favor of H_1 with the observed value of the test statistic $H = .771$ is .7108.

For the large-sample approximation, we compare the value of H (because there are no ties) to the chi-square distribution with $k - 1 = 2$ degrees of freedom. Using the R command `1 - pchisq(.771, 2)`, we find that the observed value of $H = .771$ is approximately the .68 upper percentile for the chi-square distribution with two degrees of freedom. Thus, the approximate P-value for these data and test procedure (6.7) is .68. Both the exact test and the large-sample approximation indicate that there is virtually no sample evidence in support of significant differences in mucociliary clearance half-times for the three subject populations.

Comments

1. *More General Setting.* We could replace Assumptions A1–A3 and H_0 (6.2) with the more general null hypothesis that all $N!/\left(\Pi_{j=1}^{k} n_j!\right)$ assignments of n_1 ranks to the treatment 1 observations, n_2 ranks to the treatment 2 observations, and $\ldots, n_k$ ranks to the treatment k observations are equally likely.
2. *Motivation for the Test.* Under Assumptions A1–A3 and H_0 (6.2), the rank vector $\mathbf{R}^* = (r_{11}, \ldots, r_{n_1 1}, r_{22}, \ldots, r_{n_2 2}, \ldots, r_{1k}, \ldots, r_{n_k k})$ has a uniform distribution over the set of all $N!$ permutations of the vector of integers $(1, 2, \ldots, N)$. It follows that

$$E_0(r_{ij}) = \frac{1}{N!}(N-1)! \sum_{i=1}^{N} i = \frac{N+1}{2},$$

the average rank being assigned in the joint ranking. Thus,

$$E_0(R_{.j}) = E_0\left(\frac{1}{n_j}\sum_{i=1}^{n_j} r_{ij}\right) = \frac{1}{n_j}\sum_{i=1}^{n_j} E_0(r_{ij})$$
$$= \frac{n_j(N+1)}{2n_j} = \frac{N+1}{2}, \text{ for } j = 1, 2, \ldots, k,$$

and we would expect the $R_{.j}$'s to be close to $(N+1)/2$ when H_0 is true. As the test statistic H (6.5) is a constant times a weighted sum of squared differences between the observed treatment average ranks, $R_{.j}$, and their null expected values, $E_0(R_{.j}) = (N+1)/2$, small values of H represent agreement with H_0 (6.2). When the τ's are not all equal, we would expect a portion of the associated treatment average ranks to differ from their common null expectation, $(N+1)/2$, with some tending to be larger and some smaller. The net result (after squaring the observed differences to obtain the $(R_{.j} - (N+1)/2)^2$ terms) would be a large value of H. This suggests rejecting H_0 in favor of H_1 (6.3) for large values of H and motivates procedures (6.6) and (6.7) (see also Comment 3).

3. *Connection to Normal Theory Test.* The Kruskal–Wallis test can also be motivated by considering the usual analysis of variance $\mathcal{F}$ statistic calculated using the ranks, rather than the original observations. The $\mathcal{F}$ statistic can be written as $\mathcal{F} = c(\text{SSB})/(\text{SST} - \text{SSB})$, where c is a constant depending only on the sample sizes, SST is the total sum of squares, and SSB is the between sum of squares. The statistic SSB reduces to $\sum_{j=1}^{k} n_j(R_{.j} - (N+1)/2)^2$ when applied to the ranks rather than the original observations and SST becomes a fixed constant when calculated on the ranks. Using these facts, it can be shown that when $\mathcal{F}$ is calculated for the ranks, $\mathcal{F}$ is an increasing function of H.

4. *Assumptions*. It is important to point out that Assumption A3 stipulates that the k treatment distributions $F_1, \ldots, F_k$ can differ at most in their locations (medians). In particular, Assumption A3 requires that the k underlying distributions belong to the same general family (F) and that they do not differ in scale parameters (variability). (For a discussion of methodology designed for a more general setting where differences in scale parameters are permitted, see Comment 11.)

5. *Special Case of Two Treatments.* For the case of $k = 2$ treatments, the procedures in (6.6) and (6.7) are equivalent to the exact and large-sample approximation forms, respectively, of the two-sided Wilcoxon rank sum test, as discussed in Section 4.1.

6. *Derivation of the Distribution of H under H_0 (No-Ties Case).* The null distribution of H (6.5) can be obtained using the fact that under H_0 (6.2), all $N!/\left(\Pi_{j=1}^{k} n_j!\right)$ assignments of n_1 ranks to the treatment 1 observations, n_2 ranks to the treatment 2 observations, and $\ldots, n_k$ ranks to the treatment k observations are equally likely. We illustrate how the null distribution can be derived in the particular case $k = 3$, $n_1 = n_2 = n_3 = 2$. In this case, we have $H = [\{12/[6(7)]\}\{(R_1^2 + R_2^2 + R_3^2)/2\} - 21] = [(A/7) - 21]$, where $A = R_1^2 + R_2^2 + R_3^2$. We next enumerate 15 of the total possible $\{6!/[(2!)(2!)(2!)]\} = 90$ rank assignments and their corresponding values of A and H.

(a)	I	II	III		(b)	I	II	III	
	1	3	5	$A = 179$		1	3	4	$A = 173$
	2	4	6	$H = 4.57$		2	5	6	$H = 3.71$
(c)	I	II	III		(d)	I	II	III	
	1	3	4	$A = 171$		1	2	5	$A = 173$
	2	6	5	$H = 3.43$		3	4	6	$H = 3.71$
(e)	I	II	III		(f)	I	II	III	
	1	2	4	$A = 165$		1	2	4	$A = 161$
	3	5	6	$H = 2.57$		3	6	5	$H = 2$
(g)	I	II	III		(h)	I	II	III	
	1	2	3	$A = 155$		1	2	5	$A = 171$
	4	5	6	$H = 1.14$		4	3	6	$H = 3.43$
(i)	I	II	III		(j)	I	II	III	
	1	2	3	$A = 153$		1	2	4	$A = 161$
	4	6	5	$H = .86$		5	3	6	$H = 2$
(k)	I	II	III		(l)	I	II	III	
	1	2	3	$A = 153$		1	2	3	$A = 149$
	5	4	6	$H = .86$		5	6	4	$H = .29$
(m)	I	II	III		(n)	I	II	III	
	1	2	4	$A = 155$		1	2	3	$A = 149$
	6	3	5	$H = 1.14$		6	4	5	$H = .29$
(o)	I	II	III						
	1	2	4	$A = 147$					
	6	5	3	$H = 0$					

For each of the foregoing rank configurations, there are five other configurations (corresponding to the six permutations of the names of the samples I, II, and III), which yield the same value of H. This covers the complete total of 90 possible rank assignments. Thus,

$$P_0\{H = 4.57\} = 1/15, \quad P_0\{H = 3.71\} = 2/15, \quad P_0\{H = 3.43\} = 2/15,$$

$$P_0\{H = 2.57\} = 1/15, \quad P_0\{H = 2\} = 2/15, \quad P_0\{H = 1.14\} = 2/15,$$

$$P_0\{H = .86\} = 2/15, \quad P_0\{H = .29\} = 2/15, \quad P_0\{H = 0\} = 1/15.$$

The probability, under H_0, that H is greater than or equal to 3.71, for example, is therefore

$$\begin{aligned} P_0\{H \geq 3.71\} &= P_0\{H = 3.71\} + P_0\{H = 4.57\} \\ &= \tfrac{1}{15} + \tfrac{2}{15} = .20. \end{aligned}$$

Note that we have derived the null distribution of H without specifying the common from (F) of the underlying distribution function for the X's under H_0 beyond the point of requiring that it be continuous. This is why the test procedure (6.6) based on H is called a *distribution-free procedure*. From the null distribution of H, we can determine the critical value h_α and control the probability α of

falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying continuous X distribution.

For a given number of treatments k and sample sizes $n_1, \ldots, n_k$, the R command $\texttt{cKW}(\alpha, \boldsymbol{n})$ can be used to find the available upper-tail critical values h_α for possible values of H. For a given available significance level α, the critical value h_α then corresponds to $P_0(H \geq h_\alpha) = \alpha$ and is given by $\texttt{cKW}(\alpha, \boldsymbol{n})$. Thus, for example, for $k = 5,\ n_1 = 3,\ n_2 = 2,\ n_3 = 3,\ n_4 = 2$, and $n_5 = 3$, we have $P_0(H \geq 8.044) = .0492$, so that $h_{.0492} = 8.044$ for $k = 5,\ n_1 = 3,\ n_2 = 2,\ n_3 = 3,\ n_4 = 2$, and $n_5 = 3$.

7. *Exact Conditional Distribution of H with Ties among the X-Values.* To have a test with the exact significance level even in the presence of tied X's, we need to consider all $N!/\left(\prod_{j=1}^{k} n_j!\right)$ assignments of n_1 ranks to the treatment 1 observations, n_2 ranks to the treatment 2 observations, $\ldots, n_k$ ranks to the treatment k observations, where now these joint ranks are obtained by using average ranks to break the ties. As in Comment 6, it still follows that under H_0 each of these $N!/\left(\prod_{j=1}^{k} n_j!\right)$ assignments is equally likely. For each such assignment, the value of H is computed and the results are tabulated. We illustrate this construction for $k = 3$ and $n_1 = n_2 = 2,\ n_3 = 1$ and the data $X_{11} = 1.3, X_{21} = 1.7, X_{12} = 1.3, X_{22} = 2.0$, and $X_{13} = 2.0$. Using average ranks to break the ties, the observed rank vector is $(r_{11}, r_{21}, r_{12}, r_{22}, r_{13}) = (1.5, 3, 1.5, 4.5, 4.5)$. Thus, $R_1 = 4.5,\ R_2 = 6,\ R_3 = 4.5$, and the attained value of H is

$$H = \left[\frac{12}{5(6)}\left\{\frac{(4.5)^2}{2} + \frac{(6)^2}{2} + \frac{(4.5)^2}{1}\right\} - 3(6)\right] = 1.35.$$

To assess the significance of H, we obtain its conditional null distribution by considering the [5!/(2! 2! 1!)] = 30 equally likely (under H_0) possible assignments of the observed rank vector (1.5, 3, 1.5, 4.5, 4.5) to the three treatments. These 30 assignments and associated values of H are in the following table

I	II	III		I	I	III	
1.5	4.5	1.5		1.5	4.5	1.5	
3	4.5		$H = 3.15$	3	4.5		$H = 3.15$
1.5	3	1.5		1.5	3	1.5	
4.5	4.5		$H = 1.35$	4.5	4.5		$H = 1.35$
1.5	3	1.5		1.5	3	1.5	
4.5	4.5		$H = 1.35$	4.5	4.5		$H = 1.35$
3	1.5	1.5		3	1.5	1.5	
4.5	4.5		$H = 1.35$	4.5	4.5		$H = 1.35$
3	1.5	1.5		3	1.5	1.5	
4.5	4.5		$H = 1.35$	4.5	4.5		$H = 1.35$
4.5	1.5	1.5		4.5	1.5	1.5	
4.5	3		$H = 3.15$	4.5	3		$H = 3.15$
1.5	4.5	3		1.5	1.5	3	
1.5	4.5		$H = 3.60$	4.5	4.5		$H = 0$
1.5	1.5	3		1.5	1.5	3	
4.5	4.5		$H = 0$	4.5	4.5		$H = 0$
1.5	1.5	3		4.5	1.5	3	

I	II	III		I	I	III	
4.5	4.5		$H = 0$	4.5	1.5		$H = 3.60$
1.5	3	4.5		1.5	3	4.5	
1.5	4.5		$H = 3.15$	1.5	4.5		$H = 3.15$
1.5	1.5	4.5		1.5	1.5	4.5	
3	4.5		$H = 1.35$	3	4.5		$H = 1.35$
1.5	1.5	4.5		1.5	1.5	4.5	
3	4.5		$H = 1.35$	3	4.5		$H = 1.35$
1.5	1.5	4.5		1.5	1.5	4.5	
4.5	3		$H = 1.35$	4.5	3		$H = 1.35$
1.5	1.5	4.5		1.5	1.5	4.5	
4.5	3		$H = 1.35$	4.5	3		$H = 1.35$
3	1.5	4.5		3	1.5	4.5	
4.5	1.5		$H = 3.15$	4.5	1.5		$H = 3.15$

As each of these values for H has null probability $\frac{1}{30}$, it follows that

$$P_0(H = 3.60) = \tfrac{2}{30} \qquad P_0(H = 1.35) = \tfrac{16}{30}$$
$$P_0(H = 3.15) = \tfrac{8}{30} \qquad P_0(H = 0) = \tfrac{4}{30}.$$

This distribution is called the *conditional distribution* or the *permutation distribution* of H, given the set of tied ranks {1.5, 1.5, 3, 4.5, and 4.5}. For the particular observed value $H = 1.35$, we have $P_0(H \geq 1.35) = \frac{28}{30}$, so that such a value does not indicate a deviation from H_0.

8. *Large-Sample Approximation*. Define the random variables $T_j = R_{\cdot j} - E_0(R_{\cdot j}) = R_{\cdot j} - (N + 1)/2$, for $j = 1, 2, \ldots, k$. As each $R_{\cdot j} = \sum_{i=1}^{n_j} r_{ij}/n_j$ is an average, it is not surprising (see Kruskal and Wallis (1952), e.g., for justification) that a properly standardized version of the vector $\mathbf{T}^* = (T_1, \ldots, T_{k-1})$ has an asymptotic ($\min(n_1, \ldots, n_k)$ tending to infinity) $(k - 1)$-variate normal distribution with mean vector $\mathbf{0} = (0, \ldots, 0)$ and appropriate covariance matrix $\boldsymbol{\Sigma}$ when the null hypothesis H_0 is true. (Note that $\mathbf{T}^*$ does not include $T_k = R_{\cdot k} - (N + 1)/2$, because T_k can be expressed as a linear combination of $T_1, \ldots, T_{k-1}$. This is the reason that the asymptotic normal distribution is $(k - 1)$-variate and not k-variate.) As the test statistic H (6.5) is a quadratic form in the variables $(T_1, \ldots, T_{k-1})$, it is therefore quite natural that H has an asymptotic ($\min(n_1, \ldots, n_k)$ tending to infinity) chi-square distribution with $k - 1$ degrees of freedom.
9. *Family Monotonicity*. Gabriel (1969) introduced a desirable property of a testing family called *monotonicity* and pointed out that the H statistic does not enjoy the property. We refer the interested user to Gabriel's paper, but we briefly mention here that the problem arises because it is possible that the H statistic computed for a subset can exceed the H statistic computed for a set containing the subset. Gabriel gave the following example. The sample 1 ranks are 8, 9, 10, and 11, the sample 2 ranks are 1, 2, 6, and 7, and the sample 3 ranks are 3, 4, 5, and12. Then H based on samples 1 and 2($k = 2$) is 5.33, whereas H based on samples 1, 2, and 3 ($k = 3$) is 4.77. The same anomaly can arise with the Friedman statistic (Section 7.1).

10. *k-Sample Behrens–Fisher Problem*. Two of the implicit requirements associated with Assumptions A1–A3 are that the underlying distributions belong to the same common family (F) and that they differ within this family at most in their medians. The less restrictive setting, where these assumptions are relaxed to permit the possibility of differences in scale parameters as well as medians (but still requiring the same common family F), is generally referred to as the *k-sample Behrens–Fisher problem*. (Note that this is a direct k-sample extension of the corresponding two-sample Behrens–Fisher problem considered in Section 4.4.) The Kruskal–Wallis procedure (6.6) is no longer distribution-free under these relaxed assumptions permitting unequal scale parameters. Rust and Fligner (1984) proposed a modification of the Kruskal–Wallis statistic H (6.5) to deal with this broader Behrens–Fisher setting. Their procedure is designed as a test for the less restrictive null and alternative hypotheses

$$H_0^* : [\delta_{ij} = \tfrac{1}{2} \text{ for all } i \neq j = 1, \ldots, k] \tag{6.9}$$

and

$$H_1^* : [\delta_{ij} \neq \tfrac{1}{2} \text{ for at least one } i \neq j = 1, \ldots, k], \tag{6.10}$$

respectively, where

$$\delta_{ij} = P(X_{1i} > X_{1j}), \quad \text{for } i \neq j = 1, \ldots, k.$$

The Rust–Fligner modification of the Kruskal–Wallis statistic provides a test procedure that is still exactly distribution-free under the more restrictive null hypothesis H_0 (6.2). However, their modified procedure is also asymptotically ($\min(n_1, \ldots, n_k)$ tending to infinity) distribution-free under the considerably broader null hypothesis H_0^* (6.9) so long as the underlying populations (not necessarily of the same form) are all symmetric. In the special case of $k = 2$ populations, the Rust–Fligner procedure reduces approximately to the Fligner–Policello modifications to the Mann–Whitney–Wilcoxon two-sample test procedure discussed in Section 4.4.

11. *Pairwise Rankings*. The Kruskal–Wallis statistic H (6.5) is based on the treatment rank sums $R_1, \ldots, R_k$ associated with the *joint* ranking of all N sample observations. As an alternative approach, one could just as well choose to compare the k treatments through a combination of all $k(k-1)/2$ *pairwise* rankings. Fligner (1985) proposed such a pairwise ranking analog of the Kruskal–Wallis statistic and demonstrated that the associated pairwise ranking test procedure has some nice efficiency properties. Such pairwise rankings (as opposed to joint rankings) have also proved useful in certain multiple comparison settings (see Sections 6.5 and 6.10 for more in this regard).

12. *Consistency of the H Test*. Replace Assumptions A1–A3 by the less restrictive Assumptions A1′: the X's are mutually independent and A2′ : $X_{1j}, \ldots, X_{n_j j}$ come from the same continuous population $\Pi_j, j = 1, \ldots, k$, but where $\Pi_1, \ldots, \Pi_k$ are not assumed to be identical. Then Kruskal and Wallis (1952) pointed out that (roughly speaking) the test defined by (6.6) is consistent if (and only if) "... there be at least one of the populations for which the limiting probability is not one-half that a random observation from this population is greater than an independent random member of the N sample observations."

Properties

1. *Consistency*. Under Assumptions A1–A3 and equal sample sizes ($n_1 = \cdots = n_k$), the test defined by (6.6) is consistent against the alternative for which $\tau_i \neq \tau_j$ for at least one $i \neq j = 1, \ldots, k$. For arbitrary sample sizes, see Kruskal (1952) and Comment 12.
2. *Asymptotic Chi-Squareness*. See Kruskal and Wallis (1952) and Hettmansperger (1984, pp. 184–185).
3. *Efficiency*. See Andrews (1954), Hodges and Lehmann (1956), and Section 6.10.

Problems

1. Pretherapy training of clients has been shown to have beneficial effects on the process and outcome of counseling and psychotherapy. Sauber (1971) investigated four different approaches to pretherapy training:

1. Control (no treatment).
2. Therapeutic reading (TR) (indirect learning).
3. Vicarious therapy pretraining (VTP) (videotaped, vicarious learning).
4. Group, role induction interview (RII) (direct learning).

Treatment conditions 2–4 were expected to enhance the outcome of counseling and psychotherapy as compared with a control group, those subjects receiving no prior set of structuring procedures. One of the major variables of the study was that of "psychotherapeutic attraction." The basic data in Table 6.2 consist of the raw scores for this measure according to each of the four experimental conditions. Apply procedure (6.7), with the correction for ties given by (6.8).

2. Show that the two expressions for H in (6.5) are indeed equivalent.

3. Show directly, or illustrate by means of an example, that the maximum value of H is $H_{\max} = \{N^3 - \sum_{j=1}^{k} n_j^3\}/\{N(N+1)\}$. For what rank configurations is this maximum achieved?

4. To determine the number of game fish to stock in a given system and to set appropriate catch limits, it is important for fishery managers to be able to assess potential growth and survival of game fish in that system. Such growth and survival rates are closely related to the availability of appropriately sized prey. Young-of-year (YOY) gizzard shad (*Dorosoma cepedianum*) are the primary food source for game fish in many Ohio environments. However, because of their fast growth rate, YOY gizzard shad can quickly become too large for predators to swallow.

Table 6.2 Raw Scores Indicating the Degree of Psychotherapeutic Attraction for Each Experimental Condition

Control	Reading (TR)	Videotape (VTP)	Group (RII)
0	0	0	1
1	6	5	5
3	7	8	12
3	9	9	13
5	11	11	19
10	13	13	22
13	20	16	25
17	20	17	27
26	24	20	29

Source: S. R. Sauber (1971).

Table 6.3 Length of YOY Gizzard Shad from Kokosing Lake, Ohio, Sampled in Summer, 1984 (mm)

Site I	Site II	Site III	Site IV
46	42	38	31
28	60	33	30
46	32	26	27
37	42	25	29
32	45	28	30
41	58	28	25
42	27	26	25
45	51	27	24
38	42	27	27
44	52	27	30

Source: B. Johnson (1984).

Thus to be able to predict predator growth rates in such settings, it is useful to know both the density and the size structure of the resident YOY shad populations. With this in mind, Johnson (1984) sampled the YOY gizzard shad population at four different sites in Kokosing Lake (Ohio) in summer 1984. The data in Table 6.3 are lengths (mm) for a subset of the YOY gizzard shad sampled by Johnson.

Apply procedure (6.7), with the correction for ties given in (6.8), to assess whether there are any differences between the median lengths for the YOY gizzard shad populations in the four Kokosing Lake sites.

5. Suppose $k = 3$ and $n_1 = 2$, $n_2 = n_3 = 6$. Compare the critical region for the exact level $\alpha = .050$ test of H_0 (6.2) based on H with the critical region for the corresponding nominal level $\alpha = .050$ test based on the large-sample approximation. What is the exact significance level of this .050 nominal level test based on the large-sample approximation?
6. Suppose $k = 4$, $n_1 = n_2 = n_3 = 1$, and $n_4 = 2$. Obtain the form of the exact null (H_0) distribution of H for the case of no tied X observations.
7. Suppose $k = 3$, $n_1 = n_2 = n_3 = 2$, and we observe the data $X_{11} = 2.7$, $X_{21} = 3.4$, $X_{12} = 2.7$, $X_{22} = 4.5$, $X_{13} = 4.9$, and $X_{23} = 2.7$. What is the conditional probability distribution of H under H_0 (6.2) when average ranks are used to break the ties among the X's? How extreme is the observed value of H in this conditional null distribution? Compare this fact with that obtained by taking the observed value of H to the (incorrect) unconditional null distribution of H.
8. Leukemia is a disease characterized by proliferation of the white blood cells or leukocytes. One form of chemotherapy used in the treatment of leukemia involves the administration of corticosteroids. Some researchers suggested that forms of leukemia characterized by leukocytes with a large number of glucocorticoid receptor (GR) sites per cell are more effectively controlled by corticosteroids. Other researchers questioned this relationship. In an effort to aid in the resolution of this controversy, Kontula et al. (1980) developed a method for determining more accurately the number of GR sites per cell. In this research and later work by Kontula et al. (1982), this new methodology was used to count the number of GR sites for samples of leukocyte cells from normal subjects, as well as patients with hairy-cell leukemia, chronic lymphatic leukemia, chronic myelocytic leukemia, or acute leukemia. The data in Table 6.4 are a subset of the data considered by the authors in these two publications.

   Use these data to assess whether there are any differences between the median numbers of GR sites per leukocyte cell for the population of normal subjects and the populations of patients with hairy-cell leukemia, chronic lymphatic leukemia, chronic myelocytic leukemia, or acute leukemia.

Table 6.4 Number of Glucocorticoid Receptor (GR) Sites per Leukocyte Cell

Normal subjects	Hairy-cell anemia	Chronic lymphatic leukemia	Chronic myelocytic leukemia	Acute leukemia
3,500	5,710	2,930	6,320	3,230
3,500	6,110	3,330	6,860	3,880
3,500	8,060	3,580	11,400	7,640
4,000	8,080	3,880	14,000	7,890
4,000	11,400	4,280		8,280
4,000		5,120		16,200
4,300				18,250
4,500				29,900
4,500				
4,900				
5,200				
6,000				
6,750				
8,000				

Source: K. Kontula, L. C. Andersson, T. Paavonen, G. Myllyla, L. Teerenhovi, and P. Vuopio (1980) and K. Kontula, T. Paavonen, R. Vuopio, and L. C. Andersson (1982).

9. Generate the conditional permutation distribution of H *using only the last two sample lengths from each of the four sites for the gizzard shad data in Table 6.3.* From this conditional permutation distribution of H, obtain the exact conditional P-value for a test of H_0 (6.2) versus H_1 (6.3) with this subset of data from Table 6.3. Compare this exact conditional P-value with the approximate P-value associated with taking the observed value of H to the unadjusted (for ties) unconditional null distribution of H.
10. Habitat plays an important role in fish behavior, particularly feeding, spawning, and protection/security. One of the modern methods of fisheries management is habitat modification in large, constructed reservoirs. Previous studies have shown that the type of structure introduced is an important factor in such habitat modifications. Of particular relevance in many settings is the size openings or *interstices* in the introduced structure. The data in Table 6.5 represent a subset of that obtained by Kayle (1984) from Alum Creek Lake in Westerville, Ohio, in a study to determine the relative effectiveness of three species of pine trees for habitat modification.

Table 6.5 Mean Interstitial Lengths (mm)

Scotch pine	Blue spruce	White pine
52.2	46.7	75.2
56.4	60.5	63.7
57.1	58.9	73.2
46.9	82.9	66.2
49.1	65.8	67.4
52.5	93.3	69.4
63.0	66.9	70.4
52.0	70.9	72.3
61.1	73.7	63.6
55.3	65.8	61.9
46.2	90.2	74.4
57.2	68.9	70.1

Source: K. A. Kayle (1984).

The measurements in Table 6.5 are averages (mm) of interstitial lengths (distances between midpoints) of 10 pairs of secondary branches for each of 12 scotch pine, 12 blue spruce, and 12 white pine trees. Use an appropriate procedure to test whether there are any differences in median interstitial lengths between secondary branches for the three studied species of pine.

6.2 A DISTRIBUTION-FREE TEST FOR ORDERED ALTERNATIVES (JONCKHEERE–TERPSTRA)

In many practical settings, the treatments are such that the appropriate alternatives to no differences in treatment effects (H_0) are those of increasing (or decreasing) treatment effects according to some natural labeling for the treatments. Examples of such settings include "treatments" corresponding to degrees of knowledge of performance, quality or quantity of materials, severity of disease, amount of practice, drug dosage levels, intensity of a stimulus, and temperature. We note that the Kruskal–Wallis procedure (6.6) does not utilize any such partial prior information regarding a postulated alternative ordering. The statistic H (6.5) takes on the same value for all $k!$ possible labelings of the treatments. In this section, we consider a procedure for testing H_0 (6.2) against the a priori ordered alternatives

$$H_2 : [\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k, \text{ with at least one strict inequality}]. \tag{6.11}$$

The Jonckheere (1954a, 1954b) and Terpstra (1952) test of this section is preferred to the Kruskal–Wallis test in Section 6.1 when the treatments can be labeled a priori in such a way that the experimenter expects any deviation from H_0 (6.2) to be in the particular direction associated with H_2 (6.11). We emphasize, however, that the labeling of the treatments so that the ordered alternatives (6.11) are appropriate *cannot* depend on the observed sample observations. This labeling must correspond completely to a factor(s) implicit in the nature of the *experimental design* and *not* the *observed data.*

Procedure

First, we must label the treatments so that they are in the expected order associated with the alternative H_2 (6.11). (This labeling must be done prior to data collection.) To compute the Jonckheere–Terpstra statistic, J, we calculate the $k(k-1)/2$ Mann–Whitney (see Comment 4.7) counts U_{uv} given by

$$U_{uv} = \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \phi(X_{iu}, X_{jv}), \quad 1 \leq u < v \leq k, \tag{6.12}$$

where $\phi(a,b) = 1$ if $a < b, 0$ otherwise. (Thus, U_{uv} is the number of sample u before sample v precedences.) The Jonckheere–Terpstra statistic, J, is then the sum of these $k(k-1)/2$ Mann–Whitney counts,

$$J = \sum_{u=1}^{v-1} \sum_{v=2}^{k} U_{uv}. \tag{6.13}$$

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the ordered alternative

$$H_2 : [\tau_1 \le \tau_2 \le \cdots \le \tau_k, \text{ with at least one strict inequality}],$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } J \ge j_\alpha; \text{ otherwise do not reject,} \tag{6.14}$$

where the constant j_α is chosen to make the type I error probability equal to α. The constant j_α is the upper α percentile for the null ($\tau_1 = \cdots = \tau_k$) distribution of J. Comment 17 explains how to obtain the critical value j_α for k treatments and sample sizes $n_1, \ldots, n_k$ and available levels of α.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic ($\min(n_1, n_2, \ldots, n_k)$ tending to infinity) normality of J, suitably standardized. We first need to know the expected value and variance of J when the null hypothesis is true. Under H_0, the expected value and variance of J are:

$$E_0(J) = \frac{N^2 - \sum_{j=1}^k n_j^2}{4} \tag{6.15}$$

and

$$\text{var}_0(J) = \frac{N^2(2N+3) - \sum_{j=1}^k n_j^2(2n_j+3)}{72}, \tag{6.16}$$

respectively. These expressions for $E_0(J)$ and $\text{var}_0(J)$ are verified by direct calculations in Comment 18 for the special case of $k = 3,\ n_1 = n_2 = 1,\ n_3 = 2$. General derivations of both expressions are outlined in Comment 19.

The standardized version of J is

$$J^* = \frac{J - E_0(J)}{\sqrt{\text{var}_0(J)}} = \frac{J - \left[\frac{N^2 - \sum_{j=1}^k n_j^2}{4}\right]}{\left\{\left[N^2(2N+3) - \sum_{j=1}^k n_j^2(2n_j+3)\right] \Big/ 72\right\}^{1/2}}. \tag{6.17}$$

When H_0 is true, J^* has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic $N(0,1)$ distribution (see Comment 19 for indications of the proof). The normal theory approximation for procedure (6.14) is

$$\text{Reject } H_0 \text{ if } J^* \ge z_\alpha; \quad \text{otherwise do not reject.} \tag{6.18}$$

Ties

If there are ties among the N X's, replace $\phi(a,b)$ in the calculation of the Mann–Whitney counts U_{uv} by $\phi^*(a,b) = 1, \frac{1}{2}, 0$ if $a <, =,$ or $> b$, respectively, so that for each between-sample comparison where there is a tie, the contribution to the appropriate Mann–Whitney count will be $\frac{1}{2}$. After computing J with these modified Mann–Whitney

counts, use procedure (6.14). Note, however, that this test associated with tied X's is only approximately, and not exactly, of the significance level α.

When applying the large-sample approximation, an additional factor must be taken into account. Although ties in the X's do not affect the null expected value of J, its null variance is reduced to

$$\begin{aligned}\mathrm{var}_0(J) &= \left\{ \frac{1}{72}\left[N(N-1)(2N+5) - \sum_{i=1}^{k} n_i(n_i-1)(2n_i+5) - \sum_{j=1}^{g} t_j(t_j-1)(2t_j+5) \right] \right. \\ &\quad + \frac{1}{36N(N-1)(N-2)}\left[\sum_{i=1}^{k} n_i(n_i-1)(n_i-2) \right]\left[\sum_{j=1}^{g} t_j(t_j-1)(t_j-2) \right] \\ &\quad \left. + \frac{1}{8N(N-1)}\left[\sum_{i=1}^{k} n_i(n_i-1) \right]\left[\sum_{j=1}^{g} t_j(t_j-1) \right] \right\}, \end{aligned} \tag{6.19}$$

where, in (6.19), g denotes the number of tied X groups and t_j is the size of tied group j. We note that an untied observation is considered to be a tied group of size 1. In particular, if there are no ties among the X's, then $g = N$ and $t_j = 1$, for $j = 1, \ldots, N$. In this case, each term in (6.19) that involves the factor $(t_j - 1)$ reduces to zero and (as you are asked to show in Problem 19) the variance expression in (6.19) reduces to the usual null variance of J when there are no ties, as given previously in (6.16).

As a consequence of the effect that ties have on the null variance of J, the following modification is needed to apply the large-sample approximation when there are tied X's. Compute J using the modified Mann–Whitney counts and set

$$J^* = \frac{J - \left[\frac{N^2 - \sum_{j=1}^{k} n_j^2}{4} \right]}{\{\mathrm{var}_0(J)\}^{1/2}}, \tag{6.20}$$

where $\mathrm{var}_0(J)$ is now given by display (6.19). With this modified value of J^*, the approximation (6.18) can be applied.

EXAMPLE 6.2 ***Motivational Effect of Knowledge of Performance.***

Hundal (1969) described a study designed to assess the purely motivational effects of knowledge of performance in a repetitive industrial task. The task was to grind a metallic piece to a specified size and shape. Eighteen male workers were divided randomly into three groups. The subjects in the control group, A, received no information about their output, subjects in group B were given a rough estimate of their output, and subjects in group C were given an accurate information about their output and could check it further by referring to a figure that was placed before them. The basic data in Table 6.6 consist of the numbers of pieces processed by each subject in the experimental period.

We apply the Jonckheere–Terpstra test with the notion that a deviation from H_0 is likely to be in the direction of increased output with increased degree of knowledge of

Table 6.6 Number of Pieces Processed

Control (no information)	Group B (rough information)	Group C (accurate information)
40 (5.5)[a]	38 (2.5)	48 (18)
35 (1)	40 (5.5)	40 (5.5)
38 (2.5)	47 (17)	45 (15)
43 (10.5)	44 (13)	43 (10.5)
44 (13)	40 (5.5)	46 (16)
41 (8)	42 (9)	44 (13)

Source: P. S. Hundal (1969).

[a] Although we do not need to perform the joint ranking to compute the Jonckheere–Terpstra statistic, we give these ranks here for use in Sections 6.4 and 6.7.

performance. Thus, we are interested in using procedure (6.14) with the treatment labels 1 $\equiv$ control (no information), 2 $\equiv$ group B (rough information), and 3 $\equiv$ group C (accurate information). For purpose of illustration, we take the significance level to be $\alpha = .0490$. Applying the R command `cJCK`$(\alpha, \boldsymbol{n})$, we find `cJCK(.0490,c(6,6,6))` $= 75$; that is, $P_0(J \geq 75) = .0490$, and, in the notation of (6.14) with $k = 3$, $n_1 = n_2 = n_3 = 6$, we have $j_{.0490} = 75$, and procedure (6.14) reduces to

$$\text{Reject } H_0 \text{ if } J \geq 75.$$

We now illustrate the computations leading to the sample value of J (6.13). As there are ties in the sample data, we use $\phi^*(a, b) = 1, \frac{1}{2}, 0$ if $a <, =,$ or $> b$, respectively, to compute the $3(2)/2 = 3$ Mann–Whitney counts. We obtain

$$U_{12} = 1.5 + 2.5 + 6 + 5.5 + 2.5 + 4 = 22,$$

$$U_{13} = 6 + 2.5 + 6 + 4.5 + 6 + 5.5 = 30.5,$$

and

$$U_{23} = 6 + 2 + 5 + 4 + 5 + 4.5 = 26.5.$$

From (6.13), it follows that

$$J = 22 + 30.5 + 26.5 = 79.$$

As this value of J is greater than the critical value 75, we reject H_0 at the .0490 level. In fact, from the observed value $J = 79$, we see that the R command `pJCK(motivational.effect)` that $P_0(J \geq 79) =$ `pJCK(motivational.effect)` $= .0231$. Thus, the lowest significance level at which we can reject H_0 in favor of H_2 with the observed value of $J = 79$ is the P-value .0231.

For the large-sample approximation, we need to compute the standardized form of J^* using (6.19) and (6.20), because there are ties in the data. The null expected value for J is $E_0(J) = [(18)^2 - (6^2 + 6^2 + 6^2)]/4 = 54$. For the ties-corrected null variance of J, we note that $g = 11$ and $t_1 = 1, t_2 = 2, t_3 = 4, t_4 = 1, t_5 = 1, t_6 = 2, t_7 = 3, t_8 = 1, t_9 = 1,$

$t_{10} = 1, t_{11} = 1$, for the Hundal data. Hence, using the ties correction in (6.19), we have

$$\begin{aligned}\text{var}_0(J) = \Bigg\{ &\frac{1}{72}[18(17)(41) - 3(6)(5)(17) - 2(2)(1)(9) - 3(2)(11) - 4(3)(13)] \\ &+ \frac{1}{36(18)(17)(16)}[3(6)(5)(4)][3(2)(1) + 4(3)(2)] \\ &+ \frac{1}{8(18)(17)}[3(6)(5)][2(2)(1) + 1(3)(2) + 1(4)(3)]\Bigg\} = 150.29,\end{aligned}$$

from which it follows that the ties-corrected value of J^* (6.20) is

$$J^* = \frac{79 - 54}{\{150.29\}^{1/2}} = 2.04.$$

Thus, using the approximate procedure (6.18) with the ties-corrected value of $J^* = 2.04$ and the R command `pnorm(·)`, we see that the approximate P-value for these data is $P_0(J^* \geq 2.04) \approx 1 - \texttt{pnorm}(2.04) = .0207$. Both the exact test and the large-sample approximation indicate that strong evidence in support of increased output with increase in degree of knowledge of performance for the task considered by Hundal.

Comments

13. *More General Setting*. As with the Kruskal–Wallis procedure in Section 6.1, we could replace Assumptions A1–A3 and H_0 (6.2) with the more general null hypothesis that all $N!/\left(\prod_{j=1}^{k} n_j!\right)$ assignments of n_1 joint ranks to the treatment 1 observations, n_2 joint ranks to the treatment 2 observations, $\ldots, n_k$ joint ranks to the treatment k observations are equally likely.
14. *Motivation for the Test*. Consider J (6.13) and note that the term $\sum_{u=1}^{v-1}\sum_{v=2}^{k} U_{uv}$ takes the postulated ordering into account. Consider, for simplicity, the case $k = 3$. Then $\sum_{u=1}^{v-1}\sum_{v=2}^{3} U_{uv} = U_{12} + U_{13} + U_{23}$ and if $\tau_1 < \tau_2 < \tau_3, U_{12}$ would tend to be larger than $n_1 n_2/2$ (its null expectation); U_{13} would tend to be larger than $n_1 n_3/2$; U_{23} would tend to be larger than $n_2 n_3/2$; and, consequently, $J = U_{12} + U_{13} + U_{23}$ would tend to be larger than its null expectation $(n_1 n_2 + n_1 n_3 + n_2 n_3)/2 = \{[N^2 - (n_1^2 + n_2^2 + n_3^2)]/4\}$. This serves as partial motivation for the J test.
15. *Assumptions*. It is once again (as with the Kruskal–Wallis procedure in Section 6.1) important to point out that Assumption A3 stipulates that the k treatment distributions $F_1, \ldots, F_k$ can differ at most in their locations (medians) (see also Comment 4).
16. *Special Case of Two Treatments*. When there are only two treatments, the procedures in (6.14) and (6.18) are equivalent to the exact and large-sample approximation forms, respectively, of the one-sided upper-tail Wilcoxon rank sum test, as discussed in Section 4.1.
17. *Derivation of the Distribution of J under H_0 (No-Ties Case)*. A little thought will convince the reader that J can be computed from the joint ranking of all $N = \sum_{j=1}^{k} n_j$ observations. That is, although we do not need to perform this

joint ranking in order to compute J, given the ranking, we can, without the knowledge of the actual X_{ij} values, retrieve the value of J. Thus, one way to obtain the null distribution of J is to follow the method of Comment 6; namely, use the fact that under H_0 (6.2) all $N!/\left(\Pi_{j=1}^{k} n_j!\right)$ rank assignments are equally likely, and compute the associated value of J for each possible ranking. Consider how this would work in the small-sample-size case of $k = 3, n_1 = 1, n_2 = 1$, and $n_3 = 2$. The $4!/[1!\ 1!\ 2!] = 12$ possible assignments of the joint ranks 1, 2, 3, and 4 to the three treatments and their associated values of J (6.13) are as follows:

(a)	I	II	III	(b)	I	II	III	(c)	I	II	III	
	1	2	3		2	1	3		1	3	2	
			4				4				4	
	$J = 5$				$J = 4$				$J = 4$			
(d)	I	II	III	(e)	I	II	III	(f)	I	II	III	
	3	1	2		1	4	2		4	1	2	
			4				3				3	
	$J = 3$				$J = 3$				$J = 2$			
(g)	I	II	III	(h)	I	II	III	(i)	I	II	III	
	2	3	1		3	2	1		2	4	1	
			4				4				3	
	$J = 3$				$J = 2$				$J = 2$			
(j)	I	II	III	(k)	I	II	III	(l)	I	II	III	
	4	2	1		3	4	1		4	3	1	
			3				2				2	
	$J = 1$				$J = 1$				$J = 0$			

Thus, the null distribution for J for $n_1 = 1, n_2 = 1, n_3 = 2$, and $k = 3$ is given by

$$P_0\{J = 0\} = \tfrac{1}{12}, \quad P_0\{J = 1\} = \tfrac{2}{12}, \quad P_0\{J = 2\} = \tfrac{3}{12},$$
$$P_0\{J = 3\} = \tfrac{3}{12}, \quad P_0\{J = 4\} = \tfrac{2}{12}, \quad P_0\{J = 5\} = \tfrac{1}{12}.$$

The probability, under H_0, that J is greater than or equal to 4, for example, is therefore

$$P_0\{J \geq 4\} = P_0\{J = 4\} + P_0\{J = 5\}$$
$$= \tfrac{1}{12} + \tfrac{2}{12} = .25.$$

Note that we have derived the null distribution of J without specifying the common form (F) of the underlying distribution function for the X's under H_0 beyond the requirement that it be continuous. This is why the test procedure (6.14) based on J is called a *distribution-free procedure*. From the null distribution of J, we can determine the critical value j_α and control the probability α of

falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying continuous X distribution.

For a given number of treatments k and sample sizes $n_1, \ldots, n_k$, the R command cJCK$(\alpha, \boldsymbol{n})$ can be used to find the available upper-tail critical values j_α for possible values of J. For a given available significance level α, the critical value j_α then corresponds to $P_0(J \geq j_\alpha) = \alpha$ and is given by cJCK$(\alpha, \boldsymbol{n})$. Thus, for example, for $k = 3,\ n_1 = 6,\ n_2 = 5$, and $n_3 = 7$, we have $P_0(J \geq 79) = .0204$, so that $j_{.0204} = 79$ for $k = 3,\ n_1 = 6,\ n_2 = 5$, and $n_3 = 7$.

18. *Calculation of the Mean and Variance of J under the Null Hypothesis H_0.* In displays (6.15) and (6.16), we presented formulas for the mean and variance of J when the null hypothesis is true. In this comment, we illustrate a direct calculation of $E_0(J)$ and $\text{var}_0(J)$ in the particular case of $k = 3$ and $n_1 = n_2 = 1, n_3 = 2$ and no tied observations, using the null distribution of J obtained in Comment 17. (Later, in Comment 19, we present arguments for the general derivations of $E_0(J)$ and $\text{var}_0(J)$.) The null mean, $E_0(J)$, is obtained by multiplying each possible value of J with its probability under H_0. Thus,

$$E_0(J) = 0\left(\tfrac{1}{12}\right) + 1\left(\tfrac{2}{12}\right) + 2\left(\tfrac{3}{12}\right) + 3\left(\tfrac{3}{12}\right) + 4\left(\tfrac{2}{12}\right) + 5\left(\tfrac{2}{12}\right) = 2.5.$$

This is in agreement with what we obtain using (6.15), namely,

$$E_0(J) = \frac{4^2 - \{1^2 + 2^2 + 1^2\}}{4} = 2.5.$$

A check on the expression for $\text{var}_0(J)$ is also easy, using the well-known fact that

$$\text{var}_0(J) = E_0(J^2) - \{E_0(J)\}^2.$$

The value of $E_0(J^2)$, the second moment of the null distribution of J, is again obtained by multiplying possible values (in this case, of J^2) by the corresponding probabilities under H_0. We find

$$E_0(J^2) = 0^2\left(\tfrac{1}{12}\right) + 1^2\left(\tfrac{2}{12}\right) + 2^2\left(\tfrac{3}{12}\right) + 3^2\left(\tfrac{3}{12}\right) + 4^2\left(\tfrac{2}{12}\right) + 5^2\left(\tfrac{1}{12}\right) = \tfrac{49}{6}.$$

Thus,

$$\text{var}_0(J) = \tfrac{49}{6} - (2.5)^2 = \tfrac{23}{12} = 1.92,$$

which agrees with what we obtain using (6.16) directly, namely,

$$\text{var}_0(J) = \frac{\{4^2(2(4)+3) - [1^2(2(1)+3) + 1^2(2(1)+3) + 2^2(2(2)+3)]\}}{72} = 1.92.$$

19. *Large-Sample Approximation.* From the definition of J (6.13) and U_{uv} (6.12), we see that

$$\begin{aligned} E(J) &= E\left[\sum_{u=1}^{v-1}\sum_{v=2}^{k} U_{uv}\right] = E\left[\sum_{u=1}^{v-1}\sum_{v=2}^{k}\sum_{i=1}^{n_u}\sum_{j=1}^{n_v}\phi(X_{iu}, X_{jv})\right] \\ &= \sum_{u=1}^{v-1}\sum_{v=2}^{k}\sum_{i=1}^{n_u}\sum_{j=1}^{n_v} E[\phi(X_{iu}, X_{jv})] \\ &= \sum_{u=1}^{v-1}\sum_{v=2}^{k}\sum_{i=1}^{n_u}\sum_{j=1}^{n_v} P(X_{iu} < X_{jv}) \\ &= \sum_{u=1}^{v-1}\sum_{v=2}^{k} n_u n_v P(X_{1u} < X_{1v}). \end{aligned} \tag{6.21}$$

Under the null hypothesis H_0 (6.2), $P_0(X_{1u} < X_{1v}) = \frac{1}{2}$ for every $1 \le u < v \le k$. It follows that

$$\begin{aligned} E_0(J) &= \sum_{u=1}^{v-1}\sum_{v=2}^{k}\frac{(n_u n_v)}{2} = \frac{1}{4}\sum_{\substack{u=1\\u\neq v}}^{k}\sum_{v=1}^{k} n_u n_v \\ &= \frac{1}{4}\left[\sum_{u=1}^{k}\sum_{v=1}^{k} n_u n_v - \sum_{t=1}^{k} n_t^2\right] \\ &= \frac{1}{4}\left[N^2 - \sum_{t=1}^{k} n_t^2\right], \end{aligned}$$

which agrees with the general expression stated in (6.15).

It also follows from (6.12) and (6.13) that

$$\begin{aligned} \operatorname{var}(J) &= \operatorname{var}\left(\sum_{u=1}^{v-1}\sum_{v=2}^{k} U_{uv}\right) \\ &= \sum_{u=1}^{v-1}\sum_{v=2}^{k}\operatorname{var}(U_{uv}) + \sum_{\substack{u=1\\(u,v)\neq(s,t)}}^{v-1}\sum_{v=2}^{k}\sum_{s=1}^{t-1}\sum_{t=2}^{k}\operatorname{cov}(U_{uv}, U_{st}). \end{aligned} \tag{6.22}$$

Under H_0 (6.2), it can be shown (we will not here) that

$$\operatorname{var}_0(U_{uv}) = \frac{n_u n_v(n_u + n_v + 1)}{12}, \quad \text{for } 1 \le u < v \le k, \tag{6.23}$$

$$\operatorname{cov}_0(U_{uv}, U_{st}) = 0, \quad \text{for all distinct } u, v, s, t \text{ in } \{1, \ldots, k\}, \tag{6.24}$$

$$\text{cov}_0(U_{uv}, U_{ut}) = \frac{n_u n_v n_t}{12}, \quad \text{for } 1 \le u < v, t \le k,\ v \ne t, \tag{6.25}$$

$$\text{cov}_0(U_{uv}, U_{su}) = \frac{-n_s n_u n_v}{12}, \quad \text{for } 1 \le s < u < v \le k, \tag{6.26}$$

$$\text{cov}_0(U_{uv}, U_{vt}) = \frac{-n_u n_v n_t}{12}, \quad \text{for } 1 \le u < v < t \le k, \tag{6.27}$$

$$\text{cov}_0(U_{uv}, U_{sv}) = \frac{n_u n_v n_s}{12}, \quad \text{for } 1 \le u, s < v \le k,\ u \ne s. \tag{6.28}$$

Combining the results in (6.23)–(6.27), and (6.28) with the expression for $\text{var}(J)$ in (6.22), it follows after significant algebraic manipulation that

$$\text{var}_0(J) = \frac{N^2(2N+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)}{72},$$

which agrees with the general expression stated in (6.16).

The null asymptotic normality of the standardized form

$$J^* = \frac{J - E_0(J)}{\{\text{var}_0(J)\}^{1/2}} = \frac{J - \left[\frac{N^2 - \sum_{j=1}^{k} n_j^2(2n_j+3)}{4}\right]}{\left\{\left[N^2(2N+1) - \sum_{t=1}^{k} n_t^2(2n_t+3)\right]\Big/72\right\}^{1/2}}$$

follows from the fact that J can be expressed as a sum of certain mutually independent combined-samples Mann–Whitney statistics and standard theory for such sums of mutually independent, but not necessarily identically distributed, random variables (see, e.g., Terpstra (1952) or Section 12.1 of Randles and Wolfe (1979)). Asymptotic normality results for J under general alternatives to H_0 are obtainable from standard results in the k-sample U-statistics theory (see, e.g., Lehmann (1975, pp. 401–402)).

20. *Power of the Jonckheere–Terpstra Test.* The Jonckheere–Terpstra procedures (6.14) and (6.18) are quite superior to the Kruskal–Wallis procedures in (6.6) and (6.7) when the conjectured ordering of the treatment effects ($\tau_1 \le \tau_2 \le \cdots \le \tau_k$) is, indeed, appropriate. In addition, small violations in the conjectured ordering for τ_i and τ_j do not seriously affect the power of the Jonckheere–Terpstra tests if i and j correspond to treatment labels near the middle of the conjectured orderings. However, if i and j are both near 1 or k, the effect of such violations can be rather substantial, especially if the magnitude of the difference $|\tau_j - \tau_i|$ is fairly large. Mack and Wolfe (1981) presented the results of a small-sample power study that illustrates this phenomenon about the power of the Jonckheere–Terpstra procedures. In Section 6.3, we will discuss test procedures designed to deal with this possibility of early or late violations of the conjectured orderings $\tau_1 \le \tau_2 \le \cdots \le \tau_k$. The Jonckheere–Terpstra procedures will turn out to be special cases of this class of tests designed for the more general form of alternatives $\tau_1 \le \tau_2 \le \cdots \le \tau_{p-1} \le \tau_p \ge \tau_{p+1} \ge \cdots \ge \tau_k$, known in the literature as *umbrella orderings* for the pictorial shape of the graphed treatment effects.

21. *k-Sample Behrens–Fisher Problem.* Two of the implicit requirements associated with Assumptions A1–A3 are that the underlying distributions belong to the same

common family (F) and that they differ within this family at most in their medians. The less restrictive setting where these assumptions are relaxed to permit the possibility of differences in scale parameters as well as medians within the common family F is referred to as the *k-sample Behrens–Fisher problem*. The Jonckheere–Terpstra procedure (6.14) is no longer distribution-free under this more relaxed Behrens–Fisher setting. Chen and Wolfe (1990a) suggested a modification of the Jonckheere–Terpstra statistic J (6.13) to deal with this less restrictive setting. Their approach is similar to that used by Rust and Fligner (1984) to modify the Kruskal–Wallis statistic H for the same setting (see Comment 10).

22. *Consistency of the J Test*. Replace Assumptions A1–A3 by the less restrictive Assumptions A1′: the X's are mutually independent and A2′ : $X_{1j}, \ldots, X_{n_j j}$ come from the same continuous population $\Pi_j, j = 1, \ldots, k$. The populations $\Pi_1, \ldots, \Pi_k$ need not be identical, but we do assume that

$$\delta_{ij} = P(X_{1j} > X_{1i}) \geq \tfrac{1}{2}, \quad \text{for } 1 \leq i < j \leq k.$$

Then, roughly speaking, the test defined by (6.14) is consistent if and only if there is at least one pair (i, j), with $i < j$, such that $\delta_{ij} > \frac{1}{2}$.

Properties

1. *Consistency*. The condition n_j/N tends to $\lambda_j, 0 < \lambda_j < 1, j = 1, \ldots, k$, is sufficient to ensure that the test defined by (6.14) is consistent against the H_2 (6.11) alternatives. For a more general consistency statement, see Terpstra (1952) and Comment 22.
2. *Asymptotic Normality*. See Randles and Wolfe (1979, pp. 396–397) and Lehmann (1975, pp. 401–402).
3. *Efficiency*. See Puri (1965) and Section 6.10.

Problems

11. Apply the Jonckheere–Terpstra test to the psychotherapeutic attraction data of Table 6.2 using the postulated ordering $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$. Compare and contrast this result with that obtained for the Kruskal–Wallis test in Problem 1.

12. The statistic J can be computed either from (a) the joint ranking of the $N = \sum_{j=1}^{k} n_j$ observations or from (b) $k(k-1)/2$ "two-sample" rankings. Explain.

13. What are the minimum and maximum values for J? Justify your answer.

14. Suppose $k = 3$ and $n_1 = 4$, $n_2 = 7$, $n_3 = 8$. Compare the critical region for the exact level $\alpha = .0444$ test of H_0 (6.2) based on J with the critical region for the corresponding nominal level $\alpha = .0444$ test based on the large-sample approximation. What is the nominal probability of a type I error assigned by the large-sample approximation to the exact level $\alpha = .0444$ critical region?

15. Suppose $k = 4$, $n_1 = n_2 = n_3 = 1$, and $n_4 = 2$. Obtain the form of the exact null (H_0) distribution of J for the case of no tied observations.

16. Use (6.23)–(6.27), and (6.28) to show that the expression for $\text{var}_0(J)$ in (6.16) follows, under H_0, from the general expression for var (J) in (6.22).

Table 6.7 Average Basal Area Increment (BAI) Values for Oak Stands in Southeastern Ohio

Growing site index interval				
66–68	69–71	72–74	75–77	78–80
1.91	2.44	2.45	2.52	2.78
1.53		2.04	2.36	2.88
2.08		1.60	2.73	2.10
1.71		2.37		1.66

Source: M. Dale (1984).

17. In a project designed to study stand density (i.e., number of trees in a fixed area) and its relationship to other important features of a timber area such as tree growth, wood quality, and total wood production, Dale (1984) collected data on a quantity (related to yearly growth increment in a tree) known as *basal area increment* (*BAI*) for 16 stands of mixed species of oak trees in southeastern Ohio. The 16 stands were grouped according to the value of a second factor called *growing site index*. This index ranges in value from the low 50s to 100s for oak species, and as the value of the site index increases, the growing environment becomes more favorable for a stand of trees. The data in Table 6.7 are a subset of the data obtained by Dale and represent average BAI values for the 16 stands in his study. The BAI data are grouped into five distinct categories according to the associated growing site index values.

Use an appropriate test procedure to evaluate the conjecture that the average basal area increment for a given stand of oak trees is an increasing function of the value of the stand's growing site index.

18. Apply the Kruskal–Wallis test to the knowledge of performance data in Table 6.6. Compare and contrast this result with that obtained by the Jonckheere–Terpstra test in Example 6.2.

19. Show that the expression given in (6.19) for the null variance of J in the case of tied X observations reduces to the usual null variance of J when there are no ties, as given in (6.16).

6.3 DISTRIBUTION-FREE TESTS FOR UMBRELLA ALTERNATIVES (MACK–WOLFE)

In Section 6.2, we introduced the idea of designing test procedures to be especially effective against a restricted class of alternatives. There we considered the special class of monotonically ordered alternatives. In this section, we extend that idea to a broader class of alternatives, which includes the ordered alternatives of Section 6.2 as a special case.

Let $p \in \{1, 2, \ldots, k\}$ be a fixed treatment label. In this section, we consider procedures for testing H_0 (6.2) against the class of umbrella alternatives corresponding to

$$H_3 : [\tau_1 \leq \tau_2 \leq \cdots \leq \tau_{p-1} \leq \tau_p \geq \tau_{p+1} \geq \cdots \geq \tau_k, \text{ with at least one strict inequality}]. \tag{6.29}$$

(The label *umbrella* was given to these alternatives by Mack and Wolfe (1981) because of the pictorial configuration of the τ's.) The umbrella in (6.29) is said to have a peak at population p. (Note that the ordered alternatives of Section 6.2 are simply a special case of umbrella alternatives with peak at $p = k$.) These umbrella alternatives are one-way layout analogs to a quadratic regression setting and are appropriate, for example, in

evaluating marginal gain in performance efficiency as a function of time, crop yield as a function of fertilizer applied, reaction to increasing drug dosage levels where a downturn in effect may occur after the optimal dosage is exceeded, effect of age on responses to certain stimuli, etc. (These umbrella alternatives can be effectively used in place of ordered alternatives when one is concerned about possible violations of the monotonic ordering at either the beginning or the end of the sequence of treatment effects. See Comment 20 for further discussions along these lines.)

In Section 6.3A, we present a procedure specifically designed to test H_0 (6.2) against the umbrella alternatives H_3 (6.29), where the peak, p, of the conjectured umbrella is known *prior* to data collection. This procedure is preferred to the general alternatives Kruskal–Wallis test in Section 6.1 when the treatments can be labeled a priori in such a way that the experimenter expects any deviation from H_0 (6.2) to be in the particular direction of H_3 (6.29) with known p. In Section 6.3B, we extend the idea of umbrella alternatives to the more practical setting where it is not necessary to specify the peak, p, of the umbrella configuration prior to data collection. Here, we present a procedure designed to test H_0 (6.2) against the class of umbrella alternatives with peak (p) unspecified, namely,

$$H_4 : [\tau_1 \leq \cdots \leq \tau_{p-1} \leq \tau_p \geq \tau_{p+1} \geq \cdots \geq \tau_k, \text{ with at least one strict inequality, for some } p \in \{1, 2, \ldots, k\}]. \quad (6.30)$$

The Mack–Wolfe procedure in Section 6.3B is preferred to the *peak-known* procedure presented in Section 6.3A for the more common settings when umbrella alternatives are appropriate but where there is some uncertainty about the treatment at which the maximum effect is expected to occur if H_0 (6.2) is not true.

As, with the ordered alternatives in Section 6.2, we emphasize that the labeling of the treatments so that either of the umbrella alternatives H_3 (6.29) or H_4 (6.30) is appropriate *cannot* depend on the observed sample values. This labeling must correspond to a factor (s) associated with the *experimental design* and *not* on the *sample data*. In Section 6.2B, however, the peak of the conjectured umbrella needs not be specified prior to data collection.

6.3A A DISTRIBUTION-FREE TEST FOR UMBRELLA ALTERNATIVES, PEAK KNOWN (MACK–WOLFE)

In this subsection, we present a procedure for testing H_0 (6.2) against the peak-known (at p) umbrella alternatives given by H_3 (6.29).

Procedure

First, we must label the treatments so that they are in the prescribed ordered relationships to the known peak, p, corresponding to the umbrella configuration in H_3 (6.29). To calculate the known-peak umbrella statistic, A_p, we first compute the $p(p-1)/2$ Mann–Whitney counts U_{uv} (6.12) for every pair of treatments with labels less than or equal to the hypothesized peak (i.e., for $1 \leq u < v \leq p$). In addition, we compute the $(k - p + 1)(k - p)/2$ reverse Mann–Whitney counts U_{vu} (6.12) for every pair of treatments with labels greater than or equal to the hypothesized peak (i.e., for $p \leq u < v \leq k$). (Thus, U_{vu} is the number of sample v before sample u precedences. Note that if there

are no ties between the uth sample and vth sample observations, $p \le u < v \le k$, then $U_{vu} = n_u n_v - U_{uv}$.) The Mack–Wolfe peak-known statistic, A_p, is then the sum of the Mann–Whitney counts to the left of the peak and the reverse Mann–Whitney counts to the right of the peak (as appropriate for the umbrella alternatives H_3 (6.29)), namely,

$$A_p = \sum_{u=1}^{v-1} \sum_{v=2}^{p} U_{uv} + \sum_{u=p}^{v-1} \sum_{v=p+1}^{k} U_{vu}. \tag{6.31}$$

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the peak-known (at $p \in \{1, \ldots, k\}$) umbrella alternative

$$H_3 : [\tau_1 \le \tau_2 \le \cdots \le \tau_{p-1} \le \tau_p \ge \tau_{p+1} \ge \cdots \ge \tau_k, \text{ with at least one strict inequality}],$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } A_p \ge a_{p,\alpha}; \quad \text{otherwise do not reject,} \tag{6.32}$$

where the constant $a_{p,\alpha}$ is chosen to make the type I error probability equal to α. The constant $a_{p,\alpha}$ is the upper α percentile for the null ($\tau_1 = \cdots = \tau_k$) distribution of A_p. Comment 25 explains how to obtain the critical value $a_{p,\alpha}$ for k treatments, known peak p, and sample sizes $n_1, \ldots, n_k$ and available levels of α.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic ($\min(n_1, \ldots, n_k)$ tending to infinity) normality of A_p, suitably standardized. For this purpose, we need to know the expected value and variance of A_p when the null hypothesis is true. Under H_0, the expected value and variance of A_p are

$$E_0(A_p) = \frac{N_1^2 + N_2^2 - \sum_{i=1}^{k} n_i^2 - n_p^2}{4} \tag{6.33}$$

and

$$\text{var}_0(A_p) = \frac{1}{72}\left\{2(N_1^3 + N_2^3) + 3(N_1^2 + N_2^2) - \sum_{i=1}^{k} n_i^2(2n_i + 3) - n_p^2(2n_p + 3) + 12n_p N_1 N_2 - 12n_p^2 N\right\}, \tag{6.34}$$

respectively, with $N_1 = \sum_{i=1}^{p} n_i$ and $N_2 = \sum_{i=p}^{k} n_i$. (Note that $N = N_1 + N_2 - n_p$, because the observations in the peak treatment p are counted in both N_1 and N_2.) These expressions for $E_0(A_p)$ and $\text{var}_0(A_p)$ are verified by direct calculations in Comment 27 for the special case of $k = 4$, $p = 3$, $n_1 = n_2 = n_4 = 1$, $n_3 = 2$. General derivations of both expressions are outlined in Comment 28.

The standardized version of A_p is

$$A_p^* = \frac{A_p - E_0(A_p)}{\sqrt{\text{var}_0(A_p)}}$$
$$= \left\{ A_p - \left[\frac{N_1^2 + N_2^2 - \sum_{i=1}^{k} n_i^2 - n_p^2}{4} \right] \right\}$$
$$\div \left\{ \left[2(N_1^3 + N_2^3) + 3(N_1^2 + N_2^2) - \sum_{i=1}^{k} n_i^2(2n_i + 3) \right.\right.$$
$$\left.\left. - n_p^2(2n_p + 3) + 12n_pN_1N_2 - 12n_p^2N \right] \Big/ 72 \right\}^{1/2}. \tag{6.35}$$

When H_0 is true, A_p^* has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic $N(0, 1)$ distribution (see Comment 28 for indications of the proof). The normal theory approximation to procedure (6.32) is

$$\text{Reject } H_0 \text{ if } A_p^* \geq z_\alpha; \text{ otherwise do not reject.} \tag{6.36}$$

Ties

If there are ties among either the N_1 X's in treatments $1, \ldots, p$ or the N_2 X's in treatments $p, \ldots, k$, replace $\phi(a, b)$ in the calculations of the appropriate Mann–Whitney counts U_{uv} or reverse Mann–Whitney counts U_{vu} by $\phi^*(a, b) = 1, \frac{1}{2}, 0$ if $a <, =, \text{or} > b$, respectively, so that for each between-sample comparison where there is a tie, the contribution to the appropriate Mann–Whitney or reverse Mann–Whitney count will be $\frac{1}{2}$. After computing A_p with these modified counts, use procedure (6.32) with this tie-modified value of A_p. Note, however, that this test associated with tied X's is only approximately, and not exactly, of the significance level α.

When applying the large-sample approximation, an additional factor should be taken into account. Although ties in the X's do not affect the null expected value of A_p, its true null variance is smaller in the case of ties than the numerical value given by the expression in (6.34). However, the appropriate expression for the exact variance of A_p in the case of ties is not available. Therefore, in the case of tied X's and large-sample sizes, we recommend computing A_p using the modified Mann–Whitney counts and then A_p^* via (6.35). With this modified value of A_p^*, the approximation (6.36) can be applied. However, the associated approximate P-value will be larger than what we would obtain if the appropriate expression for the ties-corrected null variance of A_p was available to use in the computation of A_p^*.

EXAMPLE 6.3 ***Fasting Metabolic Rate of White-Tailed Deer.***

Seasonal energy requirements of deer are an important consideration when evaluating wildlife plans for certain habitats. Both nutritional quality of the range and the physiological demands of the deer must be studied in order to prevent starvation during critical seasons and to select optimum harvest strategies. Some aspects of the energy demand were considered by Silver et al. (1969) as they studied the fasting metabolic rate (FMR)

Table 6.8 Fasting Metabolic Rate (FMR) for White-Tailed Deer (kcal/kg/day)

Two-Month Period					
January–February	March–April	May–June	July–August	September–October	November–December
36.0	39.9	44.6	53.8	44.3	31.7
33.6	29.1	54.4	53.9	34.1	22.1
26.9	43.4	48.2	62.5	35.7	30.7
35.8		55.7	46.6	35.6	
30.1		50.0			
31.2					
35.3					

Source: H. Silver, N. F. Colovos, J. B. Holter, and H. H. Hayes (1969).

of white-tailed deer. In particular, one of the questions of interest was whether or not FMR is an increasing function of environmental temperature, for which they collected the data in Table 6.8.

For these data, we expect any deviation from H_0 (6.2) to be in the direction of increasing FMR values from the January–February period up through the warmest 2-month period, July–August, with declining FMR values from July–August through the November–December period. Thus, we are interested in testing H_0 against the peak-known umbrella alternatives (6.29) with treatment labels $1 \equiv$ January–February, $2 \equiv$ March–April, $3 \equiv$ May–June, $4 \equiv$ July–August, $5 \equiv$ September–October, $6 \equiv$ November–December, and known umbrella peak at $p = 4$, corresponding to the warmest (July–August) 2-month period. For the purpose of illustration, we take the significance level to be $\alpha = .0101$. Applying the `R` command `cUmbrPK`$(\alpha, \boldsymbol{n}, p)$, we find `cUmbrPK(.0101,c(7,3,5,4,4,3),4)` $= 125$; that is, $P_0(A_4 \geq 125) = .0101$, and, in the notation of (6.32) with $k = 6$, $p = 4, n_1 = 7$, $n_2 = 3$, $n_3 = 5$, $n_4 = 4$, $n_5 = 4$, and $n_6 = 3$, we have $a_{4,.0101} = 125$ and procedure (6.32) reduces to

$$\text{reject } H_0 \text{ if } A_4 \geq 125.$$

We now illustrate the computations leading to the sample value of A_4 (6.31). For this purpose, we first need to compute the $4(3)/2 = 6$ Mann–Whitney counts U_{uv}, for $1 \leq u < v \leq 4$, and the $3(2)/2 = 3$ reverse Mann–Whitney counts U_{vu}, for $4 \leq u < v \leq 6$. We obtain

$$U_{12} = 7 + 1 + 7 = 15, \qquad U_{13} = 7 + 7 + 7 + 7 + 7 = 35,$$
$$U_{14} = 7 + 7 + 7 + 7 = 28, \quad U_{23} = 3 + 3 + 3 + 3 + 3 = 15,$$
$$U_{24} = 3 + 3 + 3 + 3 = 12, \quad U_{34} = 3 + 3 + 5 + 1 = 12,$$
$$U_{54} = 4 + 4 + 4 + 4 = 16, \quad U_{64} = 3 + 3 + 3 + 3 = 12,$$
$$U_{65} = 3 + 3 + 3 + 3 = 12.$$

From (6.31), it follows that

$$A_4 = U_{12} + U_{13} + U_{14} + U_{23} + U_{24} + U_{34} + U_{65} + U_{64} + U_{54}$$
$$= 15 + 35 + 28 + 15 + 12 + 12 + 12 + 12 + 16 = 157.$$

As this value of A_4 is greater than the critical value $a_{4,.0101} = 125$, we reject H_0 at the $\alpha = .0101$ level. In fact, from the observed value $A_4 = 157$, we see, using the R command `pUmbrPK(metabolic.rate, 4)` that $P_0(A_4 \geq 157) =$ `pUmbrPK(metabolic.rate, 4)`. Thus, the lowest significance level at which we can reject H_0 in favor of H_3 with the observed value of the test statistic $A_4 = 157$ is $<< .0001$.

For the large-sample approximation, we have $n_1 = 7, n_2 = 3, n_3 = 5, n_4 = 4, n_5 = 4$, and $n_6 = 3$, so that $N_1 = (7 + 3 + 5 + 4) = 19, N_2 = 3 + 4 + 4 = 11$, and $N = (7 + 3 + 5 + 4 + 4 + 3) = 26$. Using these figures in expressions (6.33) and (6.34) for $E_0(A_4)$ and $\text{var}_0(A_4)$, respectively, we see that

$$E_0(A_4) = \frac{(19)^2 + (11)^2 - [(7)^2 + (3)^2 + (5)^2 + (4)^2 + (4)^2 + (3)^2 + (4)^2]}{4}$$
$$= 85.5$$

and

$$\begin{aligned}\text{var}_0(A_4) = \tfrac{1}{72}\{&2[(19)^3 + (11)^3] + 3[(19)^2 + (11)^2] \\ &-[(7)^2(2(7) + 3) + (3)^2(2(3) + 3) \\ &+(5)^2(2(5) + 3) + (4)^2(2(4) + 3) \\ &+(4)^2(2(4) + 3) + (3)^2(2(3) + 3)] \\ &-(4)^2(2(4) + 3) + 12(4)(19)(11) - 12(4)^2(26)\} \\ = \frac{21{,}018}{72} &= 291.92.\end{aligned}$$

Thus, from (6.35), we obtain

$$A_4^* = \frac{A_4 - E_0(A_4)}{\sqrt{\text{var}_0(A_4)}} = \frac{157 - 85.5}{\sqrt{291.92}} = 4.18.$$

Using the R command `pnorm(·)`, the smallest approximate level at which we can reject H_0 in favor of H_3 with the observed value of $A_4^* = 4.18$ (i.e., the approximate P-value) is then given by $P_0(A_4^* \geq 4.18) \approx 1 -$ `pnorm(4.18)` $= 1 - .99999 = .00001$. Both the exact test and the large-sample approximate test provide very strong evidence in support of the claim that FMR for white-tailed deer is an increasing function of environmental temperature. (We note that the Jonckheere–Terpstra procedure from Section 6.2 would not be appropriate for these FMR data even with relabeled treatments, because, e.g., it would be difficult to properly order the temperatures of the March–April and September–October periods.)

Comments

23. *Motivation for the Test.* Notice that the statistic A_p can be viewed as the simple sum of two Jonckheere–Terpstra statistics, one (J_{up}) on treatments 1 through p with the postulated ordering $\tau_1 \leq \cdots \leq \tau_p$ and the second (J_{down}) on treatments

k through p with the postulated reverse ordering $\tau_k \leq \tau_{k-1} \leq \cdots \leq \tau_p$. Thus, the statistic $A_p = J_{\text{up}} + J_{\text{down}}$ will be large if either J_{up} or J_{down} (or both) is large. In view of Comment 14, this serves as partial motivation for the A_p test.

24. *Special Case of Three Treatments*. When there are only $k = 3$ treatments, the umbrella statistic A_p can be viewed in a special way. If $p = 3$, then $A_3 = U_{12} + U_{13} + U_{23}$ is just the usual Jonckheere–Terpstra statistic for the ordered alternatives $\tau_1 \leq \tau_2 \leq \tau_3$. If $p = 1$, we have $A_1 = U_{31} + U_{32} + U_{21}$, which is the Jonckheere–Terpstra statistic for the reverse ordered alternatives $\tau_3 \leq \tau_2 \leq \tau_1$. In either of these cases, all the properties of the Jonckheere–Terpstra test procedure (including null distribution and critical values) discussed in Section 6.2 apply directly to tests based on A_1 or A_3, add as appropriate. For the third umbrella setting with $p = 2$, we see that $A_2 = U_{12} + U_{32}$, which is the same as a *single* Mann–Whitney statistic comparing the peak sample (treatment 2) with the combined set of data from treatments 1 and 3. (Thus, A_2 is the number of sample 1 *or* sample 3 before sample 2 precedences.) As a result, if $p = 2$ and $k = 3$, the procedures in (6.32) and (6.36) for sample sizes n_1, n_2, and n_3 are equivalent to the exact and large-sample approximation forms, respectively, of the one-sided upper-tail two-sample Wilcoxon rank sum test (as discussed in Section 4.1) for sample sizes $m = n_1 + n_3$ and $n = n_2$.

25. *Derivation of the Distribution of A_p under H_0 (No Ties).* As with the Jonckheere–Terpstra statistic J (see Comment 17), it is clear that the umbrella peak-known statistic A_p can be computed from the joint ranking of all $N = \sum_{i=1}^{k} n_i$ observations. Thus, one way to obtain the null distribution of A_p is to follow the method of Comments 6 and 17, namely, to compute the value of A_p for each of the $N!/(\prod_{j=1}^{k} = n_j!)$ equally likely (under H_0) rank assignments. We illustrate how this works in the small-sample-size case of $k = 4$, $p = 3$, $n_1 = n_2 = n_4 = 1, n_3 = 2$. The $5!/[1!\ 1!\ 1!\ 2!] = 60$ possible assignments of the joint ranks 1, 2, 3, 4, and 5 to the four treatments and their associated values of A_3 (6.31) are as follows:

1.	I	II	III	IV		2.	I	II	III	IV
	1	2	4	3			2	1	4	3
			5						5	
				$A_3 = 7$						$A_3 = 6$
3.	I	II	III	IV		4.	I	II	III	IV
	1	3	4	2			3	1	4	2
			5						5	
				$A_3 = 7$						$A_3 = 6$
5.	I	II	III	IV		6.	I	II	III	IV
	2	3	4	1			3	2	4	1
			5						5	
				$A_3 = 7$						$A_3 = 6$
7.	I	II	III	IV		8.	I	II	III	IV
	1	2	3	4			2	1	3	4
			5						5	
				$A_3 = 6$						$A_3 = 5$

9.	I	II	III	IV		10.	I	II	III	IV
	1	4	3	2			4	1	3	2
			5						5	
				$A_3 = 6$						$A_3 = 5$
11.	I	II	III	IV		12.	I	II	III	IV
	2	4	3	1			4	2	3	1
			5						5	
				$A_3 = 6$						$A_3 = 5$
13.	I	II	III	IV		14.	I	II	III	IV
	1	3	2	4			3	1	2	4
			5						5	
				$A_3 = 5$						$A_3 = 4$
15.	I	II	III	IV		16.	I	II	III	IV
	1	4	2	3			4	1	2	3
			5						5	
				$A_3 = 5$						$A_3 = 4$
17.	I	II	III	IV		18.	I	II	III	IV
	3	4	2	1			4	3	2	1
			5						5	
				$A_3 = 5$						$A_3 = 4$
19.	I	II	III	IV		20.	I	II	III	IV
	2	3	1	4			3	2	1	4
			5						5	
				$A_3 = 4$						$A_3 = 3$
21.	1	II	III	IV		22.	I	II	III	IV
	2	4	1	3			4	2	1	3
			5						5	
				$A_3 = 4$						$A_3 = 3$
23.	I	II	III	IV		24.	I	II	III	IV
	3	4	1	2			4	3	1	2
			5						5	
				$A_3 = 4$						$A_3 = 3$
25.	I	II	III	IV		26.	I	II	III	IV
	1	2	3	5			2	1	3	5
			4						4	
				$A_3 = 5$						$A_3 = 4$
27.	I	II	III	IV		28.	I	II	III	IV
	1	5	3	2			5	1	3	2
			4						4	
				$A_3 = 5$						$A_3 = 4$
29.	I	II	III	IV		30.	I	II	III	IV
	2	5	3	1			5	2	3	1
			4						4	
				$A_3 = 5$						$A_3 = 4$

31.

I	II	III	IV
1	3	2	5
		4	

$A_3 = 4$

32.

I	II	III	IV
3	1	2	5
		4	

$A_3 = 3$

33.

I	II	III	IV
1	5	2	3
		4	

$A_3 = 4$

34.

I	II	III	IV
5	1	2	3
		4	

$A_3 = 3$

35.

I	II	III	IV
3	5	2	1
		4	

$A_3 = 4$

36.

I	II	III	IV
5	3	2	1
		4	

$A_3 = 3$

37.

I	II	III	IV
2	3	1	5
		4	

$A_3 = 3$

38.

I	II	III	IV
3	2	1	5
		4	

$A_3 = 2$

39.

I	II	III	IV
2	5	1	3
		4	

$A_3 = 3$

40.

I	II	III	IV
5	2	1	3
		4	

$A_3 = 2$

41.

I	II	III	IV
3	5	1	2
		4	

$A_3 = 3$

42.

I	II	III	IV
5	3	1	2
		4	

$A_3 = 2$

43.

I	II	III	IV
1	4	2	5
		3	

$A_3 = 3$

44.

I	II	III	IV
4	1	2	5
		3	

$A_3 = 2$

45.

I	II	III	IV
1	5	2	4
		3	

$A_3 = 3$

46.

I	II	III	IV
5	1	2	4
		3	

$A_3 = 2$

47.

I	II	III	IV
4	5	2	1
		3	

$A_3 = 3$

48.

I	II	III	IV
5	4	2	1
		3	

$A_3 = 2$

49.

I	II	III	IV
2	4	1	5
		3	

$A_3 = 2$

50.

I	II	III	IV
4	2	1	5
		3	

$A_3 = 1$

51.

I	II	III	IV
2	5	1	4
		3	

$A_3 = 2$

52.

I	II	III	IV
5	2	1	4
		3	

$A_3 = 1$

53.	I	II	III	IV	54.	I	II	III	IV
	4	5	1	2		5	4	1	2
			3					3	
				$A_3 = 2$					$A_3 = 1$

55.	I	II	III	IV	56.	I	II	III	IV
	3	4	1	5		4	3	1	5
			2					2	
				$A_3 = 1$					$A_3 = 0$

57.	I	II	III	IV	58.	I	II	III	IV
	3	5	1	4		5	3	1	4
			2					2	
				$A_3 = 1$					$A_3 = 0$

59.	I	II	III	IV	60.	I	II	III	IV
	4	5	1	3		5	4	1	3
			2					2	
				$A_3 = 1$					$A_3 = 0$

Thus, the null distribution for A_3 when $k = 3, n_1 = n_2 = n_4 = 1$, and $n_3 = 2$ is given by

$$P_0\{A_3 = 0\} = \tfrac{3}{60}, \quad P_0\{A_3 = 1\} = \tfrac{6}{60}, \quad P_0\{A_3 = 2\} = \tfrac{9}{60}$$
$$P_0\{A_3 = 3\} = \tfrac{12}{60}, \quad P_0\{A_3 = 4\} = \tfrac{12}{60}, \quad P_0\{A_3 = 5\} = \tfrac{9}{60}$$
$$P_0\{A_3 = 6\} = \tfrac{6}{60}, \quad P_0\{A_3 = 7\} = \tfrac{3}{60}.$$

The probability, under H_0, that A_3 is greater than or equal to 5, for example, is

$$P_0\{A_3 \geq 5\} = P_0\{A_3 = 5\} + P_0\{A_3 = 6\} + P_0\{A_3 = 7\} = \frac{9 + 6 + 3}{60} = .3.$$

Note that we have derived the null distribution of A_3 without specifying the common form (F) of the underlying distribution function for the X's under H_0 beyond the requirement that it be continuous. This is why the test procedure (6.32) based on A_p is called a *distribution-free procedure*. From the null distribution of A_p, we can determine the critical value $a_{p,\alpha}$ and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying continuous X distribution.

For a given number of treatments k, peak p, and sample sizes $n_1, \ldots, n_k$, the R command `cUmbrPK`$(\alpha, \boldsymbol{n}, p)$ can be used to find the available upper-tail critical values $a_{p,\alpha}$ for possible values of A_p. For a given available significance level α, the critical value $a_{p,\alpha}$ then corresponds to $P_0(A_p \geq a_{p,\alpha}) = \alpha$ and is given by `cUmbrPK`$(\alpha, \boldsymbol{n}, p)$. Thus, for example, for $k = 5,\ p = 3,\ n_1 = n_2 = n_3 = n_4 = n_5 = 4$, we have $P_0(A_3 \geq 68) = .0475$, so that $a_{3,.0475} = 68$ for $k = 5,\ p = 3$, and $n_1 = n_2 = n_3 = n_4 = n_5 = 4$.

26. *Calculation of the Mean and Variance of A_p under the Null Hypothesis H_0.* In displays (6.33) and (6.34), we presented formulas for the mean and variance

of A_p when the null hypothesis is true. In this comment, we provide a direct calculation of $E_0(A_p)$ and $\text{var}_0(A_p)$ in the specific case of $k = 4, p = 3, n_1 = n_2 = n_4 = 1, n_3 = 2$ and no tied observations using the null distribution of A_3 obtained in Comment 25. (Later, in Comment 27, we discuss general derivations of $E_0(A_p)$ and $\text{var}_0(A_p)$.) From the null distribution provided in Comment 25, we see that

$$\begin{aligned} E_0(A_3) &= \left[0\left(\tfrac{3}{60}\right) + 1\left(\tfrac{6}{60}\right) + 2\left(\tfrac{9}{60}\right) + 3\left(\tfrac{12}{60}\right) + 4\left(\tfrac{12}{60}\right)\right. \\ &\qquad \left. +5\left(\tfrac{9}{60}\right) + 6\left(\tfrac{6}{60}\right) + 7\left(\tfrac{3}{60}\right)\right] \\ &= 3.5. \end{aligned}$$

This is in agreement with what we obtain using (6.33), namely,

$$\begin{aligned} E_0(A_3) &= \frac{\{(1+1+2)^2 + (2+1)^2 - [1^2 + 1^2 + 2^2 + 1^2 + 2^2]\}}{4} \\ &= \frac{16 + 9 - 11}{4} = 3.5. \end{aligned}$$

Again using the null distribution in Comment 25, we have

$$\begin{aligned} E_0(A_3^2) &= \left[0^2\left(\tfrac{3}{60}\right) + 1^2\left(\tfrac{6}{60}\right) + 2^2\left(\tfrac{9}{60}\right) + 3^2\left(\tfrac{12}{60}\right) + 4^2\left(\tfrac{12}{60}\right)\right. \\ &\qquad \left. +5^2\left(\tfrac{9}{60}\right) + 6^2\left(\tfrac{6}{60}\right) + 7^2\left(\tfrac{3}{60}\right)\right] \\ &= 15.5. \end{aligned}$$

Using the well-known expression for $\text{var}_0(A_3)$, it follows that

$$\text{var}_0(A_3) = E_0(A_3^2) - \{E_0(A_3)\}^2 = 15.5 - (3.5)^2 = 3.25,$$

which agrees with what we obtain using (6.34) directly, namely,

$$\begin{aligned} \text{var}_0(A_3) &= \{2(4^3 + 3^3) + 3(4^2 + 3^2) + [(3)(1)^2(2(1) + 3) + 2(2)^2(2(2) + 3)] \\ &\qquad +12(2)(4)(3) - 12(2)^2(5)\}/72 \\ &= \frac{182 + 75 - 71 + 288 - 240}{72} = 3.25. \end{aligned}$$

27. *Large-Sample Approximation.* As noted in Comment 23, the umbrella statistic A_p can be expressed as $A_p = J_{\text{up}} + J_{\text{down}}$, where J_{up} is the Jonckheere–Terpstra statistic on treatments 1 through p with the postulated ordering $\tau_1 \le \cdots \le \tau_p$ and J_{down} is the Jonckheere–Terpstra statistic on treatments k through p with the postulated ordering $\tau_k \le \tau_{k-1} \le \cdots \le \tau_p$. Thus, using the previous development

for the Jonckheere–Terpstra statistic in Comment 19, we see that

$$\begin{aligned} E_0(A_p) &= E_0(J_{\text{up}}) + E_0(J_{\text{down}}) \\ &= \frac{1}{4}\left[N_1^2 - \sum_{t=1}^{p} n_t^2\right] + \frac{1}{4}\left[N_2^2 - \sum_{t=p}^{k} n_t^2\right] \\ &= \frac{1}{4}\left[N_1^2 + N_2^2 - \sum_{t=1}^{k} n_t^2 - n_p^2\right], \end{aligned}$$

which agrees with the general expression stated in (6.33).

It also follows from the representation $A_p = J_{\text{up}} + J_{\text{down}}$ that

$$\begin{aligned} \text{var}_0(A_p) &= \text{var}_0(J_{\text{up}} + J_{\text{down}}) \\ &= \text{var}_0(J_{\text{up}}) + \text{var}_0(J_{\text{down}}) + 2\text{cov}_0(J_{\text{up}}, J_{\text{down}}). \end{aligned} \tag{6.37}$$

Now,

$$\begin{aligned} \text{cov}_0(J_{\text{up}}, J_{\text{down}}) &= \text{cov}_0\left(\sum_{u=1}^{v-1}\sum_{v=2}^{p} U_{uv}, \sum_{s=p}^{t-1}\sum_{t=p+1}^{k} U_{ts}\right) \\ &= \text{cov}_0\left(\sum_{u=1}^{v-1}\sum_{v=2}^{p-1} U_{uv} + \sum_{u=1}^{p-1} U_{up}, \sum_{s=p+1}^{t-1}\sum_{t=p+2}^{k} U_{ts} + \sum_{t=p+1}^{k} U_{tp}\right) \\ &= \left[\text{cov}_0\left(\sum_{u=1}^{v-1}\sum_{v=2}^{p-1} U_{uv}, \sum_{s=p+1}^{t-1}\sum_{t=p+2}^{k} U_{ts}\right)\right. \\ &\quad + \text{cov}_0\left(\sum_{u=1}^{v-1}\sum_{v=2}^{p-1} U_{uv}, \sum_{t=p+1}^{k} U_{tp}\right) \\ &\quad + \text{cov}_0\left(\sum_{u=1}^{p-1} U_{up}, \sum_{s=p+1}^{t-1}\sum_{t=p+2}^{k} U_{ts}\right) \\ &\quad \left. + \text{cov}_0\left(\sum_{u=1}^{p-1} U_{up}, \sum_{t=p+1}^{k} U_{tp}\right)\right]. \end{aligned} \tag{6.38}$$

The term $\sum_{u=1}^{v-1}\sum_{v=2}^{p-1} U_{uv}$ involves only X observations from the first $(p-1)$ samples, whereas the terms $\sum_{s=p+1}^{t-1}\sum_{t=p+2}^{k} U_{ts}$ and $\sum_{t=p+1}^{k} U_{tp}$ involve only X observations from samples $p+1, p+2, \ldots, k$ and $p, p+1, \ldots, k$, respectively. As the X observations are mutually independent, it follows that

$$\text{cov}_0\left(\sum_{u=1}^{v-1}\sum_{v=2}^{p-1} U_{uv}, \sum_{s=p+1}^{t-1}\sum_{t=p+2}^{k} U_{ts}\right) = \text{cov}_0\left(\sum_{u=1}^{v-1}\sum_{v=2}^{p-1} U_{uv}, \sum_{t=p+1}^{k} U_{tp}\right) = 0. \tag{6.39}$$

Similarly, the term $\sum_{u=1}^{p-1} U_{up}$ involves only X observations from the first p samples, and the term $\sum_{s=p+1}^{t-1} \sum_{t=p+2}^{k} U_{ts}$ involves only X observations from samples $p+1, p+2, \ldots, k$, leading to

$$\operatorname{cov}_0 \left(\sum_{u=1}^{p-1} U_{up} \sum_{s=p+1}^{t-1} \sum_{t=p+2}^{k} U_{ts} \right) = 0. \tag{6.40}$$

(Note that (6.39) and (6.40) are a consequence of the fact that the sample observations from the peak treatment p are the only data used in both J_{up} and J_{down}.) Combining (6.38), (6.39), and (6.40) with a well-known result about covariances of sums, we obtain

$$\begin{aligned} \operatorname{cov}_0(J_{\text{up}}, J_{\text{down}}) &= \operatorname{cov}_0 \left(\sum_{u=1}^{p-1} U_{up}, \sum_{t=p+1}^{K} U_{tp} \right) \\ &= \sum_{u=1}^{p-1} \sum_{t=p+1}^{k} \operatorname{cov}_0(U_{up}, U_{tp}). \end{aligned} \tag{6.41}$$

From (6.28), it follows that

$$\begin{aligned} \operatorname{cov}_0(J_{\text{up}}, J_{\text{down}}) &= \frac{n_p}{12} \sum_{u=1}^{p-1} \sum_{t=p+1}^{k} n_u n_t = \frac{n_p}{12} \left(\sum_{u=1}^{p-1} n_u \right) \left(\sum_{t=p+1}^{k} n_t \right) \\ &= \frac{n_p (N_1 - n_p)(N_2 - n_p)}{12}. \end{aligned} \tag{6.42}$$

Combining (6.37) and (6.42), we see that

$$\operatorname{var}_0(A_p) = \operatorname{var}_0(J_{\text{up}}) + \operatorname{var}_0(J_{\text{down}}) + \frac{n_p (N_1 - n_p)(N_2 - n_p)}{6}. \tag{6.43}$$

Using the expression in (6.16) for both $\operatorname{var}_0(J_{\text{up}})$ and $\operatorname{var}_0(J_{\text{down}})$, it follows from (6.43) after some algebraic manipulation (see Problem 28) that

$$\begin{aligned} \operatorname{var}_0(A_p) = \frac{1}{72} \Bigg\{ & 2(N_1^3 + N_2^3) + 3(N_1^2 + N_2^2) - \sum_{i=1}^{k} n_i^2 (2n_i + 3) \\ & - n_p^2 (2n_p + 3) + 12 n_p N_1 N_2 - 12 n_p^2 N \Bigg\}, \end{aligned}$$

which agrees with the general expression stated in (6.34).

The null asymptotic normality of the standardized form

$$A_p^* = \frac{A_p - E_0(A_p)}{\{\operatorname{var}_0(A_p)\}_{1/2}}$$

follows from the fact that A_p can be expressed as a sum of certain mutually independent combined-samples Mann–Whitney statistics and standard theory for

such sums of mutually independent, but not necessarily identically distributed, random variables (see, e.g., Mack and Wolfe (1981)). Asymptotic normality results for A_p under general alternatives to H_0 follow directly from work by Archambault, Mack, and Wolfe (1977) on a large class of k-sample statistics.

28. *k-Sample Behrens–Fisher Problem.* Two of the implicit requirements associated with Assumptions A1–A3 are that the underlying distributions belong to the same common family (F) and that they differ within this family at most in their medians. The less restrictive setting where these assumptions are relaxed to permit the possibility of differences in scale parameters as well as medians within the common family F is referred to as the *k-sample Behrens–Fisher problem*. The Mack–Wolfe procedure (6.32) is no longer distribution-free under this more relaxed Behrens–Fisher setting. Chen and Wolfe (1990a) suggested a modification of the Mack–Wolfe statistic A_p (6.31) to deal with this less restrictive setting. Their approach is similar to that used by Rust and Fligner (1984) to modify the Kruskal–Wallis statistic H for the same setting (see Comment 10).

29. *Consistency of the A_p Test.* Replace Assumptions A1–A3 by the less restrictive Assumptions A1′: the X's are mutually independent and A2′ : $X_{1j}, \ldots, X_{n_j j}$ come from the same continuous population $\Pi_j, j = 1, \ldots, k$. The populations $\Pi_1, \ldots, \Pi_k$ need not be identical, but they are restricted to conform with the umbrella alternatives. Letting $\delta_{ij} = P(X_{1j} > X_{1i})$, for $1 \le i < j \le k$, we do assume that

$$\begin{aligned} \delta_{ij} &\ge \tfrac{1}{2}, \quad \text{for } 1 \le i < j \le p \\ \delta_{ij} &\le \tfrac{1}{2}, \quad \text{for } p \le i < j \le k, \end{aligned} \tag{6.44}$$

with no restrictions on δ_{ij} for $i < p$ and $j > p$. Under these conditions on $\Pi_1, \ldots, \Pi_k$, the test defined by (6.32) is, roughly speaking, consistent if and only if at least one of the inequalities in (6.44) is strict.

Properties

1. *Consistency.* The condition n_j/N tends to $\lambda_j, 0 < \lambda_j < 1, j = 1, \ldots, k$, is sufficient to insure that the test defined by (6.32) is consistent against the umbrella alternatives H_3 (6.29). For a more general consistency statement, see Mack and Wolfe (1981) and Comment 29.
2. *Asymptotic Normality.* See Mack and Wolfe (1981).
3. *Efficiency.* See Mack and Wolfe (1981) and Section 6.10.

Problems

20. Survival of stocked tiger muskellunge (*Esox masquinongy*), like other stocked sportfish, is variable to poor in Ohio reservoirs. Previous research with this species suggests three possible reasons for poor survival: (i) predation by largemouth bass, (ii) inability to forage, and (iii) stress-related mortality associated with the stocking process. Among other things, Mather (1984) studied the effect on mortality of three components of the stocking process: netting, confinement, and temperature increase. One portion of her study dealt with the glucose response to the stress of an increase in temperature. A sample of 40 tiger muskellunge were transferred

Table 6.9 Plasma Glucose (mg%)

Hours after 12 °C temperature increase				
0	1	4	24	96
61.08	95.45	205.96	67.74	61.76
86.21	169.19	82.55	79.84	69.12
90.15	216.16	116.60	78.23	77.45
72.91	141.92	107.23	90.23	73.45
83.74	116.16	103.83	64.92	71.08
76.35	172.22	96.60	65.73	52.45
91.63	126.26	112.77	49.60	71.57
56.65	177.78	140.85	77.42	54.90

Source: M. Mather (1984).

from a 15 °C holding tank into a test tank (also held at 15 °C) and allowed 24 h to recover. (This is the period of time that previous experimenters have found to be necessary for the fish's plasma glucose level to return to normal after a dipnet stressor.) Then, a random sample of eight fish were removed from the tank, anesthetized, blood collected, and plasma glucose determined. These data serve as a baseline or control sample. Next, the stressor (a 12 °C temperature increase) was applied to the test tank and blood samples were collected (in the way previously described) for random samples of eight additional fish at each of the time periods 1, 4, 24, and 96 h after the temperature increase. These plasma glucose measurements (mg%) are given in Table 6.9 for the 40 fish in the study.

In anticipation that a 24-h period is also necessary for a tiger muskellunge's plasma glucose level to recover from the 12 °C temperature increase stressor, test the hypothesis of interest using a significance level of .048. What is the P-value for these data?

21. (a) The statistic A_p can be computed from the joint ranking of all N observations. Explain.
(b) The statistic A_p can also be computed from pairwise *two-sample* rankings. Explain.
(c) How many different two-sample rankings are required in (b) to compute A_p?

22. In Example 6.2, we used the Jonckheere–Terpstra procedure to analyze the knowledge of performance data. It is quite reasonable to postulate that "too much information" (e.g., supervisor looking over your shoulder commenting at each step of the process) might actually lead to a downturn in the number of satisfactory pieces produced. Suppose that the following data were collected under such a too much information scenario.

Group D (too much information)
38
41
37
46
39
42

Use both the Jonckheere–Terpstra procedure and the Mack–Wolfe procedure with $p = 3$ to analyze the performance data with these added group D data. Discuss your findings.

23. What are the minimum and maximum values for A_p? Justify your answers.

24. Notice that the statistic A_p (6.31) does not include any Mann–Whitney comparisons between samples from pairs of treatments on opposite sides of the peak treatment p. Discuss the pros and cons of this fact in relation to the Mack–Wolfe test procedure based on A_p.

25. Consider the umbrella statistic A_p for k treatments.

(a) Which value(s) of p requires computation of the maximum number of Mann–Whitney statistics? How many Mann–Whitney statistics are required?

(b) Which value(s) of p requires computation of the fewest number of Mann–Whitney statistics? How many Mann–Whitney statistics are required?

26. Suppose $k = 4$, $n_1 = n_3 = n_4 = 1$, and $n_2 = 2$. Obtain the form of the exact null (H_0) distribution of A_2 for the case of no tied observations. Compare the null distribution of A_2 for $k = 4$, $n_1 = n_3 = n_4 = 1$, $n_2 = 2$ with the null distribution of A_3 for $k = 4$, $n_1 = n_2 = n_4 = 1$, $n_3 = 2$, as obtained in Comment 25. Discuss the differences.

27. Suppose $k = 4$, $n_1 = n_4 = 1$, and $n_2 = n_3 = 2$. Obtain the form of the exact null (H_0) distribution of A_2 for the case of no tied observations.

28. Show that the expression for the null variance (no ties) of A_p given in (6.43) is indeed the same as that stated in (6.34).

29. In many settings, a dose–response relationship needs not be monotonic in the dosage. In *in vitro* mutagenicity assays, for example, experimental organisms may not survive the toxic side effects of high doses of the test agent, thereby actually reducing the number of organisms at risk of mutation and leading to a downturn (i.e., umbrella pattern) in the dose–response curve. The date in Table 6.10 are a subset of the data considered by Simpson and Margolin (1986) in a discussion of the analysis of Ames test results. Plates containing Salmonella bacteria of strain TA98 were exposed to various doses of Acid Red 114. The tabled observations are the numbers of visible revertant colonies on the 18 plates in the study.

Test the null hypothesis H_0 (6.2) against the alternative that the peak of the dose–response curve for Salmonella bacteria of strain TA98 exposure to Acid Red 114 occurs at dosage level 1000 μg/ml.

30. For the Salmonella bacteria strain TA98 data in Table 6.10, test the null hypothesis H_0 (6.2) against the alternative that the peak of the dose–response curve for Salmonella bacteria of strain TA98 exposure to Acid Red 114 occurs at dosage level 333 μg/ml. Compare the result with that from Problem 29.

31. For the Salmonella bacteria strain TA98 data in Table 6.10, test the null hypothesis H_0 (6.2) against the alternative that the number of revertant colonies of the bacteria is a monotone increasing function of the dose level of Acid Red 114 over the range of exposure in Table 6.10. Compare this result with those obtained in Problems 29 and 30.

32. For the Salmonella bacteria strain TA98 data in Table 6.10, use the Kruskal–Wallis procedure to test H_0 (6.2) against the general alternatives H_1 (6.3). Compare this result with those obtained in Problems 29, 30, and 31.

Table 6.10 Number of Revertant Colonies of Salmonella Bacteria of Strain TA98 under Exposure to Various Doses of Acid Red 114, with Hamster Liver Activation

Dose, μg/ml					
0	100	333	1000	3333	10,000
22	60	98	60	22	23
23	59	78	82	44	21
35	54	50	59	33	25

Source: D. G. Simpson and B. H. Margolin (1986).

6.3B A DISTRIBUTION-FREE TEST FOR UMBRELLA ALTERNATIVES, PEAK UNKNOWN (MACK–WOLFE)

In this section, we present a procedure for testing H_0 (6.2) against the general peak-unknown umbrella alternatives H_4 (6.30).

Procedure

We label the treatments so that they are in the proper umbrella relationship to the unknown peak treatment p. To calculate the Mack–Wolfe statistic for this unknown peak setting, we first use the sample data to estimate which of the treatments is most likely to correspond to the peak of the umbrella; that is, we first estimate p from the sample data. To accomplish this, we calculate k combined-samples Mann–Whitney statistics

$$U_{.q} = \sum_{i \neq q} U_{iq}, \quad \text{for } q = 1, \ldots, k, \tag{6.45}$$

where U_{iq} = (number of ith sample observations that precede qth sample observations) is the usual Mann–Whitney statistic for the ith and qth samples. Thus, $U_{.q}$ is itself simply a single Mann–Whitney statistic computed between the qth sample and the remaining $(k-1)$ samples combined (i.e., it equals the number of times an observation from the qth sample exceeds an observation from the other $(k-1)$ combined samples). Next, we standardize each of the $U_{.q}$'s by subtracting off its expected value under the null hypothesis H_0 (6.2) and dividing by its null standard deviation (see Comment 35) to obtain

$$U^*_{.q} = \frac{U_{.q} - E_0(U_{.q})}{\{\text{var}_0(U_{.q})\}^{1/2}} = \frac{U_{.q} - [n_q(N-n_q)/2]}{\left\{\dfrac{n_q(N-n_q)(N+1)}{12}\right\}^{1/2}}, \quad q = 1, \ldots, k. \tag{6.46}$$

Let r equal the number of treatments that are tied for having the maximum $U^*_{.q}$ value and let B be the subset of $\{1, 2, \ldots, k\}$ that corresponds to the r treatments tied for the maximum $U^*_{.q}$ value. (As $U^*_{.1}, \ldots, U^*_{.k}$ are discrete random variables, there are sample size configurations for which the probability is positive that r will be greater than 1. See also Comment 31 and Problem 35). The Mack–Wolfe peak-unknown statistic is then given by

$$A^*_{\hat{p}} = \frac{1}{r} \sum_{j \in B} \left[\frac{A_j - E_0(A_j)}{\{\text{var}_0(A_j)\}^{1/2}} \right], \tag{6.47}$$

where A_j (6.31) is the peak-known statistic with the peak at the jth treatment group and $E_0(A_j)$ and $\text{var}_0(A_j)$ are the null expected value and null variance of A_j given by (6.33) and (6.34), respectively. (Thus, $A^*_{\hat{p}}$ is equal to the average of the r standardized peak-known statistics corresponding to peaks at each of the r samples tied for the maximum $U^*_{.q}$. In most cases, $r = 1$ and $A^*_{\hat{p}}$ is equal to the single standardized peak-known statistic with the peak at the indicated treatment group.)

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the peak-unknown umbrella alternatives

$$H_4 : [\tau_1 \leq \cdots \leq \tau_{p-1} \leq \tau_p \geq \tau_{p+1} \geq \cdots \geq \tau_k,$$
$$\text{with at least one strict inequality, for some } p \in \{1, 2, \ldots, k\}]$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } A^*_{\hat{p}} \geq a^*_{\hat{p},\alpha}; \quad \text{otherwise do not reject,} \tag{6.48}$$

where the constant $a^*_{\hat{p},\alpha}$ is chosen to make the type I error probability equal to α. The constant $a^*_{\hat{p},\alpha}$ is the upper α percentile for the null ($\tau_1 = \cdots = \tau_k$) distribution of $A^*_{\hat{p}}$. Comment 36 explains how to obtain the critical value $a^*_{\hat{p},\alpha}$ for k treatments, and sample sizes $n_1, \ldots, n_k$ and available levels of α.

Ties

If there are ties among the N X's, replace $\phi(a, b)$ in the calculation of the associated Mann–Whitney counts U_{uv} or reverse Mann–Whitney counts U_{vu} by $\phi^*(a, b) = 1, \frac{1}{2}, 0$ if $a <, =,$ or $> b$, respectively, so that for each between-sample comparison where there is a tie, the contribution to the appropriate Mann–Whitney or reverse Mann–Whitney count will be $\frac{1}{2}$. After computing the $U_{.q}$'s (6.46) and $A^*_{\hat{p}}$ (6.47) with these modified counts, use procedure (6.48) with this tie-modified value of $A^*_{\hat{p}}$. Note, however, that this test associated with tied X's is only approximately, and not exactly, of the significance level α.

EXAMPLE 6.4 ***Learning Comprehension and Age.***

It is generally believed that the ability to comprehend ideas and learn is an increasing function of age up to a certain point, and then it declines with increasing age. The data in Table 6.11 are values in the range typically obtained on the Wechsler Adult Intelligence Scale (WAIS) by males of various ages. (Actually the averages of the five samples agree with the corresponding age group means in Norman (1966).)

With $k = 5$ and $n_1 = \cdots = n_5 = 3$, we wish to test

$$H_0(6.2) \text{ versus } H_4 : [\tau_1 \leq \cdots \leq \tau_p \geq \tau_{p+1} \geq \cdots \geq \tau_5,$$
$$\text{with at least one strict inequality, for some } p \in \{1, 2, \ldots, 5\}],$$

Table 6.11 The Wechsler Adult Intelligence Scale (WAIS) Values

Age group				
16–19	20–34	35–54	55–69	≥ 70
8.62	9.85	9.98	9.12	4.80
9.94	10.43	10.69	9.89	9.18
10.06	11.31	11.40	10.57	9.27

Source: R. D. Norman (1966).

where the five age groups are numbered as treatments in order of increasing age. For the purpose of illustration, we consider the significance level $\alpha = .0495$. Applying the R command `cUmbrPU(`α`,`$\boldsymbol{n}$`)`, we find `cUmbrPU(.0495, c(3, 3, 3, 3, 3)) =` 2.226; that is, $P_0(A_{\hat{p}}^* \geq 2.226) = .0495$, and, in the notation of (6.48) with $k = 5$ and $n_1 = n_2 = n_3 = n_4 = n_5 = 3$, we have $a_{\hat{p},.0495}^* = 2.226$ and procedure (6.48) reduces to

$$\text{reject } H_0 \text{ if } A_{\hat{p}}^* \geq 2.226.$$

We now illustrate the computations leading to the sample value of $A_{\hat{p}}^*$ (6.47). First, we compute all of the $5(4)/2 = 10$ possible Mann–Whitney statistics, obtaining

$$\begin{aligned}
&U_{12} = 1+3+3 = 7,\ U_{13} = 2+3+3 = 8,\ U_{14} = 1+1+3 = 5,\\
&U_{15} = 0+1+1 = 2,\ U_{23} = 1+2+3 = 6,\ U_{24} = 0+1+2 = 3,\\
&U_{25} = 0+0+0 = 0,\ U_{34} = 0+0+1 = 1,\\
&U_{35} = 0+0+0 = 0,\ U_{45} = 0+1+1 = 2.
\end{aligned}$$

In order to estimate the age group at which WAIS values peak, we next need to compute the combined-samples Mann–Whitney statistics $U_{.q}$ (6.45), for $q = 1, \ldots, 5$. Using the fact that $U_{vu} = n_u n_v - U_{uv}$ (because there are no ties in the data), for $u, v = 1, \ldots, 5$, we find that

$$\begin{aligned}
U_{.1} &= U_{21} + U_{31} + U_{41} + U_{51}\\
&= \{[3(3) - U_{12}] + [3(3) - U_{13}] + [3(3) - U_{14}] + [3(3) - U_{15}]\}\\
&= (9-7) + (9-8) + (9-5) + (9-2) = 14.\\
U_{.2} &= U_{12} + U_{32} + U_{42} + U_{52}\\
&= U_{12} + [3(3) - U_{23}] + [3(3) - U_{24}] + [3(3) - U_{25}]\\
&= 7 + (9-6) + (9-3) + (9-0) = 25,\\
U_{.3} &= U_{13} + U_{23} + U_{43} + U_{53}\\
&= U_{13} + U_{23} + [3(3) - U_{34}] + [3(3) - U_{35}]\\
&= 8 + 6 + (9-1) + (9-0) = 31,\\
U_{.4} &= U_{14} + U_{24} + U_{34} + U_{54}\\
&= U_{14} + U_{24} + U_{34} + [3(3) - U_{45}]\\
&= 5 + 3 + 1 + (9-2) = 16,
\end{aligned}$$

and

$$U_{.5} = U_{15} + U_{25} + U_{35} + U_{45} = 2 + 0 + 0 + 2 = 4.$$

For this study, we have equal sample sizes $n_1 = \cdots = n_5 = 3$. This implies that each of the combined-samples Mann–Whitney statistics has the same null mean and null variance; that is, for $q = 1, \ldots, 5$, we have

$$E_0(U_{.q}) = \frac{3(15-3)}{2} = 18, \quad \text{var}_0(U_{.q}) = \frac{3(15-3)(15+1)}{12} = 48.$$

As a result, for this equal-sample-sizes setting, we do not need to compute the standardized forms $U_{.q}^*$ (6.46), as the treatment group with the largest $U_{.q}$ value will also be the one with the largest $U_{.q}^*$ value (see also Comment 31). Therefore, the third age group (35–54) is estimated to be the unique peak group (i.e., $\hat{p} = 3$ and $r = 1$), because

$$U_{.3} = \max\{U_{.1}, U_{.2}, U_{.3}, U_{.4}, U_{.5}\} = 31.$$

The Mack–Wolfe peak-unknown statistic $A_{\hat{p}}^*$ (6.47) with $r = 1$ and $\hat{p} = 3$ becomes

$$A_{\hat{p}}^* = \frac{A_3 - E_0(A_3)}{\{\text{var}_0(A_3)\}^{1/2}}.$$

Using the computational formula (6.35) for the peak-known setting in Section 6.3A, we obtain

$$A_3 = 45, \quad E_0(A_3) = 27, \quad \text{var}_0(A_3) = 58.5,$$

which yields

$$A_{\hat{p}}^* = \frac{45 - 27}{\sqrt{58.5}} = 2.353.$$

As this value is greater than the critical value $a_{\hat{p},.0495}^* = 2.226$, we reject H_0 at the .0495 level and conclude that there is sufficient evidence in support of the claim that the ability to comprehend ideas and learn is an increasing function of age up through the age group 35–54, from which point it declines with further age. In fact, from the observed value $A_{\hat{p}}^* = 2.353$, we see, using the R command `pUmbrPU(wechsler)`, that $P_0(A_{\hat{p}}^* \geq 2.353) =$ `pUmbrPU(wechsler)` $=.034$. Thus, the smallest significance level at which we can reject H_0 in favor of H_4 with the observed value of the test statistic $A_{\hat{p}}^* = 2.353$ is .034.

Comments

30. *Motivation for the Test.* The combined-samples Mann–Whitney statistic $U_{.q}$ represents the number of times an observation from the qth sample exceeds an observation from the other $(k - 1)$ combined samples. If the sample sizes are all equal and $\tau_1 < \tau_2 < \cdots < \tau_{p-1} < \tau_p > \tau_{p+1} > \cdots > \tau_k$, then we would expect $U_{.p}$ to be the largest of the k combined-samples Mann–Whitney statistics. Such an outcome would lead to the selection of the pth treatment as the peak group and to $A_{\hat{p}}^* = [A_p - E_0(A_p)]/\{\text{var}_0(A_p)\}^{1/2}$. In view of Comment 23, this provides partial motivation for the $A_{\hat{p}}^*$ test when we have equal sample sizes (see also Comment 31.)

31. *Equal versus Unequal Sample Sizes.* The number of individual comparisons required to produce the value of $U_{.q}$ (6.45) is $n_q(N - n_q)$. If the sample sizes are not all equal, then we will have differing numbers of comparisons leading to the various $U_{.q}$ values. This leads to the undesirable situation where even under the null hypothesis (H_0) those treatments with more sample observations are more likely to be selected as the estimated peak if we use the $U_{.q}$ statistics directly. One way to address this problem is to first standardize the $U_{.q}$'s by subtracting off their null expected values and then dividing by their null standard derivations. The use of these standardized $U_{.q}^*$ statistics to select the peak results

in each treatment having as nearly as possible an equal chance of being selected as the peak under H_0.

If the sample sizes are all equal, say $n_1 = \cdots = n_k = n^*$, then we have

$$E_0(U_{.q}) = \frac{n^*(N - n^*)}{2} \quad \text{and} \quad \text{var}_0(U_{.q}) = \frac{n^*(N - n^*)(N + 1)}{12}$$

for every $q = 1, \ldots, k$. Thus, in order to obtain the standardized $U^*_{.q}$ in such a setting, we would be subtracting the same quantity from each $U_{.q}$ and dividing each of the resulting differences by the same value. The rank order of the resulting $U^*_{.q}$'s would be identical with the rank order of the original $U_{.q}$'s; that is, if we have equal sample sizes and t is such that

$$U_{.t} = \text{maximum}\{U_{.1}, \ldots, U_{.k}\}$$

then it is also true that

$$U^*_{.t} = \text{maximum}\{U^*_{.1}, \ldots, U^*_{.k}\}.$$

As a result, the standardization to obtain the $U^*_{.q}$'s is not necessary in the case of all equal sample sizes as the $U_{.q}$'s themselves can be directly used to select the peak $\hat{p}$.

32. *More General Setting*. As with the other procedures in this chapter, we could replace Assumptions A1–A3 and H_0 (6.2) for the Mack–Wolfe umbrella procedures (both peak-known and peak-unknown) with the more general null hypothesis that all $N!/(\prod_{j=1}^{k} n_j!)$ assignments of n_1 joint ranks to the treatment 1 observations, n_2 joint ranks to the treatment 2 observations, $\ldots, n_k$ joint ranks to the treatment k observations are equally likely.
33. *Assumptions*. As with the other procedures in this chapter, it is important to point out that for the Mack–Wolfe umbrella procedures (both the peak-known and peak-unknown) the k treatment distributions $F_1, \ldots, F_k$ can differ at most in their locations (medians) (see also Comment 4).
34. *Estimation of the Umbrella Peak*. In situations where there is a unique, single treatment label, say t, for which

$$U^*_{.t} = \text{maximum}\{U^*_{.1}, \ldots, U^*_{.k}\},$$

then $r = 1$ in (6.47) and

$$A^*_{\hat{p}} = \frac{A_t - E_0(A_t)}{\{\text{var}_0(A_t)\}^{1/2}}.$$

In this setting, t also provides us with a point estimator for the unknown peak p (i.e., $\hat{p} = t$).

Pan (1996) developed a distribution-free confidence procedure designed to identify those treatments that yield the optimal effects in a one-way layout with umbrella configuration. It utilizes the theory of U-statistics and isotonic regression to provide a random confidence subset of the treatments that contains all the unknown peaks (optimal treatments) within an umbrella ordering with prespecified confidence level.

35. *Null Mean and Variance of Combined-Samples Mann–Whitney Statistics*. The combined-samples statistic $U_{\cdot q}$ (6.45) can be viewed as a single Mann–Whitney statistic between the q-th sample with n_q observations and the remaining $k-1$ samples combined with $N-n_q$ observations. Thus, from the standard formulas for the null mean and null variance of a Mann–Whitney statistic (see the derivation in Comment 19, e.g., particularly (6.23)), we see that

$$E_0(U_{\cdot q}) = \frac{n_q(N-n_q)}{2} \quad \text{and} \quad \text{var}_0(U_{\cdot q}) = \frac{n_q(N-n_q)(N+1)}{12},$$

which agree with the expressions used in (6.46).

36. *Derivation of the Distribution of $A_{\hat{p}}^*$ under H_0 (No Ties)*. As with the peak-known statistic A_p, the peak-unknown statistic $A_{\hat{p}}^*$ can also be computed from the joint ranking of all $N = \sum_{i=1}^{k} n_i$ observations. Thus, one way to obtain the null distribution of $A_{\hat{p}}^*$ is to follow the method of Comments 6, 17, and 25, namely, to compute the value of $A_{\hat{p}}^*$ for each of the $N!/\left(\prod_{j=1}^{k} n_j!\right)$ equally likely (under H_0) rank assignments. We illustrate the development in the specific case of $k=3$, $n_1=1$, $n_2=2$, and $n_3=1$. The $4!/[1!\ 2!\ 1!] = 12$ possible assignments of the joint ranks 1, 2, 3, and 4 to the three treatments and the associated values of $A_{\hat{p}}^*$ are as follows:

1.	I	II	III	2.	I	II	III
	1	2	4		4	2	1
		3				3	
			$A_{\hat{p}}^* = 1.806$				$A_{\hat{p}}^* = 1.806$
3.	I	II	III	4.	I	II	III
	I	2	3		3	2	1
		4				4	
			$A_{\hat{p}}^* = 0.775$				$A_{\hat{p}}^* = 0.775$
5.	I	II	III	6.	I	II	III
	3	1	4		4	1	3
		2				2	
			$A_{\hat{p}}^* = 0.361$				$A_{\hat{p}}^* = 0.361$
7.	I	II	III	8.	I	II	III
	2	1	4		4	1	2
		3				3	
			$A_{\hat{p}}^* = 1.084$				$A_{\hat{p}}^* = 1.084$
9.	I	II	III	10.	I	II	III
	2	1	3		3	1	2
		4				4	
			$A_{\hat{p}}^* = 0.361$				$A_{\hat{p}}^* = 0.361$
11.	I	II	III	12.	I	II	III
	1	3	2		2	3	1
		4				4	
			$A_{\hat{p}}^* = 1.549$				$A_{\hat{p}}^* = 1.549$

Thus, the null distribution for $A_{\hat{p}}^*$ when $k = 3$, $n_1 = n_3 = 2$, and $n_2 = 1$ is given by

$$P_0\{A_{\hat{p}}^* = 0.361\} = \tfrac{4}{12}, \quad P_0\{A_{\hat{p}}^* = 0.775\} = \tfrac{2}{12}, \quad P_0\{A_{\hat{p}}^* = 1.084\} = \tfrac{2}{12}$$
$$P_0\{A_{\hat{p}}^* = 1.549\} = \tfrac{2}{12}, \quad P_0\{A_{\hat{p}}^* = 1.806\} = \tfrac{2}{12}.$$

The probability, under H_0, that $A_{\hat{p}}^*$ is greater than or equal to 1.549, for example, is therefore

$$P_0\{A_{\hat{p}}^* \geq 1.549\} = P_0\{A_{\hat{p}}^* = 1.549\} + P_0\{A_{\hat{p}}^* = 1.806\} = \tfrac{2+2}{12} = \tfrac{1}{3}.$$

Note that we have derived the null distribution of $A_{\hat{p}}^*$ without specifying the common form (F) of the underlying distribution function for the X's under H_0 beyond the requirement that it be continuous. This is why the test procedure (6.48) based on $A_{\hat{p}}^*$ is called a *distribution-free procedure*. From the null distribution of $A_{\hat{p}}^*$, we can determine the critical value $a_{\hat{p},\alpha}^*$ and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying continuous X distribution.

For a given number of treatments k and sample sizes $n_1, \ldots, n_k$, the R command `cUmbrPU`$(\alpha, \boldsymbol{n})$ can be used to find the available upper-tail critical values $a_{\hat{p},\alpha}^*$ for possible values of $A_{\hat{p}}^*$. For a given available significance level α, the critical value $a_{\hat{p},\alpha}^*$ then corresponds to $P_0(A_{\hat{p}}^* \geq a_{\hat{p},\alpha}^*) = \alpha$ and is given by `cUmbrPU`$(\alpha, \boldsymbol{n})$. Thus, for example, for $k = 5$, $n_1 = 3$, $n_2 = 2$, $n_3 = 4$, $n_4 = 3$, and $n_5 = 3$, we have $P_0(A_{\hat{p}}^* \geq 2.216) = .0483$, so that $a_{\hat{p},.0483}^* = 2.216$ for $k = 5$, $n_1 = 3$, $n_2 = 2$, $n_3 = 4$, $n_4 = 3$, and $n_5 = 3$.

37. *Powers of the Mack–Wolfe Umbrella Tests.* The Mack–Wolfe unknown-peak umbrella procedure (6.48) based on $A_{\hat{p}}^*$ is generally much superior to the Kruskal–Wallis procedures in (6.6) and (6.7) when the treatment effects do, indeed, follow an umbrella pattern. When the peak is known a priori to be at treatment p, then the peak-known test (6.32) based on A_p has even better power properties. However, if there is serious uncertainty concerning the location of the true peak, the $A_{\hat{p}}^*$ procedure is preferable because the power of the A_p test can be somewhat diminished when p is not the correct peak. Mack and Wolfe (1981) presented the results of a small-sample power study comparing the relative performances of the Kruskal–Wallis, the Jonckheere–Terpstra, and the two Mack–Wolfe procedures for settings where umbrella alternatives pertain.

38. *Inverted Umbrella Alternatives.* The Mack–Wolfe procedures in this section can easily be adapted to provide tests for "inverted umbrella" alternatives of the form $\tau_1 \geq \cdots \geq \tau_{p-1} \geq \tau_p \leq \tau_{p+1} \leq \cdots \leq \tau_k$, with at least one strict inequality, for both p-known and p-unknown situations. To test for such inverted umbrella alternatives, simply redefine the peak-known statistics to be

$$A_p = \sum_{u=1}^{v-1} \sum_{v=2}^{p} U_{vu} + \sum_{u=p}^{v-1} \sum_{v=p+1}^{k} U_{uv}, \quad \text{for } p = 1, \ldots, k,$$

and for the peak-unknown case, redefine the peak selectors to be "valley" selectors of the form

$$U_{q\cdot} = \sum_{i \neq q} U_{qi}, \quad q = 1, 2, \ldots, k.$$

Everything else remains unchanged, including the necessary null distribution tables.

39. *k-Sample Behrens–Fisher Problem*. Two of the implicit requirements associated with Assumptions A1–A3 are that the underlying distributions belong to the same common family (F) and that they differ within this family at most in their medians. The less restrictive setting where these assumptions are relaxed to permit the possibility of differences in scale parameters as well as medians within the common F is referred to as the *k-sample Behrens–Fisher problem*. The Mack–Wolfe peak-unknown procedure (6.48) is no longer distribution-free under this more relaxed Behrens–Fisher setting. Chen and Wolfe (1990a) proposed a modification of the Mack–Wolfe statistic A_p^* (6.47) to deal with this less restrictive setting. Their approach is similar to that used by Rust and Fligner (1984) to modify the Kruskal–Wallis statistic H for the same setting (see Comment 10).

40. *Ordered versus Umbrella Alternatives*. In this section and Section 6.2, we have considered procedures for testing the null hypothesis H_0 (6.2) of no differences in treatment effects against either ordered or, more generally, umbrella alternatives. In some settings, however, what is actually of interest is the ability to distinguish *directly between* a strictly upward trend (ordered alternatives) and an early upward trend with an eventual downturn (umbrella alternatives). This is frequently the case with dose–response data. Simpson and Margolin (1986) proposed a recursive procedure based on the Jonckheere–Terpstra statistic for dealing with such problems.

41. *An Alternative Approach Based on Maximums*. The Mack–Wolfe approach to the setting of umbrella alternatives with unknown peak is to first use the data to estimate the unknown peak and then to base the test of H_0 (6.2) on the peak-known statistic with peak at this estimated value. An alternative approach would be to bypass the first step of estimating the unknown peak and simply assess directly which of the treatments provides the most evidence of an umbrella alternative. To this effect, Chen and Wolfe (1990b) studied competitor test procedures to procedure (6.48) based on the extreme statistic $A_{\max} = \max\{A_1^*, \ldots, A_k^*\}$, with A_p^* given by (6.35). Hettmansperger and Norton (1987) considered similar competitors to (6.48) based on the maximum of certain linear rank statistics. The results of a substantial small-sample power study of these competitors (as well as the Simpson and Margolin (1986) procedure mentioned in Comment 40) are provided in Chen and Wolfe (1990b).

Problems

33. Consider the tiger muskellunge data in Table 6.9. Test the hypothesis of no differences in the plasma glucose values over time against a general umbrella alternative using an approximate significance level of .01. Compare your result with that obtained in Problem 20.

34. Consider the fasting metabolic rate (FMR) data on white-tailed deer in Table 6.8. Test the hypothesis of no difference in FMR over the 2-month periods against a general umbrella alternative. Use an approximate significance level of .01. Compare your result with that obtained in Example 6.3.

35. (a) The statistic $A^*_{\hat{p}}$ can be computed from the joint ranking of all N observations. Explain.
(b) The statistic $A^*_{\hat{p}}$ can also be computed from pairwise two-sample rankings. Explain.

36. Suppose $k = 3, n_1 = n_2 = 1$, and $n_3 = 2$. Obtain the form of the exact null (H_0) distribution of $A^*_{\hat{p}}$ for the case of no tied observations. Compare this null distribution with that of $A^*_{\hat{p}}$ for $k = 3,\ n_1 = n_3 = 1$, and $n_2 = 2$, as obtained in Comment 36.

37. Construct a set of data with no tied observations for which $r > 1$ in the definition of $A^*_{\hat{p}}$ (6.47). Discuss the implications this has for estimation of the umbrella peak.

38. Consider the Acid Red 114 revertant colonies data in Table 6.10. Test the hypothesis of no differences in the number of revertant colonies over the dosage levels against a general umbrella alternative. Use a significance level of .05. Compare this result with those obtained in Problems 29, 30, 31, and 32.

6.4 A DISTRIBUTION-FREE TEST FOR TREATMENTS VERSUS A CONTROL (FLIGNER–WOLFE)

In this section, we discuss a test procedure specifically designed for the setting where one of the treatments corresponds to a control or baseline set of conditions and we are interested in assessing which, if any, of the treatments is better than the control. Without loss of generality, we label the treatments so that the control corresponds to treatment 1. In this setting, the null hypothesis of interest is still H_0 (6.2), but now it corresponds to the statement that none of the treatments $2, \ldots, k$ is different from the control (treatment 1). This is usually expressed as

$$H_0 : [\tau_i = \tau_1,\ i = 2, \ldots, k]. \tag{6.49}$$

(Note that the expression in (6.49) is, indeed, equivalent to the original expression for H_0 (6.2).)

Procedure

To compute the Fligner–Wolfe statistic FW, we first combine all N observations from the k samples and order them from least to greatest. Letting r_{ij} denote the rank of X_{ij} in this joint ranking, the Fligner–Wolfe statistic FW is then the sum of these joint ranks for the noncontrol treatments, namely,

$$\text{FW} = \sum_{j=2}^{k} \sum_{i=1}^{n_j} r_{ij}. \tag{6.50}$$

a. *One-Sided Upper-Tail Test.* To test

$$H_0 : [\tau_i = \tau_1,\ \text{for } i = 2, \ldots, k]$$

versus

$$H_5 : [\tau_i \geq \tau_1,\ \text{for } i = 2, \ldots, k,\ \text{with at least one strict inequality}], \tag{6.51}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if FW} \geq f_\alpha; \text{ otherwise do not reject,} \tag{6.52}$$

where the constant f_α is chosen to make the type I error probability equal to α. In order to determine the critical value f_α, we note that the statistic FW can be viewed as a two-sample Wilcoxon rank sum statistic (see Section 4.1) computed for the n_1 control treatment observations (playing the role of the X's in the two-sample setting) and the $N^* = \sum_{j=2}^{k} n_j$ combined observations from treatments $2, \ldots, k$ (playing the role of the Y's in the two-sample setting). As a result, the null distribution of FW is the same as that of the Wilcoxon rank sum statistic W with sample sizes $m = n_1$ and $n = N^*$. Thus, the critical value f_α is just the upper αth percentile w_α for the null distribution of the Wilcoxon rank sum statistic with sample sizes $m = n_1$ and $n = N^*$. Values of $f_\alpha = w_\alpha$ in this case can be obtained using the `R` command `pwilcom`, as indicated in Comment 4.3.

b. *One-Sided Lower-Tail Test.* To test

$$H_0 : [\tau_i = \tau_1, \text{ for } i = 2, \ldots, k]$$

versus

$$H_6 : [\tau_i \leq \tau_1, \text{ for } i = 2, \ldots, k, \text{ with at least one strict inequality}], \tag{6.53}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if FW} \leq N^*(N+1) - f_\alpha; \quad \text{otherwise do not reject.} \tag{6.54}$$

Large-Sample Approximation

As previously noted, when H_0 is true, the statistic FW has the same probability distribution as the null distribution of the two-sample Wilcoxon rank sum statistic W with sample sizes $m = n_1$ and $n = N^*$. Hence, it follows directly from the Large-Sample Approximation discussion of Section 4.1 that the standardized version of FW, namely,

$$\text{FW}^* = \frac{\text{FW} - E_0(\text{FW})}{\{\text{var}_0(\text{FW})\}^{1/2}} = \frac{\text{FW} - \{N^*(N+1)/2\}}{\{n_1 N^*(N+1)/12\}^{1/2}} \tag{6.55}$$

has, as $\min(n_1, N^*)$ tends to infinity, an asymptotic $N(0,1)$ distribution when H_0 is true. The normal theory approximation for procedure (6.52) is

$$\text{Reject } H_0 \text{ if FW}^* \geq z_\alpha; \quad \text{otherwise do not reject,} \tag{6.56}$$

and the normal theory approximation for procedure (6.54) is

$$\text{Reject } H_0 \text{ if FW}^* \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{6.57}$$

Ties

If there are ties among the X's, assign each of the observations in a tied group the average of the integer ranks that are associated with the tied group. After computing FW with these average ranks, use procedure (6.52) or (6.54) with this tie-averaged value of FW. Note, however, that this test associated with tied X's is only approximately, and not exactly, of the significance level α. (To get an exact level α test even in this tied setting, see Comment 45.)

When applying the large-sample approximation, an additional factor must be taken into account. Although ties in the X's do not affect the null expected value of FW, its null variance is reduced to

$$\text{var}_0(\text{FW}) = \frac{n_1 N^*}{12}\left[N + 1 - \frac{\sum_{j=1}^{g} t_j(t_j-1)(t_j+1)}{N(N-1)}\right], \tag{6.58}$$

where g denotes the number of tied groups and t_j is the size of the tied group j. We note that an united observation is considered to be a tied group of size 1. In particular, if there are no ties among the X's, then $g = N$ and $t_j = 1$ for $j = 1, \ldots, N$. In this case, each term in (6.58) of the form $t_j(t_j - 1)(t_j + 1)$ reduces to zero and the variance expression in (6.58) reduces to the usual null variance of FW when there are no ties, as given previously in (6.55). Note that the term $[n_1 N^*/12N(N-1)]\sum_{j=1}^{g} t_j(t_j - 1)(t_j + 1)$ represents the reduction in the null variance of FW due to the presence of the tied X's.

As a consequence of the effect that ties have on the null variance of FW, the following modification is needed to apply the large-sample approximation when there are tied X's. Compute FW using average ranks and set

$$\text{FW}^* = \frac{\text{FW} - \left\{\frac{N^*(N+1)}{2}\right\}}{\{\text{var}_0(\text{FW})\}^{1/2}}, \tag{6.59}$$

where $\text{var}_0(\text{FW})$ is now given by display (6.58). With this modified value of FW*, approximation (6.56) or (6.57) can be applied.

EXAMPLE 6.5 ***Motivational Effect of Knowledge of Performance—Example 6.2 Continued.***

For Hundal's (1969) study to assess the motivational effects of knowledge of performance, the no information category clearly serves as a control population, and it is very natural to ask if additional performance information of either type (rough or accurate) leads to improved performance as measured by an increase in the number of pieces processed. Thus, we will apply the Fligner–Wolfe procedure (6.51) to the data in Table 6.6 to assess whether there is a deviation from H_0 in the direction of $\tau_2 > \tau_1$ and/or $\tau_3 > \tau_1$, where the treatment numbers are the same as those taken in Example 6.2. For the purpose of illustration, we take the significance level to be $\alpha = .0415$. With $m = N^* = n_2 + n_3 = 6 + 6 = 12$ and $n = n_1 = 6$, we find using the `R` command `pwilcox` (see Comment 4.3) that $f_{.0415} = 133$ and procedure (6.52) reduces to

$$\text{Reject } H_0 \text{ if } \text{FW} \geq 133.$$

Using the joint ranks provided in parentheses beside the data in Table 6.6, we see that

$$\text{FW} = [2.5 + 5.5 + 17 + 13 + 5.5 + 9 + 18 + 5.5 + 15 + 10.5 + 16 + 13] = 130.5.$$

As this value of FW is smaller than the critical value 133, we do not reject H_0 at the .0415 level. (This example illustrates the added power of the Jonckheere–Terpstra procedure relative to that of the Fligner–Wolfe procedure when we are able to utilize the additional piece of information that $\tau_3 \geq \tau_2$. From Example 6.2, the P-value for the Jonckheere–Terpstra procedure applied to these Hundal data is .0231, indicating rejection of H_0 at $\alpha = .0415$.)

For the large-sample approximation, we need to compute the standardized form of FW* using (6.59) because there are ties in the data. The null expected value for FW is $E_0(\text{FW}) = 12(18 + 1)/2 = 114$. For the ties-corrected null variance of FW, we note that $g = 11$ and $t_1 = 1,\ t_2 = 2,\ t_3 = 4,\ t_4 = 1,\ t_5 = 1,\ t_6 = 2,\ t_7 = 3,\ t_8 = 1,\ t_9 = 1,\ t_{10} = 1$, and $t_{11} = 1$ for the Hundal data. Hence, using the ties correction in (6.58), we have that

$$\begin{aligned}\text{var}_0(\text{FW}) &= \frac{6(12)}{12}\left\{18 + 1 - \left[\frac{2(2)(1)(3) + 3(2)(4) + 4(3)(5)}{18(17)}\right]\right\} \\ &= 6\left(19 - \frac{16}{51}\right) = 112.12,\end{aligned}$$

from which it follows that the ties-corrected value of FW* (6.59) is

$$\text{FW}^* = \frac{130.5 - 114}{\{112.12\}^{1/2}} = 1.56.$$

Thus, using the approximate procedure (6.56) with the ties-corrected value of $\text{FW}^* = 1.56$, we see that the approximate P-value for these data is $P_0(\text{FW}^* \geq 1.56) \approx 1 - \texttt{pnorm(1.56)} = 1 - .9406 = .0594$. Thus, we have marginal evidence from the Fligner–Wolfe treatments-versus-control procedure that additional performance knowledge (either rough or accurate) leads to an increase in the number of pieces produced.

Comments

42. *More General Setting*. As with the other procedures of this chapter, we could replace Assumptions A1–A3 and H_0 (6.2) with the more general null hypothesis that all $N!/\left(\prod_{j=1}^{k} n_j!\right)$ assignments of n_1 joint ranks to the control observations, n_2 joint ranks to the treatment 2 observations, $\ldots, n_k$ joint ranks to the treatment k observations are equally likely.
43. *Motivation for the Test*. The statistic FW (6.50) is the sum of the joint ranks assigned to the noncontrol treatments. When some of the τ_i's are strictly greater than the control effect τ_1, we would expect the joint ranks for the observations from those treatments to be larger than the joint ranks for the control observations. The net result would be a larger value of FW. This suggests rejecting H_0 in favor of H_5 (6.51) for large values of FW and motivates procedures (6.52) and (6.56). A similar motivation leads to procedures (6.54) and (6.57). (See also Comment 47.)

44. *Assumptions*. As with the other test procedures of this chapter, Assumption A3 requires that the control and the $(k-1)$ treatment distributions $F_1, \ldots, F_k$ can differ at most in their locations (medians). (See also Comments 4 and 50.)

45. *Exact Conditional Distribution of FW with Ties among the Data*. To get an exact level α test in the presence of ties, we rely on the fact that the null distribution of FW conditional on the observed configuration of joint tied ranks is the same as the corresponding conditional tied ranks null distribution of the Wilcoxon rank sum statistic W with sample sizes $m = n_1$ and $n = N^*$. Therefore, the approach discussed and illustrated in Comment 4.5 can be used to get the exact conditional null distribution of FW and associated exact level α test in the case of ties among the data.

46. *Two-Sided Test*. We note that we have not discussed a test based on the FW statistic that is designed for a two-sided alternative. The "natural" two-sided alternative for this treatment versus control setting corresponds to [either $\tau_i \geq \tau_1$ for all $i = 2, \ldots, k$ or $\tau_i \leq \tau_1$ for all $i = 2, \ldots, k$, with at least one strict inequality]. We feel that it is rather unlikely that we would find ourselves in such a setting where either all the treatments are better than the control or all the treatments are worse than the control, but we have no idea which of the two cases pertains. As a result, a two-sided test based on FW is not presented in this section.

47. *Limitations*. The test procedures in (6.52) and (6.54) deal with very restricted alternatives where *all* the treatments are either at least as good as the control (i.e., $\tau_i \geq \tau_1$ for all $i = 2, \ldots, k$) or *all* the treatments are no better than the control (i.e., $\tau_i \leq \tau_1$ for all $i = 2, \ldots, k$), respectively. They are not appropriate tests when the possibility exists that some of the treatments might be better ($\tau_i > \tau_1$) and some might be worse ($\tau_i < \tau_1$) than the control. For such mixed alternatives, one would need to use the general alternatives Kruskal–Wallis procedure presented in Section 6.1.

48. *Comparisons Between Treatments*. The Fligner–Wolfe procedure (6.52) is a test designed to decide if *any* of treatments $2, \ldots, k$ are better (i.e., $\tau_i > \tau_1$) than the control. It involves no direct comparisons between the various treatments observations themselves. In order to reach conclusions about whether there are any differences among the treatment effects $\tau_2, \ldots, \tau_k$, one would need to apply the Kruskal–Wallis procedure of Section 6.1 (or, if appropriate, the Jonckheere–Terpstra ordered alternatives or Mack–Wolfe umbrella alternatives procedures of Sections 6.2 and 6.3, respectively) to the sample data from treatments $2, \ldots, k$. Under the null hypothesis H_0 (6.2), the Fligner–Wolfe statistic FW is independent of the Kruskal–Wallis statistic H (and also of the Jonckheere–Terpstra statistic J and the Mack–Wolfe statistics A_p and A_p^*). This implies, for example, that if we conduct the Fligner–Wolfe test (6.52) at a significance level α_1 and the Kruskal–Wallis test (6.6) (or the Jonckheere–Terpstra test (6.14), Mack–Wolfe peak-known test (6.32), or Mack–Wolfe peak-unknown test (6.48)) on treatments $2, \ldots, k$ at the significance level α_2, then the probability of incorrectly rejecting H_0 when it is true with at least one of the two tests is exactly $\alpha_1 + \alpha_2 - \alpha_1\alpha_2$. A similar comment applies to procedure (6.54).

49. *Multiple Comparisons*. If test procedure (6.52) leads to rejection of H_0 (6.2), we are led to the conclusion that at least one treatment has a greater effect than the control. However, procedure (6.52) does not address the question of exactly how many treatment effects are greater than that of the control, or does

it provide us information as to which specific treatments are better than the control. For answers to such questions, we turn to treatments-versus-control multiple comparison procedures, as discussed in Sections 6.7 and 6.8. Similar comments apply to the lower-tail test procedure in (6.54).

50. *The Treatments-versus-Control Behrens–Fisher Problem.* Two of the implicit requirements imposed by Assumptions A1–A3 are that the underlying distributions belong to the same common family (F) and that they differ within this family at most in their medians. The less restrictive setting where these assumptions are relaxed to permit the possibility of differences in scale parameters as well as medians within the common family F is referred to as the *k-sample treatments*-versus-control Behrens–Fisher problem. The Fligner–Wolfe procedures (6.52) and (6.54) are no longer distribution-free under this more general Behrens–Fisher setting. If we replace Assumption A3 by the less restrictive Assumption A3*: [The treatments' distribution functions $F_2, \ldots, F_k$ are connected through the relationship

$$F_i(t) = F^*(t - \tau_j), \quad -\infty < t < \infty,$$

for $i = 2, \ldots, k$, where F^* is a distribution function for a continuous distribution that is symmetric about its median θ and, in addition, the control distribution (F_1) is continuous and symmetric about its median $\theta + \tau_1$.], then the Fligner–Policello two-sample robust rank procedure discussed in Section 4.4 can be adapted to provide distribution-free tests of H_0 (6.2) against either H_5 (6.51) or H_6 (6.53) under these more general treatments-versus-control Behrens–Fisher Assumptions A1, A2, and A3*.

51. *Treatments-versus-Control under Umbrella Configurations.* In many settings where we are interested in comparing a number of treatments with a control, we will have additional a priori information regarding the relative magnitude of the treatment effects. One such piece of information might be that the treatment effects are known to follow an umbrella pattern (see Section 6.3) $\tau_1 \leq \cdots \leq \tau_{p-1} \leq \tau_p \geq \tau_{p+1} \geq \cdots \geq \tau_k$ with either known or unknown peak p. (Remember that the ordered pattern of Section 6.2 corresponds to $p = k$ or 1.) In a drug study, for instance, increasing dosage levels may be compared with a zero-dose control. If the treatment effects are not identical to that of the control, then it is often reasonable to assume that the higher the dose of the drug applied, the better (say, higher) will be the resulting effect on a patient, corresponding to monotonically ordered treatment effects. However, it may also be the case that a subject might potentially succumb to toxic effects at high doses, thereby actually decreasing the associated treatment effects. Such a setting would correspond to an ordering in the treatment effects that is monotonically increasing up to a point, followed by a monotonic decrease; that is, an umbrella pattern on the treatment effects. Chen and Wolfe (1993) considered a test procedure designed specifically to compare a number of treatments with a single control under this basic umbrella pattern for the treatment effects. Their test requires an equal number of observations in each of the treatments (i.e., $n_2 = \cdots = n_k = n^*$), but permits a differing number (n_1) of observations from the control setting. The necessary null distribution critical values are provided for a variety of k, n_1, and n^* combinations, and the results of a substantial Monte Carlo simulation power study are presented.

(An example of the type of data for which this Chen–Wolfe procedure would be appropriate is provided by the muskellunge plasma glucose data in Table 6.9.)

52. *Consistency of the FW test*. Replace Assumptions A1–A3 by the less restrictive Assumptions A1′: The X's are mutually independent and A2′ : $X_{1j}, \ldots, X_{n_j j}$ come from the same continuous population Π_j, $j = 1, \ldots, k$. The populations $\Pi_1, \ldots, \Pi_k$ need not be identical, but we do assume that

$$\delta_{ij} = P(X_{1j} > X_{11}) \geq \tfrac{1}{2}, \quad \text{for } j = 2, \ldots, k.$$

Then, roughly speaking, the test defined by (6.52) is consistent if and only if there is at least one $j \in \{2, 3 \ldots, k\}$ for which $\delta_{1j} > \frac{1}{2}$.

Properties

1. *Consistency*. The condition n_j/N tends to λ_j, $0 < \lambda_j < 1$, $j = 1, \ldots, k$, is sufficient to ensure that the tests defined by (6.52) and (6.54) are consistent against the H_5 (6.51) and H_6 (6.53) alternatives, respectively. For a more general consistency statement, see Comment 52.
2. *Asymptotic Normality*. See Fligner and Wolfe (1982).
3. *Efficiency*. See Fligner and Wolfe (1982) and Section 6.10.

Problems

39. Apply the appropriate Fligner–Wolfe test to the psychotherapeutic attraction data of Table 6.2. Compare and contrast this result with that obtained for the Kruskal–Wallis test in Problem 1.

40. Apply the appropriate Fligner–Wolfe procedure to the glucocorticoid receptor data for the leukemia patients in Table 6.4, using the normal subjects as the control. Compare and contrast with the result obtained from the Kruskal–Wallis test in Problem 8.

41. Apply the appropriate Fligner–Wolfe test to the muskellunge plasma glucose data in Table 6.9. Compare and contrast with the result obtained from the Mack–Wolfe test in Problem 20. (See also Comment 51.)

RATIONALE FOR MULTIPLE COMPARISON PROCEDURES

In Sections 6.1–6.4 of this chapter, we have discussed procedures designed to test the null hypothesis H_0 (6.2) against a variety of alternative hypotheses. Upon rejection of H_0 with one of these test procedures for a given set of data, our conclusions range from the general statement that there are some unspecified differences among the treatment effects (associated with the Kruskal–Wallis procedure discussed in Section 6.1) to the more informative relationships between the treatment effects associated with test procedures designed for the ordered or umbrella alternatives or the treatments-versus-control setting. However, in none of these test procedures are our conclusions pair-specific; that is, the tests in Sections 6.1–6.4 are not designed to enable us to reach conclusions about specific pairs of treatment effects. The relative sizes of the specific treatment effects τ_1 and τ_2, for example, cannot be inferred from the conclusions reached by any of the test procedures in Sections 6.1–6.4. To elicit such pairwise specific information, we turn to the class of multiple comparison procedures. In Section 6.5 we present such two-sided

all-treatments multiple comparison procedures for the omnibus setting corresponding to the general alternatives H_1 (6.3). In Section 6.6 we deal with one-sided all-treatments multiple comparison procedures associated with the restricted ordered alternatives H_2 (6.11). Finally, in Section 6.7 we discuss an approach for making treatments-versus-control multiple comparison decisions.

6.5 DISTRIBUTION-FREE TWO-SIDED ALL-TREATMENTS MULTIPLE COMPARISONS BASED ON PAIRWISE RANKINGS—GENERAL CONFIGURATION (DWASS, STEEL, AND CRITCHLOW–FLIGNER)

In this section we present a multiple comparison procedure based on pairwise two-sample rankings that is designed to make decisions about individual differences between pairs of treatment effects (τ_i, τ_j), for $i < j$, in a setting where general alternatives H_1 (6.3) are of interest. Thus, the multiple comparison procedure of this section would generally be applied to one-way layout data *after* rejection of H_0 (6.2) with the Kruskal–Wallis procedure from Section 6.1. In this setting, it is important to reach conclusions about all $\binom{k}{2} = k(k-1)/2$ pairs of treatment effects, and these conclusions are naturally two-sided in nature.

Procedure

For each pair of treatments (i, j), let

$$W_{ij} = \sum_{b=1}^{n_j} R_{ib}, \quad \text{for } 1 \le i < j \le k, \tag{6.60}$$

where $R_{i1}, \ldots, R_{in_j}$ are the ranks of $X_{1j}, \ldots, X_{n_j j}$, respectively, among the combined ith and jth samples; that is, W_{ij} is the Wilcoxon rank sum of the jth sample ranks in the joint two-sample ranking of the ith and jth sample observations. Compute

$$W_{ij}^* = \sqrt{2}\left[\frac{W_{ij} - E_0(W_{ij})}{\{\text{var}_0(W_{ij})\}^{1/2}}\right] = \frac{W_{ij} - \frac{n_j(n_i+n_j+1)}{2}}{\{n_i n_j (n_i + n_j + 1)/24\}^{1/2}}, \quad \text{for } 1 \le i < j \le k. \tag{6.61}$$

(Thus, W_{ij}^* is the standardized (under H_0) version of W_{ij} multiplied by $\sqrt{2}$.)

At an experimentwise error rate of α, the Steel–Dwass–Critchlow–Fligner two-sided all-treatments multiple comparison procedure reaches its $k(k-1)/2$ pairwise decisions, corresponding to each (τ_u, τ_v) pair $1 \le u < v \le k$, by the criterion

$$\text{Decide } \tau_u \ne \tau_v \text{ if} |W_{uv}^*| \ge w_\alpha^*; \quad \text{otherwise decide } \tau_u = \tau_v, \tag{6.62}$$

where the constant w_α^* is chosen to make the experimentwise error rate equal to α; that is, w_α^* satisfies the restriction

$$P_0(|W_{uv}^*| < w_\alpha^*, u = 1, \ldots, k-1; v = u+1, \ldots, k) = 1 - \alpha, \tag{6.63}$$

where the probability $P_0(\cdot)$ is computed under H_0 (6.2). Equation (6.63) stipulates that the $k(k-1)/2$ inequalities $|W_{uv}^*| < w_\alpha^*$, corresponding to all pairs (u, v) of treatments

with $u < v$, hold simultaneously with probability $1 - \alpha$ when H_0 (6.2) is true. Comment 55 explains how to obtain the critical value w_α^* for k treatments, sample sizes $n_1, \ldots, n_k$, and available experimentwise error rates α.

Large-Sample Approximation

When H_0 is true, the $[k(k-1)/2]$-component vector $(W_{12}^*, W_{13}^*, \ldots, W_{k-1,k}^*)$ has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic multivariate normal distribution with mean vector $\mathbf{0}$. It then follows (see Comment 58 for indications of the proof) that w_α^* can be approximated for large-sample sizes by q_α, where q_α is the upper αth percentile point for the distribution of the range of k independent $N(0,1)$ variables. Thus, the large-sample approximation for procedure (6.62) is

$$\text{Decide } \tau_u \neq \tau_v \text{ if } |W_{uv}^*| \geq q_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v. \tag{6.64}$$

To find q_α for k treatments, we use the R command `cRangekNorm`(α, k). For example, to find $q_{.05}$ for $k = 6$ treatments, we apply `cRangekNorm(.05, 6)` and obtain $q_{.05} = 4.031$.

Ties

If there are ties among the X observations, use average ranks in computing the individual Wilcoxon rank sum statistics W_{ij} (6.60). In addition, replace the term $\text{var}_0(W_{ij})/2 = n_i n_j (n_i + n_j + 1)/24$ in the denominator of W_{ij}^* (6.61) by

$$\frac{\text{var}_0(W_{ij})}{2} = \frac{n_i n_j}{24}\left[n_i + n_j + 1 - \frac{\sum_{b=1}^{g_{ij}} (t_b - 1)t_b(t_b + 1)}{(n_i + n_j)(n_i + n_j - 1)}\right], \tag{6.65}$$

where, for $1 \leq i < j \leq k$, g_{ij} denotes the number of tied groups in the joint ranking of the ith and jth sample observations and t_b is the size of tied group b in this joint ranking. Furthermore, an untied observation is considered to be a tied group of size 1. In particular, if there are no tied observations in the ith and jth combined samples, then $g_{ij} = n_i + n_j$ and $t_b = 1$ for $b = 1, \ldots, n_i + n_j$, in which case each term of the form $(t_b - 1)t_b(t_b + 1)$ reduces to 0 and $\text{var}_0(W_{ij})/2$ reduces to $n_i n_j (n_i + n_j + 1)/24$, the appropriate expression when there are no ties in the ith and jth combined samples.

EXAMPLE 6.6 ***Length of YOY Gizzard Shad.***

Consider the length of YOY gizzard shad data discussed in Problem 4. Applying the Kruskal–Wallis procedure to the length data from the four sites in Kokosing Lake yields highly significant differences between the median YOY lengths at the four sites. To examine which particular sites differ in median YOY lengths, we apply the approximate procedure (6.64) with the appropriate corrections for ties given in (6.65). For this study, we have $k = 4$, $n_1 = n_2 = n_3 = n_4 = 10$, and we must compute $k(k-1)/2 = 4(3)/2 = 6$ standardized W_{ij}^* statistics. For the sake of illustration, we take our experimentwise error

rate to be $\alpha = .01$. With $k = 4$, we find $q_{.01} =$ `cRangekNorm(.01, 4)` = 4.404 and procedure (6.64) reduces to

$$\text{Decide } \tau_u \neq \tau_v \text{ if } |W_{uv}^*| \geq 4.404.$$

Next, we compute the six W_{ij}^* statistics. For the sample observations from sites I and II (populations 1 and 2, respectively), the combined-samples ranking yields the sum of ranks for the site II data to be

$$W_{12} = 9.5 + 20 + 3.5 + 9.5 + 13.5 + 19 + 1 + 17 + 9.5 + 18 = 120.5.$$

For this pair of samples, there are tied observations, and we have $g_{12} = 14$ and $t_1 = t_2 = 1,\ t_3 = 2,\ t_4 = t_5 = t_6 = 1,\ t_7 = 4,\ t_8 = 1,\ t_9 = t_{10} = 2,$ and $t_{11} = t_{12} = t_{13} = t_{14} = 1$. From (6.65), we find

$$\begin{aligned}\frac{\text{var}_0(W_{12})}{2} &= \frac{10(10)}{24}\left[10 + 10 + 1 - \frac{3(1)(2)(3) + (3)(4)(5)}{(10+10)(10+10-1)}\right] \\ &= \frac{25}{6}\left[\frac{7,980 - 78}{380}\right] = 86.64.\end{aligned}$$

Using this result in (6.61), we obtain

$$W_{12}^* = \frac{[120.5 - 10(21)/2]}{\sqrt{86.64}} = 1.67.$$

For the other five population pairs, similar calculations yield the following:
Site I and Site III

$$\begin{aligned}W_{13} &= 13.5 + 11 + 2.5 + 1 + 8 + 8 + 2.5 + 5 + 5 + 5 = 61.5, \\ g_{13} &= 13, \quad t_1 = 1, \quad t_2 = 2, \quad t_3 = t_4 = 3, \quad t_5 = t_6 = t_7 = 1, \quad t_8 = 2, \\ t_9 &= t_{10} = t_{11} = t_{12} = 1, \quad t_{13} = 2, \\ \text{var}_0(W_{13}) &= \frac{10(10)}{24}\left[10 + 10 + 1 - \frac{3(1)(2)(3) + 2(2)(3)(4)}{(10+10)(10+10-1)}\right] = 86.78, \\ W_{13}^* &= \frac{[61.5 - 10(21)/2]}{\sqrt{86.78}} = -4.67.\end{aligned}$$

Site I and Site IV

$$\begin{aligned}W_{14} &= 11 + 9 + 4.5 + 7 + 9 + 2.5 + 2.5 + 1 + 4.5 + 9 = 60, \\ g_{14} &= 15, \quad t_1 = 1, \quad t_2 = t_3 = 2, \quad t_4 = t_5 = 1, \quad t_6 = 3, \\ t_7 &= t_8 = t_9 = t_{10} = t_{11} = t_{12} = t_{13} = t_{14} = 1, \quad t_{15} = 2, \\ \text{var}_0(W_{14}) &= \frac{10(10)}{24}\left[10 + 10 + 1 - \frac{3(1)(2)(3) + 2(3)(4)}{(10+10)(10+10-1)}\right] = 87.04, \\ W_{14}^* &= \frac{[60 - 10(21)/2]}{\sqrt{87.04}} = -4.82.\end{aligned}$$

Site II and Site III

$$W_{23} = 12 + 11 + 2.5 + 1 + 8.5 + 8.5 + 2.5 + 5.5 + 5.5 + 5.5 = 62.5,$$
$$g_{23} = 13, \quad t_1 = 1, \quad t_2 = 2, \quad t_3 = 4, \quad t_4 = 2, \quad t_5 = t_6 = t_7 = 1, \quad t_8 = 3,$$
$$t_9 = t_{10} = t_{11} = t_{12} = t_{13} = 1,$$
$$\text{var}_0(W_{23}) = \frac{10(10)}{24}\left[10 + 10 + 1 - \frac{2(1)(2)(3) + 2(3)(4) + 3(4)(5)}{(10 + 10)(10 + 10 - 1)}\right] = 86.45,$$
$$W_{23}^* = \frac{[62.5 - 10(21)/2]}{\sqrt{86.45}} = -4.57.$$

Site II and Site IV

$$W_{24} = 11 + 9 + 5 + 7 + 9 + 2.5 + 2.5 + 1 + 5 + 9 = 61,$$
$$g_{24} = 13, \quad t_1 = 1, \quad t_2 = 2, \quad t_3 = 3, \quad t_4 = 1, \quad t_5 = 3, \quad t_6 = t_7 = 1, \quad t_8 = 3,$$
$$t_9 = t_{10} = t_{11} = t_{12} = t_{13} = 1,$$
$$\text{var}_0(W_{24}) = \frac{10(10)}{24}\left[10 + 10 + 1 - \frac{1(2)(3) + 3(2)(3)(4)}{(10 + 10)(10 + 10 - 1)}\right] = 86.64,$$
$$W_{24}^* = \frac{[61 - 10(21)/2]}{\sqrt{86.64}} = -4.73.$$

Site III and Site IV

$$W_{24} = 18 + 16 + 9 + 14 + 16 + 3 + 3 + 1 + 9 + 16 = 105,$$
$$g_{34} = 10, \quad t_1 = 1, \quad t_2 = 3, \quad t_3 = 2, \quad t_4 = 5, \quad t_5 = 2, \quad t_6 = 1,$$
$$t_7 = 3, \quad t_8 = t_9 = t_{10} = 1,$$
$$\text{var}_0(W_{34}) = \frac{10(10)}{24}\left[10 + 10 + 1 - \frac{2(1)(2)(3) + 2(2)(3)(4) + 4(5)(6)}{(10 + 10)(10 + 10 - 1)}\right] = 85.53,$$
$$W_{34}^* = \frac{[105 - 10(21)/2]}{\sqrt{85.53}} = 0.$$

Taking absolute values and referring them to the critical value $q_{.01} = 4.403$, we see that

$$|W_{12}^*| = 1.67 < 4.404 \quad \Longrightarrow \quad \text{decide } \tau_1 = \tau_2,$$
$$|W_{13}^*| = 4.67 > 4.404 \quad \Longrightarrow \quad \text{decide } \tau_1 \neq \tau_3,$$
$$|W_{14}^*| = 4.82 > 4.404 \quad \Longrightarrow \quad \text{decide } \tau_1 \neq \tau_4,$$
$$|W_{23}^*| = 4.57 > 4.404 \quad \Longrightarrow \quad \text{decide } \tau_2 \neq \tau_3,$$
$$|W_{24}^*| = 4.73 > 4.404 \quad \Longrightarrow \quad \text{decide } \tau_2 \neq \tau_4,$$
$$|W_{34}^*| = 0 < 4.404 \quad \Longrightarrow \quad \text{decide } \tau_3 = \tau_4.$$

Thus, at an experimentwise error rate of .01, the six multiple comparison decisions can be summarized by the statement $(\tau_1 = \tau_2) \neq (\tau_3 = \tau_4)$. This multiple comparison procedure provides more detailed information about the lengths of the YOY gizzard shad population in Kokosing Lake. We now know that sites I and II may be viewed as providing similar living environments for gizzard shad. The same conclusion holds for sites III and IV. However, we also know that the common living environment at sites I and II is significantly different from the common living environment at sites III and IV.

Comments

53. *Rationale for Multiple Comparison Procedures*. We think of the methods of this section as multiple comparison procedures. The aim of applying such procedures goes beyond the point of deciding whether the treatments are equivalent to the (often more important) problem of selecting which, if any, treatments differ from one another. Thus, the user makes $k(k-1)/2$ decisions, one for each pair of treatments. Equation (6.63) states that the probability of making all correct decisions when H_0 is true is controlled to be $1-\alpha$. That is, when using procedure (6.62), the probability of at least one incorrect decision, when H_0 is true, is controlled to be α. This error rate is derived under the assumption that H_0 is true, but it does not depend on the particular underlying distribution F. This is why we call (6.62) a distribution-free multiple comparison procedure.

    Multiple comparison procedures can be interpreted as hypothesis tests. If we consider the test that rejects H_0 if the inequality of (6.62) holds for at least one (u, v) pair and accepts H_0 if, for every (u, v) pair, the inequality of (6.62) is not satisfied, this is a distribution-free test of size α for H_0 (6.2).

54. *Experimentwise Error Rate*. The use of an experimentwise error rate represents a very conservative approach to multiple comparisons. We are insisting that the probability of making only correct decisions be $1-\alpha$ when the hypothesis H_0 (6.2)) of treatment equivalence is true. Thus, we have a high degree of protection when H_0 is true, but we often apply such techniques when we have evidence (perhaps based on a priori information or perhaps obtained by applying the Kruskal–Wallis test, as in Example 6.6) that H_0 is not true.

    This protection under H_0 also makes it harder for the procedure to judge treatments as differing significantly when in fact H_0 is false, and this difficulty becomes more severe as k increases. We justify our use of an experimentwise error rate in much the same way as Kurtz et al. (1965). The rate provides a precise measure of a level of uncertainty, and statements at higher or lower levels are readily obtained.

    Anscombe (1965), although not advocating the use of such rates, mentioned an interesting hypothetical situation (which he attributed to Richard Olshen) in defense of such a conservative approach. Anscombe was commenting on simultaneous confidence intervals proposed by Kurtz et al., but his statements would also apply to multiple comparison procedures of the type discussed here. We quote from his comments. "A panacea manufacturer advertises on television that trials have shown his product to be more effective than any other leading brand. Such trials (if they are not a downright fabrication) certainly seem to present

a situation of the third type.* Their objective is not to help the manufacturer reach a decision, but hopefully to permit him to make a multiple comparison statement that will impress the public and boost sales. He could appropriately use the simultaneous confidence intervals of this paper; indeed, the Food and Drug Administration could appropriately require him to do so. The more equally ineffective other leading brands there were, the harder would it be for him to obtain the evidence he needed, and the more trials he would have to conduct and suppress before achieving a favorable one. Thus would Statistics and Economics go hand in hand to protect the public."

55. *Critical Values* w_α^*. The w_α^* critical values can be obtained by using the fact that under H_0 (6.2), all $N!/(\prod_{j=1}^k n_j!)$ *joint* (of all N sample observations) rank assignments of n_1 ranks to the treatment 1 observations, n_2 ranks to the treatment 2 observations, $\ldots, n_k$ ranks to the treatment k observations are equally likely. (Although the standardized pairwise Wilcoxon statistics W_{ij}^* (6.61) are formally defined in terms of pairwise two-sample ranks, it is clear that all $k(k-1)/2$ W_{ij}^* values can also be computed from the joint ranks of all N observations.) Thus, to obtain the probability, under H_0, that $|W_{uv}^*| < c$, for all $u < v$, we simply count the number of configurations for which the event $A = \{|W_{uv}^*| < c, \text{ for all } u < v\}$ occurs, and divide this count by $N!/\left[\prod_{j=1}^k n_j!\right]$. For an illustration, we return to Comment 6 and use the 15 joint rank configurations displayed there for the case $k = 3$ and $n_1 = n_2 = n_3 = 2$. (Again, we can reduce the number of configurations that need to be considered from 90 to 15 by the same reasoning as in Comment 6.) For each of these 15 configurations, we now display the values of $|W_{12}^*|$, $|W_{13}^*|$, and $|W_{23}^*|$.

(a)	$\|W_{12}^*\| = 2.1909$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 2.1909$	(b)	$\|W_{12}^*\| = 2.1909$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 1.0954$	(c)	$\|W_{12}^*\| = 2.1909$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 0$
(d)	$\|W_{12}^*\| = 1.0954$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 2.1909$	(e)	$\|W_{12}^*\| = 1.0954$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 1.0954$	(f)	$\|W_{12}^*\| = 1.0954$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 0$
(g)	$\|W_{12}^*\| = 1.0954$ $\|W_{13}^*\| = 1.0954$ $\|W_{23}^*\| = 1.0954$	(h)	$\|W_{12}^*\| = 0$ $\|W_{13}^*\| = 2.1909$ $\|W_{23}^*\| = 2.1909$	(i)	$\|W_{12}^*\| = 1.0954$ $\|W_{13}^*\| = 1.0954$ $\|W_{23}^*\| = 0$
(j)	$\|W_{12}^*\| = 0$ $\|W_{13}^*\| = 1.0954$ $\|W_{23}^*\| = 2.1909$	(k)	$\|W_{12}^*\| = 0$ $\|W_{13}^*\| = 1.0954$ $\|W_{23}^*\| = 1.0954$	(l)	$\|W_{12}^*\| = 1.0954$ $\|W_{13}^*\| = 0$ $\|W_{23}^*\| = 0$
(m)	$\|W_{12}^*\| = 0$ $\|W_{13}^*\| = 0$ $\|W_{23}^*\| = 2.1909$	(n)	$\|W_{12}^*\| = 0$ $\|W_{13}^*\| = 0$ $\|W_{23}^*\| = 1.0954$	(o)	$\|W_{12}^*\| = 0$ $\|W_{13}^*\| = 0$ $\|W_{23}^*\| = 0$

* The term *third type* is used by Anscombe to refer to experiments intended to give fundamental knowledge or insight into some phenomenon but not to aid in a particular job of decision making.

Thus, for example,

$$\begin{aligned} P_0\{|w^*_{uv}| < 2.1909, (u,v) = (1,2),(1,3),(2,3)\} \\ &= P_0\{|W^*_{12}| < 2.1909; |W^*_{13}| < 2.1909; |W^*_{23}| < 2.1909\} \\ &= \tfrac{6}{15} = 1 - .6, \end{aligned}$$

because for 6 of the 15 configurations—[(g), (i), (k), (1), (n), and (o)]—the event $\{|W^*_{12}| < 2.1909; |W^*_{13}| < 2.1909; |W^*_{23}| < 2.1909\}$ occurs. Similarly, $P_0\{|W^*_{uv}| < 1.0954, (u,v) = (1,2),(1,3),(2,3)\} = \frac{1}{15} = 1 - .9333$, as the event $\{|W^*_{12}| < 1.0954; |W^*_{13}| < 1.0954; |W^*_{23}| < 1.0954\}$ occurs only for the single configuration (o). Hence, for $k = 3$ and $n_1 = n_2 = n_3 = 2$, we have $w^*_{.6000} = 2.1909$ and $w^*_{.9333} = 1.0954$, and the values .6000 and .9333 are the only available experimentwise error rates for the Dwass–Steel–Critchlow–Fligner procedure (6.62) in this setting.

For a given number of treatments k and sample sizes $n_1, \ldots, n_k$, the R command `cSDCFlig`$(\alpha, \boldsymbol{n})$ can be used to find the available critical values w^*_α. For a given available experimentwise error rate α, the critical value w^*_α is given by `cSDCFlig`$(\alpha, \boldsymbol{n})$. Thus, for example, for $k = 3$ and $n_1 = 3$, $n_2 = 5$, and $n_3 = 7$, we have $w^*_{.0331} =$ `cSDCFlig(.0331,c(3,5,7))` $= 3.330$.

56. *Historical Development*. The multiple comparison procedures (6.62) and (6.64) based on the Wilcoxon rank sum statistics were first proposed independently by Steel (1960, 1961) and Dwass (1960) for the setting of equal sample sizes $n_1 = \cdots = n_k$. Critchlow and Fligner (1991) presented a natural generalization of these Steel–Dwass procedures when the n_i are not all equal and provided the exact critical values w^*_α for $k = 3$ and $2 \le n_1 \le n_2 \le n_3 \le 7$.

57. *Maximum Type I Error Rate*. The multiple comparison procedure (6.62) is designed so that the experimentwise error rate (see Comment 54) is controlled to be equal to α; that is, the probability of falsely declaring any pair of treatment effects to be different, when in fact *all* of the treatment effects are the same, is equal to α. However, it also satisfies the more stringent *maximum type I error rate* requirement that the probability of falsely declaring any pair of treatment effects to be different, regardless of the values of the other $k - 2$ treatment effects, is no larger than the stated α. This requires controlling the probability of making false declarations about treatment effect differences even in situations when *not all* of the treatment effects are the same. For example, if $\tau_1 < \tau_2 = \tau_3$, the probability of incorrectly deciding that $\tau_2 \ne \tau_3$ is still controlled to be α by multiple comparison procedure (6.62). Similar comments apply to the approximate procedure in (6.64).

58. *Large-Sample Approximation.* Let $\mathbf{W}^* = \left(W^*_{12}, W^*_{13}, \ldots, W^*_{k-1,k}\right)$, where W^*_{ij} is given by (6.61) for $1 \le i < j \le k$. Then it can be shown that $\mathbf{W}^*$ has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic multivariate normal distribution with mean vector $\mathbf{0}$ and appropriate covariance matrix $\boldsymbol{\Sigma}$ (see Miller (1981a) for further details). It follows directly from this result (again, see Miller (1981a)) that the procedure in (6.64) has an asymptotic experimentwise error rate equal to α when $n_1 = n_2 = \cdots = n_k$. Critchlow and Fligner (1991) used a result by Hayter (1984) to establish the fact that the asymptotic experimentwise error

rate for procedure (6.64) is also bounded above by α when we have unequal sample sizes.

When H_0 is true and $n_1 = n_2 = \cdots = n_k$, the asymptotic correlation matrix $\mathbf{\Sigma}_1$ (say) of the $\binom{k}{2}W_{ij}$'s is the same as the correlation matrix $\mathbf{\Sigma}_2$ (say) of the $\binom{k}{2}$ differences $Z_i - Z_j$, $1 \le i < j \le k$, where $Z_1, \ldots, Z_k$ are independent $N(0, 1)$ random variables (cf. Miller (1966), pp. 155–156). It follows that the asymptotic distribution of

$$\sqrt{2} \max_{1 \le i < j \le k} \left\{ \frac{|W_{ij} - E_0(W_{ij})|}{[\text{var}_0(W_{ij})]^{1/2}} \right\} = \max_{1 \le i < j \le k} |W_{ij}^*|$$

can be approximated by the distribution of

$$\max_{1 \le i < j \le k} |Z_i - Z_j| = \text{range } (Z_1, \ldots, Z_k).$$

The $\sqrt{2}$ occurs because the variance of $Z_i - Z_j$ equals 2. This justifies the use of approximation (6.64) in the equal-sample-size case. When the sample sizes are unequal, the asymptotic correlation matrix of the $\binom{k}{2}W_{ij}$'s will not in general agree with $\mathbf{\Sigma}_2$, but (6.64)) can be justified via a Tukey–Kramer approximation (see, e.g., Tukey (1953), Kramer (1956, 1957) and pages 91–93 of Hochberg and Tamhane (1987)).

59. *Joint Ranking Approach.* The multiple comparison procedure discussed in this section is based on $k(k-1)/2$ separate two-sample rankings. However, it is also reasonable to consider all-treatments multiple comparisons based on a single joint ranking of all N observations. Let $R_{.j}$ (6.4), $j = 1, \ldots, k$, be the average rank for the jth treatment sample in the joint ranking of all N observations. The joint ranking analog to procedure (6.62) is then given by

$$\text{Decide } \tau_u \ne \tau_v \text{ if } N^*|R_{.u} - R_{.v}| \ge y_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v, \tag{6.66}$$

where N^* is the least common multiple of $n_1, \ldots, n_k$ and the constant y_α is chosen to make the experimentwise error rate equal to α; that is, y_α satisfies the restriction

$$P_0(N^*|R_{.u} - R_{.v}| < y_\alpha, u = 1, \ldots, k-1; \; v = u+1, \ldots, k) = 1 - \alpha, \tag{6.67}$$

where the probability $P_0(\cdot)$ is computed under H_0 (6.2). As with the multiple comparison procedures based on pairwise rankings, (6.67) stipulates that the $k(k-1)/2$ inequalities $N^*|R_{.u} - R_{.v}| < y_\alpha$, corresponding to all pairs (u, v) of treatments with $u < v$, hold simultaneously with probability $1 - \alpha$ when H_0 (6.2) is true.

Nemenyi (1963) first proposed procedure (6.67) for the special case of equal sample sizes, in which case $N^*|R_{.u} - R_{.v}| = |R_u - R_v|$, where R_j (6.4) is the sum of the joint ranks for the treatment j observations. The general form of (6.67) for arbitrary sample sizes was considered by Damico and Wolfe (1987).

The y_α critical values can be obtained in exactly the same way as the w_α^* values for procedure (6.63). Proceeding as in Comment 55, we simply count the number of joint rank configurations for which the event

$B = \{N^*|R_{.u} - R_{.v}| < c, \text{ for all } u < v\}$ occurs and divide this count by $N!/\left[\Pi_{j=1}^{k} n_j!\right]$ to obtain the probability, under H_0, that $N^*|R_{.u} - R_{.v}| < c$ for all $u < v$. For an illustration, we again return to Comment 6 and use the 15 joint rank configurations displayed there for the case $k = 3$ and $n_1 = n_2 = n_3 = 2$. (For this setting, $N^*|R_{.u} - R_{.v}| = |R_u - R_v|$, and we can once again reduce the number of configurations that need to be considered from 90 to 15 by the same reasoning as in Comment 6.) For each of these 15 configurations, we now display the values of $|R_1 - R_2|$, $|R_1 - R_3|$, and $|R_2 - R_3|$.

(a)	$\|R_1 - R_2\| = 4$ $\|R_1 - R_3\| = 8$ $\|R_2 - R_3\| = 4$	(b)	$\|R_1 - R_2\| = 5$ $\|R_1 - R_3\| = 7$ $\|R_2 - R_3\| = 2$	(c)	$\|R_1 - R_2\| = 6$ $\|R_1 - R_3\| = 6$ $\|R_2 - R_3\| = 0$
(d)	$\|R_1 - R_2\| = 2$ $\|R_1 - R_3\| = 7$ $\|R_2 - R_3\| = 5$	(e)	$\|R_1 - R_2\| = 3$ $\|R_1 - R_3\| = 6$ $\|R_2 - R_3\| = 3$	(f)	$\|R_1 - R_2\| = 4$ $\|R_1 - R_3\| = 5$ $\|R_2 - R_3\| = 1$
(g)	$\|R_1 - R_2\| = 2$ $\|R_1 - R_3\| = 4$ $\|R_2 - R_3\| = 2$	(h)	$\|R_1 - R_2\| = 0$ $\|R_1 - R_3\| = 6$ $\|R_2 - R_3\| = 6$	(i)	$\|R_1 - R_2\| = 3$ $\|R_1 - R_3\| = 3$ $\|R_2 - R_3\| = 0$
(j)	$\|R_1 - R_2\| = 1$ $\|R_1 - R_3\| = 4$ $\|R_2 - R_3\| = 5$	(k)	$\|R_1 - R_2\| = 0$ $\|R_1 - R_3\| = 3$ $\|R_2 - R_3\| = 3$	(l)	$\|R_1 - R_2\| = 2$ $\|R_1 - R_3\| = 1$ $\|R_2 - R_3\| = 1$
(m)	$\|R_1 - R_2\| = 2$ $\|R_1 - R_3\| = 2$ $\|R_2 - R_3\| = 4$	(n)	$\|R_1 - R_2\| = 1$ $\|R_1 - R_3\| = 1$ $\|R_2 - R_3\| = 2$	(o)	$\|R_1 - R_2\| = 0$ $\|R_1 - R_3\| = 0$ $\|R_2 - R_3\| = 0$

Thus, for example,

$$\begin{aligned} &P_0\{|R_u - R_v| < 8, (u, v) = (1,2), (1,3), (2,3)\} \\ &= P_0\{|R_1 - R_2| < 8; |R_1 - R_3| < 8; |R_2 - R_3| < 8\} \\ &= \tfrac{14}{15} = 1 - .067, \end{aligned}$$

because for 14 of the configurations—all but configuration (a)—the event $\{|R_1 - R_2| < 8; |R_1 - R_3| < 8; |R_2 - R_3| < 8\}$ occurs. Similarly, $P_0\{|R_u - R_v| < 7; (u, v) = (1,2), (1,3), (2,3)\} = \frac{12}{15} = .80$, because the event $\{|R_1 - R_2| < 7; |R_1 - R_3| < 7; |R_2 - R_3| < 7\}$ occurs for 12 of the configurations—all but (a), (b), and (d). Hence, for $k = 3$ and $n_1 = n_2 = n_3 = 2$, we have $y_{.067} = 8$ and $y_{.200} = 7$. Values of y_α are available in Damico and Wolfe (1987) for available experimentwise error rates (α) closest to but not exceeding .001, .005, .01 (.005) .05 (.01) .15 and most useful combinations of either $k = 3, 1 \le n_1 \le n_2 \le n_3 \le 6$ or $k = 4, 1 \le n_1 \le n_2 \le n_3 \le n_4 \le 6$. For the special cases of equal sample sizes, these tabled values agree with those previously given by Nemenyi (1963) and McDonald and Thompson (1967). An approximation to y_α for large common sample size is discussed in Miller (1966). A related approximate procedure based on joint ranks and appropriate for large unequal sample size is suggested by Dunn (1964).

The joint ranking multiple comparison procedure given by (6.66) is a good deal simpler computationally than the corresponding pairwise ranking multiple comparison procedure in (6.62). Both procedures maintain the designated experimentwise error rate α. However, the joint ranking procedure does not provide the additional maximum type I error rate protection level α guarantee associated with the pairwise ranking procedure (see Comment 57). A second drawback for the joint ranking procedure is the fact that the absolute differences $|R_{\cdot u} - R_{\cdot v}|$ depend on the values of the observations from the other $k - 2$ treatments, in addition to the observations from treatments u and v. Thus, in the case of $k = 3$, the decision concerning treatments 1 and 2, for example, depends on the treatment 3 observations. This difficulty is discussed in Miller (1966) and Gabriel (1969).

Properties

1. *Asymptotic Multivariate Normality.* See Hayter (1984) and Critchlow and Fligner (1991).
2. *Efficiency.* See Sherman (1965) and Section 6.10.

Problems

42. Apply procedure (6.62) to the mean interstitial length data of Table 6.5.

43. Procedure (6.62) is defined specifically in terms of the $k(k-1)/2$ pairwise two-sample rankings. However, it can be applied to settings where only the joint ranks of all N observations are available. Explain.

44. Apply procedure (6.62) to the half-time of mucociliary clearance data of Table 6.1.

45. Apply the approximate procedure (6.64) to the glucocorticoid receptor data of Table 6.4.

46. For the case $k = 3, \alpha = .05$, and $n_1 = n_2 = n_3 = 6$, compare procedures (6.62) and (6.64).

47. Apply the approximate procedure (6.64) to the psychotherapeutic attraction data of Table 6.2.

48. Find the totality of all available experimentwise error rates α and the associated critical values w_α^* for procedure (6.62) when $k = 4$, $n_1 = 1$, and $n_2 = n_3 = n_4 = 2$.

49. Consider the joint ranking procedure (6.66) discussed in Comment 59. Find the totality of all available experimentwise error rates α and the associated critical values y_α for this procedure when $k = 4$, $n_1 = 1$ and $n_2 = n_3 = n_4 = 2$.

50. Consider the YOY gizzard shad data discussed in Example 6.6. Find the smallest (available) approximate experimentwise error rate at which the most significant difference in treatment effects (i.e., that between site I and site IV) would be detected.

51. Consider the mean interstitial length data in Table 6.5. Find the smallest (available) approximate experimentwise error rate at which we would declare that the typical mean interstitial length for white pines is different from that for Scotch pines.

6.6 DISTRIBUTION-FREE ONE-SIDED ALL-TREATMENTS MULTIPLE COMPARISONS BASED ON PAIRWISE RANKINGS-ORDERED TREATMENT EFFECTS (HAYTER–STONE)

In this section, we discuss a multiple comparison procedure based on pairwise two-sample rankings that is designed to make decisions about individual differences between pairs of

treatment effects (τ_i, τ_j), for $i < j$, in a setting where ordered alternatives H_2 (6.11) are of interest. Thus, the multiple comparison procedure of this section would be appropriate for one-way layout data *after* rejection of H_0 (6.2) with the Jonckheere–Terpstra procedure from Section 6.2. As with the procedure for general alternatives discussed in Section 6.5, we will once again reach conclusions about all $\binom{k}{2} = k(k-1)/2$ pairs of treatment effects. However, here these conclusions are naturally one-sided, in accordance with the ordered alternatives setting.

Procedure

For each pair of treatments $(i,j), 1 \le i < j \le k$, let W_{ij} be defined by expression (6.60); that is, W_{ij} is the Wilcoxon rank sum of the jth sample ranks in the two-sample ranking of the ith and jth sample observations. Compute the standardized form W_{ij}^* given in (6.61) for each treatment pair combination (i,j) with $i < j$.

At an experimentwise error rate of α, the Hayter–Stone one-sided all-treatments multiple comparison procedure reaches its $k(k-1)/2$ pairwise decisions, corresponding to each (τ_u, τ_v) pair, $1 \le u < v \le k$, by the criterion

$$\text{Decide } \tau_v > \tau_u \text{ if } W_{uv}^* \ge c_\alpha^*; \quad \text{otherwise decide } \tau_u = \tau_v, \tag{6.68}$$

where the constant c_α^* is chosen to make the experimentwise error rate equal to α; that is, c_α^* satisfies the restriction

$$P_0(W_{uv}^* < c_\alpha^*, u = 1, \ldots, k-1; v = u+1, \ldots, k) = 1 - \alpha, \tag{6.69}$$

where the probability $P_0(\cdot)$ is computed under H_0 (6.2). Equation (6.69) requires that the $k(k-1)/2$ inequalities $W_{uv}^* < c_\alpha^*$, corresponding to all pairs (u, v) of treatments with $u < v$, hold simultaneously with probability $1 - \alpha$ when H_0 (6.2) is true. Comment 62 explains how to obtain the critical value c_α^* for k treatments, sample sizes $n_1, \ldots, n_k$, and available experimentwise error rates α.

Large-Sample Approximation

When H_0 is true, the $k(k-1)/2$ component vector $(W_{12}^*, W_{13}^*, \ldots, W_{k-1,k}^*)$ has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic multivariate normal distribution with mean vector $\mathbf{0}$. It then follows (see Hayter and Stone (1991), e.g., for an indication of the proof) that c_α^* can be approximated for large sample sizes by d_α, where d_α is the upper αth percentile point for the distribution of

$$D = \underset{1 \le i < j \le k}{\text{maximum}} \left[\frac{Z_j - Z_i}{\left\{ \frac{n_i + n_j}{2 n_i n_j} \right\}^{1/2}} \right],$$

where $Z_1, \ldots, Z_k$ are mutually independent and Z_i has an $N(0, 1/n_i)$ distribution, for $i = 1, \ldots, k$. Thus, the large-sample approximation for procedure (6.68) is

$$\text{Decide } \tau_v > \tau_u \text{ if } W_{uv}^* \le d_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v. \tag{6.70}$$

To find d_α for k treatments, we use the R command `cHayStonLSA`(α, k). For example, to find $d_{.05}$ for $k = 6$ treatments, we apply `cHayStonLSA(.05, 6)` and obtain $d_{.05} = 3.719$ (see also Comment 64).

Ties

If there are ties among the X observations, use average ranks in computing the individual Wilcoxon rank sum statistics W_{ij} (6.60). In addition, replace the term $\text{Var}_0(W_{ij})/2 = n_i n_j (n_i + n_j + 1)/24$ in the denominator of W_{ij}^* (6.61) by the expression in (6.65).

EXAMPLE 6.7 ***Motivational Effect of Knowledge of Performance–Example 6.2 Continued.***

For Hundal's (1969) study to assess the motivational effects of knowledge of performance, we found in Example 6.2 (using the Jonckheere–Terpstra test procedure) that there was sufficient evidence in the sample data to conclude that $\tau_1 \le \tau_2 \le \tau_3$ with at least one strict inequality. To examine which of the types of information (none, rough, or accurate) lead to differences in median numbers of pieces processed, we apply procedure (6.68) with the appropriate corrections for ties, as given in (6.65). For this study, we have $k = 3, n_1 = n_2 = n_3 = 6$, and we must compute $k(k-1)/2 = 3(2)/2 = 3$ standardized W_{ij}^* statistics. For the sake of illustration, we take our experimentwise error rate to be $\alpha = .0553$. With $k = 3$ and $n_1 = n_2 = n_3 = 6$, we find $c^*_{.0553} =$ `cHayston(.0553, c(6, 6, 6))` $= 2.9439$ and procedure (6.68) reduces to

$$\text{Decide } \tau_v > \tau_u \text{ if } W_{uv}^* \ge 2.9439.$$

Next, we compute the three W_{ij}^* statistics. For the control (no information) and group B (partial information) sample observations, the combined-samples ranking yields the sum of ranks for the group B data to be

$$W_{12} = 2.5 + 5 + 12 + 10.5 + 5 + 8 + 43.$$

For this pair of samples, there are tied observations and we have $g_{12} = 8$ and $t_1 = 1, t_2 = 2, t_3 = 3, t_4 = t_5 = t_6 = 1, t_7 = 2$, and $t_8 = 1$. From (6.65), we obtain

$$\frac{\text{var}_0(W_{12})}{2} = \frac{6(6)}{24}\left[6 + 6 + 1 - \frac{2(1)(2)(3) + 2(3)(4)}{(6+6)(6+6-1)}\right]$$
$$= \frac{3}{2}\left[\frac{1716 - 36}{132}\right] = 19.09.$$

Using this result in (6.61), we find

$$W_{12}^* = \frac{[43 - 6(13)/2]}{\sqrt{19.09}} = .92.$$

For the other two population pairs, similar calculations lead to the following.

Control (No Information) and Group C (Accurate Information)

$$W_{13} = 12 + 3.5 + 10 + 6.5 + 11 + 8.5 = 51.5,$$
$$g_{13} = 9, \quad t_1 = t_2 = 1, \quad t_3 = 2, \quad t_4 = 1, \quad t_5 = t_6 = 2, \quad t_7 = t_8 = t_9 = 1,$$
$$\text{var}_0(W_{13}) = \frac{6(6)}{24}\left[6 + 6 + 1 - \frac{3(1)(2)(3)}{(6+6)(6+6-1)}\right] = 19.30,$$
$$W_{13}^* = \frac{[51.5 - 6(13)/2]}{\sqrt{19.30}} = 2.85.$$

Group B (Partial Information) and Group C (Accurate Information)

$$W_{23} = 12 + 3 + 9 + 6 + 10 + 7.5 = 47.5,$$
$$g_{23} = 9, \quad t_1 = 1, \quad t_2 = 3, \quad t_3 = t_4 = 1, \quad t_5 = 2, \quad t_6 = t_7 = t_8 = t_9 = 1,$$
$$\text{var}_0(W_{23}) = \frac{6(6)}{24}\left[6 + 6 + 1 - \frac{1(2)(3) + 2(3)(4)}{(6+6)(6+6-1)}\right] = 19.16,$$
$$W_{23}^* = \frac{[47.5 - 6(13)/2]}{\sqrt{19.16}} = 1.94.$$

Referring these W_{ij}^* values to the critical point $c_{.0553}^* = 2.9439$, we see that

$$W_{12}^* = .92 < 2.9439 \Rightarrow \text{decide } \tau_1 = \tau_2,$$
$$W_{13}^* = 2.85 < 2.9439 \Rightarrow \text{decide } \tau_1 = \tau_3,$$
$$W_{23}^* = 1.94 < 2.9439 \Rightarrow \text{decide } \tau_2 = \tau_3.$$

Thus, at an experimentwise error rate of .0553, we have reached the conclusion that $\tau_1 = \tau_2 = \tau_3$ (i.e., there are no differences in median numbers of pieces processed between the different levels of information), in contradiction with the conclusion from the Jonckheere–Terpstra test that $\tau_1 \le \tau_2 \le \tau_3$ with at least one strict inequality. Even though the P-value for the Jonckheere–Terpstra test procedure for these data is .0231, we are not able to detect any individual differences between treatment effects with the multiple comparison procedure (6.68), even with an experimentwise error rate as high as .0553. Such occurrences are, unfortunately, rather common in practice because of the conservative nature of the multiple comparison procedures (see Comment 54). For this reason, we often conduct our multiple comparison procedure at an experimentwise error rate that is higher than a typical significance level (such as .01 or .05) for a hypothesis test. If we have previously conducted a hypothesis test (such as the Jonckheere–Terpstra test in the example) and rejected H_0, we would at least like to know the *most* significant difference between pairs of treatment effects. For this reason, it is always informative in such cases to find the smallest experimentwise error rate at which the first pairwise difference in treatment effects would become significant. For the Hundal data, that corresponds to treatments 1 (no information) and 3 (accurate information) with an observed value $W_{13}^* = 2.85$. Using the `R` command `pHaySton(motivational.effect)`, we find that the smallest experimentwise error rate (among the limited number available) at which we would decide $\tau_3 > \tau_1$ (and thus conclude that accurate information is more effective than no information) is `pHaySton(motivational.effect)$p.val [2]` = .0850.

Comments

60. *Rationale for Multiple Comparison Procedures*. The general rationale for the multiple comparison procedures of this section is the same as that given in Comment 53 for the two-sided-all-treatments multiple comparison procedures of Section 6.5. The only additional factor here is that the procedures of this section yield decisions that are one-sided by nature in line with their association with the ordered restriction ($\tau_1 \leq \cdots \leq \tau_k$) on the treatment effects.

61. *Experimentwise Error Rate*. The use of an experimentwise error rate represents a very conservative approach to multiple comparisons. We are insisting that the probability of making only correct decisions be $1 - \alpha$ when the hypothesis H_0 (6.2) of treatment equivalence is true. Thus, we have a high degree of protection when H_0 is true, but we often apply the techniques of this section when we have evidence (perhaps based on a priori information or perhaps obtained by applying the Jonckheere–Terpstra test, as in Example 6.7) that H_0 is not true. (For additional general remarks about experimentwise error rates, see Comment 54.)

62. *Critical Values* c_α^*. The c_α^* critical values can be obtained by using the fact that under H_0 (6.2), all $N!/\left(\prod_{j=1}^k n_j!\right)$ *joint* (of all N sample observations) rank assignments of n_1 ranks to the treatment 1 observations, n_2 ranks to the treatment 2 observations, $\ldots, n_k$ ranks to the treatment k observations are equally likely. (Although the standardized pairwise Wilcoxon statistics W_{ij}^* (6.61) are formally defined in terms of pairwise two-sample ranks, it is clear that all $k(k-1)/2$ W_{ij}^* statistics can also be computed from the joint ranks of all N observations.) Thus, to obtain the probability, under H_0, that $W_{uv}^* < t$, for all $u < v$, we simply count the number of configurations for which the event $A = \{W_{uv}^* < t, \text{ for all } u < u\}$ occurs and divide this count by $N!/\left[\prod_{j=1}^k n_j!\right]$. For an illustration, we return to Comment 17 and use the 12 joint rank configurations displayed there for the case $k = 3, n_1 = 1, n_2 = 1$, and $n_3 = 2$. For each of these 12 configurations, we now display the values of W_{12}^*, W_{13}^*, and W_{23}^*.

(a)	$W_{12}^* = 1.4142$ $W_{13}^* = 1.7321$ $W_{23}^* = 1.7321$	(b)	$W_{12}^* = -1.4142$ $W_{13}^* = 1.7321$ $W_{23}^* = 1.7321$	(c)	$W_{12}^* = 1.4142$ $W_{13}^* = 1.7321$ $W_{23}^* = 0$
(d)	$W_{12}^* = -1.4142$ $W_{13}^* = 0$ $W_{23}^* = 1.7321$	(e)	$W_{12}^* = 1.4142$ $W_{13}^* = 1.7321$ $W_{23}^* = -1.7321$	(f)	$W_{12}^* = -1.4142$ $W_{13}^* = -1.7321$ $W_{23}^* = 1.7321$
(g)	$W_{12}^* = 1.4142$ $W_{13}^* = 0$ $W_{23}^* = 0$	(h)	$W_{12}^* = -1.4142$ $W_{13}^* = 0$ $W_{23}^* = 0$	(i)	$W_{12}^* = 1.4142$ $W_{13}^* = 0$ $W_{23}^* = -1.7321$
(j)	$W_{12}^* = -1.4142$ $W_{13}^* = -1.7321$ $W_{23}^* = 0$	(k)	$W_{12}^* = 1.4142$ $W_{13}^* = -1.7321$ $W_{23}^* = -1.7321$	(l)	$W_{12}^* = -1.4142$ $W_{13}^* = -1.7321$ $W_{23}^* = -1.7321$

Thus, for example,

$$\begin{aligned} &P_0\{W^*_{uv} < 1.7321, (u,v) = (1,2), (1,3), (2,3)\} \\ &= P_0\{W^*_{12} < 1.7321; W^*_{13} < 1.7321; W^*_{23} < 1.7321\} \\ &= \tfrac{6}{12} = 1 - .5, \end{aligned}$$

because for 6 of the 12 configurations—(g), (h), (i), (j), (k), and (l)—the event $\{W^*_{12} < 1.7321; W^*_{13} < 1.7321; W^*_{23} < 1.7321\}$ occurs. Similarly, $P_0\{W^*_{uv} < 1.4142, (u,v) = (1,2), (1,3), (2,3)\} = \frac{3}{12} = 1 - .75$, because the event $\{W^*_{12} < 1.4142; W^*_{13} < 1.4142; W^*_{23} < 1.4142\}$ occurs only for the three configurations (h), (j), and (l). Finally, $P_0\{W^*_{uv} < 0, (u,v) = (1, 2), (1, 3), (2, 3)\} = \frac{1}{12} = 1 - .9167$, corresponding to the single configuration (l). Hence, for $k = 3$, $n_1 = 1$, $n_2 = 1$, and $n_3 = 2$, we have $c^*_{.5000} = 1.7321$, $c^*_{.7500} = 1.4142$, and $c^*_{.9167} = 0$, and the values .5000, .7500, and .9167 are the only available experimentwise error rates for the Hayter–Stone procedure (6.68) in this setting.

For a given number of treatments k and sample sizes $n_1, \ldots, n_k$, the R command `cHayston(`α`,`$\boldsymbol{n}$`)` can be used to find the available critical values c^*_α. For a given available experimentwise error rate α, the critical value c^*_α is given by `cHayston(`α`,`$\boldsymbol{n}$`)`. Thus, for example, for $k = 3$ and $n_1 = 3$, $n_2 = 4$, and $n_3 = 6$, we have $c^*_{.0295} =$ `cHayston(.0295,c(3,4,6))` $= 3.015$.

63. *Maximum Type I Error Rate*. The multiple comparison procedure (6.68) is designed so that the experimentwise error rate (see Comment 61) is controlled to be equal to α; that is, the probability of falsely declaring any pair of treatment effects to be different, when in fact *all* the treatment effects are the same, is equal to α. However, it also satisfies the more stringent *maximum type I error rate* requirement that the probability of falsely declaring any pair of treatment effects to be different, regardless of the values of the other $k - 2$ treatment effects, is no larger than the stated α. This requires controlling the probability of making false declarations about treatment effect differences even in situations when *not all* the treatment effects are the same. For example, if $\tau_1 < \tau_2 = \tau_3$ the probability of incorrectly deciding that $\tau_2 \neq \tau_3$ is still controlled to be α by multiple comparison procedure (6.68). Similar comments apply to the approximate procedure in (6.70).

64. *Large and Unequal Sample Sizes*. In order to obtain the large-sample approximate critical values d_α for use in procedure (6.70) when we have an unbalanced setting (i.e., where the sample sizes are not all equal), we must evaluate a $(k - 1)$-dimensional integral expression. In view of this difficulty (even with the availability of high-speed computers) and the fact that there is a large number of possible unequal-sample-size combinations for each fixed k and N, the evaluation of these approximate critical values is practically feasible only for a small percentage of the necessary cases. To make matters even more complicated, the use of an appropriate equal-sample-size asymptotic critical value when we actually have unequal sample sizes does not result in a conservative procedure, as it does for the Dwass–Steel–Critchlow–Fligner two-sided all-treatments multiple comparison procedure in Section 6.5 (see, e.g., Comment 58). In the case of the Hayter–Stone one-sided procedure (6.70), use of a particular equal-sample-size asymptotic critical value d_α may result in either a conservative or liberal (i.e.,

experimentwise error rate $\leq \alpha$ or $> \alpha$, respectively) procedure, depending on the particular unequal sample size configurations involved. Thus, for k and unequal $(n_1, \ldots, n_k)$ configurations beyond those for which the exact critical values c_α^* can reasonably be obtained from the R program `cHaySton`$(\alpha, \boldsymbol{n})$, Hayter and Stone (1991) recommended that computer simulation techniques be used to obtain appropriate asymptotic critical values.

Properties

1. *Asymptotic Multivariate Normality*. See Hayter (1984) and Hayter and Stone (1991).

Problems

52. Apply procedure (6.70) to the psychotherapeutic attraction data of Table 6.2 using the postulated order $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$.

53. Procedure (6.68) is defined specifically in terms of the $k(k-1)/2$ pairwise two-sample rankings. However, it can be applied to settings where only the joint ranks of all N observations are available. Explain.

54. Apply procedure (6.68) to the average basal area increment data in Table 6.7. Use only the growing site index intervals 72–74, 75–77, and 78–80 with the postulated ordering $\tau_{72-74} \leq \tau_{75-77} \leq \tau_{78-80}$.

55. For the case $k = 3$, $\alpha = .05$, and $n_1 = n_2 = n_3 = 6$, compare procedures (6.68) and (6.70).

56. Find the totality of all available experimentwise error rates α and the associated critical values c_α^* for procedure (6.68) when $k = 4$, $n_1 = 1$, and $n_2 = n_3 = n_4 = 2$.

57. Consider the psychotherapeutic attraction data of Table 6.2 with the postulated ordering $\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4$. Find the smallest (available) approximate experimentwise error rate at which the most significant difference in treatment effects would be detected.

58. Consider the average basal area increment data in Table 6.7. Using only the growing site index intervals 72–74, 75–77, and 78–80 with the postulated ordering $\tau_{72-74} \leq \tau_{75-77} \leq \tau_{78-80}$, find the smallest available experimentwise error rate at which we would declare $\tau_{78-80} > \tau_{72-74}$.

6.7 DISTRIBUTION-FREE ONE-SIDED TREATMENTS-VERSUS-CONTROL MULTIPLE COMPARISONS BASED ON JOINT RANKINGS (NEMENYI, DAMICO–WOLFE)

In this section our attention turns to a multiple comparison procedure designed to make decisions about individual differences between the median effect for a single, baseline control population and the median effects for each of the remaining $k - 1$ treatments. This treatment versus control multiple comparison procedure is based on the joint ranking of all N sample observations and can be applied to one-way layout data containing a single control sample *after* rejection of H_0 (6.2) with any of the test procedures in Sections 6.1–6.4. Its application leads to conclusions about the differences between each of the $k - 1$ treatment effects and the control effect, and these conclusions are naturally one-sided in nature.

Procedure

For simplicity of notation, we let treatment 1 assume the role of the single baseline control. In addition, let N^* be the least common multiple of the sample sizes $n_1, \ldots, n_k$. Jointly rank all N of the sample observations and let $R_{.1}, \ldots, R_{.k}$ be the averages of these joint ranks associated with treatments $1, \ldots, k$, respectively. (Thus, $R_{.1}, \ldots, R_{.k}$ are as originally defined in (6.4) in conjunction with the Kruskal–Wallis statistic.) For each of the $k-1$ noncontrol treatments, calculate the difference $R_{.u} - R_{.1}$, $u = 2, \ldots, k$.

At an experimentwise error rate of α, the Nemenyi–Damico–Wolfe one-sided treatments-versus-control multiple comparison procedure (see Comment 65) reaches its $k-1$ pairwise decisions, corresponding to each (τ_1, τ_u) pair, $u = 2, \ldots, k$, by the criterion

$$\text{Decide } \tau_u > \tau_1 \text{ if } N^*(R_{.u} - R_{.1}) \geq y_\alpha^*; \quad \text{otherwise decide } \tau_u = \tau_1, \tag{6.71}$$

where the constant y_α^* is chosen to make the experimentwise error rate equal to α; that is, y_α^* satisfies the restriction

$$P_0\{N^*(R_{.u} - R_{.1}) < y_\alpha^*,\ u = 2, \ldots, k\} = 1 - \alpha, \tag{6.72}$$

where the probability $P_0(\cdot)$ is computed under H_0 (6.2). Equation (6.72) stipulates that the $k-1$ inequalities $N^*(R_{.u} - R_{.1}) < y_\alpha^*$, corresponding to all pairs $(1, u)$ of noncontrol treatments ($u = 2, \ldots, k$) with the control treatment 1, hold simultaneously with probability $1 - \alpha$ when H_0 (6.2) is true. Comment 68 explains how to obtain the critical value y_α^* for $(k-1)$ noncontrol treatments, sample sizes $n_1, \ldots, n_k$, and available experimentwise error rates α.

Large-Sample Approximations

When H_0 is true, the $(k-1)$ component vector $(R_{.2} - R_{.1}, R_{.3} - R_{.1}, \ldots, R_{.k} - R_{.1})$ has, as $\min(n_1, \ldots, n_k)$ tends to infinity, an asymptotic $(k-1)$-variate normal distribution with mean vector $\mathbf{0}$. (For an indication of the proof, see Miller (1966).) For the special case of $n_1 = b$ and $n_2 = \cdots = n_k = n$, with both n and b large, the critical value y_α^* can be approximated by $[N(N+1)/12]^{1/2}[(1/b) + (1/n)]^{1/2} N^* m_\alpha^*$, where m_α^* is the upper αth percentile point for the distribution of the maximum of $(k-1)N(0,1)$ variables with common correlation $\rho = n/(b+n)$. Thus, the large-sample approximation for procedure (6.71) when we have equal treatment sample sizes $n_2 = \cdots = n_k = n$ (possibly different from $b = n_1$) is

$$\text{Decide } \tau_u > \tau_1 \text{ if } (R_{.u} - R_{.1}) \geq m_\alpha^* \left[\frac{N(N+1)}{12}\right]^{1/2} \left(\frac{1}{b} + \frac{1}{n}\right)^{1/2};$$
$$\text{otherwise decide } \tau_u = \tau_1, u = 2, \ldots, k. \tag{6.73}$$

To find m_α^* for $k-1$ noncontrol treatments, number of control observations b and an equal number, n, of observations from each of the noncontrol treatments, we use the R command `cMaxNorm(`α`, k - 1, n/(b+n))`. For example, to find $m_{.0310}^*$ for $k - 1 = 4$ noncontrol treatments, $b = 9$, and $n = 3$, we have $\rho = 3/(9+3) = .25$ and apply `cMaxNorm(.0310, 4, 0.25)` to obtain $m_{.0310}^* = 2.40$.

For the general setting of arbitrary (not necessarily equal) treatments sample sizes, Dunn (1964) used Bonferroni's Inequality to provide the large-sample approximation to procedure (6.71) given by

$$\text{Decide } \tau_u > \tau_1 \text{ if } (R_{.u} - R_{.1}) \geq z_{\alpha^*}\left[\frac{N(N+1)}{12}\right]^{1/2}\left(\frac{1}{n_1}+\frac{1}{n_u}\right)^{1/2};$$
$$\text{otherwise decide } \tau_u = \tau_1, u = 2, \ldots, k, \tag{6.74}$$

where $\alpha^* = \alpha/(k-1)$. (We note that this general approximate procedure can often be quite conservative in practice, as a direct result of the conservative nature of the Bonferroni Inequality.)

Ties

If there are ties among the X observations, use average ranks in computing the individual treatment sums of ranks $R_1, \ldots, R_k$.

EXAMPLE 6.8 ***Motivational Effect of Knowledge of Performance—Example 6.2 Continued.***

Once again we consider Hundal's (1969) study to assess the motivational effects of knowledge of performance. We previously found in Example 6.2 (using the Jonckheere–Terpstra test procedure) that there was sufficient evidence in the sample data to conclude that $\tau_1 \leq \tau_2 \leq \tau_3$ with at least one strict inequality. To further investigate which (if either) of the two types of additional information (rough or accurate) lead to differences in median numbers of pieces processed relative to the no information control (treatment 1), we apply procedure (6.71). Here, we have $k = 3$ and $n_1 = n_2 = n_3 = N^* = 6$. For the sake of illustration, we take our experimentwise error rate to be $\alpha = .0554$. With $k = 3$ and $n_1 = n_2 = n_3 = 6$, we find $y^*_{.0554} =$ `cNDWol(.0554, c(6, 6, 6))` $= 35$ and procedure (6.71) reduces to

$$\text{Decide } \tau_u > \tau_i \text{ if } 6(R_{.u} - R_{.1}) = (R_u - R_1) \geq 35.$$

Using the joint ranks (with average ranks to break ties among the observations) provided in parentheses beside the data in Table 6.6, we see that

$$R_1 = 5.5 + 1 + 2.5 + 10.5 + 13 + 8 = 40.5,$$

$$R_2 = 2.5 + 5.5 + 17 + 13 + 5.5 + 9 = 52.5,$$

and

$$R_3 = 18 + 5.5 + 15 + 10.5 + 16 + 13 = 78.$$

Thus, $(R_2 - R_1) = (52.5 - 40.5) = 12$ and $(R_3 - R_1) = (78 - 40.5) = 37.5$. Referring these rank sum differences to the critical point $y^*_{.0554} = 35$, we see that

$$(R_2 - R_1) = 12 < 35 \quad \Rightarrow \text{decide } \tau_2 = \tau_1,$$

$$(R_3 - R_1) = 37.5 \geq 35 \Rightarrow \text{decide } \tau_3 > \tau_1.$$

Thus at an experimentwise error rate of .0554, we have reached the conclusion that accurate information leads to significantly more pieces processed than the no information control. (We note that the smallest experimentwise error rate at which we would reach this conclusion is .0426, as $y^*_{.0426} =$ `cNDWol(.0426, c(6, 6, 6))` $= 37$ and $y^*_{.0371} =$ `cNDWol(.0371, c(6, 6, 6))` $= 38$.)

For the sake of illustration for the associated large-sample approximation (with equal sample sizes) procedure in (6.73), we note that $\rho = n/(b+n) = 6/(6+6) = \frac{1}{2}$ (which is always the case with equal sample sizes in the control and the noncontrol treatments). Using an approximate experimentwise error rate of $\alpha = .05183$ with $(k-1) = 2$, we see that `cMaxNorm(.05183, 2, 0.5)` $= m^*_{.05183} = 1.90$. Thus, we have that

$$\left[\frac{N(N+1)}{12}\right]^{1/2}\left(\frac{1}{b}+\frac{1}{n}\right)^{1/2} m^*_{.05183} = \left[\frac{18(19}{12}\right]^{1/2}\left(\frac{1}{6}+\frac{1}{6}\right)^{1/2}(1.90) = 5.856$$

and procedure (6.73) becomes

$$\text{Decide } \tau_u > \tau_1 \text{ if } (R_{.u} - R_{.1}) \geq 5.856$$

or, equivalently,

$$\text{Decide } \tau_u > \tau_1 \text{ if } (R_u - R_1) = 6(R_{.u} - R_{.1}) \geq 6(5.856) = 35.14.$$

Thus, for $k = 3$ and $n_1 = n_2 = n_3 = 6$, the exact procedure (6.71) and the large-sample approximation (6.73) are virtually identical and lead to the same conclusions that $\tau_2 = \tau_1$ and $\tau_3 > \tau_1$.

We note that the treatment-versus-control procedure (6.71) yields the conclusion that $\tau_3 > \tau_1$ at a considerably smaller experimentwise error rate (as low as .0426) than is the case with the Hayter–Stone one-sided all-treatments multiple comparison procedure (as detailed in Example 6.7), where the smallest experimentwise error rate leading to this conclusion is .0850. This situation is due primarily to the fact that the Hayter–Stone procedure is required to make an additional decision about the relative magnitude of τ_2 and τ_3, which, for these data, do not appear to be significantly different.

Comments

65. *Rationale for Treatments-versus-Control Multiple Comparison Procedures.* The general rationale for the multiple comparison procedures of this section is the same as that given in Comment 53 for the two-sided all-treatments multiple comparison procedures of Section 6.5. The only additional factor here is that the treatment-versus-control procedures of this section do not compare all treatments, but only each noncontrol treatment with the control on a directional bias. This situation arises, for example, in drug screening in the examination of many new treatments in hopes of improving on a standard, and there is no initial reason to perform between treatment comparisons. Of course, similar comparisons would be carried out later between treatments that were selected as being better than the control.

66. *Experimentwise Error Rate.* The use of an experimentwise error rate represents a very conservative approach to multiple comparisons. We are insisting that the probability of making only correct decisions be $1-\alpha$ when the hypothesis H_0 (6.2) of treatment equivalence is true. Thus, we have a high degree of protection when H_0 is true, but we often apply the techniques of this section when we have evidence (perhaps based on a priori information or perhaps obtained by applying a previous test procedure, as in Example 6.8) that H_0 is not true. (For additional general remarks about experimentwise error rates, see Comment 54.)

67. *Opposite Direction Decisions.* Procedures (6.71), (6.73), and (6.74) are designed for the one-sided case where the decisions are $\tau_u > \tau_1$ versus $\tau_u = \tau_1,\ u = 2,\ldots,k$. To handle the analogous one-sided situation where the decisions involve $\tau_u < \tau_1$ versus $\tau_u = \tau_1,\ u = 2,\ldots,k$, we use (6.71), (6.73), and (6.74) with $(R_{.u} - R_{.1})$ replaced by $(R_{.1} - R_{.u})$ for $u = 2,\ldots,k$.

68. *Critical Values* y_α^*. The y_α^* critical values can be obtained by using the fact that under H_0 (6.2), all $N!/\left[\prod_{j=1}^{k} n_j!\right]$ rank assignments are equally likely. However, in this one-sided treatments-versus-control setting, we must work a little harder than in the two-sided all-treatments case (see Comments 55 and 59 in Section 6.5) as the values $R_{.u} - R_{.1},\ u = 2,\ldots,k$, are, in general, changed when we relabel the control treatment. (In either of the previous two-sided all-treatments cases, the relevant statistic is unaffected by treatment relabelings.) As a result, we will have to take the complete enumeration approach employed in Comment 62 for the one-sided all-treatments setting, where the relevant statistic is also not invariant with respect to treatment relabelings.

    For an illustration, we return to Comment 17 and use the 12 rank configurations displayed there for the case $k = 3$, $n_1 = 1$, $n_2 = 1$, and $n_3 = 2$. (Here, $N^* = 2$.) For each of these 12 configurations, we now display the values of $2(R_{.2} - R_{.1})$ and $2(R_{.3} - R_{.1})$.

(a)	$2(R_{.2} - R_{.1}) = 2$ $2(R_{.3} - R_{.1}) = 5$	(b)	$2(R_{.2} - R_{.1}) = -2$ $2(R_{.3} - R_{.1}) = 3$	(c)	$2(R_{.2} - R_{.1}) = 4$ $2(R_{.3} - R_{.1}) = 4$
(d)	$2(R_{.2} - R_{.1}) = -4$ $2(R_{.3} - R_{.1}) = 0$	(e)	$2(R_{.2} - R_{.1}) = 6$ $2(R_{.3} - R_{.1}) = 3$	(f)	$2(R_{.2} - R_{.1}) = -6$ $2(R_{.3} - R_{.1}) = -3$
(g)	$2(R_{.2} - R_{.1}) = 2$ $2(R_{.3} - R_{.1}) = 1$	(h)	$2(R_{.2} - R_{.1}) = -2$ $2(R_{.3} - R_{.1}) = -1$	(i)	$2(R_{.2} - R_{.1}) = 4$ $2(R_{.3} - R_{.1}) = 0$
(j)	$2(R_{.2} - R_{.1}) = -4$ $2(R_{.3} - R_{.1}) = -4$	(k)	$2(R_{.2} - R_{.1}) = 2$ $2(R_{.3} - R_{.1}) = -3$	(l)	$2(R_{.2} - R_{.1}) = -2$ $2(R_{.3} - R_{.1}) = -5$

    Thus, for example,

$$
\begin{aligned}
&P_0\{2(R_{.u} - R_{.1}) < 6, u = 2, 3\} \\
&= P_0\{2(R_{.2} - R_{.1}) < 6 \text{ and } 2(R_{.3} - R_{.1}) < 6\} \\
&= \tfrac{11}{12} = 1 - .0833,
\end{aligned}
$$

because the event $\{2(R_{.2} - R_{.1}) < 6 \text{ and } 2(R_{.3} - R_{.1}) < 6\}$ occurs for all but configuration (e). Similarly, $P_0\{2(R_{.u} - R_{.1}) < 2, u = 2, 3\} = \frac{5}{12} = 1 - .5833$, because the event $\{2(R_{.2} - R_{.1}) < 2 \text{ and } 2(R_{.3} - R_{.1}) < 2\}$ occurs only for the five configurations (d), (f), (h), (j), and (l). Hence, for $k = 3$, $n_1 = 1$, $n_2 = 1$, and $n_3 = 2$, we have $y^*_{.0833} = 6$ and $y^*_{.5833} = 2$. The other possible experimentwise error rates (there are 10, including 1, all together) and the associated critical values for this setting are obtained through the same type of enumeration.

For a given number of noncontrol treatments $k - 1$ and sample sizes $n_1, \ldots, n_k$, the R command `cNDWol(`α`,`$\boldsymbol{n}$`)` can be used to find the available critical values y^*_α. For a given available experimentwise error rate α, the critical value y^*_α is given by `cNDWol(`α`,`$\boldsymbol{n}$`)`. Thus, for example, for $k = 3$ and $n_1 = n_2 = n_3 = 6$, we have $y^*_{.0795} =$ `cNDWol(.0795, c(6,6,6))` $= 32$.

69. *Interpretation as Hypothesis Tests*. Procedures (6.71), (6.73), and (6.74) can also be interpreted as hypothesis tests of H_0 (6.2). (For example, the procedure that rejects H_0 if at least one of the $k - 1$ inequalities of (6.71) holds is a distribution-free test of level α for H_0.) However, they are generally more effective at detecting differences between individual treatment effects when applied to data for which the null hypothesis H_0 has previously been rejected by one of the test procedures is Sections 6.1–6.4.

70. *Dependence on Observations from Other Noninvolved Treatments*. The differences $(R_{.u} - R_{.1})$ depend on the values of the observations from the other $k - 2$ treatments, in addition to the observations from the control and treatment u. Thus, the multiple comparison procedures in (6.71), (6.73), and (6.74) all have the disadvantage that the decision concerning treatment u and the control can be affected by changes only in the observations from one or more of the other $k - 2$ noninvolved treatments. This difficulty has been emphasized by Miller (1966) and Gabriel (1969).

71. *Two-Sided Treatments-versus-Control Multiple Comparison Procedures*. All the multiple comparison procedures of this section are one-sided by nature, resulting in decisions between $\tau_u = \tau_1$ and $\tau_u > \tau_1$ for every $u = 2, \ldots, k$ (or between $\tau_u = \tau_1$ and $\tau_u < \tau_1$ for every $u = 2, \ldots, k$, as noted in Comment 67). We view such one-sided comparisons to be the most natural approach for treatments-versus-control settings. In such situations, we are generally interested in seeing which, if any, of the proposed new treatments are better than a standard control or placebo. In most practical applications, *better* is synonymous with one-sided comparisons (all in one direction or all in the other)—thus our emphasis on such procedures in this section. However, a two-sided treatments-versus-control analog to procedure (6.71) has been developed in the literature and corresponds to the criterion

$$\text{Decide } \tau_u \neq \tau_1 \text{ if } N^*|R_{.u} - R_{.1}| \geq y^{**}_\alpha; \quad \text{otherwise decide } \tau_u = \tau_1, \quad (6.75)$$

where the constant y^{**}_α is chosen to make the experimentwise error rate equal to α; that is,

$$P_0\{N^*|R_{.u} - R_{.1}| < y^{**}_\alpha,\ u = 2, \ldots, k\} = 1 - \alpha,$$

where the probability $P_0(\cdot)$ is computed under H_0 (6.2). However, the required critical values y^{**}_α are available only in a very limited fashion. Leach (1972) has

provided such critical values y_α^{**} for the very special case of $k = 3$ and equal sample sizes $n_1 = n_2 = n_3 = 2(1)6$. Associated large-sample approximations to (6.75) for equal and unequal noncontrol treatment sample sizes have been considered by Miller (1966) and Dunn (1964), respectively. For further discussion of these two-sided treatments-versus-control multiple comparison procedures, see Miller (1966).

72. *Pairwise Ranking Approach*. The treatments-versus-control multiple comparison procedures discussed in this section are based on the joint ranking of all N of the sample observations. They suffer from the same drawbacks as do other one-way layout multiple comparison procedures based on joint rankings. For example, they do not provide the maximum type I error rate protection level α guarantee and decisions between treatment u and the control depend on the values of the observations from the other $k - 2$ treatments (for more details, see, e.g., Fligner (1984)).

    Steel (1959) developed a competitor of these Nemenyi–Damico–Wolfe procedures that takes the pairwise ranking approach discussed in Sections 6.5 and 6.6. His procedure is based on $k - 1$ separate two-sample rankings between the control sample and each of the $k - 1$ noncontrol samples and has the form

$$\text{Decide } \tau_u > \tau_1 \text{ if } W_{1u}^* \geq b_\alpha^*; \quad \text{otherwise decide } \tau_u = \tau_1, u = 2, \ldots, k, \quad (6.76)$$

    where $W_{12}^*, \ldots, W_{1k}^*$ are defined by (6.61) and b_α^* is chosen to make the experimentwise error rate equal to α. This pairwise ranking treatments-versus-control procedure has many of the nice properties of the analogous pairwise rankings all-treatments multiple comparison procedures discussed in Sections 6.5 and 6.6, including proper control of the maximum type I error rate (see Comments 57 and 63).

Properties

1. *Asymptotic Multivariate Normality*. See Miller (1966).
2. *Efficiency*. See Sherman (1965) and Section 6.10.

Problems

59. Apply the approximate procedure (6.73) to the psychotherapeutic attraction data in Table 6.2.

60. For the case $k = 3$, $\alpha = .01$, $n_1 = n_2 = n_3 = 6$, compare procedures (6.71) and (6.73).

61. For the psychotherapeutic attraction data in Table 6.2, find the smallest approximate experimentwise error rate at which we would decide $\tau_4 > \tau_1$ using procedure (6.73).

62. Consider the mucociliary clearance data in Table 6.1. Use procedure (6.71) to decide whether or not either obstructive airways disease or asbestosis (or both) lead to a deterioration (slowdown) in median mucociliary clearance half-times.

63. Apply the approximate procedure (6.74) to the glucocorticoid receptor site data in Table 6.4.

64. For the glucocorticoid receptor site data in Table 6.4, find the smallest approximate experimentwise error rate at which we would decide $\tau_5 > \tau_1$ using procedure (6.74).

65. Apply the approximate procedure (6.73) to the plasma glucose data in Table 6.9.

66. For the plasma glucose data in Table 6.9, find the smallest approximate experimentwise error rate at which we would decide $\tau_3 > \tau_1$ using procedure (6.73).

67. For the plasma glucose data in Table 6.9, find the smallest approximate experimentwise error rate at which we would decide $\tau_5 < \tau_1$ with an appropriate treatments-versus-control multiple comparison procedure (see Comment 67).

68. Apply the approximate procedure (6.73) to the revertant colonies data in Table 6.10.

69. For the revertant colonies data in Table 6.10, find the smallest approximate experimentwise error rate at which we would decide $\tau_4 > \tau_1$ using procedure (6.73).

70. Consider the revertant colonies data in Table 6.10. Find the smallest approximate experimentwise error rate at which the most significant difference in treatment (dosage) effects would be detected.

71. Find the totality of all available experimentwise error rates α and the associated critical values y_α^* for procedure (6.71) when $k = 4$, $n_1 = 1$, and $n_2 = n_3 = n_4 = 2$.

6.8 CONTRAST ESTIMATION BASED ON HODGES–LEHMANN TWO-SAMPLE ESTIMATORS (SPJøTVOLL)

In this section we discuss a method for the point estimation of certain linear combinations of treatment effects known in the literature as *contrasts*. We define such a contrast in the treatment effects $\tau_1, \ldots, \tau_k$ to be any linear combination of the form

$$\theta = \sum_{i=1}^{k} a_i \tau_i, \tag{6.77}$$

where $a_1, \ldots, a_k$ are any specified set of constants such that $\sum_{i=1}^{k} a_i = 0$. Equivalently, we can write θ in terms of the individual differences in treatment effects (known in the literature as *simple contrasts*)

$$\Delta_{hj} = \tau_h - \tau_j, \quad h = 1, \ldots, k;\ j = 1, \ldots, k, \tag{6.78}$$

by noting that

$$\theta = \sum_{h=1}^{k} \sum_{j=1}^{k} d_{hj} \Delta_{hj}, \tag{6.79}$$

where

$$d_{hj} = \frac{a_h}{k},\ h = 1, \ldots, k; \quad j = 1, \ldots, k. \tag{6.80}$$

For a given setting, decisions about which contrasts to estimate can be related either to a priori interest in particular linear combinations of the τ's or the results of one of the multiple comparison procedures discussed in Sections 6.5–6.7.

Procedure

For each pair of treatments $(h,\ j), h \neq j = 1, \ldots, k$, define the pairwise estimators

$$Z_{hj} = \text{median } \{X_{\alpha h} - X_{\beta j}, \alpha = 1, \ldots, n_h; \beta = 1, \ldots, n_j\}. \tag{6.81}$$

As $Z_{hj} = -Z_{jh}$, we need to calculate only the $k(k-1)/2$ estimators Z_{hj} corresponding to $h < j$. We refer to Z_{hj} as the raw or unadjusted estimator of the simple contrast $\Delta_{hj} = \tau_h - \tau_j$. (Note that Z_{hj} is exactly the Hodges–Lehmann two-sample estimator defined in Section 4.2, as applied to the hth sample (playing the role of the Y's) and the jth sample (playing the role of the X's). For example, Z_{13} is simply the median of the $n_1 n_3$ differences $X_{\alpha 1} - X_{\beta 3}$ obtained from the treatments 1 and 3 observations.) Next, we obtain the set $\overline{\Delta}_1, \ldots, \overline{\Delta}_k$ of individual weighted average of these unadjusted estimators Z_{hj} corresponding to

$$\overline{\Delta}_h = \sum_{j=1}^{k} \left(\frac{n_j}{N}\right) Z_{hj},\ h = 1, \ldots, k, \tag{6.82}$$

where we note that $Z_{hh} = 0$ for $h = 1, \ldots, k$.

The weighted-adjusted estimator of the contrast θ (6.77) is given by

$$\widehat{\theta} = \sum_{i=1}^{k} a_i \overline{\Delta}_i, \tag{6.83}$$

or, equivalently,

$$\widehat{\theta} = \sum_{h=1}^{k} \sum_{j=1}^{k} d_{hj} (\overline{\Delta}_h - \overline{\Delta}_j) = \sum_{h=1}^{k} \sum_{j=1}^{k} d_{hj}\ W_{hj}, \tag{6.84}$$

where

$$W_{hj} = \overline{\Delta}_h - \overline{\Delta}_j = \widehat{\Delta}_{hj} \tag{6.85}$$

is the weighted-adjusted estimator of the simple contrast $\Delta_{hj} = \tau_h - \tau_j$. We note that in the special case $n_1 = n_2 = \cdots = n_k$, $\overline{\Delta}_h$ (6.82) reduces to

$$Z_{h.} = \frac{\sum_{j=1}^{k} Z_{hj}}{k}, \quad h = 1, \ldots, k, \tag{6.86}$$

and $W_{hj} = \overline{\Delta}_h - \overline{\Delta}_j$ (6.85) can be simplified to

$$W_{hj} = Z_{h.} - Z_{j.}, \quad h \neq j = 1, \ldots, k. \tag{6.87}$$

EXAMPLE 6.9 ***Motivational Effect of Knowledge of Performance—Examples 6.2 and 6.8 Continued.***

Consider the Hundal knowledge of performance data originally presented in Example 6.2. In the application of the Nemenyi–Damico–Wolfe one-sided treatments-versus-control multiple comparison procedure (Example 6.8) to these data, we concluded that the group

receiving accurate information about their output produced significantly more (experimentwise error rate .0554) pieces than the group that received no information. Thus, it is of interest to use the knowledge of performance data in Table 6.6 to estimate the simple contrast $\theta = \tau_{\text{accurate information}} - \tau_{\text{no information}} = \tau_3 - \tau_1$, thereby providing an idea of the increased output that might be expected for this task by providing accurate information to the workers.

From Table 6.6 and (6.81), the three pairwise estimators are

$$\begin{aligned} Z_{12} &= \text{median}\{2, 0, -7, -4, 0, -2, -3, -5, -12, -9, -5, -7, 0, -2, -9, \\ &\qquad -6, -2, -4, 5, 3, -4, -1, 3, 1, 6, 4, -3, 0, 4, 2, 3, 1, -6, -3, 1, -1\} \\ &= -1.5, \\ Z_{13} &= \text{median}\{-8, 0, -5, -3, -6, -4, -13, -5, -10, -8, -11, -9, -10, -2, -7, \\ &\qquad -5, -8, -6, -5, 3, -2, 0, -3, -1, -4, 4, -1, 1, -2, 0, -7, 1, -4, -2, -5, -3\} \\ &= -4, \end{aligned}$$

and

$$\begin{aligned} Z_{23} &= \text{median}\{-10, -2, -7, -5, -8, -6, -8, 0, -5, -3, -6, -4, -1, 7, 2, \\ &\qquad 4, 1, 3, -4, 4, -1, 1, -2, 0, -8, 0, -5, -3, -6, -4, -6, 2, -3, -1, -4, -2\} \\ &= -3. \end{aligned}$$

From expression (6.82), or equivalently (as $n_1 = n_2 = n_3 = 6$) from (6.86), we have

$$\overline{\Delta}_1 = \frac{Z_{11} + Z_{12} + Z_{13}}{3} = \frac{0 - 1.5 - 4}{3} = -\frac{11}{6},$$

$$\overline{\Delta}_2 = \frac{Z_{21} + Z_{22} + Z_{23}}{3} = \frac{1.5 + 0 - 3}{3} = -.5,$$

and

$$\overline{\Delta}_3 = \frac{Z_{31} + Z_{32} + Z_{33}}{3} = \frac{4 + 3 + 0}{3} = \frac{7}{3}.$$

(Note that in calculating $\overline{\Delta}_2$ and $\overline{\Delta}_3$, we have used the fact that $Z_{21} = -Z_{12}, Z_{31} = -Z_{13}$, and $Z_{32} = -Z_{23}$.) The weighted-adjusted estimator of $\theta = \tau_3 - \tau_1$ is now obtained from (6.83) with $a_1 = -1, a_2 = 0$, and $a_3 = 1$. We find

$$\widehat{\theta} = W_{31} = \overline{\Delta}_3 - \overline{\Delta}_1 = \tfrac{7}{3} - \left(-\tfrac{11}{6}\right) = \tfrac{25}{6} = 4.17 \text{ pieces}.$$

(We note that the values of the raw estimator Z_{31} and the classical estimator $\overline{X}_3 - \overline{X}_1$ are 4.00 and 4.17, respectively, for these data.)

Comments

73. *Ambiguities with the Unadjusted Estimators*. The unadjusted estimators Z_{hj} (6.81) lead to ambiguities in contrast estimation because they do not satisfy the linear relations that are satisfied by the contrasts they estimate. For example, $\Delta_{13} = \tau_1 - \tau_3 = (\tau_1 - \tau_2) + (\tau_2 - \tau_3) = \Delta_{12} + \Delta_{23}$, but in general $Z_{13} \neq Z_{12} + Z_{23}$. Thus, the two "reasonable" estimators Z_{13} and $Z_{12} + Z_{23}$ of $\Delta_{13} = \tau_1 - \tau_3$ can give different estimates. This was pointed out by Lehmann (1963a), who called the unadjusted estimators incompatible.
74. *Compatible, but Inconsistent Estimators*. Lehmann (1963a) removed the incompatibility difficulty discussed in Comment 73 by using the estimators $W_{hj} = Z_{h\cdot} - Z_{j\cdot}$ (6.87). These estimators are obtained by minimizing the sum of squares $\sum\sum_{h\neq j}[Z_{hj} - (\tau_h - \tau_j)]^2$. Although these estimators are compatible, Lehmann also pointed out that two additional difficulties now arise. First, the estimator $Z_{h\cdot} - Z_{j\cdot}$ of $\Delta_{hj} = \tau_h - \tau_j$ depends, in addition to the observations from samples h and j, on the observations from the other $k - 2$ samples. Furthermore, in the case of $k = 3$, for example, the estimator $Z_{1\cdot} - Z_{2\cdot}$ (of $\tau_1 - \tau_2$) is not consistent when n_1 and n_2 tend to infinity unless n_3 also tends to infinity.
75. *Consistency*. Spjøtvoll (1968) removed the nonconsistency difficulty by obtaining the weighted-adjusted estimators $W_{hj} = \overline{\Delta}_h - \overline{\Delta}_j$ (6.85). These estimators minimize the sum of squares $N^{-2}\sum\sum_{h\neq j} n_h n_j [Z_{hj} - (\tau_h - \tau_j)]^2$. Spjøtvoll's estimators do, however, retain the disadvantage that the estimator of $\tau_h - \tau_j$ depends on unrelated observations from the other samples.
76. *Competitor Contrast Estimator*. Spjøtvoll (1968) also proposed weighted-adjusted estimators that minimize

$$\sum_{h\neq j}\sum \left(\frac{N}{n_h} + \frac{N}{n_j}\right)^{-1} [Z_{hj} - (\tau_h - \tau_j)]^2, \tag{6.88}$$

    using the asymptotic variances of the Z_{hj}'s as weights in the sum of squares. These estimators are more difficult to compute than the estimators W_{hj} (6.85). Furthermore, Spjøtvoll showed that the weighted-adjusted estimators W_{hj} (6.85) and those obtained by minimizing (6.88) have the same asymptotic properties when n_j tends to infinity in such a way that (n_j/N) tends to λ_j with $0 < \lambda_j < 1$, for $j = 1, \ldots, k$.
77. *Equivalence with Equal Sample Sizes*. Spjøtvoll pointed out that the estimator of Δ_{hj} obtained by minimizing (6.88) and the estimator W_{hj} (6.85) both reduce to Lehmann's estimator $Z_{h\cdot} - Z_{j\cdot}$ (6.88) when $n_1 = n_2 = \cdots = n_k$.

Properties

1. *Standard Deviation of $\widehat{\theta}$ (6.83)*. For the asymptotic standard deviation of $\widehat{\theta}$ (6.83), see Spjøtvoll (1968).
2. *Asymptotic Normality*. See Spjøtvoll (1968) and Lehmann (1963a).
3. *Efficiency*. See Spjøtvoll (1968), Lehmann (1963a), and Section 6.10.

Problems

72. Estimate the simple contrast $\theta = \tau_4 - \tau_1$ for the psychotherapeutic attraction data in Table 6.2.

73. Estimate the simple contrasts $\theta_1 = \tau_2 - \tau_1, \theta_2 = \tau_4 - \tau_1$, and $\theta_3 = \tau_5 - \tau_1$ for the glucocorticoid receptor sites data in Table 6.4.

74. Estimate all possible simple contrasts for the mean interstitial lengths data in Table 6.5.

75. Estimate the simple contrasts $\tau_2 - \tau_1, \tau_4 - \tau_1$, and $\tau_5 - \tau_1$ for the BAI data in Table 6.7.

76. Take $k = 4$ and construct a data example where $Z_{13} + Z_{24} \neq Z_{14} + Z_{23}$. (Note that $\Delta_{13} + \Delta_{24} = \Delta_{14} + \Delta_{23}$. See also Comment 73.)

77. As suggested by the application of the Dwass–Steel–Critchlow–Fligner multiple comparison procedure (in Example 6.6), estimate the contrast $\theta = [\frac{1}{2}(\tau_1 + \tau_2) - \frac{1}{2}(\tau_3 + \tau_4)]$ for the gizzard shad data in Table 6.3.

78. Estimate the contrast $\theta = [\frac{1}{3}(\tau_2 + \tau_3 + \tau_4) - \frac{1}{3}(\tau_1 + \tau_5 + \tau_6)]$ for the revertant colonies data in Table 6.10.

79. Estimate the simple contrasts $\tau_2 - \tau_1$ and $\tau_3 - \tau_1$ for the plasma glucose data in Table 6.9.

80. Estimate all contrasts found to be of interest in Problem 59 for the psychotherapeutic attraction data in Table 6.2.

6.9 SIMULTANEOUS CONFIDENCE INTERVALS FOR ALL SIMPLE CONTRASTS (CRITCHLOW–FLIGNER)

A contrast θ (6.77) is said to be a simple contrast if it involves only two treatment effects (i.e., all but two of the a_i coefficients are zero). In this section, we present a method for obtaining simultaneous confidence intervals for the entire collection, C, of all $\binom{k}{2}$ simple contrasts given by

$$C = \{\Delta_{uv} : \Delta_{uv} = \tau_v - \tau_u, 1 \leq u < v \leq k\}. \tag{6.89}$$

Procedure

For each pair of treatments $(u, v), u \neq v = 1, \ldots, k$, define the sample differences

$$D_{ij}^{uv} = X_{jv} - X_{iu}, \quad i = 1, \ldots, n_u^*; \quad j = 1, \ldots, n_v. \tag{6.90}$$

Let $D_{(1)}^{uv} \leq D_{(2)}^{uv} \leq \cdots \leq D_{(n_u n_v)}^{uv}$ denote the ordered values of the $n_u n_v$ D_{ij}^{uv} differences, for $u \neq v = 1, \ldots, k$. Let w_α^* be the upper αth percentile for the distribution of maximum $\{|W_{uv}^*|, u \neq v = 1, \ldots, k\}$ under H_0 (6.2), where W_{uv}^* is the standardized two-sample rank sum statistic (multiplied by $\sqrt{2}$) for the uth and vth samples, as defined previously in (6.61) for the two-sided all-treatments multiple comparisons setting. Comment 55 explains how to obtain the critical values w_α^* for k treatments, sample sizes $n_1, \ldots, n_k$, and available experimentwise error rates α.

For $1 \leq u < v \leq k$, set

$$a_{uv} = \frac{n_u n_v}{2} - w_\alpha^* \left[\frac{n_u n_v (n_u + n_v + 1)}{24}\right]^{1/2} + 1 \tag{6.91}$$

and

$$b_{uv} = a_{uv} - 1. \tag{6.92}$$

The simultaneous $100(1-\alpha)\%$ confidence intervals for the collection C (6.89) of all simple contrasts are then

$$\{[D^{uv}_{(\langle a_{uv}\rangle)}, D^{uv}_{(n_u n_v - \langle b_{uv}\rangle)}), 1 \le u < v \le k\}, \tag{6.93}$$

where $\langle t \rangle$ denotes the greatest integer less than or equal to t. This set of intervals satisfies the condition

$$\begin{aligned} P_{\tau_1,\ldots,\tau_k}(D^{uv}_{(\langle a_{uv}\rangle)} \le \tau_v - \tau_u < D^{uv}_{(n_u n_v - \langle b_{uv}\rangle)}, \text{ for } 1 \le u < v \le k) \\ = 1 - \alpha, \quad \text{for all } -\infty < \tau_i < \infty, i = 1, \ldots, k. \end{aligned} \tag{6.94}$$

(For simultaneous lower confidence bounds for the collection C that are appropriate under the ordered alternatives setting of Section 6.2, see Comment 79.)

Large-Sample Approximation

When H_0 is true, the $[k(k-1)/2]$-component vector $(W^*_{12}, W^*_{13}, \ldots, W^*_{k-1,k})$ has, as min $(n_1, \ldots, n_k)$ tends to infinity, an asymptotic multivariate normal distribution with mean vector $\mathbf{0}$. It then follows (see Comment 58) that w^*_α can be approximated for large sample sizes by q_α, where q_α is the upper αth percentile point for the distribution of the range of k independent $N(0,1)$ variables. Thus, the large-sample approximate simultaneous $100(1-\alpha)\%$ confidence intervals for C (6.89) are given by (6.93) with w^*_α replaced by q_α in the expressions for a_{uv} (6.91) and b_{uv} (6.92). To find q_α for k treatments, we use the R command `cRangekNorm(`α`, k)`. For example, to find $q_{.05}$ for $k = 6$ treatments, we apply `cRangekNorm(.05, 6)` and obtain $q_{.05} = 4.30$.

EXAMPLE 6.10 ***Motivational Effect of Knowledge of Performance—Examples 6.2, 6.8, and 6.9 Continued.***

Consider the Hundal knowledge of performance data originally presented in Example 6.2. In this example, we wish to find simultaneous $100(1-\alpha)\%$ confidence intervals for the $3(2)/2 = 3$ simple contrasts

$$C = \{\tau_2 - \tau_1, \tau_3 - \tau_1, \tau_3 - \tau_2\}.$$

For the sake of illustration, we take $\alpha = .1041$. Using the R program `cSDCFlig(`α`, n)` with $k = 3$ and $n_1 = n_2 = n_3 = 6$, we have $w^*_{.1041} =$ `cSDCFlig(.1041,c(6, 6, 6))` $=$ 2.9439. It follows from expressions (6.91) and (6.92) that

$$a_{12} = a_{13} = a_{23} = \frac{6(6)}{2} - 2.944\left[\frac{6(6)(6+6+1)}{24}\right]^{1/2} + 1 \approx 6.00$$

and

$$b_{12} = b_{13} = b_{23} = 6.00 - 1 = 5.00.$$

Thus, the simultaneous 89.51% confidence intervals for the simple contrasts $\Delta_{12} = \tau_2 - \tau_1, \Delta_{13} = \tau_3 - \tau_1$, and $\Delta_{23} = \tau_3 - \tau_2$ correspond to $[D^{12}_{(\langle 6\rangle)}, D^{12}_{(36-\langle 5\rangle)}) = [D^{12}_{(6)}, D^{12}_{(31)}), [D^{13}_{(6)}, D^{13}_{(31)})$, and $[D^{23}_{(6)}, D^{23}_{(31)})$, respectively. Using the individual differences already computed in Example 6.9 to obtain a point estimate of the contrast $\tau_3 - \tau_1 = \tau_{\text{accurate information}} - \tau_{\text{no information}}$, we see that the three sets of $n_u n_v = 36$ ordered $D^{uv}_{(t)}$'s are given by

$$D^{12}_{(t)} : \{-6, -5, -4, -4, -3, -3, -3, -2, -2, -1, -1, -1, 0, 0, 0, 0, 1, 1,$$
$$2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 6, 6, 7, 7, 9, 9, 12\},$$

$$D^{13}_{(t)} : \{-4, -3, -1, -1, 0, 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4,$$
$$4, 5, 5, 5, 5, 5, 6, 6, 7, 7, 8, 8, 8, 9, 10, 10, 11, 13\},$$

and

$$D^{23}_{(t)} : \{-7, -4, -4, -3, -2, -2, -1, -1, 0, 0, 0, 1, 1, 1, 2, 2, 2, 3,$$
$$3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 8, 8, 8, 10\}.$$

Hence, the simultaneous 89.51% confidence intervals for the simple contrasts $\Delta_{12} = \tau_2 - \tau_1, \Delta_{13} = \tau_3 - \tau_1$, and $\Delta_{23} = \tau_3 - \tau_2$ for the Hundal data are

$$[D^{12}_{(6)}, D^{12}_{(31)}) = [-3, 6),$$

$$[D^{13}_{(6)}, D^{13}_{(31)}) = [0, 8),$$

and

$$[D^{23}_{(6)}, D^{23}_{(31)}) = [-2, 6),$$

respectively.

For the sake of illustration for the associated large-sample approximation, we take an approximate α value of .10. With $k = 3$, we find $q_{.10} = $ `cRangeKNorm(.10, 3)` $= 2.902$. The associated approximate values for the a_{uv}'s and b_{uv}'s are

$$a_{12} = a_{13} = a_{23} \approx \frac{6(6)}{2} - 2.902\left[\frac{6(6)(6+6+1)}{24}\right]^{1/2} + 1 = 6.19$$

and

$$b_{12} = b_{13} = b_{23} \approx 6.19 - 1 = 5.19.$$

As $\langle 6.19\rangle = 6$ and $\langle 5.19\rangle = 5$, we see that the approximate 90% simultaneous confidence intervals for the simple contrasts $\tau_2 - \tau_1, \tau_3 - \tau_1$, and $\tau_3 - \tau_2$ are identical with the exact 89.51% simultaneous confidence intervals for these Hundal data. This provides some indication that the common sample size of six observations is already large enough to

enable the large-sample approximation to be effective for these simultaneous confidence intervals.

Comments

78. *Relationship of Simultaneous Confidence Intervals to Two-Sided All-Treatments Multiple Comparisons*. The simultaneous $100(1-\alpha)\%$ confidence intervals (6.93) for the collection C (6.89) of all simple contrasts are directly related to the Dwass–Steel–Critchlow–Fligner two-sided all-treatments multiple comparison procedure (6.62) discussed in Section 6.5. In fact, for *every* (u, v) pair, $1 \le u < v \le k$, the two-sided multiple comparison procedure (6.62) yields the decision $\tau_u \ne \tau_v$ at an experimentwise error rate α if and only if 0 does not belong to the corresponding simultaneous $100(1-\alpha)\%$ confidence interval $[D^{uv}_{(\langle a_{uv}\rangle)}, D^{uv}_{(n_u n_v - \langle b_{uv}\rangle)})$ for $\tau_v - \tau_u$. Thus, each of the $2^{\binom{k}{2}}$ sets of possible multiple comparison decisions associated with procedure (6.62) at an experimentwise error rate α corresponds to a collection of simultaneous $100(1-\alpha)\%$ confidence intervals (6.93) for C (6.89) for which the particular (u, v) intervals not containing the value 0 match exactly with those treatment pairs for which procedure (6.62) leads to the decision $\tau_u \ne \tau_v$.

79. *Simultaneous* $100(1-\alpha)\%$ *Lower Confidence Bounds*. In situations where an order restriction $\tau_1 \le \tau_2 \le \cdots \le \tau_k$ on the treatment effects is appropriate (see Sections 6.2 and 6.6 for further details), it is more natural to seek out simultaneous $100(1-\alpha)\%$ lower confidence bounds (rather than two-sided intervals) for the collection C (6.89) of simple contrasts. In such a setting, let c^*_α be the critical value for the Hayter–Stone one-sided all-treatments multiple comparison procedure (6.68) and set

$$h_{uv} = \frac{n_u n_v}{2} - c^*_\alpha \left[\frac{n_u n_v (n_u + n_v + 1)}{24}\right]^{1/2} + 1. \tag{6.95}$$

The simultaneous $100(1-\alpha)\%$ lower confidence bounds for the collection C (6.89) suggested by Hayter and Stone (1991) are then given by

$$\left\{[D^{uv}_{(\langle h_{uv}\rangle)}, \infty), 1 \le u < v \le k\right\}, \tag{6.96}$$

where, once again, $\langle t\rangle$ denotes the greatest integer less than or equal to t and the ordered $D^{uv}_{(t)}$'s are as defined in the Procedure of this section. When either the number of treatments exceeds 3 or $k = 3$ and one or more of the sample sizes is larger than 7, Hayter and Stone (1991) suggest approximating c^*_α in expression (6.95) by d_α, the upper αth percentile point for the distribution of

$$D = \underset{1 \le i < j \le k}{\text{maximum}} \left[\frac{Z_j - Z_i}{\left\{\frac{n_i + n_j}{2 n_i n_j}\right\}^{1/2}}\right],$$

where $Z_1, \ldots, Z_k$ are mutually independent and Z_i has an $N(0, 1/n_i)$ distribution, for $i = 1, \ldots, k$. To find d_α for k treatments, we use the R command

`cHayStonLSA(`α, k`)`. For example, to find $d_{.05}$ for $k = 6$ treatments, we apply `cHayStonLSA(.05, 6)` and obtain $d_{.05} = 3.725$.

The relationship between these simultaneous $100(1-\alpha)\%$ lower confidence bounds (6.96) for C (6.89) and the Hayter–Stone one-sided all-treatments multiple comparison procedure (6.68) at experimentwise error rate α is identical to that described in Comment 78 for the simultaneous $100(1-\alpha)\%$ confidence intervals (6.93) for C (6.89) and the two-sided all-treatments multiple comparison procedure (6.62) at experimentwise error rate α.

80. *Pairwise versus Joint Rankings*. In the latter portion of Comment 59, we discussed some of the pros and cons of pairwise rankings versus joint rankings in the one-way layout setting. The simultaneous $100(1-\alpha)\%$ confidence intervals (6.93) for the collection of all simple contrasts (and the analogous simultaneous lower confidence bounds (6.96) discussed in Comment 79) are clearly associated with pairwise rankings. This provides an additional advantage to the use of pairwise rankings, as the joint ranking approach discussed in Comment 59 does not lead directly to such simultaneous confidence intervals or bounds for C (6.89).

Properties

1. *Distribution-Freeness*. For populations satisfying Assumptions A1–A3, (6.94) holds. Hence, we can control the simultaneous coverage probability to be $1-\alpha$ without having more specific knowledge about the form of the underlying F. As a result, the intervals in (6.93) are distribution-free simultaneous confidence intervals for the collection C (6.89) of all simple contrasts over a very large class of populations.
2. *Asymptotic Multivariate Normality*. See Hayter (1984) and Critchlow and Fligner (1991).

Problems

81. Consider the length of YOY gizzard shad data discussed in Problem 4. Find a set of approximate simultaneous 95% confidence intervals for the set of all simple contrasts.

82. Consider the Hundal knowledge of performance data originally discussed in Example 6.2. Find a set of simultaneous 88.11% lower confidence bounds for the three simple contrasts $\tau_2 - \tau_1, \tau_3 - \tau_1$, and $\tau_3 - \tau_2$ (see Comment 79). Compare with the set of 89.59% simultaneous confidence intervals obtained in Example 6.10 for these same simple contrasts.

83. Consider the Acid Red 114 revertant colonies data in Table 6.10. Find a set of approximate simultaneous 90% confidence intervals for the set of all simple contrasts.

84. Consider the tiger muskellunge plasma glucose data in Table 6.9. Find a set of approximate simultaneous 95% confidence intervals for the set of all simple contrasts.

85. Consider the white-tailed deer fasting metabolic rate data in Table 6.8. Find a set of approximate simultaneous 80% confidence intervals for the set of all simple contrasts.

86. Consider the half-time of mucociliary clearance data in Table 6.1. Find a set of approximate simultaneous 91.81% confidence intervals for the set of all simple contrasts. Without further calculations, what decisions would be reached for these data by the multiple comparison procedure (6.62) at experimentwise error rate $\alpha = .0819$? (See Comment 78.)

87. Consider the average basal area increment data in Table 6.7. Find a set of approximate simultaneous 90% confidence intervals for the set of all simple contrasts. Do you have any concerns about the application of this procedure to these data?

88. Consider the Wechsler Adult Intelligence Scale data in Table 6.11. For the age groups 16–19, 20–34, 35–54, and 55–69 years only, find a set of approximate simultaneous 90% lower confidence bounds for the set of all simple contrasts for these four age groups. Without further calculations, what decisions would be reached for these data by the multiple comparison procedure (6.70) at approximate experimentwise error rate $\alpha = .10$? (See Comments 78 and 79.)

6.10 EFFICIENCIES OF ONE-WAY LAYOUT PROCEDURES

The Pitman asymptotic relative efficiencies for translation alternatives of most of the nonparametric procedures discussed in this chapter with respect to the corresponding normal theory procedures are given by the expression

$$e_F = 12\sigma_F^2 \left\{ \int_{-\infty}^{\infty} f^2(u)du \right\}^2, \tag{6.97}$$

where σ_F^2 is the variance of the common underlying (continuous) distribution F (6.1) and $f(\cdot)$ is the probability density function corresponding to F. The parameter $\int_{-\infty}^{\infty} f^2(u)\, du$ is the area under the curve associated with $f^2(\cdot)$, the square of the common probability density function. We note that this same expression (6.97) also yields the corresponding Pitman efficiencies in the one-sample and two-sample location settings (see Sections 3.11 and 4.5).

In particular, the Pitman asymptotic relative efficiency of the Kruskal–Wallis test based on H (6.5) with respect to the normal theory one-way layout $\mathcal{F}$-test was found to be e_F (6.97) by Andrews (1954). The asymptotic relative efficiency of the Jonckheere–Terpstra test for ordered alternatives based on the statistic J (6.13) with respect to a suitable normal theory competitor was found by Puri (1965) to be e_F (6.97) as well. Mack and Wolfe (1981) found the same expression to hold for the asymptotic relative efficiency of their peak-known umbrella test procedure based on A_p (6.31) relative to an analogous normal theory procedure based on sample averages. Fligner and Wolfe (1982) found the same to be case for the treatments-versus-control test based on FW (6.50).

Sherman (1965) obtained e_F (6.97) as the asymptotic relative efficiency of the two-sided all-treatments and the one-sided treatments-versus-control multiple comparison procedures discussed in Sections 6.5 and 6.7 with respect to the corresponding classical normal theory procedures based on sample means. Spjøtvoll (1968) showed that, when n_j/N tends to ρ_j, with $0 < \rho_j < 1$, the estimators W_{hj} (6.85) have the same asymptotic properties as the estimators $[Z_{h\cdot} - Z_{j\cdot}]$ (see Comment 74). It then follows from Lehmann's (1963a) results that e_F (6.97) is the asymptotic relative efficiency of the estimator $\widehat{\theta}$ (6.83) with respect to the least squares estimator $\overline{\theta} = \sum_{h=1}^{k} \sum_{j=1}^{k} d_{hj}(\overline{X}_{\cdot h} - \overline{X}_{\cdot j})$, where

$$\overline{X}_{\cdot t} = \sum_{i=1}^{n_t} \frac{X_{it}}{n_t}, \quad \text{for } t = 1, \ldots, k.$$

As noted in both Sections 3.11 and 4.5, the asymptotic relative efficiency e_F (6.97) is always greater than or equal to .864 and can be infinite. See expression (3.116) for the value of e_F (6.97) for a variety of underlying F populations.

We do not know of any results for the asymptotic relative efficiencies of the Mack–Wolfe peak-unknown umbrella test (Section 6.3B), the Hayter–Stone one-sided all-treatments multiple comparison procedure (Section 6.6), or the Critchlow–Fligner procedure for simultaneous confidence intervals for all simple contrasts (Section 6.9).

Chapter 7

The Two-Way Layout

INTRODUCTION

The procedures of this chapter are designed for statistical analyses of data collected under the auspices of an experimental design involving two factors, each at two or more levels. Our primary interest is in the relative location effects (medians) of the different levels of one of these factors, hereafter called the *treatment* factor, *within* the various levels of the second factor, hereafter called the *blocking* factor. This blocking factor is associated quite commonly with the experimental design where subjects are first divided into more homogeneous subgroups (called *blocks*) and then randomly assigned to the various treatment levels within these blocks. Such a design is called a *randomized block design*, and we will use this treatment/block terminology to describe the two-way layout setting throughout this chapter. In addition, we will refer, without loss of generality, to the k levels of a treatment as the k *treatments*. (In the case of a randomized complete block design, where the data consist of one observation on each of k treatments in each of n blocks, this represents a direct generalization of the paired replicates setting discussed in Chapter 3.)

The basic null hypothesis of interest is that of no differences in the location effects (medians) of the k treatments within each of the blocks. The alternatives considered here correspond to either general or ordered differences between the treatment effects (medians). As with the one-way layout setting in Chapter 6, we also differentiate between those cases where all of the k treatments represent "new" categories for study and those where one of the treatments corresponds to a control or a baseline category. Finally, we must deal separately with a variety of different possibilities (and correspondingly different statistical procedures) for the number of observations available from each treatment–block combination (cell), ranging from 0 (missing data), 1, to more than 1 (replications).

Sections 7.1–7.5 are devoted to the case of one observation per treatment–block cell (commonly known as a randomized complete block design). Section 7.1 presents a distribution-free test directed at general alternatives. A distribution-free test designed specifically to detect ordered differences among the k treatments is discussed in Section 7.2. Multiple comparison procedures designed to detect which, if any, treatment effects differ from one another are presented in Section 7.3 (all-treatments comparisons) and 7.4 (treatments-versus-control comparisons). In Section 7.5 we present estimators of contrasts in the treatment effects.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

Sections 7.6–7.8 deal with settings where certain treatment–block cells yield single observations, but there are also treatment–block combinations for which we have no observations; that is, we have either zero or one observation from each treatment–block cell. Sections 7.6 and 7.7 present a distribution-free hypothesis test for general alternatives and an all-treatments multiple comparison procedure, respectively, for the structured setting where the data arise from a balanced incomplete block design (BIBD). Section 7.8 discusses a distribution-free hypothesis test for general alternatives in a two-way layout with an arbitrary configuration of either zero or one observation per cell.

In Sections 7.9 and 7.10 we discuss procedures for the setting where there is at least one observation from each cell and there are some cells with multiple observations (replications). Section 7.9 presents a distribution-free hypothesis test for general alternatives for this replications setting, with an emphasis on the special case where we have an equal number (>1) of replications in each cell. An all-treatments multiple comparison procedure for this setting of an equal number of replications is detailed in Section 7.10.

All of the procedures in Sections 7.1–7.10 are associated with within-blocks rankings (known as the *Friedman ranks*) and represent direct extensions to the two-way layout of the paired replicates sign procedures discussed in Sections 3.4–3.6. The corresponding extensions to the two-way layout of the paired replicates Wilcoxon signed ranks procedures discussed in Sections 3.1–3.3 yield asymptotically (number of blocks tending to infinity) distribution-free test and multiple comparison procedures, and we present simplified conservative versions that are nearly asymptotically distribution-free. In Sections 7.11–7.15 we discuss these extensions associated with Wilcoxon signed ranks for data from a randomized complete block design with k treatments and n blocks. Section 7.11 contains a conservative signed ranks test directed at general alternatives, and Section 7.12 presents the corresponding conservative signed ranks test procedure designed for ordered alternatives. The associated approximate signed ranks multiple comparison procedures are given in Sections 7.13 (all-treatments comparisons) and 7.14 (treatments-versus-control comparisons). Section 7.15 contains the contrast estimators linked to the Wilcoxon signed ranks.

The asymptotic relative efficiencies for translation alternatives of the procedures with respect to their normal theory counterparts are discussed in Section 7.16.

	Treatments			
Blocks	1	2	...	k
1	X_{111}	X_{121}	...	X_{1k1}
	$\vdots$	$\vdots$	...	$\vdots$
	$X_{11c_{11}}$	$X_{12c_{12}}$	...	$X_{1kc_{1k}}$
2	X_{211}	X_{221}	...	X_{2k1}
	$\vdots$	$\vdots$	...	$\vdots$
	$X_{21c_{21}}$	$X_{22c_{22}}$	...	$X_{2kc_{2k}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	X_{n11}	X_{n21}	...	X_{nk1}
	$\vdots$	$\vdots$	...	$\vdots$
	$X_{n1c_{n1}}$	$X_{n2c_{n2}}$	...	$X_{nkc_{nk}}$

Data. The data consist of $N = \sum_{i=1}^{n} \sum_{j=1}^{k} c_{ij}$ observations, with c_{ij} observations from the combination of the ith block with the jth treatment (i.e., the (i, j)th cell), for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

Assumptions

A1. The N random variables $\{(X_{ij1}, \ldots, X_{ijc_{ij}}), i = 1, \ldots, n \text{ and } j = 1, \ldots, k\}$ are mutually independent.

A2. For each fixed (i, j), with $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, k\}$, the c_{ij} random variables $(X_{ij1}, \ldots, X_{ijc_{ij}})$ are a random sample from a continuous distribution with distribution function F_{ij}.

A3. The distribution functions $F_{11}, \ldots, F_{1k}, \ldots, F_{n1}, \ldots, F_{nk}$ are connected through the relationship

$$F_{ij}(u) = F(u - \beta_i - \tau_j), \quad -\infty < u < \infty, \tag{7.1}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, k$, where F is a distribution function for a continuous distribution with unknown median θ, β_i is the unknown additive effect contributed by block i, and τ_j is the unknown additive treatment effect contributed by the jth treatment.

We note that Assumptions A1–A3 correspond directly to the usual two-way layout *additive* (See Comment 6) model associated with normal theory assumptions; that is, Assumptions A1–A3 are equivalent to the representation

$$X_{ijt} = \theta + \beta_i + \tau_j + e_{ijt}, \quad i = 1, \ldots, n; \quad j = 1, \ldots, k; \quad t = 1, \ldots, c_{ij},$$

where θ is the overall median, τ_j is the treatment j effect, β_i is the block i effect, and the N e's form a random sample from a continuous distribution with median 0. (Under the additional assumption of normality, the medians θ and 0 are, of course, also the respective means.)

Hypothesis

The null hypothesis of interest in Sections 7.1, 7.2, 7.6, 7.8, 7.9, 7.11, and 7.12 is that of no differences among the additive treatment effects $\tau_1, \ldots, \tau_k$, namely,

$$H_0 : [\tau_1 = \cdots = \tau_k]. \tag{7.2}$$

The null hypothesis asserts that the underlying distributions $F_{i1}, \ldots, F_{ik}$ within block i are the same, for each fixed $i = 1, \ldots, n$; that is, $F_{i1} \equiv F_{i2} \equiv \cdots \equiv F_{ik} \equiv F_i$, for $i = 1, \ldots, n$, in (7.1).

In Sections 7.1–7.5 we consider the special case of one observation per treatment–block combination (commonly known as a *randomized complete block design*), corresponding to $c_{ij} = 1$ for every $i = 1, \ldots, n$ and $j = 1, \ldots, k$. For ease of notation in these five sections, we drop the third subscript on the X variables, since it is always equal to 1 in this setting.

7.1 A DISTRIBUTION-FREE TEST FOR GENERAL ALTERNATIVES IN A RANDOMIZED COMPLETE BLOCK DESIGN (FRIEDMAN, KENDALL-BABINGTON SMITH)

In this section we present a procedure for testing H_0 (7.2) against the general alternative that at least two of the treatment effects are not equal, namely,

$$H_1 : [\tau_1, \ldots, \tau_k \text{ not all equal}], \tag{7.3}$$

when $c_{ij} \equiv 1$, for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

Procedure

To compute the Friedman (1937) statistic S, we first order the k observations from least to greatest separately within each of the n blocks. Let r_{ij} denote the rank of X_{ij} in the joint ranking of the observations $X_{i1}, \ldots, X_{ik}$ in the ith block and set

$$R_j = \sum_{i=1}^{n} r_{ij} \quad \text{and} \quad R_{.j} = \frac{R_j}{n}. \tag{7.4}$$

Thus, for example, R_2 is the sum (over the n blocks) of the within-blocks ranks received by the treatment 2 observations and $R_{.2}$ is the average within-blocks rank for these same observations. The Friedman statistic S is then given by

$$\begin{aligned} S &= \frac{12n}{k(k+1)} \sum_{j=1}^{k} \left(R_{.j} - \frac{k+1}{2} \right)^2 \\ &= \left[\frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3n(k+1), \end{aligned} \tag{7.5}$$

where $(k+1)/2 = \sum_{i=1}^{n} \sum_{j=1}^{k} r_{ij}/nk$ is the average rank assigned via this within-blocks ranking scheme.

To test

$$H_0 = [\tau_1 = \cdots = \tau_k]$$

versus the general alternative

$$H_1 : [\tau_1, \ldots, \tau_k \text{ not all equal}],$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } S \geq s_\alpha; \quad \text{otherwise do not reject,} \tag{7.6}$$

where the constant s_α is chosen to make the type I error probability equal to α. The constant s_α is the upper α percentile for the null $(\tau_1 = \cdots = \tau_k)$ distribution of S. Comment 8 explains how to obtain the critical values s_α for k treatments, n blocks, and available levels of α.

Large-Sample Approximation

When H_0 is true, the statistic S has, as n tends to infinity, an asymptotic chi-square (χ^2) distribution with $k-1$ degrees of freedom. (See Comment l0 for indications of the proof.) The chi-square approximation for procedure (7.6) is

$$\text{Reject } H_0 \text{ if } S \geq \chi^2_{k-1,\alpha}; \quad \text{otherwise do not reject,} \tag{7.7}$$

where $\chi^2_{k-1,\alpha}$ is the upper α percentile point of a chi-square distribution with $k-1$ degrees of freedom. To find $\chi^2_{k-1,\alpha}$ we use the `R` command `qchisq`$(1-\alpha, k-1)$. For example, to find $\chi^2_{5,.05}$, we apply `qchisq(.95, 5)` and obtain $\chi^2_{5,.05} = 11.071$.

Ties

If there are ties among the k observations in a given block, assign each of the observations in a tied group the average of the within-blocks integer ranks that are associated with the tied group and compute S with these within-blocks average ranks. As a consequence of the effect that ties have on the null distribution of S, the following modification is required to apply either procedure (7.6) or the large-sample approximation in (7.7) when there are tied data values within any of the blocks. For either of these procedures, we replace S by

$$\begin{aligned} S' &= \frac{12\sum_{j=1}^{k}\left(R_j - \dfrac{n(k+1)}{2}\right)^2}{nk(k+1) - [1/(k-1)]\sum_{i=1}^{n}\left\{\left(\sum_{j=1}^{g_i} t_{i,j}^3\right) - k\right\}} \\ &= \frac{12\sum_{j=1}^{k} R_j^2 - 3n^2k(k+1)^2}{nk(k+1) - [1/(k-1)]\sum_{i=1}^{n}\left\{\left(\sum_{j=1}^{g_i} t_{i,j}^3\right) - k\right\}}, \end{aligned} \tag{7.8}$$

where g_i denotes the number of tied groups in the ith block and $t_{i,j}$ is the size of the jth tied group in the ith block. We note that an untied observation within a block is considered to be a tied group of size 1. In particular, if there are no ties among the X's in the ith block, then $g_i = k, t_{i,j} = 1$ for each $j = 1, \ldots, k$, and the correction term for the ith block becomes $\{(\sum_{j=1}^{g_i} t_{i,j}^3) - k\} = k - k = 0$. If each block is void of ties, then we have $\sum_{i=1}^{n}\{(\sum_{j=1}^{g_i} t_{i,j}^3) - k\} = 0$ and S' (7.8) reduces to S, as given in (7.5).

We note that even the small-sample procedure (7.6) is only approximately, and not exactly, of significance level α in the presence of tied X observations within any of the blocks. To get an exact level α-test in this tied setting, see Comment 9.

EXAMPLE 7.1 ***Rounding First Base.***

The data in Table 7.1 were obtained by Woodward (1970) in a study to determine which, if any, of three methods of rounding first base is best, in the sense that it minimizes, on the average, the time to reach second base. The three methods, "round out," "narrow angle," and "wide angle" are illustrated in Figure 7.1.

Table 7.1 Rounding-First-Base Times

Players	Methods		
	Round out	Narrow angle	Wide Angle
1	5.40 (1)	5.50 (2)	5.55 (3)
2	5.85 (3)	5.70 (1)	5.75 (2)
3	5.20 (1)	5.60 (3)	5.50 (2)
4	5.55 (3)	5.50 (2)	5.40 (1)
5	5.90 (3)	5.85 (2)	5.70 (1)
6	5.45 (1)	5.55 (2)	5.60 (3)
7	5.40 (2.5)	5.40 (2.5)	5.35 (1)
8	5.45 (2)	5.50 (3)	5.35 (1)
9	5.25 (3)	5.15 (2)	5.00 (1)
10	5.85 (3)	5.80 (2)	5.70 (1)
11	5.25 (3)	5.20 (2)	5.10 (1)
12	5.65 (3)	5.55 (2)	5.45 (1)
13	5.60 (3)	5.35 (1)	5.45 (2)
14	5.05 (3)	5.00 (2)	4.95 (1)
15	5.50 (2.5)	5.50 (2.5)	5.40 (1)
16	5.45 (1)	5.55 (3)	5.50 (2)
17	5.55 (2.5)	5.55 (2.5)	5.35 (1)
18	5.45 (1)	5.50 (2)	5.55 (3)
19	5.50 (3)	5.45 (2)	5.25 (1)
20	5.65 (3)	5.60 (2)	5.40 (1)
21	5.70 (3)	5.65 (2)	5.55 (1)
22	6.30 (2.5)	6.30 (2.5)	6.25 (1)
	$R_1 = 53$	$R_2 = 47$	$R_3 = 32$

Source: W. F. Woodward (1970).

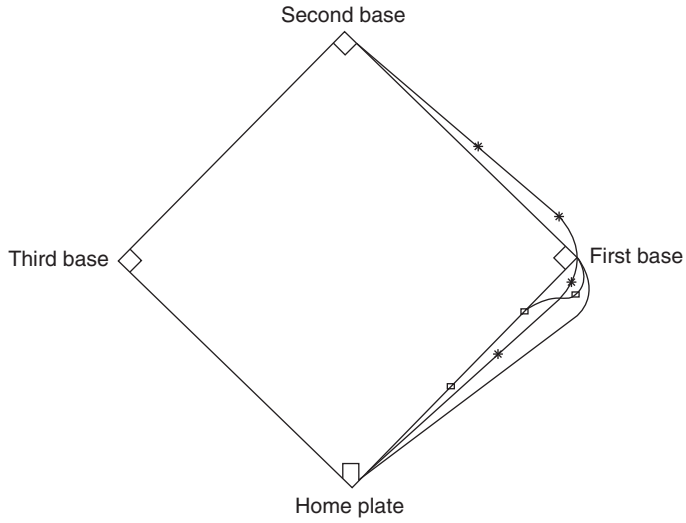

Figure 7.1 Three methods of rounding first base: ◊ path = round out method, ∗ path = narrow angle method, solid path = wide angle method.

Twenty-two baseball players participated in the study, and each of them ran from home plate to second base six times. Using a randomized order, these six trials per player were evenly divided (two each) among the three methods (round out, narrow angle, and wide angle). The entries in Table 7.1 are average times of the two runs per method from a point on the first base line 35 ft from home plate to a point 15 ft short of second base. The within-blocks (players) ranks (r_{ij}'s) of the observations are also given in Table 7.1 in parentheses after the data values (using average ranks to break the ties) and the treatment (running method) rank sums (R_1, R_2, and R_3) are provided at the bottom of the columns.

Since ties exist in blocks 7, 15, 17, and 22, we use S' (7.8). The term in braces in the denominator of (7.8) is zero for each block i in which there are no tied observations. Thus, we need to evaluate that term only for $i = 7$, 15, 17, and 22, corresponding to the blocks in which ties exist. In block 7 there is one tied group of size 2 (5.40) and one tied group of size 1 (5.35). Thus, $t_{7,1} = 2, t_{7,2} = 1$, $g_7 = 2$, and $\{(\sum_{j=1}^{g_7} t_{7,j}^3) - k\} = \{(2^3 + 1^3) - 3\} = 6$. In the same way $\{(\sum_{j=1}^{g_i} t_{i,j}^3) - k\} = 6$ for $i = 15$, 17, and 22. Hence, from (7.8) we obtain

$$S' = \frac{12[(53-44)^2 + (47-44)^2 + (32-44)^2]}{22(3)(4) - \left(\frac{1}{2}\right)(6+6+6+6)} = 11.1.$$

For the large-sample approximation, we compare the value of S' to the chi-square distribution with $k - 1 = 2$ degrees of freedom. Since `1 - pchisq (11.1, 2) = 1 -.9961 = .0039`, we see that the lowest level at which we reject H_0, using the large-sample procedure (7.7) adjusted for ties, is approximately .004. Hence, there is strong evidence here to reject the hypothesis that the methods are equivalent with respect to time to reach second base.

Comments

1. *Basic Model*. Model (7.1) is the most basic form of the two-way layout There is just one observation per cell, and we assume that there is no interaction between the block and treatment factors.
2. *More General Setting*. We could replace Assumptions A1–A3 and H_0 (7.2) with the more general null hypothesis that all possible $(k!)^n$ rank configurations for the r_{ij}'s are equally likely. Procedure (7.6) remains distribution-free for this more general hypothesis.
3. *Design Rationale*. The n blocks in this basic two-way layout design represent an effort to reduce experimental errors and prevent misleading comparisons of "apples and oranges." (We prefer to compare apples with apples.) Thus, in Example 7.1, the 22 blocks correspond to 22 different baseball players. The treatments are to be assigned at random within each block (i.e., in each block, the order in which each player is assigned to run the three different rounding-first-base methods should be decided by a random mechanism, where each of the six possible orders has equal probability of being chosen, and the assignments in the different blocks are to be independent). Note that in the Procedure, we rank only within each block. Thus, in block 1, for example, the three treatment times of player 1 are compared. This is an attempt to eliminate a nuisance effect due to player 1's intrinsic speed. It would be foolish to compare round out times of player 1 with wide angle times of player 2 if player 1 is a (slow) 200-1b catcher

and player 2 is a (speedy) 160-lb shortstop. In such a comparison, a difference in treatment effects would be confounded with the basic speed differences of the players, the latter being of little or no interest in this particular experiment.

4. *Motivation for the Test*. Under Assumptions A1–A3 and H_0 (7.2), each of the block rank vectors $\mathbf{R}_i^* = (r_{i1}, \ldots, r_{ik}), i = 1, \ldots, n$, has a uniform distribution over the set of all $k!$ permutations of the vector of integers $(1, 2, \ldots, k)$. It follows that

$$E_0(r_{ij}) = \frac{1}{k!}(k-1)! \sum_{t=1}^{k} t = \frac{k+1}{2},$$

the average rank being assigned separately in each of the blocks. Thus, we have

$$E_0(R_{\cdot j}) = E_0\left(\frac{1}{n}R_j\right) = \frac{1}{n}E_0\left(\sum_{i=1}^{n} r_{ij}\right) = \frac{1}{n}\sum_{i=1}^{n} E_0(r_{ij})$$
$$= \frac{n(k+1)}{2n} = \frac{k+1}{2}, \quad \text{for } j = 1, \ldots, k,$$

and we would expect the $R_{\cdot j}$'s to be close to $(k+1)/2$ when H_0 is true. Since the test statistic S (7.5) is a constant times a sum of squared differences between the observed treatment average ranks, $R_{\cdot j}$ and their common null expected value, $E_0(R_{\cdot j}) = (k+1)/2$, small values of S represent agreement with H_0 (7.2). When the τ's are not all equal, we would expect a portion of the associated treatment average ranks to differ from their common null expectation, $(k+1)/2$, with some tending to be smaller and some larger. The net result (after squaring the observed differences to obtain the $\left[R_{\cdot j} - (k+1)/2\right]^2$ terms) would be a large value of S. This naturally suggests rejecting H_0 in favor of H_1 (7.3) for large values of S and motivates procedures (7.6) and (7.7). (See also Comment 5.)

5. *Connection to Normal Theory Test*. The Friedman S statistic also arises naturally if we apply the usual two-way layout $\mathcal{F}$ statistic to the ranks instead of the actual observations. Then S may be written as $S = [12/k(k+1)]$ SST, where SST is the treatment sum of squares applied to the ranks.

6. *Assumptions.* We emphasize that Assumption A3 stipulates that the nk cell distributions $F_{ij}, i = 1, \ldots, n$ and $j = 1, \ldots, k$, can differ at most in their locations (medians) and that these location differences (if any) must be a result of additive block and/or treatment effects (i.e., there is no interaction between the treatment and block factors). In particular, Assumption A3 requires that the nk underlying distributions belong to the same general family (F) and that they do not differ in scale parameters (variability). We do note, that the test procedure (7.6) remains distribution-free under the less restrictive setting where Assumption A3 is replaced by the weaker condition

   A3′. The distribution functions $F_{11}, \ldots, F_{1k}, \ldots, F_{n1}, \ldots, F_{nk}$ are connected through the relationship

$$F_{ij}(u) = F_i(u - \tau_j), \quad -\infty < u < \infty,$$

   for $i = 1, \ldots, n$ and $j = 1, \ldots, k$, where $F_1, \ldots, F_n$ are arbitrary distribution functions for continuous distributions with unknown medians $\theta_1, \ldots, \theta_n$,

respectively, and, as before, τ_j is the unknown additive treatment effect contributed by the jth treatment.

Assumption A3 then corresponds to Assumption A3′ with the additional condition that $F_1 \equiv \cdots \equiv F_n$. (See also Comment 2.)

7. *Special Case of Two Treatments*. For the case of $k = 2$ treatments, the procedures in (7.6) and (7.7) are equivalent to the exact and large-sample approximation forms, respectively, of the two-sided sign test, as discussed in Section 3.4.

8. *Derivation of the Distribution of S under H_0 (No-Ties Case)*. The null distribution of S (7.5) can be obtained by using the fact that under H_0 (7.2), all possible $(k!)^n$ rank configurations for the r_{ij}'s are equally likely. We now take $k = 4$, $n = 2$ to illustrate how the null distribution can be derived. In this case, S (7.5) reduces to $S = (.3R^* - 30)$, where $R^* = R_1^2 + R_2^2 + R_3^2 + R_4^2$. We note that S does not vary with changes of the names of the blocks or with relabeling of the k samples. Thus, for example,

(a)

	I	II	III	IV
Block 1	1	2	3	4
Block 2	3	1	2	4

(b)

	I	II	III	IV
Block 1	3	1	2	4
Block 2	1	2	3	4

yield the same value of S, because (b) is obtained from (a) by reversing the roles of blocks 1 and 2. Similarly,

(c)

	I	II	III	IV
Block 1	1	2	3	4
Block 2	3	1	2	4

(d)

	I	II	III	IV
Block 1	2	1	3	4
Block 2	1	3	2	4

yield the same value of S, since (d) is obtained from (c) by reversing the roles of samples I and II. Instead of $(4!)^2$ rank configurations, therefore, we list only $4! = 24$ configurations (the 24 different configurations in block 2 corresponding to a fixed configuration 1, 2, 3, 4 in block 1) and their associated values of R^* and S.

(a)

I	II	III	IV
1	2	3	4
1	2	3	4

$R^* = 120, S = 6$

(b)

I	II	III	IV
1	2	3	4
1	2	4	3

$R^* = 118, S = 5.4$

(c)

I	II	III	IV
1	2	3	4
1	3	4	2

$R^* = 114, S = 4.2$

(d)

I	II	III	IV
1	2	3	4
1	3	2	4

$R^* = 118, S = 5.4$

(e)

I	II	III	IV
1	2	3	4
1	4	2	3

$R^* = 114, S = 4.2$

(f)

I	II	III	IV
1	2	3	4
1	4	3	2

$R^* = 112, S = 3.6$

(g)

I	II	III	IV
1	2	3	4
2	1	3	4

$R^* = 118, S = 5.4$

(h)

I	II	III	IV
1	2	3	4
2	1	4	3

$R^* = 116, S = 4.8$

(i)

I	II	III	IV
1	2	3	4
2	3	4	1

$R^* = 108, S = 2.4$

(j)

I	II	III	IV
1	2	3	4
2	3	1	4

$R^* = 114, S = 4.2$

(k)

I	II	III	IV
1	2	3	4
2	4	1	3

$R^* = 110, S = 3$

(l)

I	II	III	IV
1	2	3	4
2	4	3	1

$R^* = 106, S = 1.8$

(m)

I	II	III	IV
1	2	3	4
3	1	2	4

$R^* = 114, S = 4.2$

(n)

I	II	III	IV
1	2	3	4
3	1	4	2

$R^* = 110, S = 3$

(o)

I	II	III	IV
1	2	3	4
3	2	4	1

$R^* = 106, S = 1.8$

(p)

I	II	III	IV
1	2	3	4
3	2	1	4

$R^* = 112, S = 3.6$

(q)

I	II	III	IV
1	2	3	4
3	4	1	2

$R^* = 104, S = 1.2$

(r)

I	II	III	IV
1	2	3	4
3	4	2	1

$R^* = 102, S = .6$

(s)

I	II	III	IV
1	2	3	4
4	1	2	3

$R^* = 108, S = 2.4$

(t)

I	II	III	IV
1	2	3	4
4	1	3	2

$R^* = 106, S = 1.8$

(u)

I	II	III	IV
1	2	3	4
4	2	1	3

$R^* = 106, S = 1.8$

(v)

I	II	III	IV
1	2	3	4
4	2	3	1

$R^* = 102, S = .6$

(w)

I	II	III	IV
1	2	3	4
4	3	1	2

$R^* = 102, S = .6$

(x)

I	II	III	IV
1	2	3	4
4	3	2	1

$R^* = 100, S = 0$

Thus, we find

$$P_0\{S = 6\} = \tfrac{1}{24}, \quad P_0\{S = 5.4\} = \tfrac{3}{24}, \quad P_0\{S = 4.8\} = \tfrac{1}{24},$$

$$P_0\{S = 4.2\} = \tfrac{4}{24}, \quad P_0\{S = 3.6\} = \tfrac{2}{24}, \quad P_0\{S = 3\} = \tfrac{2}{24},$$

$$P_0\{S = 2.4\} = \tfrac{2}{24}, \quad P_0\{S = 1.84\} = \tfrac{4}{24}, \quad P_0\{S = 1.2\} = \tfrac{1}{24},$$

$$P_0\{S = .6\} = \tfrac{3}{24}, \quad P_0\{S = 0\} = \tfrac{1}{24}.$$

The probability, under H_0, that S is greater than or equal to 5.4, for example, is therefore

$$\begin{aligned} P_0\{S \geq 5.4\} &= P_0\{S = 5.4\} + P_0\{S = 6\} \\ &= \tfrac{3}{24} + \tfrac{1}{24} = \tfrac{1}{6}. \end{aligned}$$

Note that we have derived the null distribution of S without specifying the common form (F) of the underlying distribution function for the X's under H_0

beyond the point of requiring that it be continuous. This is why the test procedure (7.6) based on S is called a *distribution-free Procedure*. From the null distribution of S, we can determine the critical value s_α and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying continuous X distribution.

For a given number of treatments k and blocks n, the R command `cFrd(α,k,n)` can be used to find the available upper-tail critical values s_α for possible values of S. For a given available significance level α, the critical value s_α then corresponds to $P_0(S \geq s_\alpha) = \alpha$ and is given by `cFrd(α, k, n)` $= s_\alpha$. Thus, for example, for $k = 5$ and $n = 7$, we have $P_0(S \geq 10.40) = .0261$, so that $s_{.0261} =$ `cFrd(.0261, 5, 7)` $= 10.40$ for $k = 5$ and $n = 7$.

9. *Exact Conditional Distribution of S with Ties among the X Values*. To have a test with exact significance level even in the presence of tied X's, we need to consider all $(k!)^n$ block rank configurations, where now these within-blocks ranks are obtained by using average ranks to break the ties. As in Comment 8, it still follows that under H_0 each of the $(k!)^n$ block rank configurations (now with these tied ranks) is equally likely. For each such configuration, the value of S is computed and the results are tabulated. We illustrate this construction only for the very limited case of $k = 3$, $n = 2$, and the tied data $X_{11} = 2.4$, $X_{12} = 3.0$, $X_{13} = 3.0$, $X_{21} = 4.0$, $X_{22} = 6.0$, and $X_{23} = 3.0$. Using average ranks to break within-blocks ties, the observed rank vector is $(r_{11}, r_{12}, r_{13}, r_{21}, r_{22}, r_{23}) = (1, 2.5, 2.5, 2, 3, 1)$. Thus, $R_1 = 3, R_2 = 5.5, R_3 = 3.5$, and the attained value of S is

$$S = \left[\frac{12}{2(3)(4)}\{(3)^2 + (5.5)^2 + (3.5)^2\} - 3(2)(4)\right] = 1.75.$$

To assess the significance of S, we obtain its conditional null distribution by considering the 36 equally likely (under H_0) possible rank configurations (i.e., permutation combinations) of the observed rank vector (1, 2.5, 2.5, 2, 3, 1). These 36 configurations and associated values of S are as follows:

I	II	III		I	II	III	
1	2.5	2.5		1	2.5	2.5	
2	3	1	$S = 1.75$	2	3	1	$S = 1.75$
2.5	1	2.5		2.5	1	2.5	
2	3	1	$S = 0.25$	2	3	1	$S = 0.25$
2.5	2.5	1		2.5	2.5	1	
2	3	1	$S = 3.25$	2	3	1	$S = 3.25$
1	2.5	2.5		1	2.5	2.5	
2	1	3	$S = 1.75$	2	1	3	$S = 1.75$
2.5	1	2.5		2.5	1	2.5	
2	1	3	$S = 3.25$	2	1	3	$S = 3.25$
2.5	2.5	1		2.5	2.5	1	
2	1	3	$S = 0.25$	2	1	3	$S = 0.25$
1	2.5	2.5		1	2.5	2.5	
1	2	3	$S = 3.25$	1	2	3	$S = 3.25$
2.5	1	2.5		2.5	1	2.5	
1	2	3	$S = 1.75$	1	2	3	$S = 1.75$

I	II	III		I	II	III	
2.5	2.5	1		2.5	2.5	1	
1	2	3	$S = 0.25$	1	2	3	$S = 0.25$
1	2.5	2.5		1	2.5	2.5	
3	2	1	$S = 0.25$	3	2	1	$S = 0.25$
2.5	1	2.5		2.5	1	2.5	
3	2	1	$S = 1.75$	3	2	1	$S = 1.75$
2.5	2.5	1		2.5	2.5	1	
3	2	1	$S = 3.25$	3	2	1	$S = 3.25$
1	2.5	2.5		1	2.5	2.5	
1	3	2	$S = 3.25$	1	3	2	$S = 3.25$
2.5	1	2.5		2.5	1	2.5	
1	3	2	$S = 0.25$	1	3	2	$S = 0.25$
2.5	2.5	1		2.5	2.5	1	
1	3	2	$S = 1.75$	1	3	2	$S = 1.75$
1	2.5	2.5		1	2.5	2.5	
3	1	2	$S = 0.25$	3	1	2	$S = 0.25$
2.5	1	2.5		2.5	1	2.5	
3	1	2	$S = 3.25$	3	1	2	$S = 3.25$
2.5	2.5	1		2.5	2.5	1	
3	1	2	$S = 1.75$	3	1	2	$S = 1.75$

Since each of these values of S has null probability $\frac{1}{36}$, it follows that

$$P_0\{S = 0.25\} = P_0\{S = 1.75\} = P_0\{S = 3.25\} = \tfrac{1}{3}.$$

This distribution is called the *conditional distribution* or the *permutation distribution* of S, given the tied ranks $\{(1, 2.5, 2.5), (1, 2, 3)\}$. For the particular observed value $S = 1.75$, we have $P_0\{S \geq 1.75\} = \frac{2}{3}$.

10. *Large-Sample Approximation*. Define the random variables $T_j = R_{\cdot j} - E_0(R_{\cdot j}) = R_{\cdot j} - (k+1)/2$, for $j = 1, \ldots, k$. Since each $R_{\cdot j} = \sum_{i=1}^{n} r_{ij}/n$ is an average, it is not surprising (see, e.g., pages 388–389 of Lehmann (1975) for justification) that a properly standardized version of the vector $\mathbf{T}^* = (T_1, \ldots, T_{k-1})$ has an asymptotic (n tending to infinity) $(k-1)$-variate normal distribution with mean vector $\mathbf{0} = (0, \ldots, 0)$ and appropriate covariance matrix $\boldsymbol{\Sigma}$ when the null hypothesis H_0 is true. (Note that $\boldsymbol{T}^*$ does not include $T_k = R_{\cdot k} - (k+1)/2$, because T_k can be expressed as a linear combination of $T_1, \ldots, T_{k-1}$. This is the reason that the asymptotic normal distribution is $(k-1)$-variate and not k-variate.) Since the test statistic S (7.5) is a quadratic form in the variables $(T_1, \ldots, T_{k-1})$, it is, therefore, quite natural that S has an asymptotic (n tending to infinity) chi-square distribution with $k-1$ degrees of freedom.

11. *Competitor Based on Wilcoxon Signed Ranks*. The statistic S (7.5) utilizes the treatment observations only through comparisons within blocks. As noted in Comment 7, this provides a natural extension of the sign test procedure for paired data and it is this restriction to within-blocks comparisons that leads directly to the distribution-free nature of procedure (7.6). An alternative approach would be to extend the (generally) more powerful signed rank test procedure, as discussed in Section 3.1. This approach utilizes between-block comparisons

of the observations and is discussed further in Section 7.11. The associated test procedure utilizing between-blocks signed rank comparisons is (generally) more powerful than the Friedman test based on S (7.5). However, this two-way layout signed rank procedure is no longer exactly distribution-free for small numbers (n) of blocks and tests based on this approach require the use of a large-sample approximation.

12. *Consistency of the S Test*. Replace Assumptions A1–A3 by the less restrictive Assumption $A1': X_{ij} = \beta_i + e_{ij}$, where the e's are mutually independent, and Assumption $A2': e_{1j}, \ldots, e_{nj}$ come from the same continuous population $\prod_j, j = 1, \ldots, k$, but where $\prod_1, \ldots, \prod_k$ are not assumed to be identical. Then the test defined by (7.6) is consistent against alternatives for which $\sum_{v=1}^{k}(1 - p_{uv}) \neq \sum_{v=1}^{k} p_{uv}$ for at least one $u \in \{1, \ldots, k\}$, where $p_{uv} = P(e_{iu} < e_{iv})$ with e_{iu} a random member from $\prod_u$ and e_{iv} a random member from $\prod_v$ that is independent of e_{iu}.

Properties

1. *Consistency*. See Noether (1967a, p. 54) and Comment 12.
2. *Asymptotic Chi-Squaredness*. See Lehmann (1975, pp. 388–389).
3. *Efficiency*. See van Elteren and Noether (1959) and Section 7.16.

Problems

1. Goldsmith and Nadel (1969) have studied respiratory function following exposure to various levels of ozone for periods of 1 h. The subjects were four presumably healthy males employed by the California State Department of Public Health. The objective measurement used was airway resistance as evaluated by the body plethysmographic technique (see DuBois et al. (1956) and Comroe, Botelho, and DuBois (1959)). Goldsmith and Nadel reported average values for four consecutive measurements taken immediately prior to and again about 5 min after termination of each level of ozone exposure. Table 7.2 is based on a subset of the Goldsmith-Nadel data, where the tabled values are average airway resistance after ozone exposure minus average airway resistance prior to ozone exposure. Use procedure (7.6) to test H_0.

2. Show that the two expressions for S in (7.5) are, indeed, equivalent.

3. Could Friedman's test be applied to data from a one-way layout in which there are the same number, n, of observations from each of the k treatments? Explain. Should Friedman's test be applied to such data? Explain.

Table 7.2 Effect of Experimental Ozone Exposures on Airway Resistance (cm H_2Ol/s)

Subject	After .1 ppm	After .6 ppm	After 1.0 ppm
1	−.08	.01	.06
2	.21	.17	.19
3	.50	−.11	.34
4	.14	.07	.14

Source: J. R. Goldsmith and J. A. Nadel (1969).

4. Show directly, or illustrate by means of an example, that the maximum value of S is $S_{\max} = n(k-1)$. For what configuration is this maximum achieved?

5. Creatine phosphokinase (CPK) is a skeletal muscle isoenzyme that is often found to be elevated in the serum of acutely psychotic subjects during the initial stages of a psychotic episode. A number of variables known to affect serum CPK activity have been evaluated as possible causes of the serum CPK activity elevations observed during acute psychotic episodes. One such variable of interest is that of physical exercise, which is well known to increase serum CPK levels in normal subjects. In this regard, Goode and Meltzer (1976) studied the relationship between isometric exercises (designed to strengthen and tone muscle without lengthening and contracting the muscles themselves) and increased CPK levels in psychotic patients. In particular, they were interested in whether the elevation of CPK in the serum of psychiatric patients may be in part due to increased covert isometric motor activity. The subjects in their study were patients hospitalized on a research unit at the Illinois Psychiatric Institute. Fourteen such patients were isometrically exercised following remission of psychotic symptoms, usually 2–4 weeks after admission. The 60-min isometric exercise procedure involved stationary wall bars and required maximal use of all major muscle groups. The subjects described the exercises as extremely fatiguing and at or near the limits of their endurances.

Table 7.3 contains the plasma CPK activity (mU/l) levels for each of these 14 patients prior to the period of isometric exercises, as well as at 18 and 42 h after completion of such exercises. Also recorded for each patient is the peak plasma CPK activity exhibited during the period of psychosis immediately following admission to the Institute.

Use these data to assess whether there are any differences in CPK activity between the four patient conditions considered in Table 7.3.

6. Suppose $k = 3$ and $n = 13$. Compare the critical region for the exact level $\alpha = .025$ test of H_0 (7.2) based on S with the critical region for the corresponding nominal level $\alpha = .025$ test based on the large-sample approximation. What is the exact significance level of this .025 nominal level test based on the large-sample approximation?

7. Suppose $k = 3$ and $n = 3$. Obtain the form of the exact null (H_0) distribution of S for the case of no-tied observations.

8. Suppose $k = 4$ and $n = 8$. Compare the critical region for the exact level $\alpha = .005$ test of H_0 (7.2) based on S with the critical region for the corresponding nominal level $\alpha = .005$

Table 7.3 Effect of lsometric Exercise on Serum Creatine Phosphokinase (CPK) Activity (mU/l) in Psychotic Patients

Subject	Preexercise	19 h postexercise	42 h postexercise	Peak-psychotic period
1	27	101	82	63
2	30	112	50	78
3	24	26	68	69
4	54	89	135	1,137
5	21	30	49	57
6	36	41	48	800
7	36	29	46	105
8	16	20	8	111
9	21	26	25	61
10	26	25	31	74
11	65	60	69	190
12	25	27	28	107
13	19	18	21	306
14	48	41	28	109

Source: D. J. Goode and H. Y. Meltzer (1976).

test based on the large-sample approximation. What is the exact significance level of this .005 nominal level test based on the large-sample approximation?

9. Suppose $k = 3$ and $n = 2$, and we observe the data $X_{11} = 3.6, X_{12} = 3.6, X_{13} = 5.2, X_{21} = 4.3, X_{22} = 5.2$, and $X_{23} = 4.3$. What is the conditional probability distribution of S under H_0 (7.2) when average ranks are used to break ties among the X's? How extreme is the observed value of S in this conditional null distribution? Compare this fact with that obtained by taking the observed value of S to the (incorrect) unconditional null distribution of S.

10. Consider the CPK activity data in Table 7.3. Ignoring the patients' peak psychotic period data, assess the conjecture that isometric exercise has an effect on the CPK activity of psychotic patients.

11. Use the CPK data in Table 7.3 and an appropriate nonparametric test procedure to assess whether there is any difference between peak CPK activity during the psychotic period and peak CPK activity over the combined pre/post exercise periods.

12. Nicholls and Ling (1982) conducted a study to assess the effectiveness of a system employing hand cues in the teaching of language to severely hearing-impaired children. In particular, they considered syllables presented to hearing-impaired children under the following seven conditions: (A) audition, (L) lip reading, (AL) audition and lip reading, (C) cued speech, (AC) audition and cued speech, (LC) lip reading and cued speech, and (ALC) audition, lip reading, and cued speech. The 18 subjects in the study were all severely hearing-impaired children who had been taught through the use of cued speech for at least 4 years. Syllables were presented to the subjects under each of the seven conditions (presented in random orders) and the subjects were asked in each case to identify the consonants in each syllable by writing down what they perceived them to be. The subjects' results were scored by marking properly identified consonants in the appropriate order as correct. After tallying the responses, an overall percentage correct was assigned to each participant under each experimental condition. These correct percentage data for the 18 children in the study are given in Table 7.4.

Table 7.4 Percentage Consonants Correctly Identified under Each of the Conditions: (A) Audition, (L) Lip Reading, (AL) Audition and Lip Reading, (C) Cued Speech, (AC) Audition and Cued Speech, (LC) Lip Reading and Cued Speech, and (ALC) Audition, Lip Reading, and Cued Speech

Subject	A	L	AL	C	AC	LC	ALC
1	1.1	36.9	52.4	42.9	31.0	83.3	63.0
2	1.1	33.3	34.5	34.5	41.7	77.3	81.0
3	13.0	28.6	40.5	33.3	44.0	81.0	76.1
4	0	23.8	22.6	33.3	33.3	69.0	65.5
5	11.9	40.5	57.1	35.7	46.4	98.8	96.4
6	0	27.4	46.4	42.9	47.4	78.6	77.4
7	5.0	20.2	22.6	35.7	37.0	69.0	73.8
8	4.0	29.8	42.9	13.0	33.3	95.2	91.7
9	0	27.4	38.0	42.9	45.2	89.3	85.7
10	1.1	26.2	31.0	31.0	32.1	70.2	71.4
11	2.4	29.8	38.0	34.5	46.4	86.9	92.9
12	0	21.4	21.4	41.7	33.3	67.9	59.5
13	0	32.1	33.3	44.0	34.5	86.9	82.1
14	0	28.6	23.8	32.1	39.3	85.7	72.6
15	1.1	28.6	29.8	41.7	35.7	81.0	78.6
16	1.1	36.9	33.3	25.0	31.0	95.2	95.2
17	0	27.4	26.1	40.5	44.0	91.7	89.3
18	0	41.7	35.7	42.9	45.2	95.2	95.2

Source: G. H. Nicholls and D. Ling (1982).

Use these data to assess whether there are any differences in the effectiveness of these seven conditions for teaching severely hearing-impaired children.

13. Consider the study with severely hearing-impaired children in Problem 12. Using the percentage correctly identified data from Table 7.4, assess whether there are any differences in the effectiveness of the three stand-alone conditions A, L, and C for teaching severely hearing-impaired children.

14. Consider the study with severely hearing-impaired children in Problem 12. Using the percentage correctly identified data for only the first eight children (and the proper correction for ties), assess whether there are any differences in the teaching effectiveness from adding one or more of the factors L (lip reading) and C (cued speech) to the baseline A (audition) approach.

7.2 A DISTRIBUTION-FREE TEST FOR ORDERED ALTERNATIVES IN A RANDOMIZED COMPLETE BLOCK DESIGN (PAGE)

In many practical two-way layout settings where an additive model is appropriate, it is also the case that the treatments are such that the appropriate alternatives to no differences in treatment effects (H_0) are those of increasing (or decreasing) treatment effects according to some natural labeling for the treatments. Examples of such settings include treatments corresponding to quality or quantity of materials, severity of disease, drug dosage levels, and intensity of stimulus. We note that the Friedman procedure (7.6) does not utilize any such partial prior information regarding the postulated alternative ordering. The statistic S (7.5) takes on the same value for all possible $k!$ labelings of the treatments. In this section, we consider a procedure for testing H_0 (7.2) against the a priori ordered alternatives,

$$H_2 : [\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k, \text{ with at least one strict inequality}]. \tag{7.9}$$

The Page test of this section is preferred to the Friedman test in Section 7.1 when the treatments can be labeled a priori in such a way that the experimenter expects any deviation from H_0 (7.2) to be in the particular direction associated with H_2 (7.9). We emphasize, however, that the labeling of the treatments, so that the ordered alternatives (7.9) are appropriate, *cannot* depend on the observed sample values. This labeling must correspond completely to a factor (s) implicit in the nature of the *experimental design* and *not* the *observed data.*

Procedure

First, we must label the treatments so that they are in the expected order associated with the alternative H_2 (7.9). (This labeling must be done prior to data collection.) To compute the Page (1963) statistic L, we once again rank within blocks and compute the Friedman treatment sums of ranks $R_1, \ldots, R_k$ as defined in (7.4). The Page statistic L is then the weighted combination of these rank sums given by

$$L = \sum_{j=1}^{k} jR_j = R_1 + 2R_2 + \cdots + kR_k. \tag{7.10}$$

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the ordered alternative

$$H_2 : [\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k, \text{ with at least one strict inequality}],$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } L \geq l_\alpha; \qquad \text{otherwise do not reject,} \tag{7.11}$$

where the constant l_α is chosen to make the type I error probability equal to α. The constant l_α is the upper α percentile for the null $(\tau_1 = \cdots = \tau_k)$ distribution of L. Comment 17 explains how to obtain the critical value l_α for k treatments, n blocks, and available levels of α.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic (n tending to infinity) normality of L, suitably standardized. We first need to know the expected value and variance of L when the null hypothesis is true. Under H_0, the expected value and variance of L are

$$E_0(L) = \frac{nk(k+1)^2}{4} \tag{7.12}$$

and

$$\text{var}_0(L) = \frac{nk^2(k+1)(k^2-1)}{144}, \tag{7.13}$$

respectively. These expressions for $E_0(L)$ and $\text{var}_0(L)$ are verified by direct calculations in Comment 18 for the special case of $k = 3$ and $n = 2$. General derivations of both expressions are outlined in Comment 20.

The standardized version of L is

$$L^* = \frac{L - E_0(L)}{\sqrt{\text{var}_0(L)}} = \frac{L - \left[\dfrac{nk(k+1)^2}{4}\right]}{\left\{\dfrac{nk^2(k+1)(k^2-1)}{144}\right\}^{1/2}}. \tag{7.14}$$

When H_0 is true, L^* has, as n tends to infinity, an asymptotic $N(0,1)$ distribution (see Comment 20 for indications of the proof). The normal theory approximation for procedure (7.11) is

$$\text{Reject } H_0 \text{ if } L^* \geq_{Z_\alpha}; \quad \text{otherwise do not reject.} \tag{7.15}$$

Ties

If there are ties among the k X's within any of the n blocks, assign each of the observations in a tied group the average of the integer ranks that are associated with the tied group and compute L with these average ranks.

We note that even procedure (7.11) using these average ranks to break ties and the critical value l_α is only approximately, and not exactly, of significance level α in the presence of tied X observations within any of the blocks. To get an exact level α test in this tied setting, see Comment 19. (See also Comment 21 regarding the use of the large-sample approximation in the case of within-blocks ties.)

EXAMPLE 7.2 ***Breaking Strength of Cotton Fibers.***

An experiment in Cochran and Cox (1957, p. 108) considered the effect, in terms of breaking strength of cotton fibers, of the level of potash (K_2O) in the soil. Five levels of potash were applied ($k = 5$) in a randomized block pattern with three blocks ($n = 3$). The criterion used for the analysis was the Pressley strength index, obtained by measuring the breaking strength of a bundle of fibers of a given cross-sectional area. A single sample of cotton was taken from each plot, and four determinations were made on each sample. The main entries of Table 7.5 are the means of the four determinations and the parenthetical values are the within-block ranks. (No dimensions are associated with the data of Table 7.5, because the machine that measures the strength index is calibrated in arbitrary units.)

We are interested here in using procedure (7.11) to test the hypothesis of equivalent strengths versus the ordered alternative that specifies a trend of decreasing breaking strength with increasing levels of potash. For the purpose of illustration, we take the significance level to be $\alpha = .0097$. Applying the R command `cPage(α,k,n)` with $k = 5$ and $n = 3$, we find `cPage(.0097, 5, 3) = 155`. That is, $P_0(L \geq 155) = .0097$, and we have that $l_{.0097} = 155$ and procedure (7.11) becomes

$$\text{Reject } H_0 \text{ if } L \geq 155.$$

Now, we illustrate the computations leading to the sample value of L (7.10). Using the treatment sums of within-block ranks given in Table 7.3, we see from (7.10) that

$$\begin{aligned} L &= R_1 + 2R_2 + 3R_3 + 4R_4 + 5R_5 \\ &= 5 + 2(5) + 3(9) + 4(14) + 5(12) = 158. \end{aligned}$$

Table 7.5 Strength Index of Cotton

	Potash (lb/acre)				
Replications	144	108	72	54	36
1	7.46 (2)	7.17 (1)	7.76 (4)	8.14 (5)	7.63 (3)
2	7.68 (2)	7.57 (1)	7.73 (3)	8.15 (5)	8.00 (4)
3	7.21 (1)	7.80 (3)	7.74 (2)	7.87 (4)	7.93 (5)
	$R_1 = 5$	$R_2 = 5$	$R_3 = 9$	$R_4 = 14$	$R_5 = 12$

Source: W. G. Cochran and G. M. Cox (1957).

Since the value of L is greater than the critical value 155, we can reject H_0 at the $\alpha = .0097$ level, providing strong evidence (for the levels of potash considered) in favor of the trend of decreasing breaking strength with increasing level of potash.

For the large-sample approximation we need to compute the standardized form of L^* using (7.14). Since $k = 5$, $n = 3$, and the sample value of L is 158, we see from (7.14) that

$$L^* = \frac{158 - \left[\dfrac{3(5)(5+1)^2}{4}\right]}{\left\{\dfrac{3(5^2)(5+1)(5^2-1)}{144}\right\}^{1/2}} = \frac{158 - 135}{\sqrt{75}} = 2.66.$$

Thus, using the approximate procedure (7.15) with the value of $L^* = 2.66$ and the `R` command `pnorm(·)`, we see that the approximate P-value for these data is $P_0(L^* \geq 2.66) \approx 1 - \texttt{pnorm(2.66)} = 1 - .9961 = .0039$. This is in good agreement with our previous outcome using the exact test, even though n is only 3.

Comments

13. *More General Setting.* As with the Friedman procedure in Section 7.1, we could replace Assumptions A1–A3 and H_0 (7.2) with the more general null hypothesis that all possible $(k!)^n$ rank configurations for the r_{ij}'s are equally likely. Procedure (7.11) remains distribution-free for this more general hypothesis.
14. *Motivation for the Test.* If the ordering $\tau_1 < \tau_2 < \cdots < \tau_k$ is true, then R_v will tend to be larger than R_u for $u < v$. Note that L (7.10) weights R_v by the integer v and R_u by the integer u. Thus, L tends to be large when H_2 (7.9) is true, serving as partial motivation for the L test in (7.11).
15. *Assumptions.* As with the Friedman procedure in Section 7.1, we emphasize that Assumption A3 stipulates that the nk cell distributions $F_{ij}, i = 1, \ldots, n$ and $j = 1, \ldots, k$, can differ at most in their locations (medians) and that these location differences (if any) must be a result of additive block and/or treatment effects (i.e., there is no interaction between the treatment and block factors). In particular, Assumption A3 requires that the nk underlying distributions belong to the same general family (F) and that they do not differ in scale parameters (variability). We do note, however, that the test procedure (7.11) remains distribution-free under the less restrictive setting where Assumption A3 is replaced by the weaker condition A3′ stated in Comment 6. (See also Comment 13.)
16. *Special Case of Two Treatments.* For the case of $k = 2$ treatments, the procedures in (7.11) and (7.15) are equivalent to the exact and large-sample approximation forms, respectively, of the one-sided sign test, as discussed in Section 3.4.
17. *Derivation of the Distribution of L under H_0 (No-Ties Case).* The null distribution of L (7.10) can be obtained by using the fact that under H_0 (7.2), all $(k!)^n$ possible rank configurations are equally likely. As is the case for S (7.5) (see Comment 8), L does not vary with changes of the names of the blocks; however, unlike S, because it is directed toward a particular ordered alternative, the L values (in general) do change with changes of names of the treatments. Thus, building up the null distribution of L is more tedious than in the case of S. We illustrate this construction for the very special case of $k = 3$ and $n = 2$. In this

case, we need to consider $(3!)^2 = 36$ block–treatment rank configurations. We list these configurations and their associated values of $L = R_1 + 2R_2 + 3R_3$.

(a)

I	II	III
1	2	3
1	2	3

$L = 28$

(b)

I	II	III
1	2	3
1	3	2

$L = 27$

(c)

I	II	III
1	2	3
2	1	3

$L = 27$

(d)

I	II	III
1	2	3
2	3	1

$L = 25$

(e)

I	II	III
1	2	3
3	1	2

$L = 25$

(f)

I	II	III
1	2	3
3	2	1

$L = 24$

(g)

I	II	III
1	3	2
1	2	3

$L = 27$

(h)

I	II	III
1	3	2
1	3	2

$L = 26$

(i)

I	II	III
1	3	2
2	1	3

$L = 26$

(j)

I	II	III
1	3	2
2	3	1

$L = 24$

(k)

I	II	III
1	3	2
3	1	2

$L = 24$

(l)

I	II	III
1	3	2
3	2	1

$L = 23$

(m)

I	II	III
2	1	3
1	2	3

$L = 27$

(n)

I	II	III
2	1	3
1	3	2

$L = 26$

(o)

I	II	III
2	1	3
2	1	3

$L = 26$

(p)

I	II	III
2	1	3
2	3	1

$L = 24$

(q)

I	II	III
2	1	3
3	1	2

$L = 24$

(r)

I	II	III
2	1	3
3	2	1

$L = 23$

(s)

I	II	III
2	3	1
1	2	3

$L = 25$

(t)

I	II	III
2	3	1
1	3	2

$L = 24$

(u)

I	II	III
2	3	1
2	1	3

$L = 24$

(v)

I	II	III
2	3	1
2	3	1

$L = 22$

(w)

I	II	III
2	3	1
3	1	2

$L = 22$

(x)

I	II	III
2	3	1
3	2	1

$L = 21$

(y)

I	II	III
3	1	2
1	2	3

$L = 25$

(z)

I	II	III
3	1	2
1	3	2

$L = 24$

(aa)

I	II	III
3	1	2
2	1	3

$L = 24$

(bb)

I	II	III
3	1	2
2	3	1

$L = 22$

(cc)

I	II	III
3	1	2
3	1	2

$L = 22$

(dd)

I	II	III
3	1	2
3	2	1

$L = 21$

(ee)

I	II	III
3	2	1
1	2	3

$L = 24$

(ff)

I	II	III
3	2	1
1	3	2

$L = 23$

(gg)

I	II	III
3	2	1
2	1	3

$L = 23$

(hh)

I	II	III
3	2	1
2	3	1

$L = 21$

(ii)

I	II	III
3	2	1
3	1	2

$L = 21$

(jj)

I	II	III
3	2	1
3	2	1

$L = 20$

Thus, we find

$$P_0\{L=28\}=\frac{1}{36}, P_0\{L=27\}=\frac{4}{36}, P_0\{L=26\}=\frac{4}{36},$$

$$P_0\{L=25\}=\frac{4}{36}, P_0\{L=24\}=\frac{10}{36}, P_0\{L=23\}=\frac{4}{36},$$

$$P_0\{L=22\}=\frac{4}{36}, P_0\{L=21\}=\frac{4}{36}, P_0\{L=20\}=\frac{1}{36}.$$

The probability, under H_0, that L is greater than or equal to 27, for example, is therefore

$$P_0\{L\geq 27\}=P_0\{L=27\}+P_0\{L=28\}=\frac{4}{36}+\frac{1}{36}=\frac{5}{36}=.139.$$

Similarly, $P_0\{L\geq 28\}=P_0\{L=28\}=\frac{1}{36}=.028$.

Since the null distribution for L has been derived without specifying the common form (F) of the underlying distribution function for the X's under H_0 beyond the point of requiring that it be continuous, the test procedure (7.11) based on L is a distribution-free procedure. From the null distribution of L we can determine the critical value l_α and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying X distribution.

For a given number of treatments k and blocks n, the R command `cPage(`α`,k,n)` can be used to find the available upper-tail critical values l_α for possible values of L. For a given available significance level α, the critical value l_α then corresponds to $P_0(L\geq l_\alpha)=\alpha$ and is given by `cPage(`α, k, n`)` $= l_\alpha$. Thus, for example, for $k=4$ and $n=5$, we have $P_0(L\geq 140)=.0106$ so that $l_{.0106}=$ `cPage(.0106, 4, 5)` $=140$ for $k=4$ and $n=5$.

18. *Calculation of the Mean and variance of L under the Null Hypothesis H_0.* In displays (7.12) and (7.13), we presented formulas for the mean and variance of L, respectively, when the null hypothesis is true. In this comment we illustrate a direct calculation of $E_0(L)$ and $\text{var}_0(L)$ in the particular case of $k=3, n=2$, and no tied observations, using the null distribution of L obtained in Comment 17. (Later, in Comment 20, we present arguments for the general derivations of $E_0(L)$ and $\text{var}_0(L)$.) The expected value of the null distribution of L is obtained directly from multiplication of each possible value of L by its probability under H_0. Thus, using the null probability values from Comment 17, we obtain

$$E_0(L)=\tfrac{1}{36}(20+28)+\tfrac{4}{36}(21+22+23+25+26+27)+\tfrac{10}{36}(24)=24.$$

This is in agreement with what we obtain using (7.12), namely,

$$E_0(L)=\frac{2(3)(3+1)^2}{4}=24.$$

A direct check on the expression for $\text{var}_0(L)$ is also easy. Again using the null probabilities from Comment 17, we have

$$\begin{aligned}\text{var}_0(L) &= E_0[\{L - E_0(L)\}^2] = E_0[\{L - 24\}^2] \\ &= \Big\{ \tfrac{1}{36}[(20-24)^2 + (28-24)^2] + \tfrac{4}{36}\big[(21-24)^2 \\ &\quad + (22-24)^2 + (23-24)^2 + (25-24)^2 \\ &\quad + (26-24)^2 + (27-24)^2\big] + \tfrac{10}{36}[(24-24)^2]\Big\} \\ &= \tfrac{1}{36}(16+16) + \tfrac{4}{36}(9+4+1+1+4+9) + \tfrac{10}{36}(0) = 4,\end{aligned}$$

which agrees with what we obtain using (7.13) directly, namely,

$$\text{var}_0(L) = \frac{2(3)^2(3+1)(3^2-1)}{144} = 4.$$

19. *Exact Conditional Distribution of L under H_0 with Ties within the Blocks*. To have a test with exact significance level even in the presence of tied X's within some of the blocks, we need to consider all $(k!)^n$ possible rank configurations, where now the within-blocks ranks are obtained by using average ranks to break the ties. As in Comment 17, it still follows that under H_0 each of these $(k!)^n$ configurations is equally likely. For each such configuration, the value of L is computed and the results are tabulated. As an example, consider the case of $k = 3$, $n = 2$, and the data $X_{11} = 2.4$, $X_{12} = 3.6$, $X_{13} = 2.4$, $X_{21} = 4.0$, $X_{22} = 5.9$, and $X_{23} = 1.7$. Using average ranks to break the tie in the first block, the observed block rank vectors are $(r_{11}, r_{12}, r_{13}) = (1.5, 3, 1.5)$ and $(r_{21}, r_{22}, r_{23}) = (2, 3, 1)$. Thus, $R_1 = 3.5, R_2 = 6, R_3 = 2.5$, and the attained value of L is $3.5 + 2(6) + 3(2.5) = 23$. To assess the significance of this value of L, we would need to obtain the entire conditional null distribution of L by computing its value for each of the $(3!)^2 = 36$ equally likely (under H_0) possible configurations of the observed block rank vectors (1.5, 3, 1.5) and (2, 3, 1). This would be accomplished in exactly the same manner as is illustrated for the no-ties case in Comment 17.

20. *Large-Sample Approximation*. We can rewrite the expression for L (7.10) to obtain

$$L = \sum_{j=1}^{k} jR_j = \sum_{j=1}^{k} j\left(\sum_{i=1}^{n} r_{ij}\right) = \sum_{i=1}^{n}\left(\sum_{j=1}^{k} jr_{ij}\right) = \sum_{i=1}^{n} Q_i,$$

with $Q_i = \sum_{j=1}^{k} jr_{ij}, i = 1, \ldots, n$. Moreover, from Assumptions A1 and A3, $Q_1, \ldots, Q_n$ are mutually independent and identically distributed random variables, regardless of whether or not the null hypothesis H_0 is true. The asymptotic normality, as n tends to infinity, of the standardized form

$$L^* = \frac{L - E(L)}{[\text{var}(L)]^{1/2}} = \frac{L - nE[Q_1]}{[n\ \text{var}(Q_1)]^{1/2}} \tag{7.16}$$

then follows at once from standard central limit theory for sums of mutually independent, identically distributed random variables (cf. Randles and Wolfe (1979, p. 421)).

The computation of $E(Q_1)$ and $\text{var}(Q_1)$ is simplified by noting that

$$E(Q_1) = E\left(\sum_{j=1}^{k} j r_{1j}\right) = \sum_{j=1}^{k} j E(r_{1j}) \tag{7.17}$$

and

$$\begin{aligned}\text{var}(Q_1) &= \text{var}\left(\sum_{j=1}^{k} j r_{1j}\right) \\ &= \sum_{j=1}^{k} \text{var}(j r_{1j}) + 2\sum_{u=1}^{v-1}\sum_{v=2}^{k} \text{cov}(u r_{1u}, v r_{1v}) \\ &= \sum_{j=1}^{k} j^2 \text{var}(r_{1j}) + 2\sum_{u=1}^{v-1}\sum_{v=2}^{k} uv\ \text{cov}(r_{1u}, r_{1v}).\end{aligned} \tag{7.18}$$

In particular, when H_0 is true, $(r_{11}, \ldots, r_{1k})$ is an exchangeable random vector. Thus, under H_0, we have

$$\text{var}_0(r_{1j}) = \text{var}_0(r_{11}), \quad \text{for } j = 2, \ldots, k$$

and

$$\text{cov}_0(r_{1u}, r_{1v}) = \text{cov}_0(r_{11}, r_{12}), \quad \text{for } 1 \le u < v \le k.$$

Using these facts in (7.15) and (7.16), we obtain

$$E_0(Q_1) = E_0(r_{11}) \sum_{j=1}^{k} j = \frac{k(k+1)}{2} E_0(r_{11}) \tag{7.19}$$

and

$$\begin{aligned}\text{var}_0(Q_1) &= \text{var}_0(r_{11}) \sum_{j=1}^{k} j^2 + 2\,\text{cov}_0(r_{11}, r_{12}) \sum_{u=1}^{v-1}\sum_{v=2}^{k} uv \\ &= \frac{k(k+1)(2k+1)}{6} \text{var}_0(r_{11}) \\ &\quad + \text{cov}_0(r_{11}, r_{12}) \left[\sum_{u=1}^{k}\sum_{v=1}^{k} uv - \sum_{t=1}^{k} t^2\right] \\ &= \frac{k(k+1)(2k+1)}{6} [\text{var}_0(r_{11}) - \text{cov}_0(r_{11}, r_{12})] \\ &\quad + \left[\frac{k(k+1)}{2}\right]^2 \text{cov}_0(r_{11}, r_{12}).\end{aligned} \tag{7.20}$$

It can be shown (see Problem 22) that

$$E_0(r_{11}) = \frac{k+1}{2}, \text{var}_0(r_{11}) = \frac{k^2-1}{12} \tag{7.21}$$

and

$$\text{cov}_0(r_{11}, r_{12}) = -\frac{(k+1)}{12}. \tag{7.22}$$

Using these results in expressions (7.19) and (7.20), we obtain

$$E_0(Q_1) = \frac{k(k+1)}{2}\frac{(k+1)}{2} = \frac{k(k+1)^2}{4} \tag{7.23}$$

and

$$\text{var}_0(Q_1) = \frac{k(k+1)(2k+1)}{6}\left[\frac{k^2-1}{12} + \frac{k+1}{12}\right] + \left[\frac{k(k+1)}{2}\right]^2\left(-\frac{k+1}{12}\right),$$

which, after some straightforward algebra, yields

$$\text{var}_0(Q_1) = \frac{k^2(k+1)(k^2-1)}{144}. \tag{7.24}$$

Combining equations (7.17), (7.18), (7.23), and (7.24), we obtain

$$E_0(L) = nE_0(Q_1) = \frac{nk(k+1)^2}{4}$$

and

$$\text{var}_0(L) = \frac{nk^2(k+1)(k^2-1)}{144},$$

as stated in expressions (7.12) and (7.13), respectively. In conjunction with (7.16), this provides the justification for the approximate α level procedure in (7.15).

21. *Conservative Nature of the Large-Sample Approximation when There Are Ties within Blocks*. In applications where tied X values are observed in one or more of the blocks and average ranks are used to deal with these ties, the null variance of L based on its exact conditional null distribution (see Comment 19) is always smaller than the value obtained from expression (7.13). (This fact is illustrated in Problems 20 and 21.) As a result, the approximate level α procedure in (7.15) is conservative in the presence of within-blocks ties in the following sense: If we reject H_0 using procedure (7.15) with $\text{var}_0(L)$ obtained from expression (7.13), then we would also reject H_0 if we were to more properly use the exact conditional null variance of L in computing the value of L^* (7.14).

22. *Relation to Rank Order Correlation*. The L test is directly related to Spearman's rank order correlation coefficient r_s (8.63). Let r_i denote Spearman's correlation coefficient computed between the observed rank order and the postulated order in block i, and set $\bar{r} = \left(\sum_{i=1}^{n} r_i/n\right)$. Then, it can be shown that

$$\bar{r} = \left\{\frac{12L}{nk(k^2-1)} - \frac{3(k+1)}{(k-1)}\right\}.$$

23. *Consistency of the L Test*. Replace Assumptions A1–A3 by the less restrictive Assumption A1′ : $X_{ij} = \beta_i + e_{ij}$, where the e's are mutually independent, and Assumption A2′ : $e_{1j}, \ldots, e_{nj}$ come from the same continuous population $\prod_j, j = 1, \ldots, k$, but where $\prod_1, \ldots, \prod_k$ are not assumed to be identical. Then the test defined by (7.11) is consistent against alternatives for which $\left\{ \sum_{u<v} (v - u) p_{uv} > k(k-1)(k+1)/12 \right\}$, where $p_{uv} = P(e_{iu} < e_{iv})$ with e_{iu} a random member from $\prod_u$ and e_{iv} a random member from $\prod_v$ that is independent of e_{iu} (see Hollander (1967a)). For those situations covered by Assumptions A1–A3, this consistency statement implies the consistency statement given in Property 1.

Properties

1. *Consistency*. The test defined by (7.11) is consistent against the H_2 (7.9) alternatives. See Hollander (1967a) and Comment 23.
2. *Asymptotic Normality*. See Comment 20 and Randles and Wolfe (1979, p. 421).
3. *Efficiency*. See Hollander (1967a) and Section 7.16.

Problems

15. Brady (1969) described an experiment concerning the influence of the rhythmicity of a metronome on the speech of stutterers. The subjects were 12 severe stutterers. Each subject spoke extemporaneously for 3 min under the three conditions N, A, and R.

N: Subject spoke unaided by a metronome.

R: Subject spoke with a regular (rhythmic) metronome set at 120 ticks per minute and was instructed to pace one syllable of speech to each tick.

A: Subject spoke with an arrhythmic metronome in which the intervals between ticks ranged randomly between 0.3 and 0.7 s but with an average of 120 ticks per minute. Again the subject was instructed to pace one syllable of speech to each tick.

Table 7.6 gives the number of dysfluencies under each condition. On the basis of the conditions, and prior to looking at the data, we might expect a deviation from H_0 to be in the direction $\tau_R < \tau_A < \tau_N$. Perform Page's test using this postulated ordering.

16. Verify the relationship (see Comment 22) between L (7.10) and r_s (8.63).

Table 7.6 Influence of Rhythmicity of Metronome on Speech Fluency

	Dysfluencies under Each Condition		
Subject	*R*	*A*	*N*
1	3	5	15
2	3	3	11
3	1	3	18
4	5	4	21
5	2	2	6
6	0	2	17
7	0	2	10
8	0	3	8
9	0	2	13
10	1	0	4
11	2	4	11
12	2	1	17

Source: J. P. Brady (1969).

17. Show directly, or illustrate by means of an example, that the maximum value of L is $L_{\max} = nk(k+1)(2k+1)/6$. For what rank configuration is the maximum achieved?

18. Show directly, or illustrate by means of an example, that the minimum value of L is $L_{\min} = nk(k+1)(k+2)/6$. For what rank configuration is this minimum achieved?

19. Shelterbelts (long rows of tree plantings across the direction of the prevailing winds) have been used extensively for sometime in developed countries to protect crops and livestock from the effects of the wind. Ujah and Adeoye (1984) conducted a study to see if such shelterbelts could be used effectively to ameliorate the severe losses from droughts experienced almost annually in the arid and semiarid zones of Nigeria and considered to be a leading factor in the declining food production in Nigeria and many of its neighbors.

Ujah and Adeoye investigated the effect of shelterbelts on a variety of factors related to drought conditions, including wind velocity, air and soil temperatures, and soil moisture. The experiment was conducted at two locations about 3.5 km apart, near Dambatta. Table 7.7 presents the wind velocity data (averaged over these two locations) at various distances leeward of the shelterbelt expressed as percent wind speed reduction relative to the wind velocity on the windward side of the shelterbelt. The data are monthly (except for July, November, and December, for which the data were not available) and at leeward distances of 20, 40, 100, 150, and 200 m from the shelterbelt.

Use these data to test the hypothesis of a negative relationship between percent reduction in average wind speed and the leeward distance from a shelterbelt.

20. Consider the case of $k = 3$, $n = 2$, and the tied data set $X_{11} = 2.4$, $X_{12} = 3.6$, $X_{13} = 2.4$, $X_{21} = 4.0$, $X_{22} = 5.9$, and $X_{23} = 1.7$. What is the conditional probability distribution of L under H_0 (7.2) when average ranks are used to break within-blocks ties among the X's? (see Comment l9). How extreme is the observed value of $L = 23$ in this conditional null distribution?

21. Consider the tied data set in Problem 20 for the setting of $k = 3$ and $n = 2$. Use the conditional null probability distribution of L obtained in Problem 20 to compute the conditional null variance of L and compare this value with that of the unconditional null variance given by (7.13). Interpret these two numbers in view of the discussion in Comment 21.

22. Let $\boldsymbol{r}_1 = (r_{11}, \ldots, r_{1k})$ be a random vector of ranks that is uniformly distributed over the set of all $k!$ permutations of $(1, \ldots, k)$. Show that $E(r_{11}) = (k+1)/2$, $\text{var}(r_{11}) = (k^2-1)/12$, and $\text{cov}(r_{11}, r_{12}) = -(k+1)/12$.

23. Carry out the algebra to verify the final expression for $\text{var}_0(Q)$ in (7.24).

24. Suppose $k = 3$ and $n = 3$. Obtain the form of the exact null (H_0) distribution of L for the case of no-tied observations.

Table 7.7 Percent Reduction in Average Wind Speed at Dambatta, 1980/81

	Leeward Distance from Shelterbelt (m)				
Month	20	40	100	150	200
January	22.1	20.7	15.4	12.3	6.9
February	19.2	18.7	14.9	9.3	6.5
March	21.5	21.9	14.3	9.9	7.1
April	21.5	21.2	11.1	9.4	6.2
May	21.3	20.9	11.2	9.4	7.7
June	20.9	19.6	16.9	11.6	7.0
August	19.3	18.7	14.4	12.5	7.0
September	20.1	19.6	15.6	12.6	7.5
October	23.7	20.4	14.6	12.4	8.5

Source: J. E. Ujah and K. B. Adeoye (1984).

Table 7.8 Maximum Soil Temperature (°C) at 5-cm Depth at Dambatta, 1980/81

	Leeward Distance from Shelterbelt (m)			
Month	20	40	100	200
January	37.7	37.5	37.6	37.4
February	39.7	39.4	39.6	39.6
March	42.0	42.0	41.9	41.9
April	43.4	43.1	42.8	43.0
May	42.5	42.3	42.3	42.1
June	39.7	39.7	39.6	39.7
July	38.7	38.5	38.6	38.5
August	39.1	38.8	38.9	38.4
September	39.7	39.5	39.2	39.4
October	39.9	40.0	40.0	40.2
November	39.6	39.7	39.8	39.7

Source: J. E. Ujah and K. B. Adeoye (1984).

25. In their study of shelterbelts (see Problem 19), Ujah and Adeoye (1984) also obtained measurements of the monthly maximum soil temperature (°C) at a 5-cm depth at leeward distances of 20, 40, 100, and 200 m from the shelterbelt. These data are presented in Table 7.8.

Use these data to test the hypothesis that there is a negative relationship between maximum soil temperature at a 5-cm depth and the leeward distance from a shelterbelt.

26. For the case of $k = 2$, show that procedure (7.11) is equivalent to the exact one-sided sign test, as discussed in Section 3.4.

27. Consider the data on percentage consonants correctly identified in Table 7.4 from the study on hearing-impaired children by Nicholls and Ling (1982). From previous studies, there is reason to believe that cued speech (C) is more effective as a stand-alone method for teaching language to hearing-impaired children than lip reading (L), which, in turn, is thought to be more effective than audition (A) by itself. Find an approximate P-value using procedure (7.15) to test this conjecture.

RATIONALE FOR MULTIPLE COMPARISON PROCEDURES

In Sections 7.1 and 7.2 we have discussed procedures designed to test the null hypothesis H_0 (7.2) against either general or ordered alternatives. Upon rejection of H_0 with one of these test procedures for a given set of data, our conclusion is either that there are some unspecified differences among the treatment effects (associated with the Friedman procedure discussed in Section 7.1) or that the treatment effects follow an ordered pattern (associated with the Page procedure of Section 7.2). However, in neither of these test procedures is our conclusion pair-specific; that is, the tests in Sections 7.1 and 7.2 are not designed to enable us to reach conclusions about specific pairs of treatment effects. The relative sizes of the specific treatment effects τ_1 and τ_2, for example, cannot be inferred from the conclusions reached by either of the test procedures of Sections 7.1 or 7.2. To elicit such pairwise specific information, we turn to the class of multiple comparison procedures. In Section 7.3 we present a two-sided all-treatments multiple comparison procedure for the omnibus setting corresponding to the general alternatives H_1 (7.3). In Section 7.4 we deal with treatments-versus-control multiple comparison decisions for settings where one of the treatments plays a special role as the study control.

7.3 DISTRIBUTION-FREE TWO-SIDED ALL-TREATMENTS MULTIPLE COMPARISONS BASED ON FRIEDMAN RANK SUMS—GENERAL CONFIGURATION (WILCOXON, NEMENYI, MCDONALD-THOMPSON)

In this section we present a multiple comparison procedure based on Friedman's within-blocks ranks that is designed to make decisions about individual differences between pairs of treatment effects (τ_i, τ_j) for $i < j$, in a setting where general alternatives H_1 (7.3) are of interest. Thus, the multiple comparison procedure of this section would generally be applied to two-way layout data (with one observation per cell) *after* rejection of H_0 (7.2) with the Friedman procedure from Section 7.1. In this setting it is important to reach conclusions about all $\binom{k}{2} = k(k-1)/2$ pairs of treatment effects and these conclusions are naturally two-sided.

Procedure

Let $R_1, \ldots, R_k$ be the treatment sums of within-blocks ranks given by (7.4). Calculate the $k(k-1)/2$ absolute differences $|R_u - R_v|, 1 \le u < v \le k$. At an experimentwise error rate of α the Wilcoxon–Nemenyi–McDonald–Thompson two-sided all-treatments multiple comparison procedure reaches its $k(k-1)/2$ pairwise decisions, corresponding to each (τ_u, τ_v) pair, $1 \le u < v \le k$, by the criterion

$$\text{Decide } \tau_u \ne \tau_v \text{ if } |R_u - R_v| \ge r_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v, \tag{7.25}$$

where the constant r_α is chosen to make the experimentwise error rate equal to α; that is, r_α satisfies the restriction

$$P_0(|R_u - R_v| < r_\alpha, u = 1, \ldots, k-1; v = u+1, \ldots, k) = 1 - \alpha, \tag{7.26}$$

where the probability $P_0(.)$ is computed under H_0 (7.2). Equation (7.26) stipulates that the $k(k-1)/2$ inequalities $|R_u - R_v| < r_\alpha$ corresponding to all pairs (u, v) of treatments with $u < v$, hold simultaneously with probability $1 - \alpha$ when H_0 (7.2) is true. Comment 26 explains how to obtain the critical values r_α for k treatments, n blocks, and available experimentwise error rates α.

Large-Sample Approximation

When H_0 is true, the k-component vector $(R_1, \ldots, R_k)$ has, as n tends to infinity, an asymptotic $(k-1)$-variate normal distribution with appropriate mean vector and covariance matrix (see Comment 29 for indications of the proof). It then follows that the critical value r_α can, when the number of blocks n is large, be approximated by $[nk(k+1)/12]^{1/2} q_\alpha$, where q_α is the upper αth percentile point for the distribution of the range of k independent $N(0, 1)$ variables. Thus, the large-sample approximation for procedure (7.25) is

$$\text{Decide } \tau_u \ne \tau_v \text{ if } |R_u - R_v| \ge q_\alpha \left[\frac{nk(k+1)}{12}\right]^{1/2}; \quad \text{otherwise decide } \tau_u = \tau_v. \tag{7.27}$$

To find q_α for k treatments and a specified experimentwise error rate α, we use the R command `cRangeNor(`α, k`)`. For example, to find $q_{.05}$ for $k = 5$ treatments, we apply `cRangeNor(.05,5)` and obtain $q_{.05} = 3.858$ for $k = 5$.

Ties

If there are ties among the X observations within any of the blocks, use average ranks to break the ties and compute the individual treatment sums of ranks $R_1, \ldots, R_k$. In such cases, the experimentwise error rate associated with procedure (7.25) is only approximately equal to α.

EXAMPLE 7.3 ***Rounding First Base.***

Consider the rounding-first-base data discussed in Example 7.1. There we had found (using the large-sample approximation for the Friedman procedure) that there is strong evidence to conclude that the three methods of running to first base are not equivalent with respect to time to reach second base. To determine which of the three running methods differ in median times to second base, we apply the approximate procedure (7.27), using average ranks to break the within-runners ties in computing R_1, R_2, and R_3. Here, we have $k = 3$ and $n = 22$. For the sake of illustration, we take our approximate experimentwise error rate to be $\alpha = .01$. Using the R command `cRangeNor(`α, k`)` with $\alpha = .01$ and $k = 3$, we find `cRangeNor(.01, 3)` $= q_{.01} = 4.12$, and procedure (7.27) reduces to

$$\text{Decide } \tau_u \neq \tau_v \text{ if } |R_u - R_v| \geq (4.12)\left[\frac{22(3)(4)}{12}\right]^{1/2} = 19.3.$$

Using the treatments sums of within-runners ranks given in Table 7.1, we find that

$$|R_2 - R_1| = 6, \quad |R_3 - R_1| = 21, \quad \text{and} \quad |R_3 - R_2| = 15.$$

Referring these absolute value rank sum differences to the approximate critical value 19.3, we see that

$$|R_2 - R_1| = 6 < 19.3 \quad \Rightarrow \quad \text{decide } \tau_2 = \tau_1,$$

$$|R_3 - R_1| = 21 \geq 19.3 \quad \Rightarrow \quad \text{decide } \tau_3 \neq \tau_1,$$

and

$$|R_3 - R_2| = 15 < 19.3 \quad \Rightarrow \quad \text{decide } \tau_3 = \tau_2.$$

Thus, at an approximate experimentwise error rate of .01, we have reached the conclusion that only the round out (treatment 1) and wide angle (treatment 3) running methods yield significantly different median times to second base. (We note that the

smallest approximate experimentwise error rate at which we would reach this conclusion is obtained by first setting

$$\max_{(u,v)} |R_u - R_v| = |R_3 - R_1| = 21 = q_\alpha \left[\frac{22(3)(4)}{12}\right]^{\frac{1}{2}}$$

and solving for

$$q_\alpha = 21 \left[\frac{22(3)(4)}{12}\right]^{-\frac{1}{2}} = 4.477.$$

Using the R command pRangeNor(q_α, k), we then find α = `pRangeNor(4.48, 3)` = .0044 to be the smallest experimentwise error rate at which we would decide that the round out (treatment 1) and wide angle (treatment 3) running methods yield significantly different median times to second base.

For the sake of illustration for the exact procedure in (7.25), we consider the subset of the sample data associated with the first 15 baseball players in Table 7.1. For that subset we have $k = 3$, $n = 15$, and the three treatment sums of ranks $R_1^* = 37$, $R_2^* = 31$, and $R_3^* = 22$. With $k = 3$, $n = 15$, and experimentwise error rate $\alpha = .047$, we apply the R command cWNMT(α,k,n) and find `cWNMT(.047, 3, 15)` = 13. Thus, we have $r_{.047} = 13$ and procedure (7.25) becomes

$$\text{Decide } \tau_u \neq \tau_v \text{ if } |R_u^* - R_v^*| \geq 13.$$

Since $|R_2^* - R_1^*| = 6, |R_3^* - R_1^*| = 15$, and $|R_3^* - R_2^*| = 9$, we see that our decisions for this subset of data using procedure (7.25) would be $\tau_2 = \tau_1, \tau_3 \neq \tau_1$, and $\tau_3 = \tau_2$, in agreement with what we found using the entire set of 22 baseball players and the approximate procedure (7.27). (Note, however, that for this smaller set of data, we could no longer conclude that $\tau_3 \neq \tau_1$ at an experimentwise error rate as low as .01.)

Comments

24. *Rationale for Multiple Comparison Procedures*. We think of the methods of this section as multiple comparison procedures. The aim of applying such procedures goes beyond the point of deciding whether the treatments are equivalent to the (often more important) problem of selecting which, if any, treatments differ from one another. Thus, the user makes $k(k-1)/2$ decisions, one for each pair of treatments. Equation (7.26) states that the probability of making all correct decisions when H_0 is true is controlled to be $1 - \alpha$; that is, when using procedure (7.25), the probability of at least one incorrect decision, when H_0 is true, is controlled to be α. This error rate is derived under the assumption that H_0 is true, but it does not depend on the particular underlying distributional form F. This is why we call (7.25) a distribution-free multiple comparison procedure.

    The multiple comparison procedures of this section can also be interpreted as hypothesis tests. If we consider the procedure that rejects H_0 if and only if the inequality of (7.25) [or of (7.27)] holds for at least one (u, v) pair, $1 \leq u < v \leq k$, this is a distribution-free test of size α for H_0 (7.2).

25. *Experimentwise Error Rate*. The use of an experimentwise error rate represents a very conservative approach to multiple comparisons. We are insisting that the probability of making only correct decisions be $1 - \alpha$ when the null hypothesis H_0 (7.2) of treatment equivalence is true. Thus, although we have a high degree of protection when H_0 is true, we often apply such techniques where we have evidence (perhaps based on a priori information or perhaps obtained by applying the Friedman test, as in Example 7.3) that H_0 is not true. This protection under H_0 also makes it harder for the procedure to judge treatments as differing significantly when in fact H_0 is false, and this difficulty becomes more severe as k increases. See Comment 6.54 for additional discussion of experimentwise error rates.

26. *Critical Values* r_α. The r_α critical values can be obtained by using the fact that under H_0 (7.2), all $(k!)^n$ rank configurations are equally likely. Thus, to obtain the probability under H_0 that $|R_u - R_v| < c$ simultaneously for $u = 1, \ldots, k-1$ and $v = u+1, \ldots, k$, we can count the number of configurations for which the event $B = \{|R_u - R_v| < c, u = 1, \ldots, k-1; v = u+1, \ldots, k\}$ occurs and divide this number by $(k!)^n$. For an illustration, consider the 24 configurations of Comment 8, corresponding to the case $k = 4$, $n = 2$. (As in Comment 8, the same reasoning enables us to consider only 24 rather than $(4!)^2 = 576$ configurations.) For each configuration, we now display the values of $|R_1 - R_2|, |R_1 - R_3|, |R_1 - R_4|, |R_2 - R_3|, |R_2 - R_4|$, and $|R_3 - R_4|$.

(a)	$\|R_1 - R_2\| = 2$	(b)	$\|R_1 - R_2\| = 2$	(c)	$\|R_1 - R_2\| = 3$	(d)	$\|R_1 - R_2\| = 3$
	$\|R_1 - R_3\| = 4$		$\|R_1 - R_3\| = 5$		$\|R_1 - R_3\| = 5$		$\|R_1 - R_3\| = 3$
	$\|R_1 - R_4\| = 6$		$\|R_1 - R_4\| = 5$		$\|R_1 - R_4\| = 4$		$\|R_1 - R_4\| = 6$
	$\|R_2 - R_3\| = 2$		$\|R_2 - R_3\| = 3$		$\|R_2 - R_3\| = 2$		$\|R_2 - R_3\| = 0$
	$\|R_2 - R_4\| = 4$		$\|R_2 - R_4\| = 3$		$\|R_2 - R_4\| = 1$		$\|R_2 - R_4\| = 3$
	$\|R_3 - R_4\| = 2$		$\|R_3 - R_4\| = 0$		$\|R_3 - R_4\| = 1$		$\|R_3 - R_4\| = 3$
(e)	$\|R_1 - R_2\| = 4$	(f)	$\|R_1 - R_2\| = 4$	(g)	$\|R_1 - R_2\| = 0$	(h)	$\|R_1 - R_2\| = 0$
	$\|R_1 - R_3\| = 3$		$\|R_1 - R_3\| = 4$		$\|R_1 - R_3\| = 3$		$\|R_1 - R_3\| = 4$
	$\|R_1 - R_4\| = 5$		$\|R_1 - R_4\| = 4$		$\|R_1 - R_4\| = 5$		$\|R_1 - R_4\| = 4$
	$\|R_2 - R_3\| = 1$		$\|R_2 - R_3\| = 0$		$\|R_2 - R_3\| = 3$		$\|R_2 - R_3\| = 4$
	$\|R_2 - R_4\| = 1$		$\|R_2 - R_4\| = 0$		$\|R_2 - R_4\| = 5$		$\|R_2 - R_4\| = 4$
	$\|R_3 - R_4\| = 2$		$\|R_3 - R_4\| = 0$		$\|R_3 - R_4\| = 2$		$\|R_3 - R_4\| = 0$
(i)	$\|R_1 - R_2\| = 2$	(j)	$\|R_1 - R_2\| = 2$	(k)	$\|R_1 - R_2\| = 3$	(l)	$\|R_1 - R_2\| = 3$
	$\|R_1 - R_3\| = 4$		$\|R_1 - R_3\| = 1$		$\|R_1 - R_3\| = 1$		$\|R_1 - R_3\| = 3$
	$\|R_1 - R_4\| = 2$		$\|R_1 - R_4\| = 5$		$\|R_1 - R_4\| = 4$		$\|R_1 - R_4\| = 2$
	$\|R_2 - R_3\| = 2$		$\|R_2 - R_3\| = 1$		$\|R_2 - R_3\| = 2$		$\|R_2 - R_3\| = 0$
	$\|R_2 - R_4\| = 0$		$\|R_2 - R_4\| = 3$		$\|R_2 - R_4\| = 1$		$\|R_2 - R_4\| = 1$
	$\|R_3 - R_4\| = 2$		$\|R_3 - R_4\| = 4$		$\|R_3 - R_4\| = 3$		$\|R_3 - R_4\| = 1$
(m)	$\|R_1 - R_2\| = 1$	(n)	$\|R_1 - R_2\| = 1$	(o)	$\|R_1 - R_2\| = 0$	(p)	$\|R_1 - R_2\| = 0$
	$\|R_1 - R_3\| = 1$		$\|R_1 - R_3\| = 3$		$\|R_1 - R_3\| = 3$		$\|R_1 - R_3\| = 0$
	$\|R_1 - R_4\| = 4$		$\|R_1 - R_4\| = 2$		$\|R_1 - R_4\| = 1$		$\|R_1 - R_4\| = 4$
	$\|R_2 - R_3\| = 2$		$\|R_2 - R_3\| = 4$		$\|R_2 - R_3\| = 3$		$\|R_2 - R_3\| = 0$
	$\|R_2 - R_4\| = 5$		$\|R_2 - R_4\| = 3$		$\|R_2 - R_4\| = 1$		$\|R_2 - R_4\| = 4$
	$\|R_3 - R_4\| = 3$		$\|R_3 - R_4\| = 1$		$\|R_3 - R_4\| = 2$		$\|R_3 - R_4\| = 4$

(q)	$\lvert R_1 - R_2\rvert = 2$	(r)	$\lvert R_1 - R_2\rvert = 2$	(s)	$\lvert R_1 - R_2\rvert = 2$	(t)	$\lvert R_1 - R_2\rvert = 2$
	$\lvert R_1 - R_3\rvert = 0$		$\lvert R_1 - R_3\rvert = 1$		$\lvert R_1 - R_3\rvert = 0$		$\lvert R_1 - R_3\rvert = 1$
	$\lvert R_1 - R_4\rvert = 2$		$\lvert R_1 - R_4\rvert = 1$		$\lvert R_1 - R_4\rvert = 2$		$\lvert R_1 - R_4\rvert = 1$
	$\lvert R_2 - R_3\rvert = 2$		$\lvert R_2 - R_3\rvert = 1$		$\lvert R_2 - R_3\rvert = 2$		$\lvert R_2 - R_3\rvert = 3$
	$\lvert R_2 - R_4\rvert = 0$		$\lvert R_2 - R_4\rvert = 1$		$\lvert R_2 - R_4\rvert = 4$		$\lvert R_2 - R_4\rvert = 3$
	$\lvert R_3 - R_4\rvert = 2$		$\lvert R_3 - R_4\rvert = 0$		$\lvert R_3 - R_4\rvert = 2$		$\lvert R_3 - R_4\rvert = 0$
(u)	$\lvert R_1 - R_2\rvert = 1$	(v)	$\lvert R_1 - R_2\rvert = 1$	(w)	$\lvert R_1 - R_2\rvert = 0$	(x)	$\lvert R_1 - R_2\rvert = 0$
	$\lvert R_1 - R_3\rvert = 1$		$\lvert R_1 - R_3\rvert = 1$		$\lvert R_1 - R_3\rvert = 1$		$\lvert R_1 - R_3\rvert = 0$
	$\lvert R_1 - R_4\rvert = 2$		$\lvert R_1 - R_4\rvert = 0$		$\lvert R_1 - R_4\rvert = 1$		$\lvert R_1 - R_4\rvert = 0$
	$\lvert R_2 - R_3\rvert = 0$		$\lvert R_2 - R_3\rvert = 2$		$\lvert R_2 - R_3\rvert = 1$		$\lvert R_2 - R_3\rvert = 0$
	$\lvert R_2 - R_4\rvert = 3$		$\lvert R_2 - R_4\rvert = 1$		$\lvert R_2 - R_4\rvert = 1$		$\lvert R_2 - R_4\rvert = 0$
	$\lvert R_3 - R_4\rvert = 3$		$\lvert R_3 - R_4\rvert = 1$		$\lvert R_3 - R_4\rvert = 2$		$\lvert R_3 - R_4\rvert = 0$

Thus, for example,

$$\begin{aligned} P_0\{|R_u - R_v| < 6, u = 1,2,3; v = u+1,\ldots,4\} &= P_0\{|R_1 - R_2| < 6; |R_1 - R_3| < 6; |R_1 - R_4| < 6; \\ &\quad |R_2 - R_3| < 6; |R_2 - R_4| < 6; |R_3 - R_4| < 6\} \\ &= \tfrac{22}{24} = 1 - .083, \end{aligned}$$

because for 22 of the configurations—all but configurations (a) and (d)—the event $\{|R_1 - R_2| < 6; |R_1 - R_3| < 6; |R_1 - R_4| < 6; |R_2 - R_3| < 6; |R_2 - R_4| < 6; |R_3 - R_4| < 6\}$ occurs. This .083 probability agrees with the result obtained from using the R command `cWNMT(`α`,k,n)`; that is, `cWNMT(.083, 4, 2)` $= r_{.083} = 6$ for $k = 4$ treatments and $n = 3$ blocks.

27. *Relationship to Range of Rank Sums*. Define the range of $R_1, \ldots, R_k$ as

$$\text{Range } [R_1, \ldots, R_k] = \max[R_1, \ldots, R_k] - \min[R_1, \ldots, R_k].$$

Then, $|R_u - R_v|$ is less than c, for all $u < v$, if and only if range $[R_1, \ldots, R_k]$ is less than c. Thus, in Comment 26, instead of computing the values of $|R_1 - R_2|, |R_1 - R_3|, |R_1 - R_4|, |R_2 - R_3|, |R_2 - R_4|, |R_3 - R_4|$ for each rank configuration, we need to have calculated only range $[R_1, \ldots, R_k]$ for each configuration. That is, we can obtain the critical constants r_α by computing only the range for each possible rank configuration. We do, however, need to compute the individual absolute differences $|R_u - R_v|$ in order to apply procedure (7.25) to a set of data.

28. *Historical Development*. The basic idea behind the multiple comparison procedures (7.25) and (7.27) based on the Friedman rank sums is attributed by McDonald and Thompson (1967) to Wilcoxon who "... in 1956, in an unpublished notebook, carried out the first correct probability computation for 3 objects (treatments) and 3 judges (blocks)..." Nemenyi (1963) obtained a small number of exact critical values r_α for procedure (7.25) in his Ph.D. dissertation. McDonald and Thompson (1967) provided additional exact critical values.

29. *Large-Sample Approximation*. Let $\boldsymbol{R} = (R_1, \ldots, R_k)$ be the vector of the Friedman rank sums. Then, it can be shown that a properly standardized $\boldsymbol{R}$ has, as n tends to infinity, an asymptotic multivariate normal distribution with appropriate mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (see Miller (1966) for details). It follows directly from this result (again, see Miller (1966)) that the procedure in (7.27) has an asymptotic experimentwise error rate equal to α.
30. *Dependence on Observations from Other Noninvolved Treatments*. The absolute difference $|R_u - R_v|$ depends on the values of the observations from the other $k - 2$ treatments, in addition to the observations from treatments u and v. Thus, the multiple comparison procedures (7.25) and (7.27) both have the disadvantage that the decision concerning treatment u and treatment v can be affected by changes only in the observations from one or more of the other $k - 2$ treatments that are not directly involved. This difficulty has been emphasized by Miller (1966) and Gabriel (1969).
31. *Approximately Distribution-Free Multiple Comparison Procedures Based on Signed Ranks*. The multiple comparison procedures in (7.25) and (7.27) both utilize the within-blocks Friedman ranking schemes, and, as a result, they are related to the sign procedures for paired replicates data (see Comments 7 and 16). Competitor all-treatments multiple comparison procedures based on signed ranks that utilize between-block information are discussed in Section 7.13. These signed rank procedures are, however, only asymptotically distribution-free.

Properties

1. *Asymptotic Multivariate Normality*. See Miller (1966).
2. *Efficiency*. See Section 7.16.

Problems

28. Livesey (1967) has compared the performance of rats, Rabbits, and cats on the Hebb–Williams (1946) elevated pathway test (EPT). Table 7.9, based on a subset of the Livesey data, gives mean error scores by species for 12 problems. Using procedure (7.25), find the species (if any) that differ significantly.

29. For the case of $k = 3$, $n = 9$, and $\alpha = .05$, compare procedures (7.25) and (7.27).

30. Consider the rounding-first-base data discussed in Examples 7.1 and 7.3. Using the data for all 22 players, find the smallest approximate experimentwise error rate at which we would declare that the median time to second base for the narrow angle method of running is different from that for the wide angle method of running.

31. Apply procedure (7.25) to the ozone exposure data of Table 7.2.

32. Apply the approximate procedure (7.27) to the percentage of correctly identified consonants data of Table 7.4.

33. Find the totality of all available experimentwise error rates α and the associated critical values r_α for procedure (7.25) when $k = 3$ and $n = 3$.

34. Illustrate the difficulty discussed in Comment 30 by means of a numerical example.

35. Consider the serum CPK data of Table 7.3 in Problem 5. Find the smallest (available) experimentwise error rate at which the most significant difference in treatment effects between the time measurements would be detected.

Table 7.9 Error Scores by Species

Problem Number	Rats	Rabbits	Cats
1	1.5	1.7	0.3
2	1.1	1.5	1.0
3	1.8	8.1	3.6
4	1.9	1.3	0.0
5	4.3	4.0	0.6
6	2.0	4.6	5.5
7	8.4	4.0	1.0
8	6.6	5.1	3.1
9	2.4	2.5	0.1
10	6.5	6.9	1.6
11	2.6	2.5	4.3
12	6.5	6.8	1.0

Source: P. J. Livesey (1967).

36. Consider the serum CPK data of Table 7.3 in Problem 5. Find the smallest (available) experimentwise error rate at which we would declare that the typical serum CPK activity at 19 h postexercise is different from that at 42 h postexercise.

7.4 DISTRIBUTION-FREE ONE-SIDED TREATMENTS VERSUS CONTROL MULTIPLE COMPARISONS BASED ON FRIEDMAN RANK SUMS (NEMENYI, WILCOXON–WILCOX, MILLER)

In this section we turn our attention to a multiple comparison procedure designed to make decisions about individual differences between the median effect for a single, baseline control population and the median effects of each of the remaining $k-1$ treatments. This treatments-versus-control multiple comparison procedure can be applied *after* rejection of H_0 (7.2) with either the Friedman or the Page test discussed in Sections 7.1 and 7.2, respectively. Its application leads to conclusions about the differences between each of the $k-1$ treatment effects and the control effect, and these conclusions are naturally one-sided.

Procedure

For simplicity of notation, we let treatment 1 assume the role of the single baseline control. Let $R_1, \ldots, R_k$ be the treatment sums of the within-blocks ranks given by (7.4). Calculate the $k-1$ differences $(R_u - R_1), u = 2, \ldots, k$. At an experimentwise error rate of α, the Nemenyi–Wilcoxon–Wilcox–Miller one-sided treatments-versus-control multiple comparison procedure reaches its $k-1$ pairwise decisions, corresponding to each (τ_1, τ_u) pair, for $u = 2, \ldots, k$, by the criterion

$$\text{Decide } \tau_u > \tau_1 \text{ if } (R_u - R_1) \geq r_\alpha^*; \text{ otherwise decide } \tau_u = \tau_1, \tag{7.28}$$

where the constant r_α^* is chosen to make the experimentwise error rate equal to α; that is, r_α^* satisfies the restriction

$$P_0((R_u - R_1) < r_\alpha^*, u = 2, \ldots, k) = 1 - \alpha, \tag{7.29}$$

where the probability $P_0(.)$ is computed under H_0 (7.2). Equation (7.29) stipulates that the $k-1$ inequalities $(R_u - R_1) < r^*_\alpha$, corresponding to each treatment paired with the control, hold simultaneously with probability $1-\alpha$ when H_0(7.2) is true. Comment 35 explains how to obtain the critical value r^*_α for k treatments, n blocks, and available experimentwise error rates α. (For discussion of how to adjust procedure (7.28) for settings where it is of interest to decide whether the treatment effects are *smaller* than the control effect, see Comment 34.)

Large-Sample Approximation

When H_0 is true, the $(k-1)$-component vector $(R_2 - R_1, \ldots, R_k - R_1)$ has, as n tends to infinity, an asymptotic $(k-1)$-variate normal distribution with mean vector $\mathbf{0}$ (see Comment 37 for an indication of the proof). It then follows that the critical value r^*_α can, when the number of blocks is large, be approximated by $[nk(k+1)/6]^{1/2} m^*_{\alpha,1/2}$, where $m^*_{\alpha,1/2}$ is the upper αth percentile point for the distribution of the maximum of $(k-1)$ $N(0,1)$ random variables with common correlation $\rho = \frac{1}{2}$. Thus, the large-sample approximation for procedure (7.28) is

$$\begin{aligned} &\text{Decide } \tau_u > \tau_1 \text{ if } (R_u - R_1) \geq [nk(k+1)/6]^{1/2} m^*_{\alpha,1/2}; \\ &\qquad \text{otherwise decide } \tau_u = \tau_1. \end{aligned} \tag{7.30}$$

To find $m^*_{\alpha,1/2}$ for k treatments and a specified experimentwise error rate α, we use the R command `cMaxCorrNor(`α`,k,0.5)`. For example, to find $m^*_{.04584,1/2}$ for $k=4$ treatments, we apply `cMaxCorrNor(.04584,4,0.5)` and obtain $m^*_{.04584,1/2} = 2.19$.

Ties

If there are ties among the X observations within any of the blocks, use average ranks to break the ties and compute the individual treatment sums of ranks $R_1, \ldots, R_k$. In such cases, the experimentwise error rate associated with procedure (7.28) is only approximately equal to α.

EXAMPLE 7.4 ***Stuttering Adaptation.***

Daly and Cooper (1967) considered the rate of stuttering adaptation under three conditions. Eighteen subjects (college-age stutterers) read each of three different passages five consecutive times. In one condition, electroshock was administered during each moment of stuttering, and in another condition, electroshock was administered immediately following each stuttered word. The remaining condition was a control with no electroshock. The percentage of stuttering behavior during each reading was recorded, and Table 7.10 presents for each subject a rate of adaptation score under each condition. The score was found by using the residual measurement method suggested by Tate, Cullinan, and Ahlstrand (1961).

To determine if either of the two treatments yield improved (larger) median adaptation scores, we apply procedure (7.28), using average ranks to break the within-subjects ties in computing R_1, R_2, and R_3. Here, we have $k=3$ and $n=18$. For the sake of illustration, we take our experimentwise error rate to be $\alpha = .0492$. Using the R command

Table 7.10 Adaptation Scores for College-Age Stutterers

	Treatment		
Subject	1 (No shock)	2 (Shock following)	3 (Shock during)
1	57 (3)	38 (1)	51 (2)
2	59 (3)	48 (1)	56 (2)
3	44 (1.5)	50 (3)	44 (1.5)
4	51 (2)	53 (3)	44 (1)
5	43 (1)	53 (3)	50 (2)
6	49 (1)	56 (3)	54 (2)
7	48 (2)	37 (1)	50 (3)
8	56 (2)	58 (3)	40 (1)
9	44 (1.5)	44 (1.5)	50 (3)
10	50 (2)	50 (2)	50 (2)
11	44 (1)	58 (3)	56 (2)
12	50 (3)	48 (2)	46 (1)
13	70 (2)	60 (1)	74 (3)
14	42 (1)	58 (3)	57 (2)
15	58 (1)	60 (2)	74 (3)
16	54 (3)	38 (1)	48 (2)
17	38 (1)	48 (2.5)	48 (2.5)
18	48 (2)	56 (3)	44 (1)
	$R_1 = 33$	$R_2 = 39$	$R_3 = 36$

Source: D. A. Daly and E. B. Cooper (1967).

`cNWWM`(α,k,n) with $k = 3$ and $n = 18$, we find `cNWWM(.0492,3,18)` $= r^*_{.0492} = 12$, and procedure (7.28) reduces to

$$\text{Decide } \tau_u > \tau_1 \text{ if } (R_u - R_1) \geq 12.$$

Using the treatments sums of within-subjects ranks given in Table 7.10, we find that

$$(R_2 - R_1) = 6 \quad \text{and} \quad (R_3 - R_1) = 3.$$

Referring these rank sum differences to the critical value 12, we see that

$$(R_2 - R_1) = 6 < 12 \quad \Rightarrow \quad \text{decide } \tau_2 = \tau_1,$$
$$(R_3 - R_1) = 3 < 12 \quad \Rightarrow \quad \text{decide } \tau_3 = \tau_1.$$

Thus, at an experimentwise error rate of .0492, we find no statistical evidence that either of the two electroshock treatments lead to an increase in median adaptation scores over the control setting. (In fact, we can use the `R` command `pNWWM`($\cdot$) to obtain the smallest experimentwise error rate at which we would be able to declare a statistically significant increase in median adaptation scores for either of the two treatments. Since the largest observed difference in rank sums is $(R_2 - R_1) = 6$, we see that the smallest experimentwise error rate at which we could declare a statistically significant increase in median adaptation scores for either of the two treatments is `pNWWM(adaptation.scores)` `$p.val[1]` = .2859.

For the sake of illustration for the large-sample approximation (7.30), we simply note that $m^*_{.02002,1/2} =$ `cMaxCorrNor(.02002,3,0.5)` $= 2.30$ for $k = 3$ and $m^*_{.05410,1/2} =$ `cMaxCorrNor(.05410,6,0.5)` $= 2.20$ for $k = 6$.

Comments

32. *Rationale for Treatments-versus-Control Multiple Comparison Procedures*. The general rationale for the multiple comparison procedures of this section is the same as that given in Comment 24 for the two-sided all-treatments multiple comparison procedures of Section 7.3. The only additional factor here is that the treatments-versus-control procedures of this section do not compare all treatments, but only each noncontrol treatment with the control on a directional basis. This situation arises, for example, in drug screening in the examination of many new treatments in hopes of improving on a standard, and there is no initial reason to perform between treatment comparisons. Of course, similar comparisons between treatments that were selected as being better than the control would most likely be carried out later in a follow-up study.

    The multiple comparison procedures of this section, which involve making $k-1$ decisions, can also be interpreted as hypothesis tests. If we consider the procedure that rejects H_0 if and only if the inequality in (7.28) [or in (7.30)] holds for at least one $(1, u)$ pair, $u = 2, \ldots, k$, then this is a distribution-free test of level α for H_0 (7.2).

33. *Experimentwise Error Rate*. The use of an experimentwise error rate represents a very conservative approach to multiple comparisons. We insist that the probability of making only correct decisions be $1-\alpha$ when the hypothesis H_0 (7.2) of treatment equivalence is true. Thus we have a high degree of protection when H_0 is true, but we often apply the techniques of this section when we have evidence (perhaps based on a priori information or perhaps obtained by applying a previous test procedure) that H_0 is not true. (For additional general remarks about experimentwise error rates, see Comment 6.54.)

34. *Opposite Direction Decisions*. Procedures (7.28) and (7.30) are designed for the one-sided case where the decisions are $\tau_u > \tau_1$ versus $\tau_u = \tau_1, u = 2, \ldots, k$. To handle the analogous one-sided situation where the decisions involve $\tau_u < \tau_1$ versus $\tau_u = \tau_1, u = 2, \ldots, k$, use (7.28) and (7.30) with $(R_u - R_1)$ replaced by $(R_1 - R_u)$ for $u = 2, \ldots, k$.

35. *Critical Values r_α^**. The r_α^* critical values can be obtained by using the fact that under H_0 (7.2), all $(k!)^n$ rank configurations are equally likely. However, the computational effort is greater in this treatments-versus-control problem than in the all-treatments problem, because the values $R_u - R_1, u = 2, \ldots, k$, are in general changed when we relabel the control treatment. (In the all-treatments case, the relevant statistic range $[R_1, \ldots, R_k]$ is unaffected by treatment relabelings. (See Comment 27.)

    Let us now do an example to illustrate the nature of the necessary computations. For simplicity, we take the case $n = 3$, $k = 3$. Here, the largest possible value of $R_3 - R_1$ is 6, corresponding to the configuration

(a)

I	II	III
1	2	3
1	2	3
1	2	3,

where $R_1 = 3$, $R_3 = 9$. Similarly, the largest possible value of $R_2 - R_1$ is 6, corresponding to

(b)	I	II	III
	1	3	2
	1	3	2
	1	3	2.

Since none of the other configurations can yield an $R_u - R_1$ difference as large as 6, we have $P_0\{(R_2 - R_1) \geq 6 \text{ or } (R_3 - R_1) \geq 6\} = 2/[(3!)^3] = 2/216 = .0093$. Thus, in the notation of (7.28), we have $r^*_{.0093} = 6$, in agreement with the result obtained from using the R command `cNWWM(`α`,k,n)`; that is, `cNWWM(.0093,3,3)` $= r^*_{.0093} = 6$ for $k = 3$ treatments and $n = 3$ blocks.

36. *Historical Development*. The basic idea behind the treatments-versus-control multiple comparison procedures (7.28) and (7.30) based on the Friedman rank sums was initially discussed by Nemenyi (1963), Wilcoxon and Wilcox (1964), and Miller (1966). Windham (1971) provided the exact critical values r^*_α for procedure (7.28) for the case of $k = 3$, $n = 2(1)18$ and for $k = 4$, $n = 2(1)5$. Additional values of r^*_α were obtained by Odeh (1977) for the settings $k = 2(1)5$, $n = 2(1)8$ and $k = 6$, $n = 2(1)6$.

37. *Large-Sample Approximation.* Let $\boldsymbol{R}_d = (R_2 - R_1, \ldots, R_k - R_1)$ be the vector of differences between the treatment rank sums and the control rank sum. Then, it can be shown that a properly standardized $\boldsymbol{R}_d$ has, as n tends to infinity, an asymptotic multivariate normal distribution with mean vector $\mathbf{0}$ and appropriate covariance matrix Σ (see Miller (1966) for details). It follows directly from this result (again, see Miller (1966)) that the procedure in (7.30) has an asymptotic experimentwise error rate equal to α.

38. *Dependence on Observations from Other Noninvolved Treatments*. The treatments-versus-control multiple comparison procedures of this section suffer from the same disadvantage mentioned in Comment 30 for the corresponding all-treatments multiple comparisons. The decision between treatment $u(>1)$ and the control can be affected by changes only in the observations from one or more of the other $k - 2$ treatments that are not directly involved.

39. *Two-Sided Treatments-versus-Control Multiple Comparison Procedures*. The multiple comparison procedures of this section are both one sided by nature, resulting in decisions between $\tau_u = \tau_1$ and $\tau_u > \tau_1$ for every $u = 2, \ldots, k$ (or between $\tau_u = \tau_1$ and $\tau_u < \tau_1$ for every $u = 2, \ldots, k$, as noted in Comment 34). We view such one-sided comparisons to be the most natural approach for treatments-versus-control settings. In such situations, we are generally interested in seeing which, if any, of the proposed new treatments are better than a standard control or placebo. In most practical applications, *better* is synonymous with one-sided comparisons (all in one direction or all in the other), and thus our emphasis on such procedures in this section. However, a two-sided treatments-versus-control analog to procedure (7.28) has been developed in the literature and corresponds to the criterion

$$\text{Decide } \tau_u \neq \tau_1 \text{ if } |R_u - R_1| \geq r^{**}_\alpha; \quad \text{otherwise decide } \tau_u = \tau_1, \tag{7.31}$$

where the constant r_α^{**} is chosen to make the experimentwise error rate equal to α; that is,

$$P_0\{|R_u - R_1| < r_\alpha^{**}, u = 2, \ldots, k\} = 1 - \alpha,$$

where the probability $P_0(.)$ is computed under H_0 (7.2). Windham (1971) provided values of r_α^{**} for procedure (7.31) for the settings $k = 3$, $n = 2(1)18$ and $k = 4$, $n = 2(1)5$ (see also Hollander and Wolfe (1973)). A large-sample approximation to (7.31) is discussed in Miller (1966).

40. *Approximately Distribution-Free Multiple Comparison Procedures Based on Signed Ranks*. The multiple comparison procedures (7.28) and (7.30) both utilize the within-blocks Friedman ranking schemes, and, as a result, they are related to the sign procedures for paired replicates data (see Comments 7 and 16). Competitor treatments-versus-control multiple comparison procedures based on signed ranks that utilize between-block information are discussed in Section 7.14. These signed rank procedures are, however, only asymptotically distribution-free.

Properties

1. *Asymptotic Multivariate Normality*. See Miller (1966).
2. *Efficiency*. See Section 7.16.

Problems

37. Consider the serum CPK activity data from Problem 5. Treating preexercise as a control and ignoring the peak psychotic period data, apply procedure (7.28) to decide if there is statistical evidence of increased serum CPK activity either 19 or 42 h after exercise.

38. Apply an appropriate one-sided multiple comparison procedure (see (7.28) and Comment 34) to the rhythmicity data of Table 7.6, letting the condition N (subject spoke unaided by a metronome) serve as the control.

39. For the case $k = 3$, $n = 18$, and $\alpha \approx .01$, compare procedures (7.28) and (7.30).

40. Consider the rounding-first-base data discussed in Examples 7.1 and 7.3. Using the data for all 22 players and treating the round out method of running to second base as the control, find the smallest approximate experimentwise error rate at which we would declare that the median time to second base for the wide angle method of running is smaller than that for the round out method.

41. Illustrate the difficulty discussed in Comment 38 by means of a numerical example.

42. Find the complete list of available experimentwise error rates α and the associated r_α^* critical values for procedure (7.28) when $k = 3$ and, $n = 2$.

43. Consider the subset of the data on percentage correctly identified consonants in Table 7.4 corresponding to conditions A, AL, and AC. Treating condition A as a control, find the smallest experimentwise error rate for procedure (7.28) at which we would detect the condition (L or C) yielding the most improvement in performance when added to A in the syllable presentation.

44. Treating condition A as a control, apply procedure (7.30) to the data on percentage correctly identified consonants in Table 7.4.

45. Consider the maximum soil temperature data in Table 7.8. Apply the appropriate treatments-versus-control procedure to decide if maximum soil temperature is significantly warmer at 20, 40, or 100 m from the shelterbelt than at a distance of 200 m.

7.5 CONTRAST ESTIMATION BASED ON ONE-SAMPLE MEDIAN ESTIMATORS (DOKSUM)

In this section we discuss a method for point estimation of certain linear combinations of treatment effects known in the literature as *contrasts*. We define such a contrast in the treatment effects $\tau_1, \ldots, \tau_k$ to be any linear combination of the form

$$\theta = \sum_{i=1}^{k} a_i \tau_i, \tag{7.32}$$

where $a_1, \ldots, a_k$ are any specified constants such that $\sum_{i=1}^{k} a_i = 0$. Equivalently, we can write θ in terms of the individual differences in treatment effects (known in the literature as *simple contrasts*)

$$\Delta_{hj} = \tau_h - \tau_j, \quad h = 1, \ldots, k; \quad j = 1, \ldots, k, \tag{7.33}$$

by noting that

$$\theta = \sum_{h=1}^{k} \sum_{j=1}^{k} d_{hj} \Delta_{hj}, \tag{7.34}$$

where

$$d_{hj} = \frac{a_h}{k}, \quad h = 1, \ldots, k; \quad j = 1, \ldots, k. \tag{7.35}$$

For a given setting, decisions about which contrasts to estimate can be related to either a priori interest in particular linear combinations of the τ's or the results of one of the multiple comparison procedures discussed in Sections 7.3 and 7.4.

Procedure

For each pair of treatments $(u, v), u \neq v = 1, \ldots, k$, compute the differences

$$D_{uv}^{i} = X_{iu} - X_{iv}, i = 1, \ldots, n, \tag{7.36}$$

between the treatment u and treatment v observations for each of the n blocks. For $1 \leq u \neq v \leq k$, let

$$Z_{uv} = \text{median } \{D_{uv}^{i}, i = 1, \ldots, n\}. \tag{7.37}$$

Since $Z_{vu} = -Z_{uv}$, we need only to calculate the $k(k-1)/2$ values Z_{uv} corresponding to $u < v$. We refer to Z_{uv} as the "unadjusted" estimator of the simple contrast $\Delta_{uv} = \tau_u - \tau_v$. (Note that Z_{uv} is just the median estimator of Section 3.5, applied here to the $X_{iu} - X_{iv}$ differences. For example, Z_{23} is the median of the $X_{i2} - X_{i3}$ differences, $i = 1, \ldots, n$, and is the "unadjusted" estimator of the simple contrast $\tau_2 - \tau_3$.) Next, we compute

$$Z_{u.} = \sum_{j=1}^{k} \frac{Z_{uj}}{k}, \quad u = 1, \ldots, k, \tag{7.38}$$

where we note that $Z_{uu} = 0$ for $u = 1, \ldots, k$. Setting

$$\tilde{\Delta}_{uv} = Z_{u.} - Z_{v.}, \tag{7.39}$$

the adjusted estimator of θ is given by

$$\tilde{\theta} = \sum_{j=1}^{k} a_j Z_{j.}, \tag{7.40}$$

or, equivalently,

$$\tilde{\theta} = \sum_{h=1}^{k} \sum_{j=1}^{k} d_{hj} \tilde{\Delta}_{hj}. \tag{7.41}$$

EXAMPLE 7.5 ***Rounding First Base.***

Consider the rounding-first-base data originally presented in Table 7.1 of Example 7.1. We illustrate the Doksum contrast estimator $\tilde{\theta}$ (7.40) on the simple contrast $\theta = \tau_{\text{roundout}} - \tau_{\text{wide angle}} = \tau_1 - \tau_3$. In Example 7.3, we found the round out and wide angle methods differed significantly at the .01 experimentwise error rate. An estimate of $\tau_1 - \tau_3$ provides us with an idea of the time saved by running wide angle as opposed to round out.

From Table 7.11 and (7.37), we obtain $Z_{12} = .05$, $Z_{13} = .125$, and $Z_{23} = .10$. From (7.38), we have

$$Z_{1.} = \frac{Z_{11} + Z_{12} + Z_{13}}{3} = \frac{0 + .05 + .125}{3} = .058,$$

$$Z_{2.} = \frac{Z_{21} + Z_{22} + Z_{23}}{3} = \frac{-.05 + 0 + .10}{3} = .017,$$

$$Z_{3.} = \frac{Z_{31} + Z_{32} + Z_{33}}{3} = \frac{-.125 - .10 + 0}{3} = -.075.$$

Note that for calculating $Z_{2.}$ and $Z_{3.}$, we have used the fact that $Z_{uv} = -Z_{vu}$.

The adjusted estimator of $\theta = \tau_1 - \tau_3$ is now obtained using (7.32) with

$$a_1 = 1, \quad a_2 = 0, \quad a_3 = -1,$$

so that from (7.40), we have

$$\tilde{\theta} = Z_{1.} - Z_{3.} = .058 - (-.075) = .133.$$

Parenthetically, it should be noted that the equivalent form (7.34) is obtained with the identifications

$$d_{11} = d_{12} = d_{13} = \ \ \frac{1}{3},$$

$$d_{21} = d_{22} = d_{23} = \ \ 0,$$

$$d_{31} = d_{32} = d_{33} = -\frac{1}{3}.$$

Table 7.11 Values of D_{uv} Differences for Data of Table 7.1

Player i	D^i_{12}	D^i_{13}	D^i_{23}
1	−.10	−.15	−.05
2	.15	.10	−.05
3	−.40	−.30	.10
4	.05	.15	.10
5	.05	.20	.15
6	−.10	−.15	−.05
7	.00	.05	.05
8	−.05	.10	.15
9	.10	.25	.15
10	.05	.15	.10
11	.05	.15	.10
12	.10	.20	.10
13	.25	.15	−.10
14	.05	.10	.05
15	.00	.10	.10
16	−.10	−.05	.05
17	.00	.20	.20
18	−.05	−.10	−.05
19	.05	.25	.20
20	.05	.25	.20
21	.05	.15	.10
22	.00	.05	.05

Comments

41. *Unadjusted Estimator*. The unadjusted estimator Z_{uv} (7.37) of Δ_{uv} (7.33) is simply the estimator associated with the sign test and previously discussed in Section 3.5. However, the Doksum adjusted estimator $\tilde{\theta}$ (7.40) is quite often different from this simple unadjusted estimator Z_{uv}. This is the case in Example 7.5, for instance, where $Z_{13} = .125$, but $\tilde{\theta} = \widetilde{\tau_1 - \tau_3} = .133$.
42. *Ambiguities with the Unadjusted Estimators*. The unadjusted estimators Z_{uv} (7.37) lead to ambiguities in contrast estimation because they do not satisfy the linear relations that are satisfied by the contrasts they estimate. For example, $\Delta_{13} = \tau_1 - \tau_3 = (\tau_1 - \tau_2) + (\tau_2 - \tau_3) = \Delta_{12} + \Delta_{23}$, but, in general, $Z_{13} \neq Z_{12} + Z_{23}$. Thus, the two "reasonable" estimators Z_{13} and $Z_{12} + Z_{23}$ of $\Delta_{13} = \tau_1 - \tau_3$ can give different estimates. We refer to this property as the incompatibility of the unadjusted estimators Z_{uv}.
43. *Efficiency*. The adjusted estimators $\tilde{\Delta}_{uv}$ (7.39) of Δ_{uv} (7.33) are always at least as efficient as the unadjusted ones and they are compatible. They do, however, have the disadvantage that the estimator of $\Delta_{uv} = \tau_u - \tau_v$ depends on the observations from the other $k - 2$ treatments.
44. *Contrast Estimator Associated with Signed Ranks*. As noted in Comment 41, the contrast estimator $\tilde{\theta}$ (7.40) is related to paired replicates estimators associated with the sign statistic, as discussed in Sections 3.4 and 3.5. A competitor contrast estimator related to the Hodges–Lehmann paired replicates estimator associated with the signed rank statistic and utilizing between-block information is discussed in Section 7.15.

Properties

1. *Standard Deviation of $\tilde{\theta}$ (7.40)*. For the asymptotic standard deviation of $\tilde{\theta}$ (7.40), see Doksum (1967).
2. *Asymptotic Normality*. See Doksum (1967).
3. *Efficiency*. See Doksum (1967) and Section 7.16.

Problems

46. Estimate $2\tau_N - \tau_A - \tau_R$ for the metronome data of Table 7.6.

47. Illustrate, using a numerical example, the incompatibility of the unadjusted estimators Z_{uv} (see Comment 42).

48. Estimate the simple contrasts $\theta_1 = \tau_2 - \tau_1, \theta_2 = \tau_3 - \tau_1$, and $\theta_3 = \tau_3 - \tau_2$ for the CPK activity data in Table 7.3.

49. Estimate the contrast $3\tau_{ALC} - \tau_{AL} - \tau_{AC} - \tau_{LC}$ for the percentage consonants correctly identified data in Table 7.4.

50. Using the data of Table 7.4, estimate the simple contrast that represents the benefit from adding lip reading to audition in teaching severely hearing-impaired children.

51. Estimate the contrast $\tau_{AC} + \tau_{LC} - 2\tau_C$ for the percentage consonants correctly identified data in Table 7.4.

52. Estimate all contrasts found to be of interest in Problem 45 for the maximum soil temperature data in Table 7.8.

53. Estimate all possible simple contrasts for the ozone exposure data in Table 7.2.

54. Consider the percent average wind speed reduction data in Table 7.7. Use an appropriate all-treatments multiple comparison procedure (see Section 7.3) to decide which distances from the shelterbelt have significantly different reductions in average wind speed. Estimate all contrasts suggested to be important from this multiple comparison analysis.

55. Estimate the contrast $\tau_{\text{rats}} - \tau_{\text{cats}}$ for the Livesey EPT error score data of Table 7.9.

INCOMPLETE BLOCK DATA—TWO-WAY LAYOUT WITH ZERO OR ONE OBSERVATION PER TREATMENT–BLOCK COMBINATION

In two-way layout settings the most common form of data collection corresponds to the case of a single observation for every treatment–block combination. However, it is not uncommon to deal with two-way layout problems where certain treatment–block cells yield single observations, but where there are also treatment–block combinations for which we have no observations. This could be the result of a deliberate design to deal with data collection problems where it is not feasible (economically, time constraints, etc.) to collect data from every treatment–block combination or it could be simply a result of missing data from what was intended to be a complete block design.

In the next three sections we discuss procedures developed for such incomplete block data sets. In Sections 7.6 and 7.7 we present a distribution-free hypothesis test for general alternatives and an all-treatments multiple comparison procedure, respectively, for the most commonly used design specifically structured to yield less than complete block data, namely, the BIBD. In Section 7.8 we detail a distribution-free hypothesis

test for general alternatives that is applicable for two-way layout data representing an arbitrary configuration of either zero or one observation per cell.

Throughout these three sections, we continue to operate under the general conditions of Assumptions A1–A3. However, in these sections we impose the additional constraint that each c_{ij} is either 0 or 1 and $N = \sum_{i=1}^{n} \sum_{j=1}^{k} c_{ij} \neq kn$; that is, we have incomplete block data. We again drop the third subscript on the X variables in Sections 7.6–7.8. This will not be problematic, however, as there are no cells with more than one observation.

7.6 A DISTRIBUTION-FREE TEST FOR GENERAL ALTERNATIVES IN A RANDOMIZED BALANCED INCOMPLETE BLOCK DESIGN (BIBD) (DURBIN–SKILLINGS–MACK)

In this section we present a procedure for testing H_0 (7.2) against the general alternatives H_1 (7.3) for incomplete block data that arise from a very structured randomized BIBD. Such a BIBD corresponds to a setting where we observe $s(< k)$ treatments in each of the n blocks, every pair of treatments occurs together in the same number, λ, of blocks, and each of the k treatments is observed for a total of p times. These parameters of a BIBD must satisfy the restriction that $p(s-1) = \lambda(k-1)$, which, of course, forces additional constraints on the c_{ij}'s (see Problems 57 and 60).

Procedure

To compute the Durbin–Skillings–Mack statistic for such a balanced incomplete block design setting, we first order the available s observations from least to greatest separately within each of the n blocks. Let r_{ij} be this within-block rank of X_{ij} if there is an observation for the ith block–jth treatment combination; otherwise, let $r_{ij} = 0$. Set

$$R_j = \sum_{i=1}^{n} r_{ij}, \quad \text{for } j = 1, \ldots, k. \tag{7.42}$$

Thus, for example, R_3 is the sum (over the n blocks) of the within-blocks ranks received by the p available treatment 3 observations. (Note that each R_j will be the sum of exactly p nonzero within-blocks ranks.) The Durbin–Skillings–Mack statistic is then given by

$$\begin{aligned} D &= \left[\frac{12}{\lambda k(s+1)}\right] \sum_{j=1}^{k} \left\{R_j - \frac{p(s+1)}{2}\right\}^2 \\ &= \left\{\left[\frac{12}{\lambda k(s+1)}\right] \sum_{j=1}^{k} R_j^2\right\} - \frac{3(s+1)p^2}{\lambda}, \end{aligned} \tag{7.43}$$

where $(s+1)/2 = \sum_{j=1}^{k} r_{ij}/k$ is the average within-blocks rank assigned for each of the n blocks. Since each treatment is observed p times, it follows that $p(s+1)/2$ is the expected sum of within-blocks ranks for each of the k treatments when H_0 (7.2) is true.

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the general alternative

$$H_1 : [\tau_1, \ldots, \tau_k \text{ not all equal}],$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } D \geq d_{\alpha,s}; \quad \text{otherwise do not reject,} \tag{7.44}$$

where the constant $d_{\alpha,s}$ is chosen to make the type I error probability equal to α. The constant $d_{\alpha,s}$ is the upper α percentile for the null $(\tau_1 = \cdots = \tau_k)$ distribution of D. Comment 49 explains how to obtain the critical values $d_{\alpha,s}$ for k treatments, n blocks, and available values of α.

Large-Sample Approximation

When H_0 is true, the statistic D has, as n tends to infinity, an asymptotic chi-square (χ^2) distribution with $k - 1$ degrees of freedom. (See Comment 50 for indications of the proof.) The chi-square approximation for procedure (7.44) is

$$\text{Reject } H_0 \text{ if } D \geq \chi^2_{k-1,\alpha}; \quad \text{otherwise do not reject,} \tag{7.45}$$

where $\chi^2_{k-1,\alpha}$ is the upper α percentile of a chi-square distribution with $k - 1$ degrees of freedom. To find $\chi^2_{k-1,\alpha}$, we use the `R` command `qchisq`$(1 - \alpha, k - 1)$. For example, to find $\chi^2_{6,.025}$, we apply `qchisq(.975,6)` and obtain $\chi^2_{6,.025} = 14.45$.

Skillings and Mack (1981) noted that this approximate procedure (7.45) can be quite conservative when α is small (say, $\leq .01$) and either the number of blocks n or the number of common occurrences λ is small. In particular, they suggest that the approximation is conservative whenever λ is not at least 3. In such cases, they strongly recommend the use of the exact values of $d_{\alpha,s}$ whenever possible.

Ties

If there are ties among the X observations within any of the blocks, use average ranks to break the ties and compute the individual treatment sums of ranks $R_1, \ldots, R_k$. In such cases, the significance level associated with procedure (7.44) is only approximately equal to α. (See Comment 51 for discussion of how to construct a conditionally distribution-free test of H_0 even when there are ties within some of the blocks.)

EXAMPLE 7.6 ***Chemical Toxicity.***

Moore and Bliss (1942) compared the toxicity of each of seven chemicals applied to *Aphis rumicis*, a black aphid found on nasturtiums. The logarithm of the dose required to kill 95% of the insects exposed to a chemical was the measurement reported. Since the experimenters could test only three chemicals in any given day, they used a balanced incomplete block design requiring 7 days for completion of the experiment. The toxicities for the studied chemicals are shown in Table 7.12.

Table 7.12 Logarithm of Toxic Dosages

	Chemical						
Day	A	B	C	D	E	F	G
1	.465	.343		.396			
2	.602		.873		.634		
3			.875	.325			.330
4	.423					.987	.426
5		.652	1.142			.989	
6		.536			.409		.309
7				.609	.417	.931	

Source: W. Moore and C. I. Bliss (1942).

This experiment constitutes a BIBD with $k = 7$ treatments, of which $s = 3$ are observed in each of the $n = 7$ blocks, every pair of treatments occur together in $\lambda = 1$ of the blocks, and each of the treatments is observed for a total of $p = 3$ times. We are interested in assessing whether there are any differences in the toxicities of the seven chemicals relative to *A. rumicis*. We will use the approximate procedure (7.45) to test if there are any differences in the toxicities of the seven chemicals. For the sake of illustration, we take the approximate significance level to be $\alpha = .05$. Using the R command `qchisq`$(1-\alpha,\ k)$, we find the value $\chi^2_{6,.05} =$ `qchisq(.95, 6)` $= 12.59$ and procedure (7.45) reduces to

$$\text{Reject } H_0 \text{ if } D \geq 12.59.$$

Now, we illustrate the computations leading to the sample value of D (7.43). Ranking from 1 to 3 within each of the seven blocks (days) and summing across the blocks for each of the chemicals, we obtain the following treatment sums of ranks:

$$R_1 = 3+1+1 = 5, \quad R_2 = 1+1+3 = 5, \quad R_3 = 3+3+3 = 9, \quad R_4 = 2+1+2 = 5,$$

$$R_5 = 2+2+1 = 5, \quad R_6 = 3+2+3 = 8, \quad R_7 = 2+2+1 = 5.$$

Hence, from (7.43), we find that

$$D = \left\{\left[\frac{12}{1(7)(4)}\right](5^2+5^2+9^2+5^2+5^2+8^2+5^2)\right\} - \frac{3(4)(3^2)}{1}$$

$$= \left\{\frac{3(270)}{7}\right\} - 108 = 7.71.$$

Comparison of this observed value of 7.71 with the approximate critical value $\chi^2_{6,.05} = 12.59$ leads us to conclude that there is not strong sample evidence to indicate any significant difference between the seven studied chemicals with respect to their toxicities for *A. rumicis*. In fact, the observed value of $D = 7.71$ is approximately the .26 upper percentile for the chi-square distribution with 6 degrees of freedom (i.e., `qchisq(.74, 6)` ≈ 7.71). Thus, the approximate P-value for these data and procedure

(7.45) is .26, providing further evidence of the similarity of the studied chemicals with respect to their toxic effects on *A. rumicis*.

Comments

45. *More General Setting*. We could replace Assumptions A1–A3 and H_0 (7.2) with the more general null hypothesis that all possible $(s!)^n$ rank configurations for the nonzero r_{ij}'s are equally likely. Procedure (7.44) remains distribution-free for this more general hypothesis.
46. *Design Rationale*. In a two-way layout setting with no replications within block–treatment combinations, it is best to use a randomized complete block design (as discussed in Sections 7.1–7.5) whenever possible. However, there are times when experimental constraints such as fixed costs or limited time or facilities make it impossible to obtain an observation for every treatment–block combination. When this is the case, the use of a balanced incomplete block design is often a good alternative. Such a BIBD provides for sufficient data to be collected to permit comparison of each treatment with every other one. Moreover, the fact that the BIBD imposes a rigid structure on the missing observations within and across the blocks enables the associated data analysis to be both relatively simple and efficient.
47. *Motivation for the Test*. Under Assumptions A1–A3 and H_0 (7.2), each of the block rank vectors $\boldsymbol{R}_i^*$ for those s observations present in the ith block, $i = 1, \ldots, n$, has a uniform distribution over the set of all $s!$ permutations of the vector of integers $(1, 2, \ldots, s)$. If r_{ij} is nonzero, it follows that $E_0(r_{ij}) = (1/s!)[(s-1)!] \sum_{t=1}^{k} t = (s+1)/2$, the average rank being assigned separately to the partial data in each of the blocks. Thus, $E_0(R_j) = \sum_{i=1}^{n} E_0(r_{ij}) = p(s+1)/2$, for $j = 1, \ldots, k$, because there are observations in exactly p of the blocks for each of the k treatments. Therefore, we would expect each of the R_j's to be close to $p(s+1)/2$ when H_0 is true. Since the test statistic D (7.43) is a constant times a sum of squared differences between the observed treatment sums of ranks, R_j, and their common null expected value, $E_0(R_j) = p(s+1)/2$, small values of D represent agreement with H_0 (7.2). When the τ's are not all equal, we would expect a portion of the associated treatment sums of ranks to differ from their common null expectation, $p(s+1)/2$, with some tending to be smaller and some larger. The net result (after squaring the observed differences to obtain the $[R_j - p(s+1)/2]^2$ terms) would be a large value of D. This quite naturally suggests rejecting H_0 in favor of H_1 (7.3) for large values of D and motivates procedures (7.44) and (7.45).
48. *Assumptions.* We emphasize that Assumption A3 stipulates that the *ns* cell distributions F_{ij} for those treatment–block combinations where observations are collected can differ at most in their locations (medians) and that these location differences (if any) must be a result of additive block and/or treatment effects (i.e., there is no interaction between the treatment and block factors). In particular, Assumption A3 requires that the *ns* underlying distributions belong to the same general family (F) and that they do not differ in scale parameters (variability). We do note, however, that the test procedure (7.44) remains distribution-free under the less restrictive setting where Assumption A3 is replaced by the weaker condition

A3′. The distribution functions $F_{11}, \ldots, F_{1k}, \ldots, F_{n1}, \ldots, F_{nk}$ are connected through the relationship

$$F_{ij}(u) = F_i(u - \tau_j), -\infty < u < \infty,$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, k$, where $F_1, \ldots, F_n$ are arbitrary distribution functions for continuous distributions with unknown medians $\theta_1, \ldots, 0_n$, respectively, and, as before, τ_j is the unknown additive treatment effect contributed by the jth treatment.

Assumption A3 then corresponds to Assumption A3′ with the additional condition that $F_1 \equiv \cdots \equiv F_n$ (see also Comment 45).

49. *Derivation of the Distribution of D under H_0 (No-Ties Case).* The null distribution of D (7.43) can be obtained using the fact that under H_0 (7.2), all possible $(s!)^n$ rank configurations for the nonzero r_{ij}'s are equally likely. We now take the simplest (but not very useful in practice) balanced incomplete block design corresponding to $k = 3, s = 2, n = 3, p = 2$, and $\lambda = 1$ to illustrate how the null distribution can be derived. In this case, D (7.43) reduces to $D = (\frac{4}{3})R^* - 36$, where $R^* = R_1^2 + R_2^2 + R_3^2$.

The value of D for each of the $(2!)^3 = 8$ possible rank configurations for this setting are presented below.

I	II	III
1	2	
1		2
	1	2

$R^* = 29, D = 2.67$

I	II	III
2	1	
1		2
	1	2

$R^* = 29, D = 2.67$

I	II	III
1	2	
2		1
	1	2

$R^* = 27, D = 0$

I	II	III
2	1	
2		1
	1	2

$R^* = 29, D = 2.67$

I	II	III
1	2	
1		2
	2	1

$R^* = 29, D = 2.67$

I	II	III
1	2	
2		1
	2	1

$R^* = 29, D = 2.67$

I	II	III
2	1	
1		2
	2	1

$R^* = 27, D = 0$

I	II	III
2	1	
2		1
	2	1

$R^* = 29, D = 2.67$

Thus, we find

$$P_0\{D = 0\} = .25 \quad \text{and} \quad P_0\{D = 2.67\} = .75.$$

Note that we have derived the null distribution of D without specifying the common form (F) of the underlying distribution function for the X's under H_0 beyond the point of requiring that it be continuous. This is why the test procedure (7.44) based on D is called a *distribution-free Procedure*. From the

null distribution of D, we can determine the critical value $d_{\alpha,s}$ and control the probability α of falsely rejecting H_0 when it is true, and this error probability does not depend on the specific form of the common underlying continuous X distribution.

For a given BIBD design with incidence matrix **obs.mat**, the R command cDurSkiMa(α, **obs.mat**) can be used to find the available upper-tail critical values $d_{\alpha,s}$ for possible values of D. The incidence matrix will be an $n \times k$ matrix of ones and zeroes, which indicate where the data are observed and unobserved, respectively. Methods for finding the incidence matrix for various BIBD designs are given in the literature. While the incidence matrix will not be unique for a given (k, n, s, λ, p) combination, the distribution of D under H_0 will be the same for any of the possible incidence matrices. For a given available significance level α, the critical value $d_{\alpha,s}$ then corresponds to $P_0(D \geq d_{\alpha,s}) = \alpha$ and is given by cDurSkiMa(α, **obs.mat**) $= d_{\alpha,s}$. Thus, for example, for the BIBD combination $(k, n, s, \lambda, p) = (3, 3, 2, 1, 2)$, one possibility would be $\textbf{obs.mat} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$, which would yield cDurSkiMa(.75, **obs.mat**) $= d_{.75,s} = 2.67$, as noted previously in this comment. As a second more practical example, we can consider the case of $(k, n, s, \lambda, p) = (6, 15, 4, 6, 10)$. A possible incidence matrix is given by

$$\textbf{obs.mat} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \end{bmatrix},$$

which yields cDurSkiMa(.0487, **obs.mat**) $= d_{.0487,s} = 10.80$.

50. *Large-Sample Approximation*. Define the random variables $T_j = R_j - E_0(R_j) = R_j - p(s+1)/2$, for $j = 1, \ldots, k$. Since each R_j is a sum, it is not surprising (see, e.g., Skillings and Mack (1981) for formal justification) that a properly standardized version of the vector $\boldsymbol{T}^* = (T_1, \ldots, T_{k-1})$ has an asymptotic (n tending to infinity) $(k-1)$-variate normal distribution with mean vector $\mathbf{0} = (0, \ldots, 0)$ and appropriate covariance matrix Σ when the null hypothesis H_0 is

true. (Note that $\boldsymbol{T}^*$ does not include $T_k = R_k - p(s+1)/2$, because T_k can be expressed as a linear combination of $T_1, \ldots, T_{k-1}$. This is the reason that the asymptotic normal distribution is $(k-1)$-variate and not k-variate.) Since the test statistic D (7.43) is a quadratic form in the variables $(T_1, \ldots, T_{k-1})$, it is therefore quite natural that D has an asymptotic (n tending to infinity) chi-square distribution with $k-1$ degrees for freedom.

51. *Exact Conditional Distribution of D with Ties among the Observed X Values within Blocks*. To have a test with exact significance level even in the presence of tied X's within some of the blocks, we need to consider all $(s!)^n$ block rank configurations, where now these within-blocks ranks are obtained using average ranks to break ties. As in Comment 49, it still follows that under H_0 each of the $(s!)^n$ block rank configurations (now with these tied ranks) is equally likely. For each such configuration, the value of D is computed and the results are tabulated. We illustrate this for the same setting as was used in the untied example in Comment 49 (namely, $k=3$, $s=2$, $n=3$, $p=2$, and $\lambda=1$), except here we assume that the two observations within block one are tied in value. Thus, the block ranks we are dealing with here are (1.5, 1.5), (1, 2), and (1, 2) for blocks 1, 2, and 3, respectively.

As in Comment 49, D (7.43) reduces to $D = (\frac{4}{3})R^* - 36$, where $R^* = R_1^2 + R_2^2 + R_3^2$. Since the rank configuration for the first block is always (1.5, 1.5), we need only compute the value of D for $(2!)^2 = 4$ possible rank configurations. The values of D for these four configurations are as follows:

I	II	III
1.5	1.5	
1		2
	1	2

$R^* = 28.5, D = 2$

I	II	III
1.5	1.5	
1		2
	2	1

$R^* = 27.5, D = 0.67$

I	II	III
1.5	1.5	
2		1
	1	2

$R^* = 27.5, D = 0.67$

I	II	III
1.5	1.5	
2		1
	2	1

$R^* = 28.5, D = 2$

Thus, we find

$$P_0\{D = 0.67\} = .50 \quad \text{and} \quad P_0\{D = 2\} = .50.$$

This distribution is called the *conditional distribution* or the *permutation distribution* of D, given the set of tied within-blocks ranks (1.5, 1.5), (1, 2), and (1, 2).

52. *Historical Development*. The test procedures (7.44) and (7.45) based on D were first proposed by Durbin (1951). Later, Skillings and Mack (1981) studied a more general procedure for arbitrary incomplete block data (see Section 7.8)

and, for the first time, made available the exact critical values $d_{\alpha,s}$ for procedure (7.44) for a reasonable set of balanced incomplete block designs.

Properties

1. *Consistency*. See van Elteren and Noether (1959).
2. *Asymptotic Chi-Squaredness*. See Durbin (1951), Benard and van Elteren (1953) or Skillings and Mack (1981).
3. *Efficiency*. See van Elteren and Noether (1959) and Section 7.16.

Problems

56. Mendenhall (1968) discusses an experiment that was conducted to compare the effects of seven different chemical substances on the skin of male rats. The necessity to use relatively homogeneous patches of a rat's skin for the study restricted the experimenter to three experimental units (patches of skin) per animal. However, to avoid the confounding effect of rat-to-rat variability in the comparison of the seven chemicals, the experimenter was obligated to block on rats and any given rat could be treated with only three of the seven chemicals. This resulted in the use of a balanced incomplete block design with parameters $k = 7$, $n = 7$, $s = 3$, $p = 3$, and $\lambda = 1$. The experimental measurements for the study are presented in Table 7.13.

Apply procedure (7.45) with approximate significance level $\alpha \approx .05$ to these data to test the hypothesis of interest.

57. Verify that the relationship $p(s-1) = \lambda(k-1)$ must hold for a balanced incomplete block design.

58. Verify that the two representations for D (7.43) are, in fact, equivalent.

59. What are the maximum and minimum values for the test statistic D (7.43)? What rank configurations lead to these maximum and minimum values?

60. Consider the relationship $p(s-1) = \lambda(k-1)$ that must hold for a balanced incomplete block design. What constraints does this condition place on the c_{ij}'s for the data?

61. Kuehl (1994) described an experiment by J. Berry and A. Deutschman at the University of Arizona designed to study the effect of pressure on percent conversion of methyl glucoside to monovinyl isomers. The conversion is achieved by addition of acetylene to methyl glucoside in the presence of a base under high pressure. Five pressures were of interest in the study, but only three could be examined at any one time under identical experimental conditions because

Table 7.13 Reactions of Male Rats to Chemical Substances

	Chemical Substance						
Rat	A	B	C	D	E	F	G
1	10.2	6.9		14.2			
2			9.9	12.9		14.1	
3		12.1	11.7		8.6		
4				14.3	9.1		7.7
5		8.8				16.3	8.6
6	13.1				9.2	15.2	
7	11.3		9.7				6.2

Source: W. Mendenhall (1968).

Table 7.14 Percent Conversion of Methyl Glucoside to Monovinyl Isomers

	Pressure (psi)				
Experimental run	250	325	400	475	550
1	16	18		32	
2	19			46	45
3		26	39		61
4			21	35	55
5		19		47	48
6	20		33	31	
7	13	13	34		
8	21		30		52
9	24	10			50
10		24	31	37	

Source: R. O. Kuehl (1994).

of limited laboratory space. This necessitated the use of a balanced incomplete block design. The data obtained in the experiment and design are given in Table 7.14.

State the parameters for the BIBD employed in this chemical conversion study. Apply procedure (7.44) to these data to assess whether pressure (at the levels included in the study) has any effect on the percent conversion of methyl glucoside to monovinyl isomers.

62. Consider the BIBD corresponding to $k = 5$, $n = 10$, $s = 3$, $p = 6$, and $\lambda = 3$. Compare the critical region for the exact level $\alpha = .0499$ test of H_0 (7.2) based on D with the critical region for the corresponding nominal level $\alpha = .0499$ test based on the large-sample approximation.

63. Consider the BIBD corresponding to $k = 4$, $n = 6$, $s = 2$, $p = 3$, and $\lambda = 1$. Obtain the form of the exact null H_0 distribution of D (7.43) for the case of no-tied observations.

64. Consider the BIBD corresponding to $k = 5$, $n = 20$, $s = 2$, $p = 8$, and $\lambda = 2$. Compare the critical region for the exact level $\alpha = .0685$ test of H_0 (7.2) based on D with the critical region for the corresponding nominal level $\alpha = .0685$ test based on the large-sample approximation.

65. Consider the BIBD corresponding to $k = 4$, $n = 6$, $s = 2$, $p = 3$, and $\lambda = 1$. Suppose that the two observations in each of blocks 3 and 5 are tied. Obtain the conditional exact probability distribution of D under H_0 (7.2) when average ranks are used to break these two sets of within-blocks ties. Compare this conditional null distribution of D with the null distribution for D obtained in Problem 63 when there are no ties.

66. Consider the percentage consonants correctly identified data in Table 7.4 for conditions AL, AC, LC, and ALC only. Suggest a possible BIBD that could have been utilized in this study for these four treatments to reduce the number of conditions under which each severely hearing-impaired child had to be observed. Using a random mechanism for deciding how to apply the BIBD in question to the existing data set, analyze the corresponding data subset to assess whether there are any differences in the effectiveness of the conditions AL, AC, LC, and ALC for teaching severely hearing-impaired children.

67. Consider the percentage consonants correctly identified data in Table 7.4 for conditions A, L, and C only, Suggest a possible BIBD that could have been utilized in this study for these three treatments to reduce the number of conditions under which each severely hearing-impaired child had to be observed. Using a random mechanism for deciding how to apply the BIBD in question to the existing data set, analyze the corresponding data subset to assess whether there are any differences in the effectiveness of the conditions A, L, and C for teaching severely hearing-impaired children.

7.7 ASYMPTOTICALLY DISTRIBUTION-FREE TWO-SIDED ALL-TREATMENTS MULTIPLE COMPARISONS FOR BALANCED INCOMPLETE BLOCK DESIGNS (SKILLINGS–MACK)

In this section we present an asymptotically distribution-free multiple comparison procedure using the Friedman within-blocks ranks that is designed to make decisions about individual differences between pairs of treatment effects (τ_i, τ_j), for $i < j$, for data obtained from a balanced incomplete block design. The multiple comparison procedure of this section would generally be applied to BIBD data *after* rejection of H_0 (7.2) with the Durbin–Skillings–Mack procedure from Section 7.6. In this setting we will reach conclusions about all $k(k-1)/2$ pairs of treatment effects and these conclusions are naturally two-sided in nature.

Procedure

Let $R_1, \ldots, R_k$ be the treatment sums of within-blocks ranks given by (7.42). Calculate the $k(k-1)/2$ absolute differences $|R_u - R_v|$, $1 \le u < v \le k$.

When H_0 (7.2) is true, the $k(k-1)/2$-component vector $(R_1, \ldots, R_k)$ has, when properly standardized and as n tends to infinity, an asymptotic $(k-1)$-variate normal distribution with appropriate mean vector and covariance matrix (see Skillings and Mack (1981) for details of the proof). At an approximate experimentwise error rate of α, the Skillings–Mack two-sided all-treatments multiple comparison procedure reaches its $k(k-1)/2$ pairwise decisions, corresponding to each (τ_u, τ_v) pair, $1 \le u < v \le k$, by the criterion

$$\begin{aligned} &\text{Decide } \tau_u \neq \tau_v \text{ if } |R_u - R_v| \ge [(s+1)(ps - p + \lambda)/12]^{1/2} q_\alpha; \\ &\qquad \text{otherwise decide } \tau_u = \tau_v, \end{aligned} \tag{7.46}$$

where q_α is the upper αth percentile for the distribution of the range of k independent $N(0, 1)$ variables. To find q_α for k treatments and a specified experimentwise error rate α, we use the R command `cRangeNor(`α`,`k`)`. For example, to find $q_{.025}$ for $k = 6$ treatments, we apply `cRangeNor(.025,6)` and obtain $q_{.025} = 4.361$ for $k = 6$. (See also Comment 55.)

Ties

If there are ties among the X observations within any of the blocks, use average ranks to break the ties and compute the individual treatment sums of ranks $R_1, \ldots, R_k$.

EXAMPLE 7.7 ***Chemical Toxicity.***

For the sake of illustration, we apply procedure (7.46) to the chemical toxicity data relative to the black aphid, *A. rumicis*, as previously discussed in Example 7.6, even though the Durbin–Skillings–Mack procedure did not find any significant differences (approximate P-value of .26) between the treatment effects. Taking our approximate

experimentwise error rate to be $\alpha \approx .05$, we find `cRangeNor(.05, 7)` $= q_{.05} = 4.170$ for $k = 7$ and procedure (7.46) reduces to

$$\text{Decide } \tau_u \neq \tau_v \text{ if } |R_u - R_v| \geq [4(9 - 3 + 1)/12]^{1/2}(4.170) = 6.370.$$

Using the treatments sums of within-blocks ranks obtained in Example 7.6, we find that

$$\begin{array}{lllll} |R_2 - R_1| = 0, & |R_3 - R_1| = 4, & |R_4 - R_1| = 0, & |R_5 - R_1| = 0, & |R_6 - R_1| = 3, \\ |R_7 - R_1| = 0, & |R_3 - R_2| = 4, & |R_4 - R_2| = 0, & |R_5 - R_2| = 0, & |R_6 - R_2| = 3, \\ |R_7 - R_2| = 0, & |R_4 - R_3| = 4, & |R_5 - R_3| = 4, & |R_6 - R_3| = 1, & |R_7 - R_3| = 4, \\ |R_5 - R_4| = 0, & |R_6 - R_4| = 3, & |R_7 - R_4| = 0, & |R_6 - R_5| = 3, & |R_7 - R_5| = 0, \\ & & |R_7 - R_6| = 3. & & \end{array}$$

Since all these absolute differences are less than the critical value, 6.370, we see that the $7(6)/2 = 21$ decisions at this approximate experimentwise error rate of .05 are that $\tau_u = \tau_v$, for $1 \leq u < v \leq 7$. This is, of course, not at all surprising, because, in Example 7.6, the Durbin–Skillings–Mack test procedure found no support for rejecting H_0 (7.2) for these data.

Comments

53. *Rationale for Multiple Comparison Procedure.* The rationale behind the multiple comparison procedure of this section for data from a balanced incomplete block design is similar to that behind the two-sided multiple comparison procedures for data from a complete randomized block design. For further discussion, see Comment 24.
54. *Experimentwise Error Rate.* The use of an experimentwise error rate represents a very conservative approach to multiple comparisons. We are insisting that the probability of making only correct decisions be $1 - \alpha$ when the null hypothesis H_0 (7.2) of treatment equivalence is true. Thus, we have a high degree of protection when H_0 is true, but we often apply such techniques when we have evidence (perhaps based on a priori information or perhaps obtained by applying the Durbin–Skillings–Mack test, as in Example 7.7) that H_0 is not true. The protection under H_0 also makes it harder for the procedure to judge treatments as differing significantly when, in fact, H_0 is false, and this difficulty becomes more severe as k increases. See Comment 6.54 for additional discussion of experimentwise error rates.
55. *Conservative Procedure.* Skillings and Mack (1981) also proposed a conservative multiple comparison procedure that guarantees an upper bound on the experimentwise error rate. Let $R_1, \ldots, R_k$ be the treatment sums of within-blocks ranks given by (7.42). At an experimentwise error rate *no greater* than α, the Skillings–Mack conservative two-sided all-treatments multiple comparison procedure reaches its $k(k-1)/2$ decisions through the criterion

$$\begin{aligned} &\text{Decide } \tau_u \neq \tau_v \text{ if } |R_u - R_v| \geq [k\lambda d_{\alpha,s}(s+1)/6]^{1/2}; \\ &\qquad \text{otherwise decide } \tau_u = \tau_v, \end{aligned} \tag{7.47}$$

where $d_{\alpha,s}$ is the upper α percentile for the null distribution of the Durbin–Skillings–Mack statistic D (7.43). Skillings and Mack (1981) note that although procedure (7.47) does not require a large number of blocks, it is, nevertheless, rather conservative because it is based on the projection procedure of Scheffé; that is, the true experimentwise error rate might be considerably smaller than the bound α provided by (7.47). As a result, they recommend using the approximation (7.46) whenever the number of blocks is reasonably large.

56. *Contrast Estimators for BIBD's.* Greenberg (1966) proposed a method of contrast estimation for general (including balanced) incomplete block designs where the number of observations in a block is smaller than the number of treatments to be compared.

57. *Dependence on Observations from Other Noninvolved Treatments.* The all-treatments multiple comparison procedure of this section suffers from the same disadvantage as do the other two-way layout multiple comparison procedures of this chapter. The decision between treatment u and treatment v can be affected by changes only in the observations from one or more of the other $k - 2$ treatments that are not directly involved.

Properties

1. *Asymptotic Multivariate Normality*. See Skillings and Mack (1981).
2. *Efficiency.* See Section 7.16.

Problems

68. Apply procedure (7.46) to the chemical substance effect data of Table 7.13 in Problem 56.

69. Illustrate the difficulty discussed in Comment 57 by means of a numerical example.

70. Apply procedure (7.46) to the percent conversion data of Table 7.14 in Problem 61.

71. Consider the chemical toxicity data of Table 7.12 in Example 7.6. Find the smallest approximate experimentwise error rate at which the most significant difference(s) in black aphid (*A. rumicis*) toxicity between the studied substances would be detected by procedure (7.46).

72. Consider the chemical substance effect data of Table 7.13 in Problem 56. Find the smallest approximate experimentwise error rate at which procedure (7.46) would declare that chemical substances F and G have differing effects on the skin of male rats.

73. Consider the percent conversion data of Table 7.14 in Problem 61. Find the smallest approximate experimentwise error rate at which the most significant difference in the effects of the various pressures on the percent conversion of methyl glucoside to monovinyl isomers would be detected by procedure (7.46).

7.8 A DISTRIBUTION-FREE TEST FOR GENERAL ALTERNATIVES FOR DATA FROM AN ARBITRARY INCOMPLETE BLOCK DESIGN (SKILLINGS–MACK)

Not every set of data resulting from less than a randomized complete block design satisfies the necessary constraints (see Section 7.6) to be analyzed by the Durbin–Skillings–Mack procedure for balanced incomplete block designs. In this section we present a general

procedure for analyzing data from a two-way layout where there are either zero or one observation for each treatment–block combination but where there is not necessarily any nice pattern to the particular combinations for which we do not have observations. Such an incomplete data configuration could, of course, be intentionally designed this way, but it could also be the consequence of missing observations from an experiment that was intended to yield data for a randomized complete block design.

For this general two-way layout setting, let s_i denote the number of treatments for which an observation is present in block i, for $i = 1, \ldots, n$. (If $s_i = 1$ for block i, we remove that block from the analysis. Therefore, throughout this section, n will denote the number of blocks for which $s_i \geq 2$.) We discuss a distribution-free procedure for testing H_0 (7.2) against the general alternatives H_1 (7.3) when we are faced with such arbitrarily incomplete block data.

Procedure

To compute the Skillings–Mack statistic for arbitrarily incomplete block data, we first rank the s_i observed data values in block i from least to greatest, for each block $i = 1, \ldots, n$. Thus, in the ith block, we will be assigning ranks $1, 2, \ldots, s_i$. For $i = 1, \ldots, n$ and $j = 1, \ldots, k$, let

$$\begin{aligned} r_{ij} &= \text{rank of } X_{ij} \text{ among the observations in block } i, \quad \text{if } c_{ij} = 1, \\ &= (s_i + 1)/2, \quad \text{if } c_{ij} = 0, \end{aligned}$$

where $(s_i + 1)/2$ is the average of the ranks assigned to the observations present in the ith block. Set

$$A_j = \sum_{i=1}^{n} \left(\frac{12}{s_i + 1} \right)^{1/2} \left(r_{ij} - \frac{s_i + 1}{2} \right), \quad j = 1, \ldots, k. \tag{7.48}$$

Thus, A_j is the weighted sum of centered (around the block average) ranks for the observations from the jth treatment, with the block weighting factor $[12/(s_i + 1)]^{1/2}$ being inversely proportional to the square root of the number of observations present in the block (see Comment 63). Set

$$\boldsymbol{A} = (A_1, \ldots, A_{k-1}). \tag{7.49}$$

(Without loss of generality, we have chosen to omit A_k from the vector $\boldsymbol{A}$. The A_j's are linearly dependent, because a weighted linear combination of all k of them is a constant. We could omit any one of the A_j's in the definition of $\boldsymbol{A}$ (7.49), and the procedure we now describe would lead to the same value of the test statistic. For further discussion, see Skillings and Mack (1981).)

The covariance matrix for $\boldsymbol{A}$ under H_0 (7.2) is given by

$$\Sigma_0 = \begin{bmatrix} \sum_{t=2}^{k} \lambda_{1t} & -\lambda_{12} & -\lambda_{13} & \cdots & -\lambda_{1,k-1} \\ -\lambda_{12} & \sum_{t \neq 2}^{k} \lambda_{2t} & -\lambda_{23} & \cdots & -\lambda_{2,k-1} \\ \vdots & \vdots & \vdots & & \vdots \\ -\lambda_{1,k-1} & -\lambda_{2,k-1} & -\lambda_{3,k-1} & \cdots & \sum_{t \neq k-1}^{k} \lambda_{k-1,t} \end{bmatrix} \tag{7.50}$$

where, for $t \neq q = 1, \ldots, k$,

$$\lambda_{qt} = \lambda_{tq} = [\text{number of blocks in which both treatments } q \text{ and } t \text{ are observed}]. \tag{7.51}$$

Let Σ_0^- be any (see Comment 62) generalized inverse for Σ_0. The Skillings–Mack statistic is then given by

$$SM = \boldsymbol{A}\Sigma_0^- \boldsymbol{A}'. \tag{7.52}$$

(We note that if $\lambda_{qt} > 0$ for all $q \neq t$, then the rank of the covariance matrix Σ_0 (7.50) is $k - 1$, and we can simply use the ordinary inverse Σ_0^{-1} in the definition of SM (7.52).)

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the general alternative

$$H_1 : [\tau_1, \tau_2, \ldots, \tau_k \text{ not all equal}],$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } SM \geq sm_\alpha; \quad \text{otherwise do not reject,} \tag{7.53}$$

where the constant sm_α is chosen to make the type I error probability equal to α. Comment 64 explains how to obtain the critical values sm_α for a two-way layout configuration with k treatments, n blocks, and observation indicators $c_{ij}, i = 1, \ldots, n; j = 1, \ldots, k$.

Large-Sample Approximation

When H_0 (7.2) is true and $\lambda_{qt} > 0$ for every $q \neq t = 1, \ldots, k$ (i.e., every pair of treatments occur together in at least one block), the statistic SM has, as n tends to infinity, an asymptotic chi-square (χ^2) distribution with $k - 1$ degrees of freedom (see Comment 65 for indications of the proof). The chi-square approximation for procedure (7.53) is

$$\text{Reject } H_0 \text{ if } SM \geq \chi^2_{k-1,\alpha}; \quad \text{otherwise do not reject,} \tag{7.54}$$

where $\chi^2_{k-1,\alpha}$ is the upper α percentile of a chi-square distribution with $k - 1$ degrees of freedom. To find $\chi^2_{k-1,\alpha}$, we use the R command `qchisq`$(1 - \alpha, k - 1)$. For example, to find $\chi^2_{5,.05}$, we apply `qchisq(.95,5)` and obtain $\chi^2_{5,.05} = 11.071$.

As with the BIBD procedure discussed in Section 7.6, Skillings and Mack (1981) have pointed out that this approximate procedure (7.54) can be quite conservative when α is small (say, $\leq .01$) and the number of blocks, n, is not large. In such cases, it is preferable to generate the exact critical value sm_α and use procedure (7.53). (See Comment 64.) (We should also point out that the approximate procedure (7.54) is simply not applicable if there are at least two treatments that do not have observations together in any of the blocks; that is, if $\lambda_{qt} = 0$ for at least one pair $q \neq t = 1, \ldots, k$.)

Table 7.15 Subset of Data on the Influence of Rhythmicity of Metronome on Speech Fluency

	Dysfluencies under each condition		
Subject	*R*	*A*	*N*
1	3(1)	5(2)	15(3)
2	1(1)	3(2)	18(3)
3	5(2)	4(1)	21(3)
4	2(1)	—	6(2)
5	0(1)	2(2)	17(3)
6	0(1)	2(2)	10(3)
7	0(1)	3(2)	8(3)
8	0(1)	2(2)	13(3)

Source: J. P. Brady (1969).

Ties

If there are ties among the X observations within any of the blocks, use average ranks to break the ties and compute the individual treatment weighted sums of centered ranks $A_1, \ldots, A_k$. In such cases, the significance level associated with procedure (7.53) is only approximately equal to α. (See Comment 66 for discussion of how to construct a conditionally distribution-free test of H_0 even when there are tied observations within some of the blocks.)

EXAMPLE 7.8 ***Effect of Rhythmicity of a Metronome on Speech Fluency.***

Consider Table 7.15, which contains a subset (with subject labels renumbered) of the data in Table 7.6 obtained by Brady (1969) in his study of the influence of rhythmicity of a metronome on the speech of stutterers, where the missing observation for subject 4 might be due to a malfunction in the arrhythmic metronome during her evaluation.

Here we have a two-way layout with $k = 3$, $n = 8$, and all $c_{ij's} = 1$ except $c_{42} = 0$. We illustrate the computations leading to the sample value of *SM*. The within-blocks (subjects) ranks (r_{ij}'s) for the observations present are also given in Table 7.15 in parentheses after the data values. With respect to the missing value for subject 4 under condition *A*, the average rank $(2 + 1)/2 = 1.5$ for subject 4 is assigned as the value for r_{42}. From (7.48), we compute the weighted sums of centered ranks to be

$$A_1 = \left\{ \left[\frac{12}{(3+1)} \right]^{1/2} [(1-2) + (1-2) + (2-2) + (1-2) + (1-2) + (1-2) + (1-2)] + \left[\frac{12}{(2+1)} \right]^{1/2} (1-1.5) \right\}$$

$$= 1.732(-6) + 2(-.5) = -11.392,$$

$$A_2 = \left\{ \left[\frac{12}{(3+1)} \right]^{1/2} [(2-2) + (2-2) + (1-2) + (2-2) + (2-2)

$$+ (2-2) + (2-2)] + \left[\frac{12}{(2+1)}\right]^{1/2} (1.5 - 1.5)\Bigg\}$$

$$= 1.732(-1) + 2(0) = -1.732,$$

and

$$A_3 = \Bigg\{\left[\frac{12}{(3+1)}\right]^{1/2} [(3-2) + (3-2) + (3-2) + (3-2) + (3-2)$$

$$+ (3-2) + (3-2)] + \left[\frac{12}{(2+1)}\right]^{1/2} (2 - 1.5)\Bigg\}$$

$$= 1.732(7) + 2(.5) = 13.124.$$

Thus, we obtain $\boldsymbol{A} = (-11.392, -1.732)$.

With the single missing observation for subject 4 under condition A, the combination counts λ_{qt} (7.51) are $\lambda_{12} = 7$, $\lambda_{13} = 8$, and $\lambda_{23} = 7$. Hence, from representation (7.50), the null covariance matrix Σ_0 has form

$$\Sigma_0 = \begin{bmatrix} 15 & -7 \\ -7 & 14 \end{bmatrix}.$$

Since each of λ_{12}, λ_{13}, and λ_{23} is positive, the rank of Σ_0 is 2, and its ordinary inverse is

$$\Sigma_0^{-1} = \left(\tfrac{1}{161}\right) \begin{bmatrix} 14 & 7 \\ 7 & 15 \end{bmatrix} = \begin{bmatrix} .0870 & .0435 \\ .0435 & .0932 \end{bmatrix}.$$

From (7.52), we obtain

$$\begin{aligned} SM &= \boldsymbol{A}\Sigma_0^{-1}\boldsymbol{A}' \\ &= (-11.392, -1.732) \begin{bmatrix} .0870 & .0435 \\ .0435 & .0932 \end{bmatrix} \begin{pmatrix} -11.392 \\ -1.732 \end{pmatrix} \\ &= 13.287. \end{aligned}$$

For a given arbitrary incomplete block design with incidence matrix **obs.mat**, the R command `cSkilMack`(α,**obs.mat**) can be used to find the available upper-tail critical values sm_α for possible values of SM. The incidence matrix is an $n \times k$ matrix of ones and zeroes, which indicate where the data are observed and unobserved, respectively. For a given available significance level α, the critical value sm_α then corresponds to $P_0(SM \geq sm_\alpha) = \alpha$ and is given by `cSkilMack`(α, **obs.mat**) $= sm_\alpha$. For this example,

we define the incidence matrix as

$$\texttt{obs.mat} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

and find `cSkilMack(.0097,obs.mat)` $= sm_{.0097} = 8.528$. Since we have observed $SM = 13.287 > sm_{.0097} = 8.528$. the P-value for this test procedure is smaller than .0097. (In fact, using `pSkilMack(x)` where **x** is the metronome data, we find that the P-value for this test is .00006.) Thus, there is strong evidence to support the hypothesis that the rhythmicity of a metronome does, indeed, influence the speech of stutterers (see also Comment 60 and Problem 76).

Comments

58. *More General Setting.* We could replace Assumptions A1–A3 and H_0 (7.2) with the more general null hypothesis that all possible $\prod_{i=1}^{n} s_i!$ rank configurations for the within-blocks ranks, r_{ij}, of the *observed* data are equally likely. Procedure (7.53) remains distribution-free for this more general hypothesis.

59. *Motivation for the Test.* Under Assumptions A1–A3 and H_0 (7.2), the block rank vector R_i^* for those s_i observations present in the ith block has a uniform distribution over the set of all $s_i!$ permutations of the vector of integers $(1, 2, \ldots, s_i)$, and this is true for all blocks, $i = 1, \ldots, n$. For those r_{ij} corresponding to observations present in the collected data set, it is then the case that $E_0(r_{ij}) = (s_i + 1)/2$, the average rank being assigned to the partial data present in the ith block, for every block $i = 1, \ldots, n$. Thus, it follows from (7.48) and the definition of r_{ij} for an empty cell that $E_0(A_j) = 0$ for every $j = 1, \ldots, k$. Therefore, we would expect each of the A_j's to be close to zero when H_0 is true. Since the test statistic SM (7.52) is a quadratic form in $A_1, \ldots, A_k$, small values of SM represent agreement with H_0 (7.2). When the τ's are not all equal, we would expect a portion of the A_j's to differ from their common null expectation of zero, with some tending to be positive and some tending to be negative. The net effect would be a large value of the quadratic form SM. This quite naturally suggests rejecting H_0 in favor of H_1 (7.3) for large values of SM and motivates procedures (7.53) and (7.54).

60. *Special Cases.* When the configuration of observed data in each of the blocks satisfies the constraints of a BIBD (see Section 7.6), the procedures in (7.53) and (7.54) are equivalent to the exact and large-sample approximation forms, respectively, of the Durbin–Skillings–Mack test procedure given in (7.44) and (7.45), respectively Section 7.6. Moreover, when we have an observation present in every treatment–block combination (i.e., we have a randomized complete block design), the Skillings–Mack procedures in (7.53) and (7.54) are equivalent to the exact and large-sample approximation forms, respectively, of the Friedman test procedure presented in (7.6) and (7.7), respectively, of Section 7.1. Thus,

the Skillings–Mack procedures in (7.53) and (7.54) represent natural extensions of the most commonly used nonparametric procedures for randomized complete block and balanced incomplete block designs to the setting of arbitrary incompleteness of the blocks.

We note that there is also an alternative closed-form expression for the test statistic *SM* (7.52) when only a single treatment has missing data. Without loss of generality, suppose that treatment k is missing observations in blocks $t+1$, $t+2,\ldots,n$ (i.e., treatment k has only t observations). Then, it can be shown (see Problem 76) that *SM* can be written as

$$SM = [t + (k-1)n]^{-1} \left\{ \sum_{j=1}^{k-1} A_j^2 + \left[\frac{nA_k^2}{t} \right] \right\}. \tag{7.55}$$

61. *Assumptions.* We emphasize that Assumption A3 stipulates that the *ns* cell distributions F_{ij} for those treatment–block combinations where observations are collected can differ at most in their locations (medians) and that these location differences (if any) must be a result of additive block and/or treatment effects (i.e., there is no interaction between the treatment and block factors). In particular, Assumption A3 requires that the *ns* underlying distributions belong to the same general family (F) and that they do not differ in scale parameters (variability). We do note, however, that the test procedure (7.53) remains distribution-free under the less restrictive setting where Assumption A3 is replaced by the weaker condition Assumption A3′ stated in Comment 43. Assumption A3 then corresponds to Assumption A3′ with the additional condition that $F_1 \equiv \cdots \equiv F_n$. (See also Comment 58.)

62. *Use of Generalized Inverse.* We noted in the body of the text that *any* generalized inverse $\sum_0^-$ can be used in the computation of *SM* (7.52). Skillings and Mack (1981) have shown that the value of *SM* is invariant with respect to the choice of generalized inverse, so that there is no ambiguity in the computation of *SM* and the associated test procedures (7.53) and (7.54) even if $\sum_0$ (7.50) is not of full rank. Of course, as we also noted previously, if $\lambda_{qt} > 0$ for all $q \neq t = 1,\ldots,k$, then the rank of the covariance matrix $\sum_0$ is, in fact, $k-1$ and we can simply use the ordinary inverse $\sum_0^{-1}$ in the definition of *SM*.

63. *Weighting Factor.* In the computation of the weighted sums of centered ranks $A_1,\ldots,A_k$, Skillings and Mack (1981) chose to weight the within-blocks centered ranks $[r_{ij} - (s_i+1)/2]$ by the factor $[12/(s_i+1)]^{1/2}$. They noted that this weighting factor has several advantages over other alternatives. First, it leads to a simple null covariance structure $\sum_0$ (7.50), which is useful for computational purposes. Second, because the range of the $[r_{ij} - (s_i+1)/2]$ values is less for blocks having fewer observations than for complete blocks, it is quite reasonable to adjust for this fact by using larger weights in those blocks with fewer observations. This has the effect of equalizing the contribution of each block when computing the A_j's. Prentice (1979) showed that a similar weighting scheme (using weights of $[s_i+1]^{-1}$) leads to increased power over use of the unweighted forms of the treatment sums. Skillings and Mack (1981) also note that use of the alternative simple weights s_i^{-1} would minimize the null variance of the weighted sums of centered ranks $A_1,\ldots,A_k$. However, the use of these simple weights would also alter the simplicity of the null covariance matrix $\sum_0$,

and computation of the associated test statistic would be much more difficult than is the case for *SM* (7.52). Other weighting schemes have been considered by Benard and van Elteren (1953) and Brunden and Mohberg (1976).

64. *Derivation of the Distribution of SM under H_0 (No-Ties Case).* The null distribution of *SM* (7.52) can be obtained by using the fact that under H_0 (7.2), all possible $s_1!s_2!\ldots s_k!$ rank configurations for the within-blocks ranks of the *observed* data are equally likely. We would simply compute the value of *SM* for each of these $s_1!s_2!\ldots s_k!$ block rank configurations and then tabulate the collective distribution of the values. Since the specifics of generating such a null distribution for *SM* are virtually identical with those for the Durbin–Skillings–Mack statistic *D* (7.43) for balanced incomplete block designs, the reader is referred to Comment 49 for illustration of the details of the process.

For a given arbitrary incomplete block design with incidence matrix **obs.mat**, the R command `cSkilMack(`α`, obs.mat)` can be used to find the available upper-tail critical values sm_α for possible values of *SM*. The incidence matrix is an $n \times k$ matrix of ones and zeroes, which indicate where the data are observed and unobserved, respectively. For a given available significance level α, the critical value sm_α then corresponds to $P_0(SM \geq sm_\alpha) = \alpha$ and is given by `cSkilMack(`α`, obs.mat)` = sm_α. Thus, for example, for $k = 3$, $n = 5$, all $c_{ij} = 1$ except for $c_{23} = c_{41} = c_{52} = 0$, the incidence matris is given by

$$\textbf{obs.mat} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

and we find `cSkilMack(.0208, obs.mat)` = $sm_{.0208}$ = 6.6347, so that $P_0(SM \geq 6.6347) = .0208$ for this missing data configuration.

65. *Large-Sample Approximation.* Let $\boldsymbol{A}$ be the vector defined in (7.49). Since each A_j is a weighted sum of centered ranks, it is not surprising (see Skillings and Mack (l981) for more details) that a properly standardized version of $\boldsymbol{A}$ has an asymptotic (n tending to infinity) $(k-1)$-variate normal distribution with mean vector $\mathbf{0} = (0, \ldots, 0)$ and covariance matrix $\sum_0$ (7.50) when the null hypothesis H_0 is true and $\lambda_{qt} > 0$, for every $q \neq t = 1, \ldots, k$. (Note once again that $\boldsymbol{A}$ does not include A_k, because A_k can be expressed as a weighted linear combination of $A_1, \ldots, A_{k-1}$. This is the reason that the asymptotic normal distribution is $(k-1)$-variate and not k-variate.) Since the test statistic *SM* (7.52) is a quadratic form in the variables $A_1, \ldots, A_{k-1}$, it is therefore quite natural that *SM* has an asymptotic (n tending to infinity) chi-square distribution with $k-1$ degrees of freedom when the null hypothesis H_0 is true and $\lambda_{qt} > 0$, for every $q \neq t = 1, \ldots, k$.

66. *Exact Conditional Null Distribution of SM with Ties among the Observed X Values within Blocks.* To have a test with exact significance level even in the presence of tied X's within some of the blocks, we need to consider all $s_1!s_2!\ldots s_k!$ block rank configurations for the *observed* data, where now these within-blocks ranks are obtained using average ranks to break ties. As in Comment 64, it still follows that under H_0 each of the $s_1!s_2!\ldots s_k!$ observed block rank configurations (now with these tied ranks) is equally likely. For

each such configuration, the value of SM (7.52) is computed and the results are tabulated. Since the specifics of generating such a conditional null distribution for SM in the case of tied within-blocks observations are virtually identical with those for the case of tied observations with the Durbin–Skillings–Mack statistic D (7.43) for balanced incomplete block designs, the reader is referred to Comment 51 for illustration of the details of the process.

67. *Settings Where the Chi-Square Approximation Is Not Applicable.* We note that the sole condition (other than the number of blocks becoming large) for the chi-square approximation to be applicable is that each of the λ_{qt}'s, $q \neq t = 1, \ldots, k$, must be positive. In settings where at least one of the λ_{qt}'s is zero, the approximate procedure (7.54) is not applicable. For such cases, one could still use procedure (7.53) by generating the (exact or simulated) null distribution of SM (7.52) and obtaining the appropriate critical values. On the other hand, if $\lambda_{qt} = 0$ for a particular pair of treatments q and t (so that q and t never appear together in a block), then procedure (7.53) would not necessarily be effective in testing H_0 (7.2) even when τ_q and τ_t are quite different.

68. *Historical Development.* The test procedures (7.53) and (7.54) were proposed and studied by Skillings and Mack (1981). They provided some exact null distribution critical values for the special case of BIBDs (see Section 7.6) and for a second special case where we are only a single observation short of having complete block data.

Properties

1. *Asymptotic Chi-Squaredness.* See Skillings and Mack (1981).

Problems

74. In the data on percent reduction in average wind speed due to shelterbelts discussed in Problem 19, the month of November was omitted from the data in Table 7.7 because the percent reduction observation at 20 m was missing. In Table 7.16, we again present these data with the month of November included.

Table 7.16 Percent Reduction in Average Wind Speed at Dambatta, 1980/81

	Leeward Distance from Shelterbelt m				
Month	20	40	100	150	200
January	22.1	20.7	15.4	12.3	6.9
February	19.2	18.7	14.9	9.3	6.5
March	21.5	21.9	14.3	9.9	7.1
April	21.5	21.2	11.1	9.4	6.2
May	21.3	20.9	11.2	9.4	7.7
June	20.9	19.6	16.9	11.6	7.0
August	19.3	18.7	14.4	12.5	7.0
September	20.1	19.6	15.6	12.6	7.5
October	23.7	20.4	14.6	12.4	8.5
November	19.5	18.4	13.8	8.4	—

Source: J. E. Ujah and K. B. Adeoye (1984).

Table 7.17 Assembly Times (min)

	Assembly Methods			
Workers	A	B	C	D
1	3.2	4.1	3.8	4.2
2	3.1	3.9	3.4	4.0
3	4.3	3.5	4.6	4.8
4	3.5	3.6	3.9	4.0
5	3.6	4.2	3.7	3.9
6	4.5	4.7	3.7	—
7	—	4.2	3.4	—
8	4.3	4.6	4.4	4.9
9	3.5	—	3.7	3.9

Source: J. H. Skillings and G. A. Mack (1981).

Use the Skillings–Mack procedure (7.53) to test the hypothesis that there is a difference in percent reduction in average wind speed over the five leeward distances from a shelterbelt. (Note that the conclusion of the Skillings–Mack procedure applied to these data is not directional, as was the case with the decision in Problem 19 using the Page ordered alternatives procedure and the randomized complete design data without the month of November. A corresponding ordered alternatives analog to the Skillings–Mack procedure for arbitrary missing data is not available in the literature.)

75. Skillings and Mack (1981) consider the experiment evaluating four methods of assembling a product, where the blocking factor corresponds to the individual assembly workers. The data for this experiment are presented in Table 7.17, where the observations are assembly times in minutes. The missing observations are due to machinery breakdowns or employee absenteeism. Use these data to assess whether there are any differences among the assembly methods with regard to median time for assembly of the product.

76. Verify that expression (7.55) is an alternative way to compute the statistic SM (7.52) when only a single treatment has missing data.

77. Consider Table 7.18, which gives a subset of the Rounding-first-base data in Table 7.1 obtained by Woodward (1970) in his study of the best method of rounding first base to minimize the time to second base. (The missing observations might be due to injury during one of the other runs.) Use these data to assess whether there are any differences among median times to second base for these three ways of rounding first base.

78. We noted that the value of the test statistic SM (7.52) does not depend on the form of the particular generalized inverse $\sum_0^-$ used in the calculation. Illustrate this fact by computing SM using two different generalized inverses for a setting where the rank of $\sum_0$ is not $k - 1$.

79. Consider Table 7.19, which gives a subset of the serum CPK activity data in Table 7.3 obtained by Goode and Meltzer (1976) in their study of the effect of isometric exercise on serum CPK levels. Use these data to assess whether there are any differences in median serum CPK activity (in mU/l) for the three measurement periods.

80. Verify for the data in Table 7.15 of Example 7.8 that the value of $SM = 13.287$ would also be obtained if we take $\boldsymbol{A}$ to be (A_1, A_3) or (A_2, A_3) and make the corresponding changes in the definition of the null covariance matrix $\sum_0$ (7.50).

81. Verify that the Skillings–Mack statistic SM (7.52) simplifies to the closed-form expression for the Friedman statistic S (7.5) when we have data from a randomized complete block design.

82. Verify that the Skillings–Mack statistic SM (7.52) simplifies to the closed-form expression for the Durbin–Skillings–Mack statistic D (7.43) when we have data from a balanced incomplete block design.

Table 7.18 Rounding-First-Base Times

Players	Methods		
	Round out	Narrow angle	Wide angle
1	5.40	5.50	5.55
2	5.85	5.70	5.75
3	5.20	5.60	5.50
4	5.55	5.50	—
5	5.90	5.85	5.70
6	5.45	5.55	5.60
7	5.40	5.40	5.35
8	—	5.50	5.35
9	5.25	5.15	5.00
10	5.85	—	5.70
11	5.25	5.20	5.10
12	5.65	5.55	—
13	5.60	5.35	5.45
14	5.05	—	4.95
15	5.50	5.50	5.40
16	—	5.55	5.50
17	5.55	5.55	—
18	5.45	5.50	5.55
19	5.50	5.45	5.25
20	5.65	5.60	5.40
21	5.70	5.65	5.55
22	6.30	6.30	6.25

Source: W. F. Woodward (1970).

Table 7.19 Effect of Isometric Exercise on Serum Creatine Phosphokinase (CPK) Activity (mU/l) in Psychotic Patients

Subject	Preexercise	19-h Postexercise	42-h Postexercise
1	27	101	82
2	30	112	50
3	24	26	68
4	54	89	—
5	21	30	49
6	36	41	48
7	36	29	46
8	16	20	8
9	21	26	25

Source: D. J. Goode and H. Y. Meltzer (1976).

83. Consider the setting corresponding to $k = 4$ and $n = 10$ where we have only a single missing observation in one of the blocks. Compare the critical region for the exact level $\alpha = .0501$ test of H_0 (7.2) based on SM with the critical region for the corresponding nominal level $\alpha = .0501$ test based on the large-sample approximation.

84. Consider the setting corresponding to $k = 6$ and $n = 5$ where we have only a single missing observation in one of the blocks. Compare the critical region for the exact level $\alpha = .0499$ test of H_0 (7.2) based on SM with the critical region for the corresponding nominal level $\alpha = .0499$ test based on the large-sample approximation.

85. Consider the incomplete block data setting corresponding to $k = 3$, $n = 3$, $s_1 = s_2 = 3$, and $s_3 = 2$. Obtain the form of the exact null H_0 distribution of SM (7.52) for the case of no-tied observations.

86. Consider the incomplete block data setting corresponding to $k = 3$, $n = 3$, $s_1 = s_2 = 3$, and $s_3 = 2$. Suppose the three observations in block 2 are all tied at a single value, but there are no tied observations in any of the other blocks. Obtain the conditional exact probability distribution of SM (7.52) under H_0 (7.2) when average ranks are used to break this set of within-block ties. Compare this conditional null distribution of SM with the null distribution of SM obtained in Problem 85 when there are no ties.

REPLICATIONS—TWO-WAY LAYOUT WITH AT LEAST ONE OBSERVATION FOR EVERY TREATMENT–BLOCK COMBINATION

It is often the case in two-way layout settings that we have more than one observation for some of the treatment–block combinations. These multiple observations in a given cell are referred to as replications for that treatment–block combination. Of course, permitting such replications opens the possibility of a much wider variety of data configurations for our two-way layout. There could be some cells with no observations, some with one observation, and some with multiple observations. In the next two sections we emphasize nonparametric procedures for general alternatives in the setting where we have a common, equal number $c > 1$ of replications for every treatment–block combination. Direct extensions of these general alternatives procedures to less restrictive settings where the number of replications need not be equal but there are no empty cells are discussed in Comment 77. Nonparametric procedures that are valid for the most general two-way layout settings where there may be a mix of cells with more than one observation (i.e., replications), cells with a single observation, and empty cells with no observations are discussed in Comment 78 for the cases of general and ordered alternatives.

In Section 7.9 we present a distribution-free hypothesis test for general alternatives when we have an equal number (>1) of replications for every treatment–block combination. In Section 7.10 we discuss an all-treatments multiple comparison procedure for the same setting.

Throughout these two sections, we continue to operate under the general conditions of Assumptions A1–A3. However, in Sections 7.9 and 7.10, we impose the additional constraint that each c_{ij} is equal to $c(>1)$ and, thus, that

$$N = \sum_{i=1}^{n} \sum_{j=1}^{k} c_{ij} = nkc.$$

7.9 A DISTRIBUTION-FREE TEST FOR GENERAL ALTERNATIVES IN A RANDOMIZED BLOCK DESIGN WITH AN EQUAL NUMBER $C(>1)$ OF REPLICATIONS PER TREATMENT–BLOCK COMBINATION (MACK–SKILLINGS)

In this section we present a procedure for testing H_0 (7.2) against the general alternatives H_1 (7.3) for block data where we have an equal number $c > 1$ replications for each of the treatment–block combinations. Here, the total number of observations is $N = nkc$.

Procedure

To compute the Mack–Skillings statistic for this equal replications setting, we first rank the observations from least to greatest separately within each of the n blocks. Let r_{ijq} be the within-block rank of X_{ijq} (the qth replication from the jth treatment in the ith block) among the kc total observations present in the ith block, for $i = 1, \ldots, n$. Set

$$S_j = \sum_{i=1}^{n} \left[\sum_{q=1}^{c} r_{ijq}/c \right], \quad \text{for } j = 1, \ldots, k. \tag{7.56}$$

Thus, S_j is the sum (across blocks) of the cellwise averages of the within-blocks ranks assigned to the c observations from treatment j, for $j = 1, \ldots, k$. The Mack–Skillings statistic for equal replications is then given by

$$\begin{aligned} MS &= \left[\frac{12}{k(N+n)} \right] \sum_{j=1}^{k} \left[S_j - \frac{N+n}{2} \right]^2, \\ &= \left[\frac{12}{k(N+n)} \right] \left\{ \sum_{j=1}^{k} S_j^2 \right\} - 3(N+n), \end{aligned} \tag{7.57}$$

where $(N+n)/2n = (kc+1)/2 = \sum_{j=1}^{k} \sum_{q=1}^{c} r_{ijq}/kc$ is the average within-blocks rank assigned for each of the n blocks. It follows that $n(N+n)/2n = (N+n)/2$ is the expected sum (across blocks) of the cellwise averages for each of the k treatments when H_0 (7.2) is true; that is, $(N+n)/2$ is the expected value of S_j, for each $j = 1, \ldots, k$, when the null hypothesis H_0 is true.

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the general alternative

$$H_1 : [\tau_1, \ldots, \tau_k \text{ not all equal}],$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } MS \geq ms_\alpha; \quad \text{otherwise do not reject,} \tag{7.58}$$

where the constant ms_α is chosen to make the type I error probability equal to α. The constant ms_α is the upper α percentile for the null $(\tau_1 = \cdots = \tau_k)$ distribution of MS. Comment 73 explains how to obtain the critical values ms_α for k treatments, n blocks, c replications for each treatment–block combination, and available values of α.

Large-Sample Approximation

When H_0 (7.2) is true, the statistic MS has, as the common number of observations on each treatment, nc, tends to infinity, an asymptotic chi-square (χ^2) distribution with $k-1$ degrees of freedom. (See Comment 74 for indications of the proof.) The chi-square

approximation for procedure (7.58) is

$$\text{Reject } H_0 \text{ if } MS \geq \chi^2_{k-1,\alpha}; \quad \text{otherwise do not reject,} \tag{7.59}$$

where $\chi^2_{k-1,\alpha}$ is the upper α percentile of a chi-square distribution with $k-1$ degrees of freedom. To find $\chi^2_{k-1,\alpha}$, we use the R command `qchisq`$(1-\alpha, k-1)$. For example, to find $\chi^2_{6,.025}$, we apply `qchisq(.975, 6)` and obtain $\chi^2_{6,.025} = 14.45$.

Mack and Skillings (1980) have pointed out that this chi-square approximation is adequate when the significance level α is at least .05 and the number of replications c is at least 4, even though it is slightly conservative when the level nears .05. However, for significance levels as low as .01, they note that the conservative nature of the approximate procedure (7.59) can be somewhat severe unless the common number of replications c is rather large. Whenever possible, they recommend the use of the exact procedure (7.58) for such small significance levels.

Ties

If there are ties among the X observations within any of the blocks, use average ranks to break the ties and compute the individual sums of cellwise averages of the within-blocks ranks $S_1, \ldots, S_k$. In such cases, the significance level associated with procedure (7.58) is only approximately equal to α. (See Comment 75 for discussion of how to construct an exact conditionally distribution-free test of H_0 even when there are tied observations within some of the blocks.)

EXAMPLE 7.9 ***Determination of Niacin in Bran Flakes.***

In a study to investigate the precision and homogeneity of a procedure for assessing the amount of niacin in bran flakes, Campbell and Pelletier (1962) prepared homogenized samples of bran flakes enriched with 0, 4, or 8 mg niacin per 100 g of cereal. Portions of the homogenized samples were sent to different laboratories, which were asked to carry

Table 7.20 Amount of Niacin in Enriched Bran Flakes

	Amount of niacin enrichment (milligrams per 100 g bran flakes)		
Laboratory	0	4	8
1	7.58 (3)	11.63 (7)	15.00 (2)
	7.87 (8)	11.87 (11)	15.92 (9)
	7.71 (6)	11.40 (3)	15.58 (4)
2	8.00 (9.5)	12.20 (12)	16.60 (12)
	8.27 (12)	11.70 (8.5)	16.40 (11)
	8.00 (9.5)	11.80 (10)	15.90 (7)
3	7.60 (4)	11.04 (2)	15.87 (6)
	7.30 (1)	11.50 (5.5)	15.91 (8)
	7.82 (7)	11.49 (4)	16.28 (10)
4	8.03 (11)	11.50 (5.5)	15.10 (3)
	7.35 (2)	10.10 (1)	14.80 (1)
	7.66 (5)	11.70 (8.5)	15.70 (5)

Source: J. A. Campbell and O. Pelletier (1962).

out the specified procedure for each of three separate samples. The resulting data (in milligrams per 100 g bran flakes) for a subset (4 out of 12) of the laboratories included in the study are presented in Table 7.20.

Of primary interest here is the precision of the laboratory procedure for determining niacin content in bran flakes. The actual amount of niacin enrichment in the prepared bran flakes serves only as a "nuisance" blocking factor in our evaluation of the consistency of the results across the four laboratories for which data are included in Table 7.20. Hence, we have data from a two-way layout with $k = 4$ treatments (laboratories), $n = 3$ blocks (amounts of niacin enrichment), and $c = 3$ replications (individual bran flake samples) per laboratory/enrichment combination. For the purpose of illustration, we consider the significance level $\alpha = .0501$. Applying the `R` command `cMackSkil(`α`,k,n,c)` with $k = 4$, $n = 3$, and $c = 3$, we see that `cMackSkil(.0501,4,3,3)` $= ms_{.0501} = 7.479$ and procedure (7.58) becomes

$$\text{Reject } H_0 \text{ if } MS \geq 7.479.$$

Now, we illustrate the computations leading to the sample value of MS. The numbers in parentheses after the data values in Table 7.20 are the within-enrichment-levels (i.e., blocks) ranks (using average ranks to break ties) of the niacin content measurements obtained from the four laboratories. Using these block ranks, we obtain the following sums of cellwise averages for the four laboratories:

$$S_1 = \frac{3+8+6+7+11+3+2+9+4}{3} = 17.67,$$

$$S_2 = \frac{9.5+12+9.5+12+8.5+10+12+11+7}{3} = 30.5,$$

$$S_3 = \frac{4+1+7+2+5.5+4+6+8+10}{3} = 15.83,$$

and

$$S_4 = \frac{11+2+5+5.5+1+8.5+3+1+5}{3} = 14.$$

Hence, with $k = 4$, $n = 3$, and $N = 36$, we find from (7.57) that

$$MS = \left[\frac{12}{4(36+3)}\right]\{(17.67)^2 + (30.5)^2 + (15.83)^2 + (14)^2\} - 3(36+3)$$

$$= \left[\frac{1}{13}\right]\{312.23 + 930.25 + 250.59 + 196\} - 117 = 12.93.$$

Since the observed value of MS is greater than the critical value 7.479, we can reject H_0 at the $\alpha = .0501$ level, providing rather strong evidence that the studied process for assessing niacin content in bran flakes does not produce consistent results across a variety of laboratories and is therefore not reliable as an evaluative procedure. In fact, from the observed value of $MS = 12.93$, we can use the `R` command `pMackSkil(`**niacin**`)` to find that $P_0(MS \geq 12.93) =$ `pMackSkil(`**niacin**`)` $= .0023$. Thus, the smallest significance level at which we can reject H_0 in favor of H_1 with the observed value of the test statistic $MS = 12.93$ is .0023.

We should note in passing that there also appears to be an even more basic problem with this studied procedure for assessing niacin content in bran flakes and that is

the accuracy (in addition to the lack of reproducibility) of the numerical values of the measurements. For example, for those samples enriched with 4 mg niacin per 100 g bran flakes, the values obtained by applying this procedure to the sample bran flakes ranged from 10.10 to 12.20 mg per 100 g bran flakes, well over the preestablished niacin content. (Similar comments apply to the 0- and 8-mg enrichment samples.) This clearly indicates a rather severe basic calibration problem with the assessment procedure, in addition to the lack of portability across laboratories detected by our application of the Mack–Skillings procedure to these data.

Comments

69. *More General Setting.* We could replace Assumptions A1–A3 and H_0 (7.2) with the more general null hypothesis that all possible $[(ck)!]^n$ configurations for the permutations of the within-blocks ranks (r_{ijq}'s) are equally likely. Procedure (7.58) remains distribution-free for this more general null hypothesis.

70. *Motivation for the Test.* Under Assumptions A1–A3 and H_0 (7.2), each of the block rank vectors $\boldsymbol{R}_i^* = (r_{i11}, \ldots, r_{i1c}, r_{i21}, \ldots, r_{i2c}, \ldots, r_{ik1}, \ldots, r_{ikc})$, $i = 1, \ldots, n$, has a uniform distribution over the set of all $(ck)!$ permutations of the vector of integers $(1, 2, \ldots, ck)$ and this is true, independently, for each of the n blocks. It is then the case that $E_0(r_{ijq}) = (ck + 1)/2$ for every $i = 1, \ldots, n$; $j = 1, \ldots, k$; and $q = 1, \ldots, c$. It follows from (7.56) that $E_0(S_j) = nc(ck + 1)/2c = (nck + n)/2 = (N + n)/2$. Since the test statistic MS (7.57) is a constant times a sum of squared differences between the observed treatment sums of cellwise average ranks, S_j, and their common null expected value, $E_0(S_j) = (N + n)/2$, small values of MS represent agreement with H_0 (7.2). When the τ's are not all equal, we would expect a portion of the associated treatment sums of cellwise average ranks, S_j, to differ from their common null expectation, $(N + n)/2$, with some tending to be smaller and some larger. The net result (after squaring the observed differences to obtain the $[S_j - (N + n)/2]^2$ terms) would be a large value of MS. This quite naturally suggests rejecting H_0 in favor of H_1 (7.3) for larger values of MS and motivates procedures (7.58) and (7.59).

71. *Special Case of $c = 1$.* When we have a single observation for every treatment–block combination (i.e., $c = 1$), we are dealing with data from a complete randomized block design. In this setting, the Mack–Skillings statistic MS (7.57) is equivalent to the Friedman statistic S (7.5). Thus, the Mack–Skillings procedures (7.58) and (7.59) represent natural extensions of the Friedman procedures (7.6) and (7.7), respectively, to the case of an equal number $c > 1$ of replications per cell.

72. *Assumptions.* We emphasize that Assumption A3 stipulates that the nk cell distributions F_{ij} can differ at most in their locations (medians) and that these location differences (if any) must be a result of additive block and/or treatment effects (i.e., there is no interaction between the treatment and block factors). In particular, Assumption A3 requires that the ns underlying distributions belong to the same general family (F) and that they do not differ in scale parameters (variability). We do note, however, that the test procedure (7.58) remains distribution-free under the less restrictive setting where Assumption A3 is replaced by the weaker condition Assumption A3′ stated in Comment 43.

Assumption A3 then corresponds to Assumption A3′ with the additional condition that $F_1 \equiv \cdots \equiv F_n$. (Also see Comment 69.)

73. *Derivation of the Distribution of MS under H_0 (No-Ties Case).* The null distribution of *MS* (7.57) can be obtained by using the fact that under H_0 (7.2), all possible $[(ck)!]^n$ configurations for the permutations of the within-blocks ranks (r_{ijq}'s) are equally likely. Thus, to obtain the exact null distribution of *MS*, we compute its value for each of these $[(ck)!]^n$ block rank configurations and then tabulate the collected outcomes. We must point out, of course, that the number $[(ck)!]^n$ of configurations for which we need to compute the value of *MS* can get large rather quickly, as either k or c is moderately increased. Since the specifics of generating such a null distribution for *MS* are virtually identical with those for the Durbin–Skillings–Mack statistic *D* (7.43) for balanced incomplete block designs, the reader is referred to Comment 49 for illustration of the details of the process.

    For a given number of treatments k, blocks n, and c replications for each treatment–block combination, the `R` command `cMackSkil(`α`,k,n,c)` can be used to find the available upper-tail critical values ms_α for possible values of *MS*. For a given available significance level α, the critical value ms_α then corresponds to $P_0(MS \geq ms_\alpha) = \alpha$ and is given by `cMackSkil(`α`,k,n,c)` $= ms_\alpha$. Thus, for example, for $k = 4$, $n = 4$, and $c = 3$, we have $P_0(MS \geq 7.667) = .0502$, so that $ms_{.0502}$ = `cMackSkil(.0502,4,4,3)` $= 7.667$ for $k = 4$, $n = 4$, and $c = 3$.

74. *Large-Sample Approximation.* Define the centered treatment sums of cellwise average ranks $S_j^* = S_j - E_0(S_j) = S_j - (N + n)/2$, for $j = 1, \ldots, k$, and set $\boldsymbol{S}^* = (S_1^*, \ldots, S_{k-1}^*)$. Since each S_j is an average, it is not surprising (see Mack and Skillings (1980) for more details) that a properly standardized version of $\boldsymbol{S}^*$ has an asymptotic (*nc* tending to infinity) $(k-1)$-variate normal distribution with mean vector $\mathbf{0} = (0, \ldots, 0)$ and appropriate covariance matrix Σ^* when the null hypothesis H_0 is true. (Note that $\boldsymbol{S}^*$ does not include S_k^*, because S_k^* can be expressed as a weighted linear combination of $S_1^*, \ldots, S_{k-1}^*$. This is the reason that the asymptotic normal distribution is $(k-1)$-variate and not k-variate.) Since the test statistic *MS* (7.57) is a quadratic form in the variables (S_1^*, …, S_{k-1}^*), it is therefore quite natural that *MS* has an asymptotic (*nc* tending to infinity) chi-square distribution with $k - 1$ degrees of freedom when the null hypothesis H_0 is true.

75. *Exact Conditional Null Distribution of MS with Ties among the X Values within Blocks.* To have a test with exact significance level even in the presence of tied *X*'s within some of the blocks, we need to consider all $[(ck)!]^n$ block rank configurations for the observed data, where now these within-blocks ranks are obtained using average ranks to break ties. As in Comment 73, it still follows that under H_0 each of the $[(ck)!]^n$ observed block rank configurations (now with these tied ranks) is equally likely. For each such configuration, the value of *MS* (7.57) is computed and the results are tabulated. Since the specifics of generating such a conditional null distribution for *MS* in the case of tied within-blocks observations are virtually identical with those for the case of tied observations with the Durbin–Skillings–Mack statistic *D* (7.43) for balanced incomplete block designs, the reader is referred to Comment 51 for illustration of the details of the process.

76. *Simple Competitor Procedure when the Number of Replications Is the Same for Every Cell.* As an alternative to the Mack–Skillings procedure (7.58), we could first compute the median of the c replications separately in each of the nk cells and then apply either the Friedman procedure (7.6) or the Page procedure (7.11), whichever is appropriate for the alternatives of interest, to these nk cell medians (which now represent data from a complete randomized block design). In general, this approach could result in substantial loss of information, especially when the number of replications per cell, c, is large. However, it is simple and does provide the only available nonparametric procedure for dealing specifically with ordered alternatives when we have an equal number (>1) of replications per cell. (We note, in passing, that any appropriate measure of central tendency, such as the cell means or the medians of the Walsh averages (see Comment 3.17), for the individual cell data, could be used instead of the cell medians to summarize the data prior to application of the Friedman or the Page procedure.)

77. *Extension to Arbitrary Replication* (≥ 1) *Configurations.* We have described the Mack–Skillings procedure in detail for the setting where we have the same number of replications c (≥ 1) for each of the treatment–block combinations. However, in their original work, Mack and Skillings (1980) proposed a more general test procedure that is appropriate for any two-way layout setting for which we have at least one replication for every treatment–block combination (i.e., there are no empty cells). We now present their procedure for this more general setting where the only stipulation is that $c_{ij} > 0$ for every $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

    For $i = 1, \ldots, n$, let $q_i = \sum_{j=1}^{k} c_{ij}$ be the total number of observations present in the ith block. Once again we rank the observations from least to greatest within each of the blocks and let r_{iju} denote the rank of X_{iju} within the q_i observations present in the ith block, for $u = 1, \ldots, c_{ij}$; $i = 1, \ldots, n$; and $j = 1, \ldots, k$. For each treatment, compute the sum of cellwise weighted average ranks

$$V_j = \sum_{i=1}^{n} \sum_{u=1}^{c_{ij}} \frac{r_{iju}}{q_i}, \quad j = 1, \ldots, k. \tag{7.60}$$

    Define the vector

$$\begin{aligned} \boldsymbol{V} &= (V_1 - E_0[V_1], \ldots, V_{k-1} - E_0[V_{k-1}]) \\ &= \left(V_1 - \sum_{i=1}^{n} \left[\frac{c_{i1}(q_i + 1)}{2q_i} \right], \ldots, V_{k-1} - \sum_{i=1}^{n} \left[\frac{c_{i,k-1}(q_i + 1)}{2q_i} \right] \right). \end{aligned} \tag{7.61}$$

    Thus, the components of $\boldsymbol{V}$ are the sums of cellwise weighted average ranks centered about their expected values under H_0. (Without the loss of generality, we have chosen to omit the centered V_k from the vector $\boldsymbol{V}$. The V_j's are linearly dependent, because a weighted linear combination of all k of them is a constant. We could omit any one of the V_j's in the definition of $\boldsymbol{V}$ and the procedure we now describe would lead to the same value of the test statistic. For further discussion, see Mack and Skillings (1980).)

The covariance matrix for $\boldsymbol{V}$ under H_0 (7.2) has the form $\sum_{\boldsymbol{V},\mathbf{0}} = ((\sigma_{s,t}))$, where

$$\sigma_{s,t} = \sum_{i=1}^{n} \left[\frac{c_{is}(q_i - c_{is})(q_i + 1)}{12q_i^2} \right], \quad \text{for } s = t = 1, \ldots, k-1$$

$$= -\sum_{i=1}^{n} \left[\frac{c_{is}c_{it}(q_i + 1)}{12q_i^2} \right], \quad \text{for } s \neq t = 1, \ldots, k-1. \tag{7.62}$$

The rank of the matrix $\boldsymbol{\Sigma}_{\boldsymbol{V},\mathbf{0}}$ is $k-1$. Letting $\boldsymbol{\Sigma}_{\boldsymbol{V},\mathbf{0}}^{-1}$ denote the inverse of $\boldsymbol{\Sigma}_{\boldsymbol{V},\mathbf{0}}$, the Mack–Skillings test statistic for this general setting of unequal, but positive, numbers of replications in the treatment–block combinations, is given by

$$MS_g = \boldsymbol{V}\,\boldsymbol{\Sigma}_{\boldsymbol{V},\mathbf{0}}^{-1}\boldsymbol{V}'. \tag{7.63}$$

To test

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the general alternative

$$H_1 : [\tau_1, \tau_2, \ldots, \tau_k \text{ not all equal}],$$

at the α level of significance, the Mack–Skillings general procedure is then to

$$\text{Reject } H_0 \text{ if } MS_g \geq ms_{g,\alpha}; \quad \text{otherwise do not reject,} \tag{7.64}$$

where the constant $ms_{g,\alpha}$ is chosen to make the type I error probability equal to α.

The critical values $ms_{g,\alpha}$ are available in the literature only for the setting where we have an equal number, c, of replications in each cell, in which case the general Mack–Skillings test procedure (7.64) is equivalent to the equal replications version given in (7.58). However, when H_0 (7.2) is true, the general form statistic MS_g, has, as N tends to infinity in such a way that c_{ij}/N tends to $\rho_{ij} > 0$ for every $i = 1, \ldots, n$ and $j = 1, \ldots, k$, an asymptotic chi-square (χ^2) distribution with $k-1$ degrees of freedom. Thus, when N is large, the chi-square approximation for the general Mack–Skillings procedure (7.64) is

$$\text{Reject } H_0 \text{ if } MS_g \geq \chi^2_{k-1,\alpha}; \quad \text{otherwise do not reject,} \tag{7.65}$$

where $\chi^2_{k-1,\alpha}$ is the upper α percentile of a chi-square distribution with $k-1$ degrees of freedom.

78. *Competitor Procedures Applicable for Most General Two-Way Layout Settings Where There Are Both Replications and Empty Cells.* Thus far in this chapter we have discussed procedures that are appropriate either for settings where we have 0 or 1 observation for every treatment–block combination or for settings where we have at least one observation in every cell. None of these procedures are appropriate for the most general settings that represent a combination of these two structures, namely, those data sets where we have replications ($c_{ij} > 1$) for

some treatment–block combinations and no observations ($c_{ij} = 0$) for others. We briefly discuss now two test procedures for such general two-way layout settings, one designed for general alternatives to H_0 (7.2) and the second specifically oriented toward detecting ordered alternatives.

Let k_i be the number of treatments in the ith block for which $c_{ij} > 0$, for $i = 1, \ldots, n$. (Once again, we discard any block i for which $k_i = 1$, as such a block contains no information relative to possible differences in the treatment effects. Notationally, then, n represents the number of blocks remaining after discarding blocks with observations on only a single treatment.)

General Alternatives. We first compute the one-way layout Kruskal–Wallis statistic H (6.5) separately in each of the n blocks. Letting H_i denote this Kruskal–Wallis statistic for the ith block, $i = 1, \ldots, n$, the statistic considered by Mack (1981) for this most general two-way layout setting is given by

$$H_{\text{tot}} = \sum_{i=1}^{n} H_i. \tag{7.66}$$

The level α test of H_0 (7.2) versus the general alternatives H_1 (7.3) studied by Mack (1981) is

$$\text{Reject } H_0 \text{ if } H_{\text{tot}} \geq h_\alpha^*; \quad \text{otherwise do not reject,} \tag{7.67}$$

where the constant h_α^* is chosen to make the type I error probability equal to α. Values of h_α^* are available in Mack (1981) for $k = 3$, $n = 2$, 3, and all combinations of replications $0 \leq c_{ij} \leq 3$, as well as for $k = 3, n = 4$, 5, and all combinations of replications $0 \leq c_{ij} \leq 2$. Additional values of h_α^* can be found in DeKroon and Van der Laan (1981) for $\alpha = .01$, .05, and various combinations of k, n, and equal number of replications c in the ranges $2 \leq k \leq 4$, $1 \leq n \leq 10$, and $2 \leq c \leq 4$. (We note, in passing, that procedure (7.67) can be particularly sensitive to a large degree of interaction between the treatment and the block factors. In the presence of such extensive interaction, it is possible that a rejection of H_0 with procedure (7.67) could be a direct consequence of this interaction, rather than because of any significant differences in the treatment effects $\tau_1, \ldots, \tau_k$.)

When H_0 (7.2) is true, the statistic H_{tot} has, as min (nonzero c_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, k$) tends to infinity, an asymptotic chi-square (χ^2) distribution with $d = (k_1 + k_2 + \cdots + k_n - n)$ degrees of freedom (see Mack (1981) for details). Thus, when the minimum nonzero c_{ij} is large, the chi-square approximation for procedure (7.67) is

$$\text{Reject } H_0 \text{ if } H_{\text{tot}} \geq \chi^2_{d,\alpha}; \quad \text{otherwise do not reject,} \tag{7.68}$$

where $\chi^2_{d,\alpha}$ is the upper α percentile point of a chi-square distribution with d degrees of freedom.

Ordered Alternatives. If we are interested in ordered alternatives, H_2 (7.9), we first compute the one-way layout Jonckheere–Terpstra statistic J (6.13) separately in each of the n blocks. Letting J_i denote this Jonckheere–Terpstra statistic for the ith block, $i = 1, \ldots, n$, the statistic proposed by Skillings and

Wolfe (1977, 1978) for this most general two-way layout ordered alternatives setting is given by

$$J_{\text{tot}} = \sum_{i=1}^{n} J_i. \tag{7.69}$$

The level α test of H_0 (7.2) versus the ordered alternatives H_2 (7.9) suggested by Skillings and Wolfe (1977, 1978) is

$$\text{Reject } H_0 \text{ if } J_{\text{tot}} \geq j_\alpha^*; \quad \text{otherwise do not reject,} \tag{7.70}$$

where the constant j_α^* is chosen to make the type I error probability equal to α. Values of j_α^* are available in Skillings (1980) for $k = 2(1)6$, $n = 2(1)5$ and selected configurations of the c_{ij}'s such that $c_{ij} = C_i$, for $i = 1, \ldots, n$ and $j = 1, \ldots, k$ (i.e., within a given block, each treatment has the same number of observations C_i, but C_1, C_2, ..., C_n need not all be equal). (We note that procedure (7.70) does not have the same sensitivity to the presence of extensive interaction as does the general alternatives procedure (7.67). Rejection of H_0 with procedure (7.70) will always be indicative of the presence of an ordered structure on the treatment effects $\tau_1, \ldots, \tau_k$.)

When H_0 (7.2) is true, the standardized form

$$J_{\text{tot}}^* = \frac{J_{\text{tot}} - E_0(J_{\text{tot}})}{[\text{var}_0(J_{\text{tot}})]^{1/2}} \tag{7.71}$$

has, as min (nonzero $c_{ij}, i = 1, \ldots, n; j = 1, \ldots, k$) tends to infinity, an asymptotic $N(0, 1)$ distribution (see Skillings and Wolfe (1977, 1978) for details), where

$$E_0(J_{\text{tot}}) = \frac{\sum_{i=1}^{n} \left[q_i^2 - \sum_{j=1}^{k} c_{ij}^2 \right]}{4} \tag{7.72}$$

and

$$\text{var}_0(J_{\text{tot}}) = \frac{\sum_{i=1}^{n} \left[q_i^2(2q_i + 3) - \sum_{j=1}^{k} c_{ij}^2(2c_{ij} + 3) \right]}{72}, \tag{7.73}$$

are the expected value and variance, respectively, of J_{tot} (7.69) under the null hypothesis H_0 and $q_i = c_{i1} + \cdots + c_{ik}$ is the total number of observations present in the ith block, $i = 1, \ldots, n$. Thus, when the minimum nonzero c_{ij} is large, the normal theory approximation for procedure (7.70) is

$$\text{Reject } H_0 \text{ if } J_{\text{tot}}^* \geq z_\alpha; \quad \text{otherwise do not reject.} \tag{7.74}$$

79. *Historical Development.* Mack and Skillings (1980) proposed and studied a general test procedure for an arbitrary two-way layout setting where we have at least one observation for every treatment–block combination (see Comment 77). For the special case of an equal number of replications, c, in every cell, their general test procedure simplifies to the expression in (7.58) based on the test statistic *MS*. They also provided some exact null distribution critical values sm_α in this equal replications setting for a variety of combinations of k, n, and c.

Properties

1. *Asymptotic Chi-Squaredness*. See Mack and Skillings (1980).
2. *Efficiency.* See Mack and Skillings (1980) and Section 7.16.

Problems

87. Rice (1988) considered an experiment to determine whether two forms of iron, Fe^{2+} and Fe^{3+}, are retained differently, with the goal of comparing their potentials for use as dietary supplements. A total of 108 mice were randomly divided into six groups of 18 mice each. Three of these groups were given Fe^{2+} in the different concentrations, 10.2, 1.2, and .3 mM, and three groups were given Fe^{3+} in the same concentrations. The iron was radioactively labeled so that a counter could be used to accurately measure the initial amount given, and it was administered orally to the mice. At a later time, a second count was obtained on each mouse, and the percentage of iron retained was recorded. The data in Table 7.21 are the percentages retained by each of the 108 mice.

Use the Mack–Skillings large-sample procedure (7.59) to test the hypothesis that there is a difference across the concentrations studied between the two forms of iron Fe^{2+} and Fe^{3+} in percentage iron retained.

88. Let V_j be as defined in expression (7.60), for $j = 1, \ldots, k$. Show that

$$E_0[V_j] = \sum_{i=1}^{n} \left[\frac{c_{ij}(q_i + 1)}{2q_i} \right],$$

as noted in expression (7.61), where q_i is the number of observations present in the ith block, for $i = 1, \ldots, n$.

Table 7.21 Percentage of Iron Retained

Concentration	Form of iron					
	Fe^{2+}			Fe^{3+}		
.3 millimolar	2.71	5.43	6.38	2.25	3.93	5.08
	6.38	8.32	9.04	5.82	5.84	6.89
	9.56	10.01	10.08	8.50	8.56	9.44
	10.62	13.80	15.99	10.52	13.46	13.57
	17.90	18.25	19.32	14.76	16.41	16.96
	19.87	21.60	22.25	17.56	22.82	29.13
1.2 millimolar	4.04	4.16	4.42	2.20	2.93	3.08
	4.93	5.49	5.77	3.49	4.11	4.95
	5.86	6.28	6.97	5.16	5.54	5.68
	7.06	7.78	9.23	6.25	7.25	7.90
	9.34	9.91	13.46	8.85	11.96	15.54
	18.40	23.89	26.39	15.89	18.30	18.59
10.2 millimolar	2.20	2.69	3.54	0.71	1.66	2.01
	3.75	3.83	4.08	2.16	2.42	2.42
	4.27	4.53	5.32	2.56	2.60	3.31
	6.18	6.22	6.33	3.64	3.74	3.74
	6.97	6.97	7.52	4.39	4.50	5.07
	8.36	11.65	12.45	5.26	8.15	8.24

Source: J. A. Rice (1988).

89. Show that for the special case of one replication per cell (i.e., $c = 1$), the Mack–Skillings procedures (7.58) and (7.59) are equivalent to the Friedman procedures (7.6) and (7.7), respectively. (See Comment 71.)

90. Anderson and McLean (1974) considered the data from an experiment measuring the strength of a weld in steel bars. The two factors of interest in the experiment were the total time of the automatic weld cycle and the distance the weld die travels during the automatic weld cycle. Two weld-strength observations were collected at each combination of five different weld cycle times and three different weld die travel distances (gage bar settings). These weld-strength data are given in Table 7.22.

Use the Mack–Skillings procedure to test the hypothesis that weld cycle time has an effect on the strength of a weld, at least over the weld die travel distances considered in the study.

91. For the weld-strength data in Table 7.22, compute the median of the two observations in each of the gage bar setting/weld cycle time combinations. Apply the Friedman procedure (7.6) to the resulting medians to test the hypothesis that weld cycle time has an effect on the strength of a weld, at least over the weld die travel distances in the study. Compare with the result obtained in Problem 90. (See also Comment 76.)

92. Consider the Mack–Skillings statistic MS_g (7.63) for the most general two-way layout setting with at least one replication for every treatment–block combination, as discussed in Comment 77. Show that the test procedure (7.64) based on MS_g is equivalent to the equal replications test procedure (7.58) based on MS (7.57) when, in fact, we have an equal number, c, of replications for every treatment–block combination.

93. One method for the determination of coal acidity is based on the use of ethanolic NaOH. In an effort to assess the effect of the ethanolic NaOH concentration on the obtained acidity values, Sternhell (1958) studied three different NaOH concentrations (.404N, .626N, and .786N) in conjunction with three different types of coal (Morwell, Yallourn, and Maddingley). The data in Table 7.23 are the resulting acidity values determined under each of these three concentration levels for two different samples from each type of coal.

Use the Mack–Skillings procedure to test the hypothesis that the NaOH concentration has an effect on the measured coal acidity values, at least over the three types of coal included in this study.

94. Consider the percentage retained iron data in Table 7.21. Test the hypothesis that the iron concentration affects the percentage iron retention, regardless of which form of iron is involved.

Table 7.22 Strength of Weld

	Weld cycle times				
Gage bar setting	1	2	3	4	5
1	10 12	13 17	21 30	18 16	17 21
2	15 19	14 12	30 38	15 11	14 12
3	10 8	12 9	10 5	14 15	19 11

Source: V. L. Anderson and R. A. McLean (1974).

Table 7.23 Coat Acidity Value

	NaOH concentration					
Type of coal	.404N		.626N		.786N	
Morwell	8.27	8.17	8.03	8.21	8.60	8.20
Yallourn	8.66	8.61	8.42	8.58	8.61	8.76
Maddingley	8.14	7.96	8.02	7.89	8.13	8.07

Source: S. Sternhell (1958).

95. What is the maximum value for the Mack–Skillings statistic MS (7.57) when there are c replications per cell? For what rank configuration is this maximum achieved?

96. Consider the setting corresponding to $k = 4$, $n = 5$, and $c = 3$ replications per cell. Compare the critical region for the exact level $\alpha = .0100$ test of H_0 (7.2) based on MS with the critical region for the corresponding nominal level $\alpha = .0100$ test based on the large-sample approximation.

97. Consider the setting corresponding to $k = 2$, $n = 2$, and $c = 2$ replications per cell. Obtain the form of the exact null H_0 distribution of MS (7.57) for the case of no-tied observations.

98. Consider the setting corresponding to $k = 2$, $n = 2$, and $c = 2$ replications per cell. Suppose that one of the observations in the first cell (block 1 and treatment 1) is tied in value with one of the observations in the second cell (block 1 and treatment 2). Obtain the conditional exact probability distribution of MS (7.57) under H_0 (7.2) when average ranks are used to break this within-blocks tie. Compare this conditional null distribution of MS with the null distribution of MS obtained in Problem 97 when there are no ties.

99. Consider the setting corresponding to $k = 2$, $n = 2$, and $c = 2$ replications per cell. Suppose that one of the observations in the first cell (block 1 and treatment 1) is tied in value with the other observation in the same cell. Obtain the conditional exact probability distribution of MS (7.57) under H_0 (7.2) when average ranks are used to break this within-cell tie. Compare this conditional null distribution of MS with the null distributions of MS obtained in Problems 97 and 98 when there are no ties and ties between cells, respectively.

100. Consider the setting corresponding to $k = 5$, $n = 4$, and $c = 4$ replications per cell. Compare the critical region for the exact level $\alpha = .0500$ test of H_0 (7.2) based on MS with the critical region for the corresponding nominal level $\alpha = .0500$ test based on the large-sample approximation.

101. In a study to determine the effect of light on the release of luteinizing hormone (LH), Rice (1988) compared data for male and female rats kept in constant light with similar animals exposed to a regime of 14 h of light and 10 h of darkness. Five different dosages of a luteinizing release factor (LRF) were considered in the study and the measurement obtained from the animals was the level of LH (in nanograms per milliliter of serum) in blood samples collected after exposure to one of the regimes in combination with one of the LRF dosages. We consider data for the male rats only.

Sixty male rats were randomly allocated to the various experimental settings in such a way that six rats were exposed to each of the 10 combinations of light regime and LRH dosage. The LH level data for these 60 rats are given in Table 7.24.

Table 7.24 Serum Level of LH (in Nanograms per Milliliter of Serum)

	Light regime					
LRF dosage	Constant light			14 h light/10 h dark		
0 ng (control)	72	64	78	212	27	68
	20	56	70	72	130	153
10 ng	74	82	40	32	98	148
	87	78	88	186	203	188
50 ng	130	187	133	294	306	234
	185	107	98	219	281	288
250 ng	159	167	193	515	340	348
	196	174	250	205	505	432
1250 ng	137	426	178	296	545	630
	208	196	251	418	396	227

Source: J. A. Rice (1988).

Use the Mack–Skillings large-sample procedure (7.59) to test the hypothesis that degree of exposure to light has an effect on serum levels of LH across the LRH dosages included in the study.

7.10 ASYMPTOTICALLY DISTRIBUTION-FREE TWO-SIDED ALL-TREATMENTS MULTIPLE COMPARISONS FOR A TWO-WAY LAYOUT WITH AN EQUAL NUMBER OF REPLICATIONS IN EACH TREATMENT–BLOCK COMBINATION (MACK–SKILLINGS)

In this section we present an asymptotically distribution-free multiple comparison procedure using within-blocks ranks that is designed to make two-sided decisions about individual differences between pairs of treatment effects (τ_i, τ_j), for $i < j$, for data obtained from a two-way layout design with an equal number of replications for every treatment–block combination. The multiple comparison procedure of this section would generally be applied to data from such a two-way layout with an equal number of replications *after* rejection of H_0 (7.2) with the Mack–Skillings procedure from Section 7.9. In this setting we will reach conclusions about all $k(k-1)/2$ pairs of treatment effects and these conclusions are naturally two-sided in nature.

Procedure

Let $S_1, \ldots, S_k$ be the treatment sums of cellwise averages of within-blocks ranks given by (7.56). Calculate the $k(k-1)/2$ absolute differences $|S_u - S_v|$, $1 \le u < v \le k$.

When H_0 (7.2) is true, the $k(k-1)/2$-component vector $(S_1, \ldots, S_k)$ has, when properly standardized and as N tends to infinity, an asymptotic $(k-1)$-variate normal distribution with appropriate mean vector and covariance matrix (see Mack and Skillings (1980) for details of the proof). At an approximate experimentwise error rate of α, the Mack–Skillings two-sided all-treatments multiple comparison procedure reaches its $k(k-1)/2$ pairwise decisions, corresponding to each (τ_u, τ_v) pair, $1 \le u < v \le k$, by the criterion

$$\text{Decide } \tau_u \neq \tau_v \text{ if } |S_u - S_v| \ge [k(N+n)/12]^{1/2} q_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v, \tag{7.75}$$

where q_α is the upper αth percentile for the distribution of the range of k independent $N(0,1)$ variables. To find q_α for k treatments and a specified experimentwise error rate α, we use the R command `cRangeNor`(α, k). For example, to find $q_{.01}$ for $k = 4$ treatments, we apply `cRangeNor(.01,4)` and obtain $q_{.01} = 3.240$ for $k = 4$. (See also Comment 82.)

Ties

If there are ties among the X observations within any of the blocks, use average ranks to break the ties and compute the individual sums of cellwise averages of within-blocks ranks $S_1, \ldots, S_k$.

EXAMPLE 7.10 ***Determination of Niacin in Bran Flakes.***

For the sake of illustration, we apply procedure (7.75) to the niacin determination data discussed in Example 7.9. There we had found rather strong evidence that the studied process for assessing niacin content in bran flakes does not produce consistent results across a variety of laboratories. To determine which of the laboratories differ in median detected niacin content in the bran flakes, we consider procedure (7.75) with an approximate experimentwise error rate $\alpha \approx .025$. Using the R command `cRangeNor(`α`,`k`)` with $\alpha = .025$ and $k = 4$, we find `cRangeNor(.025,4)` $= q_{.025} = 3.984$ and procedure (7.75) reduces to

$$\text{Decide } \tau_u \neq \tau_v \text{ if } |S_u - S_v| \geq [4(36+3)/12]^{1/2}(3.984) = 14.365.$$

Using the treatments sums of cellwise averages of within-blocks ranks obtained in Example 7.9, we find that

$$|S_2 - S_1| = |30.5 - 17.67| = 12.83, \quad |S_3 - S_1| = |15.83 - 17.67| = 1.84,$$

$$|S_4 - S_1| = |14 - 17.67| = 3.67, \quad |S_3 - S_2| = |15.83 - 30.5| = 14.67,$$

$$|S_4 - S_2| = |14 - 30.5| = 16.5, \quad |S_4 - S_3| = |14 - 15.83| = 1.83.$$

Referring these differences to the approximate critical value 14.365, we see that

$$\begin{aligned}
|S_2 - S_1| = 12.83 < 14.365 &\quad \Rightarrow \quad \text{decide } \tau_2 = \tau_1,\\
|S_3 - S_1| = 1.84 < 14.365 &\quad \Rightarrow \quad \text{decide } \tau_3 = \tau_1,\\
|S_4 - S_1| = 3.67 < 14.365 &\quad \Rightarrow \quad \text{decide } \tau_4 = \tau_1,\\
|S_3 - S_2| = 14.67 > 14.365 &\quad \Rightarrow \quad \text{decide } \tau_3 \neq \tau_2,\\
|S_4 - S_2| = 16.5 > 14.365 &\quad \Rightarrow \quad \text{decide } \tau_4 \neq \tau_2,\\
\text{and} \quad |S_4 - S_3| = 1.83 < 14.365 &\quad \Rightarrow \quad \text{decide } \tau_4 = \tau_3.
\end{aligned}$$

Thus, at an approximate experimentwise error rate of .025, we see that Laboratory 2 yielded significantly different median detected niacin content than either Laboratory 3 or Laboratory 4. These multiple comparison decisions help to focus the rationale for the original rejection of H_0 (7.2) by the Mack–Skillings procedure in Example 7.9, as it now seems reasonable to question the reliability of Laboratory 2 in conducting this niacin content process.

Comments

80. *Rationale for Multiple Comparison Procedure.* The rationale behind the multiple comparison procedure of this section for data from a two-way layout design with an equal number of replications is similar to that for the two-sided multiple comparison procedures for data from a complete randomized block design. For further discussion, see Comment 24.

81. *Experimentwise Error Rate.* The use of an experimentwise error rate represents a very conservative approach to multiple comparisons. We are insisting that the probability of making correct decisions be $1 - \alpha$ when the null hypothesis H_0 (7.2) of treatment equivalence is true. Thus we have a high degree of protection when H_0 is true, but we often apply such techniques when we have evidence (perhaps based on a priori information or perhaps obtained by applying the Mack–Skillings test, as in Example 7.9) that H_0 is not true. The protection under H_0 also makes it harder for the procedure to judge treatments as differing significantly when, in fact, H_0 is false, and this difficulty becomes more severe as k increases. See Comment 6.54 for additional discussion of experimentwise error rates.
82. *Conservative Procedure.* Mack and Skillings (1980) also proposed a conservative multiple comparison procedure that guarantees an upper bound on the experimentwise error rate. Let $S_1, \ldots, S_k$ be the treatment sums of cellwise averages of within-blocks ranks given by (7.56). At an experimentwise error rate *no greater* than α, the Mack–Skillings conservative two-sided all-treatments multiple comparison procedure reaches its $k(k-1)/2$ decisions through the criterion

$$\text{Decide } \tau_u \neq \tau_v \text{ if } |S_u - S_v| \geq [k(N+n)ms_\alpha/6]^{1/2};$$
$$\text{otherwise decide } \tau_u = \tau_v, \tag{7.76}$$

where ms_α is the upper α percentile for the null distribution of the Mack–Skillings statistic MS (7.57). Comment 73 describes how to obtain values of ms_α for a given number of treatments k, blocks n, and c replications for each treatment–block combination. Mack and Skillings (1980) note that although procedure (7.76) does not require a large number of blocks, it is, nevertheless, rather conservative since it is based on the projection procedure of Scheffé; that is, the true experimentwise error rate might be considerably smaller than the bound α provided by (7.76). As a result, they recommend using the approximation (7.75) whenever the number of blocks is reasonably large.
83. *Dependence on Observations from Other Noninvolved Treatments.* The all-treatments multiple comparison procedure of this section suffers from the same disadvantage as do the other two-way layout multiple comparison procedures of this chapter. The decision between treatment u and treatment v can be affected by changes only in the observations from one or more of the other $k - 2$ treatments that are not directly involved.

Properties

1. *Asymptotic Multivariate Normality.* See Mack and Skillings (1980).
2. *Efficiency.* See Section 7.16.

Problems

102. Apply procedure (7.75) to the weld-strength data of Table 7.22 in Problem 90.

103. Illustrate the difficulty discussed in Comment 83 by means of a numerical example.

104. Apply procedure (7.75) to the coal acidity data of Table 7.23 in Problem 93.

105. Consider the niacin content data of Table 7.20 in Example 7.9. Find the smallest approximate experimentwise error rate at which the most significant difference(s) in median bran flake niacin content between the four laboratories would be detected by procedure (7.75).

106. Consider the weld-strength data of Table 7.22 in Problem 90. Find the smallest approximate experimentwise error rate at which procedure (7.75) would declare that weld cycle times 1 and 3 have differing effects on the strength of a weld.

107. Consider the coal acidity data of Table 7.23 in Problem 93. Find the smallest approximate experimentwise error rate at which the most significant difference(s) in effects of the NaOH concentration on the measured coal acidity value would be detected by procedure (7.75).

108. Consider the coal acidity data of Table 7.23 in Problem 93. Find the smallest approximate experimentwise error rate at which procedure (7.75) would declare that there is a difference in median coal acidity level between the Morwell and the Yallourn types of coal.

ANALYSES ASSOCIATED WITH SIGNED RANKS

The statistical procedures discussed in Sections 7.1–7.5 (for randomized block designs with a single observation on each treatment–block combination) utilize the treatment observations only through comparisons within blocks. It is this restriction to within-blocks comparisons that leads directly to many of these procedures being strictly distribution-free, even for small sample sizes. An alternative approach is to consider accessing between-blocks information via utilization of pairwise signed ranks in the construction of appropriate statistical procedures. Hypothesis test and multiple comparison procedures based on these pairwise signed ranks will no longer be exactly distribution-free for small numbers (n) of blocks and they require the use of large-sample approximations. However, improved efficiency can result in many cases from this use of between-blocks signed ranks.

In the next five sections we assume (as done in Sections 7.1–7.5) that we have data from a randomized complete block design satisfying Assumptions A1–A3 for the case of one observation per treatment–block combination, corresponding to $c_{ij} = 1$ for every $i = 1, \ldots, n$ and $j = 1, \ldots, k$. For ease of notation in these five sections, we once again drop the third subscript on the X variables, because it is always equal to 1 in this setting.

Section 7.11 contains a conservative signed ranks test procedure directed at general alternatives for randomized block designs with a single observation on each treatment–block combination, while Section 7.12 presents the corresponding conservative signed ranks test procedure designed for ordered alternatives. The associated approximate signed ranks multiple comparison procedures are given in Sections 7.13 (all-treatments comparisons) and 7.14 (treatments-versus-control comparisons). Section 7.15 contains the contrast estimators linked to the Wilcoxon signed ranks for this setting.

7.11 A TEST BASED ON WILCOXON SIGNED RANKS FOR GENERAL ALTERNATIVES IN A RANDOMIZED COMPLETE BLOCK DESIGN (DOKSUM)

In this section we present a conservative procedure based on pairwise signed ranks for testing H_0 (7.2) against the general alternative H_1 (7.3) that at least two of the treatment effects are not equal.

Procedure

For each of the $k(k-1)/2$ pairs of treatments (u, v), with $1 \le u < v \le k$, we form the n absolute differences

$$Y_{uv}^i = |X_{iu} - X_{iv}|, \quad i = 1, \ldots, n. \tag{7.77}$$

(Note that $Y_{uv}^i = |D_{uv}^i|$, where the D_{uv}^i are the same differences given in (7.36) and used in the contrast estimator discussed in Section 7.5.) For each pair of treatments (u, v), we let R_{uv}^i be the rank of Y_{uv}^i in the ranking from least to greatest of the n values $Y_{uv}^1, \ldots, Y_{uv}^n$. To compute the Doksum (1967) statistic D, set

$$T_{uv} = \sum_{i=1}^{n} R_{uv}^i \Psi_{uv}^i \quad \text{and} \quad B_{uv} = \sum_{i=1}^{n} \Psi_{uv}^i, \tag{7.78}$$

where

$$\Psi_{uv}^i = \begin{cases} 1, & \text{if } X_{iu} < X_{iv}, \\ 0, & \text{otherwise.} \end{cases} \tag{7.79}$$

Let

$$H_{uv} = \frac{2(T_{uv} - B_{uv})}{n(n-1)}, \quad 1 \le u < v \le k. \tag{7.80}$$

(We note that the statistics H_{uv} need be calculated directly only for $u < v$, because for $u > v$, we can use the relationship $H_{vu} = 1 - H_{uv}$.) Next, we obtain the averages

$$H_{u.} = \sum_{j=1}^{k} \frac{H_{uj}}{k}, \quad u = 1, \ldots, k, \tag{7.81}$$

where we note that $H_{uu} = 0$, for $u = 1, \ldots, k$.

The common null variance of each of the $k(k-1)/2$ differences $H_{u.} - H_{v.}$, $1 \le u < v \le k$, is given by the expression

$$\text{var}_0(H_{u.} - H_{v.}) = \frac{2n - 1 + (k-2)[24(n-2)\lambda_F + 13 - 6n]}{3kn(n-1)}, \tag{7.82}$$

with

$$\lambda_F = P_0(X_1 < X_2 + X_3 - X_4 \quad \text{and} \quad X_1 < X_5 + X_6 - X_7), \tag{7.83}$$

where $X_1, X_2, \ldots, X_7$ are independent and identically distributed according to the common continuous underlying distribution F in Assumption A3. Since the value of λ_F (7.83) depends on the particular form of the continuous F, we can not use the expression in (7.82) to construct a distribution-free procedure for testing H_0 (7.2). However, Lehmann (1964) showed that $\lambda_F \le \frac{7}{24}$ for all continuous F (see Comment 87). Replacing λ_F in equation (7.82) by this upper bound of $\frac{7}{24}$ yields the expression

$$V_U = \frac{2n - 1 + (k-2)[7(n-2) + 13 - 6n]}{3kn(n-1)}. \tag{7.84}$$

The Doksum test statistic for the conservative test of H_0 (7.2) is then

$$D = \sum_{j=1}^{k} \frac{[H_{j\cdot} - \{(k-1)/2k\}]^2}{(k-1)V_U/2k}. \tag{7.85}$$

For a conservative test (see Comment 85) of

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the general alternative

$$H_1 : [\tau_1, \tau_2, \ldots, \tau_k \text{ not all equal}],$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } D \geq \chi^2_{k-1,\alpha}; \quad \text{otherwise do not reject,} \tag{7.86}$$

where $\chi^2_{k-1,\alpha}$ is the upper α percentile point of a chi-square distribution with $k-1$ degrees of freedom. To find $\chi^2_{k-1,\alpha}$, we use the R command `qchisq`$(1-\alpha, k-1)$. For example, to find $\chi^2_{3,.05}$, we apply `qchisq(.95, 3)` and obtain $\chi^2_{3,.05} = 7.815$.

Ties

For any Y^i_{uv} (7.77), $1 \leq u < v \leq k$, that is zero, compute T_{uv} and B_{uv} in (7.78) by replacing the associated Ψ^i_{uv} (7.79) with

$$\Psi^{*i}_{uv} = \begin{cases} 1, & \text{if } X_{iu} < X_{iv}, \\ \frac{1}{2}, & \text{if } X_{iu} = X_{iv}, \\ 0, & \text{if } X_{iu} > X_{iv}. \end{cases} \tag{7.87}$$

For ties among $Y^1_{uv}, \ldots, Y^n_{uv}$, use average ranks to compute T_{uv} (7.78).

EXAMPLE 7.11 ***Rounding First Base.***

Consider once again the rounding-first-base data presented in Table 7.1 and discussed in Example 7.1. The reader should already be familiar with the calculations of the paired-data signed rank statistics T_{uv} and sign statistics B_{uv} (see Comment 86) from the materials in Sections 3.1 and 3.4, respectively. We include a detailed calculation of T_{12} and B_{12} in Table 7.25 to illustrate the method for handling zero differences and ties (see Ties and Comment 88).

The statistics B_{12} and T_{12} are obtained by summing the entries in the next-to-last and last columns, respectively, of Table 7.25. We obtain B_{13}, B_{23}, T_{13}, and T_{23} in a similar manner, and the results are

$$B_{13} = 5, \quad B_{23} = 5, \quad T_{13} = 54, \quad \text{and} \quad T_{23} = 30.5.$$

Table 7.25 Calculation of T_{12} and B_{12} for Data in Table 7.1

j	$X_{j1} - X_{j2}$	Y_{12}^j	R_{12}^j	Ψ_{12}^{*j}	$R_{12}^j \Psi_{12}^{*j}$
1	−.10	.10	17	1	17
2	.15	.15	20	0	0
3	−.40	.40	22	1	22
4	.05	.05	9.5	0	0
5	.05	.05	9.5	0	0
6	−.10	.10	17	1	17
7	.00	.00	2.5	$\frac{1}{2}$	1.25
8	−.05	.05	9.5	1	9.5
9	.10	.10	17	0	0
10	.05	.05	9.5	0	0
11	.05	.05	9.5	0	0
12	.10	.10	17	0	0
13	.25	.25	21	0	0
14	.05	.05	9.5	0	0
15	.00	.00	2.5	$\frac{1}{2}$	1.25
16	−.10	.10	17	1	17
17	.00	.00	2.5	$\frac{1}{2}$	1.25
18	−.05	.05	9.5	1	9.5
19	.05	.05	9.5	0	0
20	.05	.05	9.5	0	0
21	.05	.05	9.5	0	0
22	.00	.00	2.5	$\frac{1}{2}$	1.25
				$B_{12} = 8$	$T_{12} = 97$

It then follows from (7.80) that

$$H_{12} = .385, \quad H_{13} = .212, \quad \text{and} \quad H_{23} = .110.$$

From (7.81) and the fact that $H_{uv} = 1 - H_{vu}$, we have

$$\begin{aligned} H_{1\cdot} &= \frac{H_{11} + H_{12} + H_{13}}{3} \\ &= \frac{0 + .385 + .212}{3} = .199, \\ H_{2\cdot} &= \frac{H_{21} + H_{22} + H_{23}}{3} \\ &= \frac{.615 + 0 + .110}{3} = .242, \\ H_{3\cdot} &= \frac{H_{31} + H_{32} + H_{33}}{3} \\ &= \frac{.788 + .890 + 0}{3} = .559. \end{aligned}$$

We next find V_U (7.84) to be

$$V_U = \frac{2(22) - 1 + [7(20) + 13 - 6(22)]}{3(3)(22)(21)} = .015.$$

Substituting these values for $H_{1.}$, $H_{2.}$, $H_{3.}$, and V_U into (7.43) yields

$$D = \frac{\left[.199 - \left(\frac{1}{3}\right)\right]^2 + \left[.242 - \left(\frac{1}{3}\right)\right]^2 + \left[.559 - \left(\frac{1}{3}\right)\right]^2}{2(.015)/6} = 15.5.$$

Referring this value of D to the chi-square distribution with $k - 1 = 2$ degrees of freedom, we use the `R` command `pchisq(15.5,2)` to find that the lowest significance level at which we would reject H_0 is $1 - .99957 = .00043$ (cf. Example 7.1).

Comments

84. *Motivation for the Test.* Under H_0 (7.2), the $H_{j.}$'s (7.81) tend to be near $(k - 1)/2k$, their common null expectation, and thus the numerator of D (7.85) tends to be small. When the τ's are not all equal, we expect the $H_{j.}$'s to be more disparate, and thus (at least some of) the $[H_{j.} - \{(k - 1)/2k\}]^2$ terms tend to be large, yielding a large value of D. This provides partial motivation for procedure (7.86).

85. *Conservative Nature of the Test.* The test defined by (7.86) is neither distribution-free nor asymptotically $(n \to \infty)$ distribution-free. Rather, it is conservative in the sense that (asymptotically) the actual probability of rejecting H_0 (7.2) when it is true tends to be slightly smaller than the nominal level α. This is a consequence of using an upper bound for the parameter λ_F (7.83). See also Comments 87 and 89.

86. *Pairwise Signed Rank and Sign Statistics.* For a given pair of treatments $(u, v), 1 \le u < v \le k$, the statistics T_{uv} and B_{uv} (7.78) are simply the Wilcoxon signed rank and sign statistics, respectively, as discussed in Sections 3.1 and 3.4, respectively, applied to the paired data in treatments u and v. With this relationship in mind, we note that the difference $T_{uv} - B_{uv}$ in the numerator of H_{uv} (7.80) may equivalently be calculated as the number of Walsh averages $(X_{su} - X_{sv} + X_{tu} - X_{tv})/2$, with $1 \le s < t \le n$, that are negative. (See Comment 3.17.)

87. *Bounds for the Parameter* λ_F. The null correlation between two overlapping statistics H_{uv} and H_{uw} defined by (7.80), with $u \ne v, u \ne w$, and $v \ne w$, depends on the parameter λ_F (7.83). This, combined with the fact that λ_F varies with F (Lehmann, 1964), prevents the development of a distribution-free test procedure based on the numerator of D (7.85). Lehmann (1964) showed that $\lambda_F \le \frac{7}{24} (\approx .2917)$ for all continuous F. Replacement of λ_F in expression (7.82) for the null variance of $H_{u.} - H_{v.}$ by the upper bound $\frac{7}{24}$ enables the development of the conservative procedure based on D (7.86). Spurrier (1991) established the lower bound $\lambda_F \ge \frac{89}{315} (\approx .2825)$ for all continuous F. Since the value of λ_F is so narrowly confined between .2825 and .2917 for all continuous F, replacing λ_F by its upper bound or $\frac{7}{24}$ in expression

(7.82) sacrifices little to permit the construction of the conservative test procedure (7.86).

88. *Ties.* The reader may have noted that the method we advocate in Ties for dealing with zero differences, when computing the $T_{uv}(B_{uv})$ signed rank (sign) statistics for use in procedure (7.86), differs from the corresponding directions given for the signed rank (sign) statistic in Section 3.1 (Section 3.4). In Chapter 3, we recommended reducing the sample size by the number of zero differences. This change is initiated in the calculation of D (7.85) in order to keep all of the T_{uv}'s and B_{uv}'s based on the same sample size (n).

89. *Asymptotically Distribution-Free Competitor.* As an alternative to the conservative test procedure (7.86) based on the replacement of the unknown parameter λ_F (7.83) by its upper bound $\frac{7}{24}$, we could instead choose to estimate the value of λ_F from the sample data. Use of a consistent estimator of λ_F in this manner leads to an asymptotically ($n \to \infty$) distribution-free procedure for testing H_0 (7.2), rather than the conservative procedure in (7.86). Lehmann (1964) proposed the estimator $\widehat{\lambda}_F$ of λ_F, where $\widehat{\lambda}_F$ is the proportion of sample sextuples $(\alpha, \beta, \gamma; u, v, w)$ for which the simultaneous inequalities

$$(X_{\alpha u} < X_{\beta u} + X_{\alpha v} - X_{\beta v} \text{ and } X_{\alpha u} < X_{\gamma u} + X_{\alpha w} - X_{\gamma w})$$

are satisfied. In practice, when estimating λ_F, it would normally suffice to check only a subset of the total number of such sample sextuples. Due to the closeness of the upper bound $\frac{7}{24}$ to all values of λ_F, procedure (7.86) is, for all practical purposes, virtually equivalent to Doksum's (1967) asymptotically distribution-free procedure based on estimating λ_F.

Properties

1. *Consistency.* See Doksum (1967) and Hollander and Wolfe (1973, p. 166).
2. *Asymptotic Chi-Square Distribution.* See Doksum (1967).
3. *Efficiency.* See Doksum (1967) and Section 7.16.

Problems

109. Apply procedure (7.86) to the adaptation score data of Table 7.10 (Example 7.4).

110. The Doksum test procedure (7.86) uses between-block information, whereas Friedman's test procedure (7.6) uses only within-block information. Explain.

111. Apply procedure (7.86) to the serumCPK activity data in Table 7.3, Problem 5.

112. Apply procedure (7.86) to the percentage correctly identified consonants data in Table 7.4 (Problem 12).

113. Both the Doksum (7.86) and the Friedman (7.6) procedures are appropriate for testing against general alternatives when we have data from a randomized complete block design with one observation per treatment–block combination. Discuss the relative advantages and disadvantages of the two competing procedures.

7.12 A TEST BASED ON WILCOXON SIGNED RANKS FOR ORDERED ALTERNATIVES IN A RANDOMIZED COMPLETE BLOCK DESIGN (HOLLANDER)

In this section we present a conservative procedure based on pairwise signed ranks for testing H_0 (7.2) against the a priori ordered alternatives H_2 (7.9), corresponding to $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k$, with at least one strict inequality.

Procedure

For each of the $k(k-1)/2$ pairs of treatments (u, v), with $1 \leq u < v \leq k$, we compute the signed rank statistic T_{uv}, as defined in (7.78). To compute the Hollander statistic Q, set

$$Y = \sum_{u=1}^{k-1} \sum_{v=u+1}^{k} T_{uv}. \tag{7.88}$$

The null expected value of Y is given by

$$E_0(Y) = \frac{nk(k-1)(n+1)}{8}, \tag{7.89}$$

but the null variance of Y is unknown (see Comment 93) and depends on the particular form of the underlying continuous distribution F in Assumption 3. Thus, a test of H_0 (7.2) based on Y will not be distribution-free. However, a conservative procedure can be developed by using an upper bound for this unknown null variance of Y. Using the R command `CorrUpperBound(n)`, we obtain the value of the upper bound ρ_U^n for the null correlation between two overlapping signed rank statistics based on n observations. An upper bound for the null variance of Y (7.88) is then given by

$$\text{var}_U(Y) = \frac{nk(n+1)(2n+1)(k-1)\{3 + 2(k-2)\rho_U^n\}}{144}. \tag{7.90}$$

The Hollander test statistic for the conservative test of H_0 (7.2) is then

$$Q = \frac{Y - E_0(Y)}{\{\text{var}_U(Y)\}^{1/2}}, \tag{7.91}$$

with the expressions for $E_0(Y)$ and $\text{var}_U(Y)$ given in (7.89) and (7.90), respectively. For a conservative test (see Comment 92) of

$$H_0 : [\tau_1 = \cdots = \tau_k]$$

versus the ordered alternatives

$$H_2 : [\tau_1 \leq \tau_2 \leq \cdots \leq \tau_k, \text{ with at least one strict inequality}],$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } Q \geq z_\alpha; \quad \text{otherwise do not reject.} \tag{7.92}$$

Ties

See Ties of Section 7.11 and Comment 88.

EXAMPLE 7.12 ***Effect of Weight on Forearm Tremor Frequency.***

The data in Table 7.26 are based on a subset of the data obtained by Fox and Randall (1970) in their study of forearm tremor. Each entry in the table is the mean of five experimental values of tremor frequency. We identify treatment 1 with 7.5 lb, treatment 2 with 5 lb, treatment 3 with 2.5 lb, treatment 4 with 1.25 lb, and treatment 5 with 0 lb, and use procedure (7.92) to test H_0 (7.2) versus the ordered alternatives H_2 (7.9), which specify that adding mass decreases the tremor frequency.

Calculations similar to those presented in Example 7.11 yield

$$\begin{aligned} &T_{12} = 18.5, \quad T_{13} = 21, \quad T_{14} = 21, \quad T_{15} = 21, \quad T_{23} = 20, \\ &T_{24} = 21, \quad T_{25} = 21, \quad T_{34} = 21, \quad T_{35} = 21, \quad T_{45} = 21. \end{aligned} \tag{7.93}$$

From (7.88), we obtain

$$Y = T_{12} + T_{13} + T_{14} + T_{15} + T_{23} + T_{24} + T_{25} + T_{34} + T_{35} + T_{45} = 206.5.$$

From the R command `CorrUpperBound(6)`, we find $\rho_U^6 = .452$, and evaluating (7.89) and (7.90) gives

$$E_0(Y) = \frac{5(4)(6)(7)}{8} = 105,$$

$$\mathrm{Var}_U(Y) = \frac{6(7)(13)(5)(4)\{3 + 6(.452)\}}{144} = 433.2.$$

From (7.91), we then have

$$Q = \frac{206.5 - 105}{[433.2]^{1/2}} = 4.88.$$

Table 7.26 Forearm Tremor Frequency (Hz) as a Function of Weight Applied at the Wrist

Treatment	1	2	3	4	5
			Weight (lb)		
Subject	7.5	5	2.5	1.25	0
1	2.58	2.63	2.62	2.85	3.01
2	2.70	2.83	3.15	3.43	3.47
3	2.78	2.71	3.02	3.14	3.35
4	2.36	2.49	2.58	2.86	3.10
5	2.67	2.96	3.08	3.32	3.41
6	2.43	2.50	2.85	3.06	3.07

Source: J. R. Fox and J. E. Randall (1970).

Using the R command `pnorm(·)`, we see that the lowest approximate level at which we would reject H_0 with these data is $P_0(Q \geq 4.88) \approx 1 - \texttt{pnorm(4.88)} = 1 - .99999947 = .00000053$. Thus there is very strong evidence (over the range of weights considered in the study) that the tremor frequency does decrease as the applied weight increases.

Comments

90. *Motivation for the Test.* Note that the statistic Y (7.88) is designed to guard against the postulated ordered alternatives H_2 (7.9). Consider the case $k = 3$. Then $Y = T_{12} + T_{13} + T_{23}$, and if $\tau_1 < \tau_2 < \tau_3$, each of T_{12}, T_{13}, and T_{23} would tend to be larger than $n(n+1)/4$, their common null expectation. Thus, Y would tend to be large, as desired. Contrast this with a situation in which we suspect (and design the test for) the alternative $\tau_1 < \tau_2 < \tau_3$, but in actuality, we have $\tau_3 < \tau_2 < \tau_1$. In this case, each of T_{12}, T_{13}, and T_{23} would tend to be small. This provides partial motivation for procedure (7.92).

91. *Non-Distribution-Free Property of Y (7.88).* Consider the Y (7.88) statistic for testing against ordered alternatives in the two-way layout (7.1) in relation to Jonckheere's J (6.13) statistic for testing against ordered alternatives in the one-way layout (6.1). The statistic J is the sum $\sum_{u<v}^{k} U_{uv}$ of two-sample Mann–Whitney statistics U_{uv} (or, equivalently, Wilcoxon rank sum statistics), where each U_{uv} is distribution-free under H_0 (6.2). The statistic Y is a sum $\sum_{u<v}^{k} T_{uv}$ of the paired-sample Wilcoxon signed rank statistics T_{uv}, where each T_{uv} is distribution-free under H_0 (7.2). Although J itself is also distribution-free under H_0 (6.2), Y is not distribution-free under H_0 (7.2) when $k > 2$. (For $k = 2$, Y reduces to T_{12}, which is distribution-free.) See Hollander (1967a) for details of the non-distribution-free character of Y.

92. *Conservative Nature of the Test.* The test defined in (7.91) is neither distribution-free nor asymptotically ($n \to \infty$) distribution-free. Rather, it is conservative in the sense that (asymptotically) the actual probability of rejecting H_0 (7.2) when it is true tends to be smaller than the nominal level α. This is a direct consequence of using an upper bound $\text{var}_U(Y)$ to replace the unknown null variance of Y. Also see Comment 94.

93. *Asymptotic Null Variance of Y.* Hollander (1967a) showed that the asymptotic ($n \to \infty$) null variance of Y (7.88) has the form

$$\text{var}_0(Y) = \frac{nk(n+1)(2n+1)(k-1)\{3 + 2(k-2)\rho^*\}}{144},$$

where ρ^* is the limiting ($n \to \infty$) null correlation between two overlapping signed rank statistics T_{12} and T_{13}. This limiting correlation can also be expressed as

$$\rho^* = 12\lambda_F - 3, \tag{7.94}$$

where λ_F is defined by (7.83). In forming the Q test statistic (7.91) for the conservative test procedure (7.92), we replace ρ^* by its upper bound ρ_U^n.

94. *Asymptotically Distribution-Free Competitor.* As an alternative to the conservative test procedure (7.92) based on the use of the upper bound ρ_U^n, we could instead replace ρ^* (7.94) by a consistent estimator $\widehat{\rho}$ based on the sample data. Use of a consistent estimator of ρ^* in this manner leads to an asymptotically ($n \to \infty$) distribution-free procedure for testing H_0 (7.2) rather than the conservative procedure in (7.92). Hollander suggested such an approach to this problem based on the consistent estimator $\widehat{\rho} = 12\widehat{\lambda}_F - 3$, where $\widehat{\lambda}_F$ is defined in Comment 89. Due to the closeness of the upper bound $\frac{7}{24}$ to all values of λ_F, procedure (7.92) is, for all practical purposes, virtually equivalent to Hollander's (1967a) asymptotically distribution-free procedure based on estimating λ_F.

Properties

1. *Consistency.* The test defined by (7.92) is consistent against the ordered alternatives (7.9). See Hollander (1967a) and Hollander and Wolfe (1973, p. 170).
2. *Asymptotic Normality.* See Hollander (1967a).
3. *Efficiency.* See Hollander (1967a) and Section 7.16.

Problems

114. Apply the Q (7.92) test to the metronome data of Table 7.6. Use the postulated ordering $\tau_R < \tau_A < \tau_N$.

115. The Hollander test procedure (7.92) uses between-block information, but Page's test procedure (7.11) uses only within-block information. Explain.

116. Apply procedure (7.92) to the shelterbelt data in Table 7.7 (Problem 19).

117. Apply procedure (7.92) to the cotton strength index data in Table 7.5 (Example 7.2). Compare with the result from the use of Page's test in Example 7.2.

118. Both the Hollander (7.92) and the Page (7.11) procedures are appropriate for testing against ordered alternatives when we have data from a randomized complete block design with one observation per treatment–block combination. Discuss the relative advantages and disadvantages of the two competing procedures.

7.13 APPROXIMATE TWO-SIDED ALL-TREATMENTS MULTIPLE COMPARISONS BASED ON SIGNED RANKS (NEMENYI)

In this section we present a multiple comparison procedure based on Wilcoxon signed rank statistics that is designed to make decisions about individual differences between pairs of treatment effects (τ_u, τ_v), for $u < v$, in a setting where general alternatives H_1 (7.3) are of interest. Thus, the multiple comparison procedure of this section would generally be applied to two-way layout data (with one observation per cell) *after* rejection of H_0 (7.2) with the Doksum–Lehmann procedure from Section 7.11. In this setting it is important to reach conclusions about all $k(k-1)/2$ pairs of treatment effects and these conclusions are naturally two sided in nature.

Procedure

For $1 \le u < v \le k$, let T_{uv} be the signed rank statistic (7.78) between treatments u and v. Calculate the $k(k-1)/2$ statistics

$$T'_{uv} = \max\{T_{uv}, [n(n+1)/2] - T_{uv}\}, 1 \le u < v \le k. \tag{7.95}$$

At an approximate (see Comment 95) experimentwise error rate of α, the two-sided signed rank multiple comparison procedure reaches its $k(k-1)/2$ pairwise decisions, corresponding to each (τ_u, τ_v) pair, for $1 \le u < v \le k$, by the criterion

$$\text{Decide } \tau_u \ne \tau_v \text{ if } T'_{uv} \ge t'_\alpha; \quad \text{otherwise decide } \tau_u = \tau_v, \tag{7.96}$$

where the constant t'_α is chosen to make the experimentwise error rate approximately equal to α; that is, t'_α satisfies the restriction

$$P_0\{T'_{uv} < t'_\alpha, u = 1, \ldots, k-1 \quad \text{and} \quad v = u+1, \ldots, k\} \approx 1 - \alpha, \tag{7.97}$$

where the probability $P_0(.)$ is computed under H_0 (7.2). Equation (7.97) stipulates that the $k(k-1)/2$ inequalities $T'_{uv} < t'_\alpha$, corresponding to all pairs (u, v) of treatments with $u < v$, hold simultaneously with approximate probability $1 - \alpha$ when H_0 (7.2) is true. Selected approximate values of t'_α can be found from the relationship

$$t'_\alpha \approx \left[\frac{n(n+1)}{4}\right] + \left[\frac{n(n+1)(2n+1)}{48}\right]^{1/2} q_\alpha, \tag{7.98}$$

where q_α is the upper αth percentile point for the distribution of the range of k independent $N(0, 1)$ variables. To find q_α for k treatments and a specified experimentwise error rate α, we use the R command `cRangeNor`(α, k). For example, to find $q_{.005}$ for $k = 6$ treatments, we apply `cRangeNor(.005, 6)` and obtain $q_{.005} = 5.033$ for $k = 6$.

Ties

See Ties of Section 7.11 and Comment 88.

EXAMPLE 7.13 ***Rounding First Base.***

We illustrate procedure (7.96) using the approximation (7.98) with the rounding-first-base data of Table 7.1. In Example 7.11, we found

$$T_{12} = 97, \quad T_{13} = 54, \quad \text{and} \quad T_{23} = 30.5.$$

From (7.95), we obtain

$$T'_{12} = \max\{97, 253 - 97\} = 156, \quad T'_{13} = \max\{54, 253 - 54\} = 199,$$

$$T'_{23} = \max\{30.5, 253 - 30.5\} = 222.5.$$

With an experimentwise error rate of $\alpha = .01$ and $k = 3$, we use `cRangeNor(.01,3)` to find $q_{.01} = 4.12$ for $k = 3$. Thus, with approximation (7.98), the inequality in (7.96) reduces to

$$T'_{uv} \geq t'_{.01} \approx \left[\frac{22(23)}{4}\right] + \left[\frac{22(23)(45)}{48}\right]^{1/2} (4.12) = 216.2,$$

and procedure (7.96) becomes

$$\text{Decide } \tau_u \neq \tau_v \text{ if } T'_{uv} \geq 216.2, \quad 1 \leq u < v \leq 3.$$

Since $T'_{12} < 216.2$, $T'_{13} < 216.2$, and $T'_{23} \geq 216.2$, only the narrow angle (treatment 2) and wide angle (treatment 3) running methods differ significantly at the approximate .01 experimentwise error rate using the signed rank procedure (7.96).

At this point, the reader may have noticed that, at the approximate .01 experimentwise error rate, the signed rank analysis in this example yields a conclusion different from the corresponding analysis based on the Friedman rank sums performed in Example 7.3. Since the analyses are based on different rankings and different statistics, the reader should not be shocked. It is instructive to note that if, for example, the multiple comparisons were made at an approximate .10 experimentwise error rate, the two procedures would agree in the sense that differences between treatments 2 and 3 and between treatments 1 and 3 would be declared significant under both analyses.

Comments

95. *Non-Distribution-Free Property.* Procedure (7.96), using approximation (7.98), is neither distribution-free nor asymptotically distribution-free. Nemenyi (1963) proposed this procedure under the assumptions that (a) the statistic $\max\{T'_{uv}, 1 \leq u < v \leq k\}$ is distribution-free and (b) the limiting ($n \to \infty$) null correlation between T_{12} and T_{13} (say) is close to $\frac{1}{2}$. Assumption (a) is incorrect, but the reasonableness of assumption (b) is supported by the values of λ_F, for various distributions F, obtained by Lehmann (1964), Hollander (1966), and Obenchain (1969). (See also Comments 87 and 93.)

96. *Independence from Observations for Other Noninvolved Treatments.* The value of T'_{uv}, the statistic used in the decision relating to τ_u and τ_v, does not depend on the observation values from the other $k - 2$ treatments. Thus, the signed ranks procedure (7.96) eliminates a difficulty encountered with the corresponding multiple comparison procedures (7.25) and (7.27) of Section 7.3 based on the Friedman rank sums. (See Comment 30.)

Properties

1. *Efficiency.* See Section 7.16.

Problems

119. Apply procedure (7.96) to the serum CPK activity data in Table 7.3 (Problem 5).

120. Apply procedure (7.96) to the Hebb–Williams EPT data in Table 7.9 (Problem 28).

121. Both procedures (7.27) and (7.96) are appropriate multiple comparison procedures when we have data from a randomized complete block design with one observation per treatment–block combination, and we are interested in two-sided comparisons between all treatments. Discuss the relative advantages and disadvantages of the two competing procedures.

122. Apply procedure (7.96) to the percentage correctly identified consonants data in Table 7.4 (Problem 12).

7.14 APPROXIMATE ONE-SIDED TREATMENTS-VERSUS-CONTROL MULTIPLE COMPARISONS BASED ON SIGNED RANKS (HOLLANDER)

In this section we turn our attention to a multiple comparison procedure based on the Wilcoxon signed rank statistics that is designed to make decisions about individual differences between the median effect for a single, baseline control population, and the median effects of each of the remaining $k - 1$ treatments. This treatments- versus-control multiple comparison procedure can be applied to two-way layout data (with one observation per cell) *after* rejection of H_0 (7.2) with either the Doksum–Lehmann or the Hollander procedure discussed in Sections 7.11 and 7.12, respectively. Its application leads to conclusions about the differences between each of the $k - 1$ treatment effects and the control effect and these conclusions are naturally one sided in nature.

Procedure

For simplicity of notation, we let treatment 1 assume the role of the single, baseline control. For each of the $k - 1$ treatments $u = 2, \ldots, k$, we compute the signed rank statistic T_{1u} (7.78) between the control treatment 1 and treatment u. At an approximate (see Comment 98) experimentwise error rate of α, the one-sided treatments-versus-control signed rank multiple comparison procedure reaches its $k - 1$ pairwise decisions, corresponding to each (τ_1, τ_u) pair, for $u = 2, \ldots, k$, by the criterion

$$\text{Decide } \tau_u > \tau_1 \text{ if } T_{1u} \geq t_\alpha^*; \quad \text{otherwise decide } \tau_u = \tau_1, \tag{7.99}$$

where the constant t_α^* is chosen to make the experimentwise error rate approximately equal to α; that is, t_α^* satisfies the restriction

$$P_0\{T_{1u} < t_\alpha^*, u = 2, \ldots, k\} \approx 1 - \alpha, \tag{7.100}$$

where the probability $P_0(.)$ is computed under H_0 (7.2). Equation (7.100) stipulates that the $k - 1$ inequalities $T_{1u} < t_\alpha^*$, corresponding to each treatment paired with the control, hold simultaneously with approximate probability $1 - \alpha$ when H_0 (7.2) is true. Selected approximate values of t_α^* can be found from the relationship

$$t_\alpha^* \approx \left[\frac{n(n+1)}{4}\right] + \left[\frac{n(n+1)(2n+1)}{24}\right]^{1/2} m_{\alpha,\rho^*}^*, \tag{7.101}$$

where m_{α,ρ^*}^* is the upper αth percentile point for the distribution of the maximum of $(k-1)N(0,1)$ variables with common correlation ρ^* equal to the upper bound ρ_U^n for

the null correlation between two overlapping signed rank statistics based on n observations. The upper bound ρ_U^n for signed rank statistics based on n observations is found from the R command `CorrUpperBound(n)`. To find m^*_{α,ρ^*} for k treatments and a specified experimentwise error rate α, we use the R command `cMaxCorrNor(`α`,k,rho.hat)`. For example, to find $m^*_{.02337,.3}$ for $k = 5$ treatments and correlation $\rho^* = .3$, we apply `cMaxCorrNor(.02337,5,.3)` and obtain $m^*_{.02337,.3} = 2.50$. (For a discussion of how to adjust procedure (7.99) for settings where it is of interest to decide whether the treatment effects are *smaller* than the control effect, see Comment 97.)

Ties

See Ties of Section 7.11 and Comment 88.

EXAMPLE 7.14 ***Effect of Weight on Forearm Tremor Frequency.***

We use the tremor data of Table 7.26 to illustrate procedure (7.99) using the approximation (7.101). We relabel the treatments so that the no-weight (0 lb) treatment assumes the role of the control. To make this clear in the ensuing computations, we reproduce Table 7.26 as Table 7.26′ with the new treatment designations.

We illustrate the one-sided decisions of $\tau_u = \tau_1$ versus $\tau_u < \tau_1$, $u = 2, \ldots, 5$. We see from Comment 97 that our procedure is based on (7.99) with $T_{u1} = [n(n+1)/2] - T_{1u}$ replacing T_{1u} in the left-hand side of the inequality in (7.99). From the relabeling in Table 7.26′ and the basic computations in Example 7.12, we obtain

$$T_{12} = 0, \quad T_{13} = 0, \quad T_{14} = 0, \quad \text{and} \quad T_{15} = 0,$$

which, in turn, implies

$$T_{21} = 21 - T_{12} = 21, \quad T_{31} = 21 - T_{13} = 21,$$

$$T_{41} = 21 - T_{14} = 21, \quad T_{51} = 21 - T_{15} = 21.$$

From the R command `CorrUpperBound(6)`, we find $\rho_U^6 = .452$. With an approximate experimentwise error rate of $\alpha = .10$, we then use the R command `cMaxCorrNor(.10,`

Table 7.26′ Forearm Tremor Frequency (Hz) as a Function of Weight Applied at the Wrist

Treatment	1	2	3	4	5
			Weight (lb)		
Subject	0	1.25	2.5	5	7.5
1	3.01	2.85	2.62	2.63	2.58
2	3.47	3.43	3.15	2.83	2.70
3	3.35	3.14	3.02	2.71	2.78
4	3.10	2.86	2.58	2.49	2.36
5	3.41	3.32	3.08	2.96	2.67
6	3.07	3.06	2.85	2.50	2.43

Source: J. R. Fox and J. E. Randall (1970).

5, .452) with $\rho^* = .452$ to find $m^*_{.10,.452} = 1.935$. Thus, with approximation (7.101), the inequality in (7.99) for our one-sided decisions reduces to

$$T_{u1} \geq t^*_{.10} \approx \left[\frac{6(7)}{4}\right] + \left[\frac{6(7)(13)}{24}\right]^{1/2}(1.935) = 19.73,$$

and procedure (7.99) for these one-sided decisions becomes

$$\text{Decide } \tau_u < \tau_1 \text{ if } T_{u1} \geq 19.73, \quad u = 2, \ldots, 5.$$

Since $T_{21} = T_{31} = T_{41} = T_{51} = 21 > 19.73$, the signed rank procedure (7.99) with an approximate .10 experimentwise error rate concludes that all four weight levels (treatments) yield significantly smaller forearm tremor frequencies than does the zero weight control.

Comments

97. *Opposite Direction Decisions.* Procedure (7.99) is designed for the one-sided situation in which the relevant decisions are $\tau_u = \tau_1$ versus $\tau_u > \tau_1, u = 2, \ldots, k$. To treat the analogous one-sided case of $\tau_u = \tau_1$ versus $\tau_u < \tau_1$, $u = 2, \ldots, k$, we simply replace T_{1u} by $T_{u1} = [n(n+1)/2] - T_{1u}$, in the left-hand side of the inequality in (7.99).

98. *Non-Distribution-Free Property*. Procedure (7.99), using approximation (7.101), is neither distribution-free nor asymptotically distribution-free. However, it is generally conservative in that the attained approximate experimentwise error rate associated with procedure (7.99) tends to be slightly lower than the nominally stipulated rate. Due to the closeness of the upper bound $\frac{7}{24}$ to all values of λ_F, procedure (7.99) is, for all practical purposes, virtually equivalent to Hollander's (1966) asymptotically distribution-free procedure based on estimating λ_F.

99. *Simplification of Approximation.* One of the disadvantages of the approximation to t^*_α provided in (7.101) is that it requires obtaining the value of m^*_{α,ρ^*} for common correlation $\rho^* = \rho^n_U$. To simplify matters, one could use the further approximation associated with replacing ρ^n_U by its asymptotic ($n \to \infty$) limit of $\frac{1}{2}$. The approximation in (7.101) would then use the proper value of $m^*_{\alpha,\frac{1}{2}}$ for common correlation $\rho^* = \frac{1}{2}$.

100. *Asymptotically Distribution-Free Competitor.* As an alternative to the conservative one-sided multiple comparison procedure (7.99) based on the use of the upper bound ρ^n_U in approximation (7.101), we could instead use a consistent estimator $\hat{\rho}$ of the null correlation between two overlapping signed rank statistics based on n observations. The value of m^*_{α,ρ^*} used in approximation (7.101) would then correspond to this estimate ($\hat{\rho}$) of the null correlation rather than the upper bound ρ^n_U. Use of a consistent estimator $\hat{\rho}$ in this manner leads to an asymptotically ($n \to \infty$) distribution-free one-sided multiple comparison procedure, rather than the conservative procedure in (7.99). Hollander (1966) suggested such an approach based on the consistent estimator $\hat{\rho} = 12\hat{\lambda}_F - 3$, where $\hat{\lambda}_F$ is defined in Comment 89.

101. *Two-Sided Treatments-versus-Control Multiple Comparison Procedure*. The multiple comparison procedure (7.99) of this section is one sided by nature, resulting in decisions between $\tau_u = \tau_1$ and $\tau_u > \tau_1$ for every $u = 2, \ldots, k$ (or between $\tau_u = \tau_1$ and $\tau_u < \tau_1$ for every $u = 2, \ldots, k$, as noted in Comment 97). We view such one-sided comparisons to be the most natural approach for treatments-versus-control settings. In such situations, we are generally interested in seeing which, if any, of the proposed new treatments are better than a standard control or placebo. In most practical applications, *better* is synonymous with one-sided comparisons (all in one direction or all in the other) and thus our emphasis on such procedures in this section. However, a two-sided treatments-versus-control analog to procedure (7.99) has been developed in the literature and corresponds to the criterion

$$\text{Decide } \tau_u \neq \tau_1 \text{ if } T'_{1u} \geq t^{**}_{\alpha}; \quad \text{otherwise decide } \tau_u = \tau_1, \tag{7.102}$$

where the T'_{1u}'s are defined by (7.95) and the constant t^{**}_{α} is chosen to make the experimentwise error rate approximately equal to α; that is,

$$P_0\{T'_{1u} < t^{**}_{\alpha}, u = 2, \ldots, k\} \approx 1 - \alpha,$$

where the probability $P_0(.)$ is computed under H_0 (7.2). One approximation for t^{**}_{α} sets

$$t^{**}_{\alpha} \approx \left[\frac{n(n+1)}{4}\right] + \left[\frac{n(n+1)(2n+1)}{24}\right]^{1/2} v^{*}_{\alpha}, \tag{7.103}$$

where v^{*}_{α} is the upper αth percentile of the maximum absolute value of $(k-1)N(0,1)$ random variables with common correlation $\frac{1}{2}$. Selected values of v^{*}_{α} can be obtained from Dunnett (1964).

102. *Independence from Observations for Other Noninvolved Treatments.* The value of T_{1u}, the statistic used in the decision relating to τ_u and τ_1, does not depend on the observation values from the other $k-2$ treatments. Thus, the signed ranks procedure (7.99) eliminates a difficulty encountered with the corresponding one-sided multiple comparison procedures (7.28) and (7.30) of Section 7.4 based on the Friedman rank sums. (See Comment 38.)

Properties

1. *Efficiency.* See Section 7.16.

Problems

123. Apply an appropriate one-sided signed rank multiple comparison procedure (see (7.99) and Comment 97) to the rhythmicity data of Table 7.6 in Problem 15, letting the condition *N* serve as the control.

124. Consider the serum CPK activity data from Problem 5. Treating preexercise as a control and ignoring the peak psychotic period data, apply procedure (7.99) to decide if there is statistical evidence of increased serum CPK activity either 19 or 42 h after exercise.

125. Both procedures (7.30) and (7.99) are appropriate multiple comparison procedures when we have data from a randomized complete block design with one observation per treatment–block combination and we are interested in one-sided comparisons between $(k-1)$ treatments and a single, baseline control. Discuss the relative advantages and disadvantages of the two competing procedures.

126. Treating condition A as a control, apply procedure (7.99) to the percentage correctly identified consonants data in Table 7.4 (Problem 12).

7.15 CONTRAST ESTIMATION BASED ON THE ONE-SAMPLE HODGES–LEHMANN ESTIMATORS (LEHMANN)

In this section we describe how to use the Hodges–Lehmann estimators based on the appropriate Walsh averages to construct estimators of a contrast θ (7.32 and 7.34) in the treatment effects $\tau_1, \ldots, \tau_k$. For a given setting, decisions about which contrasts to estimate can be related either to a priori interest in particular linear combinations of the τ's or to the result of one of the multiple comparison procedures discussed in Sections 7.3, 7.4, 7.13, and 7.14.

Procedure

Let θ be an arbitrary contrast (7.32 and 7.34) in the treatment effects $\tau_1, \ldots, \tau_k$. For each pair of treatments $(u, v), u \neq v = 1, \ldots, k$, compute the differences D_{uv}^i (7.36), $i = 1, \ldots, n$, between the treatment u and treatment v observations for each of the n blocks. For each (u, v) pair, obtain the values of the $n(n+1)/2$ Walsh averages for these sample differences, namely,

$$\frac{D_{uv}^i + D_{uv}^j}{2}, \quad 1 \leq i \leq j \leq n. \tag{7.104}$$

Let W_{uv} be the median of the Walsh averages associated with the $u - v$ treatment differences; that is,

$$W_{uv} = \text{median} \left\{ \frac{D_{uv}^i + D_{uv}^j}{2}, 1 \leq i \leq j \leq n \right\}, \quad u \neq v = 1, \ldots, k. \tag{7.105}$$

(Since $W_{vu} = -W_{uv}$, we need to calculate only the $k(k-1)/2$ values W_{uv} corresponding to $u < v$.) Note that each W_{uv} is a Hodges–Lehmann estimator of the form considered in Section 3.2, applied here to the $X_{iu} - X_{iv}$ differences. For example, W_{23} is the median of the $n(n+1)/2$ Walsh averages of the form $[D_{23}^i + D_{23}^j]/2$, $1 \leq i \leq j \leq n$, and can be viewed as an "unadjusted" estimator (see Comments 7.103 and 7.104) of the simple contrast $\tau_2 - \tau_3$.

Next, we compute

$$W_{u\cdot} = \sum_{j=1}^{k} \frac{W_{uj}}{k}, \quad u = 1, \ldots, k, \tag{7.106}$$

where we note that $W_{uu} = 0$ for $u = 1, \ldots, k$. Setting

$$\hat{\Delta}_{uv} = W_{u.} - W_{v.}, \tag{7.107}$$

the adjusted estimator of θ is given by

$$\hat{\theta} = \sum_{j=1}^{k} a_j W_{j.}, \tag{7.108}$$

or, equivalently,

$$\hat{\theta} = \sum_{h=1}^{k} \sum_{j=1}^{k} d_{hj} \hat{\Delta}_{hj}. \tag{7.109}$$

(See (7.35) for the relationship between the d's and the a's.)

EXAMPLE 7.15 ***Rounding First Base.***

In Example 7.5, we obtained the Doksum estimator of the contrast $\theta = \tau_{\text{roundout}} - \tau_{\text{wide angle}} = \tau_1 - \tau_3$ relating to the rounding-first-base data of Table 7.1. We now use (7.108) to obtain the Lehmann estimator of the same contrast. To evaluate W_{12}, defined by (7.105), note that $W_{12} = \text{median}\{[D_{12}^i + D_{12}^j]/2, 1 \le i \le j \le 22\}$, where D_{12}^i and D_{12}^j are defined by (7.36). The $D_{12}^1, \ldots, D_{12}^{22}$ values are exhibited in Table 7.11. Letting $F_{12}^{(1)} \le \cdots \le F_{12}^{(253)}$ denote the 253 ordered $[D_{12}^i + D_{12}^j]/2$ values, we find

$$F_{12}^{(1)} = -.4, \quad F_{12}^{(2)} = F_{12}^{(3)} = F_{12}^{(4)} = -.25 \quad F_{12}^{(5)} = F_{12}^{(6)} = -.225,$$

$$F_{12}^{(7)} = \cdots = F_{12}^{(10)} = -.2, \qquad F_{12}^{(11)} = \cdots = F_{12}^{(18)} = -.175,$$

$$F_{12}^{(19)} = F_{12}^{(20)} = -.15, \qquad F_{12}^{(21)} = -.125, F_{12}^{(22)} = \cdots = F_{12}^{(27)} = -.1,$$

$$F_{12}^{(28)} = \cdots = F_{12}^{(34)} = -.075, \qquad F_{12}^{(35)} = \cdots = F_{12}^{(49)} = -.05,$$

$$F_{12}^{(50)} = \cdots = F_{12}^{(81)} = -.025, \qquad F_{12}^{(82)} = \cdots = F_{12}^{(113)} = 0,$$

$$F_{12}^{(114)} = \cdots = F_{12}^{(152)} = .025, \qquad F_{12}^{(153)} = \cdots = F_{12}^{(198)} = .05,$$

$$F_{12}^{(199)} = \cdots = F_{12}^{(221)} = .075, \qquad F_{12}^{(222)} = \cdots = F_{12}^{(234)} = .1,$$

$$F_{12}^{(235)} = \cdots = F_{12}^{(240)} = .125, \qquad F_{12}^{(241)} = \cdots = F_{12}^{(249)} = .15,$$

$$F_{12}^{(250)} = F_{12}^{(251)} = .175, \qquad F_{12}^{(252)} = .2, F_{12}^{(253)} = .25.$$

Thus,

$$W_{12} = F_{12}^{(127)} = .025.$$

To evaluate W_{13}, we use the equation $W_{13} = \text{median}\ \{[D_{13}^i + D_{13}^j]/2, 1 \le i \le j \le 22\}$, where D_{13}^i and D_{13}^j are defined by (7.36). The $D_{13}^1, \ldots, D_{13}^{22}$ values are exhibited in

Table 7.11. Letting $F_{13}^{(1)} \le \cdots \le F_{13}^{(253)}$ denote the 253 ordered $[D_{13}^i + D_{13}^j]/2$ values, we have

$$F_{13}^{(1)} = -.3, \quad F_{13}^{(2)} = F_{13}^{(3)} = -.225, \qquad F_{13}^{(4)} = -.2, \quad F_{13}^{(5)} = -.175,$$
$$F_{13}^{(6)} = F_{13}^{(7)} = F_{13}^{(8)} = -.15, \qquad F_{13}^{(9)} = \cdots = F_{13}^{(12)} = -.125,$$
$$F_{13}^{(13)} = \cdots = F_{13}^{(19)} = -.1, \qquad F_{13}^{(20)} = \cdots = F_{13}^{(25)} = -.075,$$
$$F_{13}^{(26)} = \cdots = F_{13}^{(33)} = -.05, \qquad F_{13}^{(34)} = \cdots = F_{13}^{(46)} = -.025,$$
$$F_{13}^{(47)} = \cdots = F_{13}^{(62)} = 0, \qquad F_{13}^{(63)} = \cdots = F_{13}^{(77)} = .025,$$
$$F_{13}^{(78)} = \cdots = F_{13}^{(94)} = .05, \qquad F_{13}^{(95)} = \cdots = F_{13}^{(108)} = .075,$$
$$F_{13}^{(109)} = \cdots = F_{13}^{(131)} = .1, \qquad F_{13}^{(132)} = \cdots = F_{13}^{(157)} = .125,$$
$$F_{13}^{(158)} = \cdots = F_{13}^{(190)} = .15, \qquad F_{13}^{(191)} = \cdots = F_{13}^{(217)} = .175,$$
$$F_{13}^{(218)} = \cdots = F_{13}^{(238)} = .2, \qquad F_{13}^{(239)} = \cdots = F_{13}^{(247)} = .225,$$
$$F_{13}^{(248)} = \cdots = F_{13}^{(253)} = .25.$$

Thus,

$$W_{13} = F_{13}^{(127)} = .1.$$

In the same way, we calculate W_{23} by using the $D_{23}^1, \ldots, D_{23}^{22}$ values in Table 7.11 and the fact that $W_{23} = \text{median } \{[D_{23}^i + D_{23}^j]/2, 1 \le i \le j \le 22\}$. Letting $F_{23}^{(1)} \le \cdots \le F_{23}^{(253)}$ denote the 253 ordered $[D_{23}^i + D_{23}^j]/2$ values, we see that

$$F_{23}^{(1)} = -.1, F_{23}^{(2)} = \cdots = F_{23}^{(5)} = -.075, \qquad F_{23}^{(6)} = \cdots = F_{23}^{(15)} = -.05,$$
$$F_{23}^{(16)} = \cdots = F_{23}^{(19)} = -.025, \qquad F_{23}^{(20)} = \cdots = F_{23}^{(42)} = 0,$$
$$F_{23}^{(43)} = \cdots = F_{23}^{(73)} = .025, \qquad F_{23}^{(74)} = \cdots = F_{23}^{(98)} = .05,$$
$$F_{23}^{(99)} = \cdots = F_{23}^{(138)} = .075, \qquad F_{23}^{(139)} = \cdots = F_{23}^{(178)} = .1,$$
$$F_{23}^{(179)} = \cdots = F_{23}^{(211)} = .125, \qquad F_{23}^{(212)} = \cdots = F_{23}^{(238)} = .15,$$
$$F_{23}^{(239)} = \cdots = F_{23}^{(247)} = .175, \qquad F_{23}^{(248)} = \cdots = F_{23}^{(253)} = .2.$$

Thus,

$$W_{23} = F_{23}^{(127)} = .075.$$

From (7.106), we find

$$W_{1.} = \frac{W_{11} + W_{12} + W_{13}}{3}$$
$$= \frac{0 + .025 + .1}{3} = .0417,$$

$$W_{2.} = \frac{W_{21} + W_{22} + W_{23}}{2}$$
$$= \frac{-.025 + 0 + .075}{3} = 0.167,$$

and

$$W_{3.} = \frac{W_{31} + W_{32} + W_{33}}{3} = \frac{-.1 - .075 + 0}{3} = -.0583.$$

Note that in calculating $W_{2.}$ and $W_{3.}$, we use the relationship $W_{uv} = -W_{vu}$.

The Lehmann estimator $\hat{\theta}$ is now obtained from (7.108) by noting that $a_1 = 1$, $a_2 = 0$, and $a_3 = -1$, so that

$$\hat{\theta} = W_{1.} - W_{3.} = .0417 - (-.0583) = .10.$$

For these data, the adjusted estimator $W_{1.} - W_{3.}$ agrees with the unadjusted estimator W_{13}. However, we do note that the value of the Lehmann estimator $\hat{\theta} = .10$ differs from that of the Doksum estimator $\tilde{\theta} = .133$ (see Example 7.5) for these rounding-first-base data.

Comments

103. *Unadjusted Estimator*. The unadjusted estimator W_{uv} (7.105) of $\Delta_{uv} = \tau_u - \tau_v$ is simply the estimator associated with the signed rank test and discussed in Section 3.2.
104. *Ambiguities with the Unadjusted Estimators.* The unadjusted estimators W_{uv} (7.105) are incompatible, leading to possible ambiguities in contrast estimation because they do not satisfy the linear relations that are satisfied by the contrasts they estimate. We have encountered this difficulty before (see Comments 6.77 and 7.42). The adjusted estimators $\hat{\Delta}_{uv}$ (7.107) are compatible but have the disadvantage that the estimator of $\Delta_{uv} = \tau_u - \tau_v$ depends on the observations from the other $k - 2$ treatments.
105. *Computational Difficulty*. Example 7.15 is a glaring illustration of the labor involved in computing W_{uv} when n is moderately large. It is necessary to obtain the median of $n(n + 1)/2$ Walsh averages, whereas the estimator Z_{uv} (7.37) is based on the median of only n differences. Thus, Doksum's contrast estimator is preferred to Lehmann's contrast estimator in terms of ease of computation. On the other hand, asymptotic efficiencies generally (but not always) favor Lehmann's contrast estimator. (See Section 7.16.)

Properties

1. *Standard Deviation of $\hat{\theta}$ (7.108).* For the asymptotic standard deviation of $\hat{\theta}$ (7.108), see Lehmann (1964).
2. *Asymptotic Normality*. See Lehmann (1964).
3. *Efficiency.* See Lehmann (1964) and Section 7.16.

Problems

127. Calculate the Lehmann estimator of the contrast $2\tau_N - \tau_A - \tau_R$ for the metronome data of Table 7.6. Compare with the Doksum estimator from Problem 46.

128. Give an example illustrating the incompatibility of the unadjusted estimators W_{uv} (7.105). (See Comment 104.)

129. Calculate the Lehmann estimators for the simple contrasts $\theta_1 = \tau_2 - \tau_1$, $\theta_2 = \tau_3 - \tau_1$, and $\theta_3 = \tau_3 - \tau_2$ for the CPK activity data in Table 7.3.

130. Estimate the contrast $3\tau_{\mathrm{ALC}} - \tau_{\mathrm{AL}} - \tau_{\mathrm{AC}} - \tau_{\mathrm{LC}}$ for the percentage consonants correctly identified data in Table 7.4.

131. Using the data of Table 7.4, obtain Lehmann's estimator of the simple contrast that represents the benefit from adding lip reading to audition in teaching severely hearing-impaired children. Compare with the Doksum estimator from Problem 50.

132. Compute Lehmann's estimator for all contrasts found to be of interest in Problem 45 for the maximum soil temperature data in Table 7.8.

133. Calculate Lehmann's estimator of the contrast $\tau_{\mathrm{rats}} - \tau_{\mathrm{cats}}$ for the Livesey EPT error score of Table 7.9.

7.16 EFFICIENCIES OF TWO-WAY LAYOUT PROCEDURES

We first consider the procedures of Sections 7.1–7.5, which are associated with the Friedman rank sums for the case of one observation per treatment–block combination (i.e., a randomized complete block design). The Pitman asymptotic relative efficiencies (for translation alternatives) of these procedures with respect to the corresponding normal theory counterparts are given by the expression

$$e_F = \left[\frac{k}{(k+1)}\right]\left[12\sigma_F^2\left\{\int_{-\infty}^{\infty} f^2(u)du\right\}^2\right], \tag{7.110}$$

where σ_F^2 is the variance of the common underlying (continuous) distribution F (7.1) and $f(.)$ is the probability density function corresponding to F. The parameter $\int_{-\infty}^{\infty} f^2(u)du$ is the area under the curve associated with $f^2(.)$, the square of the common probability density function. We note that e_F (7.110) is simply $k/(k+1)$ times the corresponding Pitman efficiencies in the one-sample, two-sample, and k-sample location settings (see Sections 3.11, 4.5, and 6.10).

In particular, the Pitman asymptotic relative efficiency of the Friedman test based on S (7.5) with respect to the normal theory two-way layout F test was found to be e_F (7.110) by van Elteren and Noether (1959). The asymptotic relative efficiency of the Page test for ordered alternatives, based on the statistic L (7.10), with respect to a suitable normal theory competitor was found by Hollander (1967a) to be e_F (7.110) as well. Furthermore, methods analogous to those of Sherman (1965) lead to expression (7.110) as the asymptotic relative efficiency of both the all-treatments two-sided and the treatments-versus-control one-sided multiple comparison procedures in Sections 7.3 and 7.4, respectively, with respect to the classical normal theory procedures based on sample means. Finally, Doksum (1967) obtained (7.110) as the asymptotic relative efficiency of

the estimator $\tilde{\theta}$ (7.40) with respect to the least-squares estimator $\bar{\theta} = \sum_{j=1}^{k} a_j X_{.j}$, where $X_{.j} = \sum_{i=1}^{n} X_{ij}/n$.

The efficiency e_F (7.110) is always greater than or equal to .576 and it can be infinite. Some values of e_F for various F and k combinations are given in Table 7.27.

We next turn to the procedures in Sections 7.6–7.8 that are designed for two-way layout data with zero or one observation per treatment–block combination. The Pitman asymptotic relative efficiency of the Durbin–Skillings–Mack test based on D (7.43) with respect to the standard normal theory procedure for a balanced incomplete block design was found to be e_F (7.110) by van Elteren and Noether (1959). Once again, methods analogous to those of Sherman (1965) lead to expression (7.110) as the asymptotic relative efficiency of the all-treatments two-sided multiple comparison procedures in Section 7.7. We do not know of any results for the asymptotic relative efficiencies of the general alternatives Skillings–Mack test in Section 7.8 for data from an arbitrary incomplete block design.

For the case of two-way layout data with at least one observation for every treatment–block combination, Mack and Skillings (1980) found that under certain conditions the asymptotic relative efficiency of their test for general alternatives based on the statistic MS (7.57) with respect to a suitable normal theory competitor is, once again, given by e_F (7.110). Combining their results with methods analogous to those of Sherman (1965) yields expression (7.110) as the asymptotic relative efficiency of the all-treatments two-sided multiple comparison procedures in Section 7.10, as well.

Finally, we turn to the procedures in Sections 7.11–7.15 which are associated with Wilcoxon signed ranks. The asymptotic relative efficiencies of Doksum's conservative test of Section 7.11, based on replacing λ_F by its upper bound $\frac{7}{24}$, are very close to those of a related test proposed by Doksum (1967) in which λ_F is estimated. The expression for the asymptotic relative efficiency e_F^* of the related test, relative to the normal theory $\mathcal{F}$-test, is given by the right-hand side of (2.12) in Doksum (1967). The parameter e_F^* is always greater than .864 and can be infinite. In Table 7.28, we provide values of e_F^* for normal, uniform and exponential distributions and various numbers (k) of treatments. Similarly, the efficiencies of Hollander's conservative test of Section 7.12, based on replacing λ_F by its upper bound $\frac{7}{24}$, are very close to those of a related test proposed by Hollander (1967a), in which λ_F is estimated. The expression for the asymptotic

Table 7.27 Values of e_F for Various Distributions and Numbers (k) of Treatments

k / Distribution	2	3	4	5	10	20	50	∞
				e_F				
Normal	0.637	0.716	0.764	0.796	0.868	0.909	0.936	0.955
Uniform	0.667	0.750	0.800	0.833	0.909	0.952	0.980	1.000
Double exponential	1.000	1.125	1.200	1.250	1.364	1.429	1.471	1.500

Table 7.28 Values of e_F^* for Various Distributions and Numbers (k) of Treatments

k / Distribution	2	3	4	5	10	20	50	∞
				e_F^*				
Normal	0.955	0.966	0.972	0.975	0.983	0.987	0.989	0.990
Uniform	0.889	0.894	0.897	0.899	0.902	0.904	0.905	0.906
Exponential	1.500	1.528	1.543	1.552	1.570	1.579	1.585	1.588

Table 7.29 Values of e_F^{**} for Various Distributions and Numbers (k) of Treatments

k	2	3	4	5	10	20	50	∞
Distribution				e_F^{**}				
Normal	0.955	0.963	0.969	0.972	0.980	0.985	0.988	0.990
Uniform	0.889	0.893	0.895	0.897	0.901	0.903	0.905	0.906
Exponential	1.500	1.521	1.534	1.543	1.563	1.575	1.583	1.588

relative efficiency e_F^{**} of this related test, with respect to a normal theory t-test for ordered alternatives, is given by the right-hand side of (4.6) of Hollander (1967a). The parameter e_F^{**} is always greater than .864 and can be infinite. In Table 7.29, we provide values of e_F^{**} for normal, uniform, and exponential distributions and various numbers (k) of treatments. The efficiencies in Table 7.29 are also close approximations to the efficiencies of the conservative multiple comparison procedures of Sections 7.13 and 7.14 with respect to normal theory competitors based on sample means. Lehmann (1964) obtained the asymptotic relative efficiency (for translation alternatives) of the contrast estimator (7.108) of Section 7.15 with respect to the least-squares estimator based on the sample means. The asymptotic relative efficiency is given by e_F^* (see Table 7.28).

Chapter 8

The Independence Problem

INTRODUCTION

The data in this chapter consist of a random sample from a bivariate population. Our basic interest here is in the statistical relationship between the two variables involved in the bivariate structure. In particular, we will discuss procedures for deciding whether or not these two variables are independent and, if not independent, for assessing both the type and degree of dependency that exists between them.

In Section 8.1, we present a distribution-free test for independence that is based on signs of appropriate products of differences. Section 8.2 presents an estimator of the measure of association τ defined by (8.2). Section 8.3 contains an asymptotically distribution-free confidence interval for τ. Section 8.4 uses Efron's bootstrap method to obtain a different asymptotically distribution-free confidence interval for τ. Section 8.5 presents a distribution-free test for independence based on ranks. Section 8.6 contains a distribution-free test of independence, which is consistent against a broader class of alternatives than those classes of alternatives that can be detected by the tests of Sections 8.1 and 8.5. Section 8.7 considers the asymptotic relative efficiencies of the procedures in this chapter with respect to their normal theory counterparts.

Data. We obtain n bivariate observations $(X_1, Y_1), \ldots, (X_n, Y_n)$, one observation on each of n subjects.

Assumptions

A The n bivariate observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are a random sample from a continuous bivariate population. That is, the (X, Y) pairs are mutually independent and identically distributed according to some continuous bivariate population.

8.1 A DISTRIBUTION-FREE TEST FOR INDEPENDENCE BASED ON SIGNS (KENDALL)

Hypothesis

Let $F_{X,Y}$ be the joint distribution function for the common bivariate population of the (X, Y) pairs. Moreover, let $F_x(x)$ and $F_Y(y)$ be the distribution functions for the marginal

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

X and Y populations, respectively. The null hypothesis of interest here is that the X and Y random variables are independent. Formally stated, this null hypothesis is

$$H_0 : [F_{X,Y}(x,y) \equiv F_X(x)F_Y(y), \text{ for all } (x,y) \text{ pairs}]. \tag{8.1}$$

The alternative hypothesis to (8.1) will be a function of the type of dependence between the X and Y variables that is of principal interest. In this section, we concentrate on a type of dependence measured by the Kendall population correlation coefficient

$$\tau = 2P\{(Y_2 - Y_1)(X_2 - X_1) > 0\} - 1. \tag{8.2}$$

We note that the event $\{(Y_2 - Y_1)(X_2 - X_1) > 0\}$ occurs if and only if either the event $\{X_2 > X_1 \text{ and } Y_2 > Y_1\}$ or the event $\{X_2 < X_1 \text{ and } Y_2 < Y_1\}$ occurs. These latter two events are mutually exclusive, therefore

$$\begin{aligned} P\{(Y_2 - Y_1)(X_2 - X_1) > 0\} = {} & P(X_2 > X_1, Y_2 > Y_1) \\ & + P(X_2 < X_1, Y_2 < Y_1). \end{aligned} \tag{8.3}$$

If X and Y are independent, it follows that

$$P(X_2 > X_1, Y_2 > Y_1) = P(X_2 > X_1)P(Y_2 > Y_1) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{4}, \tag{8.4}$$

because X_1, X_2 are independent and identically distributed variables, as are Y_1, Y_2 (although not necessarily, of course, with the same distribution as the X's). Similarly, if X and Y are independent, we also have

$$P(X_2 < X_1, Y_2 < Y_1) = \frac{1}{4}.$$

Combining this result with (8.3) and (8.4), we see that the Kendall population correlation coefficient $\tau = 2\left(\frac{1}{4} + \frac{1}{4}\right) - 1 = 0$ if X and Y are independent. (It is important to point out that this is not an if and only if statement because $\tau = 0$ does not necessarily imply that X and Y are independent. See Comment 2 for more on this relationship.)

Procedure

To compute the Kendall sample correlation statistic K, we first calculate the values of the $n(n-1)/2$ paired sign statistics $Q((X_i, Y_i), (X_j, Y_j))$, for $1 \le i < j \le n$, where

$$Q((a,b),(c,d)) = \begin{cases} 1, & \text{if } (d-b)(c-a) > 0, \\ -1, & \text{if } (d-b)(c-a) < 0. \end{cases} \tag{8.5}$$

That is, for each pair of subscripts (i,j) with $i < j$, score 1 if $(Y_j - Y_i)(X_j - X_i)$ is positive and score -1 if $(Y_j - Y_i)(X_j - X_i)$ is negative. The Kendall statistic K is then

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q((X_i, Y_i), (X_j, Y_j)), \tag{8.6}$$

corresponding to adding up the 1's and -1's from the paired sign statistics.

a. *One-Sided Upper-Tail Test.* To test the null hypothesis of independence, namely,

$$H_0 : [F_{X,Y}(x, y) \equiv F_X(x)F_Y(y), \text{ for all}(x, y) \text{ pairs}]$$

(which implies $\tau = 0$) versus the alternative that X and Y are positively correlated (see Comment 2) corresponding to

$$H_1 : \tau > 0, \tag{8.7}$$

at the α-level of significance,

$$\text{Reject } H_0 \text{ if } \overline{K} \geq k_\alpha; \quad \text{otherwise do not reject,} \tag{8.8}$$

where the constant k_α is chosen to make the type I error probability equal to α and $\overline{K} = K/(n(n-1)/2)$, the average of the paired sign statistics Q. Values of k_α are found using the command `qKendall` (Wheeler (2009)).

b. *One-Sided Lower-Tail Test.* To test

$$H_0 : [F_{X,Y}(x, y) \equiv F_X(x)F_Y(y), \text{ for all } (x, y) \text{ pairs}]$$

versus the alternative that X and Y are negatively correlated (see Comment 2) corresponding to

$$H_2 : \tau < 0,$$

at the α-level of significance,

$$\text{Reject } H_0 \text{ if } \overline{K} \leq -k_\alpha; \quad \text{otherwise do not reject.} \tag{8.9}$$

c. *Two-Sided Test.* To test

$$H_0 : [F_{X,Y}(x, y) \equiv F_X(x)F_Y(y), \text{ for all } (x, y) \text{ pairs}]$$

versus the general alternative that X and Y are dependent variables corresponding to

$$H_3 : \tau \neq 0,$$

at the α-level of significance,

$$\text{Reject } H_0 \text{ if } |\overline{K}| \geq k_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{8.10}$$

This two-sided procedure is the two-sided symmetric test with $\alpha/2$ probability in each tail of the null distribution of $\overline{K}$.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of K, suitably standardized. For this standardization, we need to know the expected value and variance of K when the null hypothesis of independence is true. Under H_0, the expected value and variance of K are

$$E_0(K) = 0 \tag{8.11}$$

and

$$\text{var}_0(K) = \frac{n(n-1)(2n+5)}{18}, \tag{8.12}$$

respectively. These expressions for $E_0(K)$ and $\text{var}_0(K)$ are verified by direct calculations in Comment 7 for the special case of $n = 4$. General derivations of both expressions are presented in Comment 10.

The standardized version of K is

$$K^* = \frac{K - E_0(K)}{\{\text{var}_0(K)\}^{1/2}} = \frac{K}{\{n(n-1)(2n+5)/18\}^{1/2}}. \tag{8.13}$$

When H_0 is true, K^* has, an n tends to infinity, an asymptotic $N(0,1)$ distribution (see Comment 10 for indications of the proof). The normal theory approximation for procedure (8.8) is

$$\text{Reject } H_0 \text{ if } K^* \geq z_\alpha; \quad \text{otherwise do not reject,} \tag{8.14}$$

the normal theory approximation for procedure (8.9) is

$$\text{Reject } H_0 \text{ if } K^* \leq -z_\alpha; \quad \text{otherwise do not reject,} \tag{8.15}$$

and the normal theory approximation for procedure (8.10) is

$$\text{Reject } H_0 \text{ if } |K^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{8.16}$$

Ties

If there are ties among the n X observations and/or separately among the n Y observations, replace the function $Q((a,b),(c,d))$ in the definition of K (8.6) by

$$Q^*((a,b),(c,d)) = \begin{cases} 1, & \text{if } (d-b)(c-a) > 0, \\ 0, & \text{if } (d-b)(c-a) = 0, \\ -1, & \text{if } (d-b)(c-a) < 0. \end{cases} \tag{8.17}$$

(Thus, in the case of tied X values and/or tied Y values, zeros are assigned to the associated paired sign statistics.) After computing K with these modified paired sign statistics, use procedure (8.8), (8.9), or (8.10). Note, however, that this test associated with tied X's and/or Y's is only approximately, and not exactly, of significance level α.

When applying the large-sample approximation, however, the loss in variability due to the tied X's and/or tied Y's must also be taken into account. While these ties do not affect the null expected value of K, its null variance is reduced to

$$\begin{aligned}\text{var}_0(K) = &\frac{\left\{n(n-1)(2n+5) - \sum_{i=1}^{g} t_i(t_i-1)(2t_i+5) - \sum_{j=1}^{h} u_j(u_j-1)(2u_j+5)\right\}}{18} \\ &+ \frac{\left\{\sum_{i=1}^{g} t_i(t_i-1)(t_i-2)\right\}\left\{\sum_{j=1}^{h} u_j(u_j-1)(u_j-2)\right\}}{9n(n-1)(n-2)} \\ &+ \frac{\left\{\sum_{i=1}^{g} t_i(t_i-1)\right\}\left\{\sum_{j=1}^{h} u_j(u_j-1)\right\}}{2n(n-1)} \end{aligned} \tag{8.18}$$

in the presence of such ties, where in (8.18) g denotes the number of tied X groups, t_i is the size of tied X group i, h is the number of tied Y groups, and u_j is the size of tied Y group j. We note that an untied $X(Y)$ observation is considered to be a tied $X(Y)$ "group" of size 1. In particular, if neither the collection of n X nor the collection of n Y observations contains tied observations, we have $g = h = n$, $t_i = u_j = 1$, $i = 1, \ldots, n$, and $j = 1, \ldots, n$. In this case of no tied X's and no tied Y's, each term involving either $(t_i - 1)$ or $(u_j - 1)$ or both reduces to zero and the variance expression in (8.18) reduces to the usual null variance of K, as given previously in (8.12).

As a consequence of the effect that ties have on the null variance of K, the following modification is needed to apply the large-sample approximation when there are tied X observations and/or tied Y observations. Compute K with the modified paired sign statistic using (8.17) and set

$$K^* = \frac{K}{\{\text{var}_0(K)\}^{1/2}}, \tag{8.19}$$

where $\text{var}_0(K)$ is now given by display (8.18). With this modified form of K^*, approximation (8.14), (8.15), or (8.16) can be applied.

EXAMPLE 8.1 ***Tuna Lightness and Quality.***

The data in Table 8.1 are a subset of the data obtained by J. Rasekh, A. Kramer, and R. Finch (1970) in a study designed to ascertain the relative importance of the various factors contributing to tuna quality and to find objective methods for determining quality parameters and consumer preference. Table 8.1 gives values of the Hunter L measure of lightness, along with panel scores for nine lots of canned tuna. The original consumer panel scores of excellent, very good, good, fair, poor, and unacceptable were converted to the numerical values of 6, 5, 4, 3, 2, and 1, respectively. The panel scores in Table 8.1 are averages of 80 such values. (The Y random variable is thus discrete, and hence, the continuity portion of Assumption A is not satisfied. Nevertheless, because each Y is an average of 80 values, we need not be nervous about this departure from Assumption A.)

It is suspected that the Hunter L value is positively associated with the panel score. Thus, we will apply procedure (8.8) to test H_0 (8.1) versus $\tau > 0$. Consider the significance level $\alpha = .10$. Using `qKendall(p=.10, N=9, lower.tail=T)` gives a value of $-.333$. By the symmetry of $\overline{K}$, the critical value is therefore .333.

Table 8.1 Hunter L Values and Consumer Panel Scores for Nine Lots of Canned Tuna

Lot	Hunter L value (X)	Panel score (Y)
1	44.4	2.6
2	45.9	3.1
3	41.9	2.5
4	53.3	5.0
5	44.7	3.6
6	44.1	4.0
7	50.7	5.2
8	45.2	2.8
9	60.1	3.8

Source: J. Rasekh, A. Kramer, and R. Finch (1970).

Table 8.2 $Q((X_i, Y_i), (X_j, Y_j))$ Values for Canned Tuna Data

$j \backslash i$	1	2	3	4	5	6	7	8
2	1							
3	1	1						
4	1	1	1					
5	1	−1	1	1				
6	−1	−1	1	1	−1			
7	1	1	1	−1	1	1		
8	1	1	1	1	−1	−1	1	
9	1	1	1	−1	1	−1	−1	1

We illustrate the computations of the paired sign statistics in (8.5) leading to the sample value of K (8.6) in Table 8.2.

Summing the $+1$ and -1 values in Table 8.2 we see that

$$K = \sum_{i=1}^{8} \sum_{j=i+1}^{9} Q((X_i, Y_i), (X_j, Y_j)) = 26 - 10 = 16,$$

and $\overline{K} = 16/36$.

This value of $\overline{K}$ is greater than the critical value .333, so we reject H_0 in favor of $\tau > 0$ at the $\alpha = .10$ level. Note that the critical value given by R results in a significance level of $\alpha = .13$, not $\alpha = .10$.

Since the one-sided P-value for these data is the lowest significance level at which we can reject H_0 in favor of $\tau > 0$ with the observed value of the test statistic $\overline{K} = 16/36$. The P-value for these data is $P_0(\overline{K} \geq 16/36) = P_0(\overline{K} \leq -16/36)$. The P-value is `pKendall(-16/36, N=9, lower.tail=T) =.060`. Thus, there is some evidence (although not overwhelming) that the Hunter L lightness values and the panel scores are positively correlated.

The R command `cor.test` will perform this test without the need to use `qKendall` or `pKendall`. The analysis above can be replicated by

```
cor.test (x, y, method="kendall", alt="greater")
```

where `x` is the Hunter L value and `y` is the panel score from Table 8.1. This results in the output

```
 Kendall's rank correlation tau

data: x and y
T = 26, p-value = 0.05972
alternative hypothesis: true tau is greater than 0
sample estimates:
tau
0.4444444.
```

The value of the test statistic is given here is $T = 26$. R sums (8.6) only over those values of Q giving a positive 1. Since the number of pairs is $n(n-1)/2$, there must be $n(n-1)/2 - T$ negative 1 values in (8.6). To convert T to K, one uses the relation $K = 2T - n(n-1)/2$. For the data in Table 8.1, $T = 26$ is equivalent to $K = 2 \cdot 26 - 9 \cdot 8/2 = 16$.

For the large-sample approximation, we find (since there are no ties in the data) from (8.13) that

$$K^* = \frac{16}{\{9(8)(23)/18\}^{1/2}} = 1.67.$$

Thus, the smallest significance level at which we can reject H_0 in favor of $\tau > 0$ using the normal theory approximation is .0475, since $z_{.0475} = 1.67$. This is in good agreement with the exact P-value of .060 found previously.

Comments

1. *Motivation for the Test.* The null hypothesis of this section is that the X and Y random variables are independent, which implies (see the discussion in Procedure) that the Kendall population correlation coefficient τ is equal to 0. However, the alternatives are stated directly in terms of $\tau (>, <, \text{or } \neq 0)$. When τ is greater than 0 (and thus $P((Y_2 - Y_1)(X_2 - X_1) > 0) > \frac{1}{2}$), there will tend to be a large number of positive paired sign statistics and fewer negative paired sign statistics. Hence, when τ is greater than 0, we would expect the sample to lead to a big, positive value for K. This suggests rejecting H_0 in favor of $\tau > 0$ for large values of K and motivates procedures (8.8) and (8.14). Similar rationales lead to procedures (8.9), (8.10), (8.15), and (8.16).
2. *Interpretation of* τ. The Kendall correlation coefficient τ can also be written as $\tau = [P((Y_2 - Y_1)(X_2 - X_1) > 0) - P((Y_2 - Y_1)(X_2 - X_1) < 0)]$. We have already noted that if X and Y are independent, then $\tau = 0$. On the other hand, if $\tau > 0$, then it is more likely that $\{X_2 > X_1 \text{ and } Y_2 > Y_1\}$ or $\{X_2 < X_1 \text{ and } Y_2 < Y_1\}$ occurs than either of the complementary events $\{X_2 > X_1 \text{ and } Y_2 < Y_1\}$ or $\{X_2 < X_1 \text{ and } Y_2 > Y_1\}$. Thus, if $\tau > 0$, it is more likely that the change from X_1 to X_2 has the same (rather than opposite) sign as that from Y_1 to Y_2. It is reasonable to interpret this type of relationship between X and Y as indicative of a positive association (as measured by τ). Similarly, $\tau < 0$ may reasonably be interpreted as indicative of a negative association (as measured by τ) between X and Y.
3. *Concordant/Discordant Pairs.* Call the $(X_i, Y_i), (X_j, Y_j)$ pairs *concordant* if $(X_i - X_j)(Y_i - Y_j) > 0$ and *discordant* if $(X_i - X_j)(Y_i - Y_j) < 0$. Thus, (X_i, Y_i) and (X_j, Y_j) are concordant if either (a) $X_i > X_j$ and $Y_i > Y_j$ or (b) $X_i < X_j$ and $Y_i < Y_j$. Similarly, (X_i, Y_i) and (X_j, Y_j) are discordant if either (c) $X_i < X_j$

and $Y_i > Y_j$ or (d) $X_i > X_j$ and $Y_i < Y_j$. Now K (8.6) can be expressed as $K = K' - K''$, where

$$K' = \text{number of concordant pairs},$$

$$K'' = \text{number of discordant pairs},$$

and the count is taken over the $n(n-1)/2$ sets of pairs $(X_i, Y_i), (X_j, Y_j)$ with $i < j$. Note that $(X_i, Y_i), (X_j, Y_j)$ are concordant if the ordering of X_i, X_j agrees with that of Y_i, Y_j. We have discordance when these orderings do not agree. Thus, $K/\{n(n-l)/2\}$ can be viewed as an average measure of agreement between the X's and the Y's, where agreement refers to order.

4. *Equivalent Expression When There Are No Ties.* Let K' = (number of concordant pairs) and K'' = (number of discordant pairs), as defined in Comment 3. If there are no ties among the X's and no ties among the Y's, then $K' + K'' = n(n-1)/2$. Thus, with no ties, we have $K = K' - K'' = K' - [n(n-1)/2 - K'] = 2K' - \{n(n-1)/2\}$. To illustrate, consider the tuna data in Example 8.1. Summing the 1's in Table 8.2 (corresponding to concordant pairs), we obtain $K' = 7 + 5 + 6 + 3 + 2 + 1 + 1 + 1 = 26$. Adding the 0's in Table 8.2 (corresponding to discordant pairs), we have $K'' = 1 + 2 + 0 + 2 + 2 + 2 + 1 + 0 = 10$. It follows that $K = K' - K'' = 26 - 10 = 2K' - \{n(n-1)/2\} = [2(26) - 9(8)/2] = 16$, in agreement with the value obtained directly in Example 8.1.

5. *Convenience Through Ordering.* It is convenient to compute the number of concordant pairs, K' by first rearranging the (X_i, Y_i) pairs so that the (new) X's are in increasing order. Then, after rearrangement, K' is equal to the number of pairs for which the corresponding Y's are in increasing order. For example, suppose our observations are

i	1	2	3	4	5
X_i	4.1	−2.4	−2.2	−5.6	5.5
Y_i	2.3	3.7	1.1	2.2	3.8

We arrange these so that the X's are in increasing order and obtain the following:

X	−5.6	−2.4	−2.2	4.1	5.5
Y	2.2	3.7	1.1	2.3	3.8

Then, proceeding from left to right, we find the Y pairs that are in increasing order to be (2.2, 3.7), (2.2, 2.3), (2.2, 3.8), (3.7, 3.8), (1.1, 2.3). (1.1, 3.8), and (2.3, 3.8). Thus, $K' = 7$ and $K = 2K' - \{5(4)/2\} = 4$.

6. *Derivation of Distribution of K under H_0 (No-Ties Case).* Let R_i be the rank X_i in the joint ranking of $X_1, \ldots, X_n$ and let S_i be the rank of Y_i in the joint ranking of $Y_1, \ldots, Y_n$. It is clear that knowledge of the R's and S's is sufficient to calculate K (8.6). (See Problem 2.) We use this fact to illustrate how the null distribution of K can be obtained. Without loss of generality, we take

$R_1 = 1, \ldots, R_n = n$; then, as under H_0 (8.1), all possible $n!$ $(S_1, S_2, \ldots, S_n)$ Y-rank configurations are equally likely, implying each has probability $(1/n!)$.

Let us consider the case $n = 4$. In the following table, we display the $4! = 24$ possible (S_1, S_2, S_3, S_4) configurations, the associated values of K, and the corresponding null probabilities.

(R_1, R_2, R_3, R_4)	(S_1, S_2, S_3, S_4)	Null probability	K
(1, 2, 3, 4)	(1, 2, 3, 4)	$\frac{1}{24}$	6
(1, 2, 3, 4)	(1, 2, 4, 3)	$\frac{1}{24}$	4
(1, 2, 3, 4)	(1, 3, 2, 4)	$\frac{1}{24}$	4
(1, 2, 3, 4)	(1, 3, 4, 2)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(1, 4, 2, 3)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(1, 4, 3, 2)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(2, 1, 3, 4)	$\frac{1}{24}$	4
(1, 2, 3, 4)	(2, 1, 4, 3)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(2, 3, 1, 4)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(2, 3, 4, 1)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(2, 4, 1, 3)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(2, 4, 3, 1)	$\frac{1}{24}$	−2
(1, 2, 3, 4)	(3, 1, 2, 4)	$\frac{1}{24}$	2
(1, 2, 3, 4)	(3, 1, 4, 2)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(3, 2, 1, 4)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(3, 2, 4, 1)	$\frac{1}{24}$	−2
(1, 2, 3, 4)	(3, 4, 1, 2)	$\frac{1}{24}$	−2
(1, 2, 3, 4)	(3, 4, 2, 1)	$\frac{1}{24}$	−4
(1, 2, 3, 4)	(4, 1, 2, 3)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(4, 1, 3, 2)	$\frac{1}{24}$	−2
(1, 2, 3, 4)	(4, 2, 1, 3)	$\frac{1}{24}$	−2
(1, 2, 3, 4)	(4, 2, 3, 1)	$\frac{1}{24}$	−4
(1, 2, 3, 4)	(4, 3, 1, 2)	$\frac{1}{24}$	−4
(1, 2, 3, 4)	(4, 3, 2, 1)	$\frac{1}{24}$	−6

Thus, for example, the probability is $\frac{5}{24}$ under H_0 that K is equal to 2, because $K = 2$ when any of the five outcomes $(S_1, S_2, S_3, S_4) = (1, 3, 4, 2)$, (1, 4, 2, 3), (2, 1, 4, 3), (2, 3, 1,4), or (3, 1, 2, 4) occurs and each of these outcomes has null probability $\frac{1}{24}$. Simplifying, we obtain the null distribution

Possible value of K	Probability under H_0
-6	$\frac{1}{24}$
-4	$\frac{3}{24}$
-2	$\frac{5}{24}$
0	$\frac{6}{24}$
2	$\frac{5}{24}$
4	$\frac{3}{24}$
6	$\frac{1}{24}$

The probability, under H_0, that K is greater than or equal to 2, for example, is therefore

$$\begin{aligned} P_0(K \geq 2) &= P_0(K = 2) + P_0(K = 4) + P_0(K = 6) \\ &= \frac{5}{24} + \frac{3}{24} + \frac{1}{24} = \frac{3}{8} = .375. \end{aligned}$$

This agrees with the upper-tail probability for $n = 4$ and the value $K = 2$ when using `pKendall`.

Note that we have derived the null distribution of K without specifying the form of the underlying independent X and Y populations under H_0 beyond the point of requiring that they be continuous. That is why the test procedures based on K are called *distribution-free procedures*. From the null distribution of K, we can determine the critical value k_α and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific forms of the underlying continuous and independent X and Y distributions.

7. *Calculation of the Mean and Variance of K under the Null Hypothesis.* Displays (8.11) and (8.12) present formulas for the mean and variance of K when the null hypothesis is true. In this comment, we illustrate a direct calculation of $E_0(K)$ and $\text{var}_0(K)$ in the particular case of $n = 4$, using the null distribution of K obtained in Comment 6. (Later, in Comment 10, we present general derivations of $E_0(K)$ and $\text{var}_0(K)$.) The null mean, $E_0(K)$, is obtained by multiplying each possible value of K with its probability under H_0 and summing the products. Thus,

$$\begin{aligned} E_0(K) &= -6\left(\frac{1}{24}\right) - 4\left(\frac{3}{24}\right) - 2\left(\frac{5}{24}\right) + 0\left(\frac{6}{24}\right) + 2\left(\frac{5}{24}\right) \\ &\quad + 4\left(\frac{3}{24}\right) + 6\left(\frac{1}{24}\right) \\ &= 0. \end{aligned}$$

This is in agreement with the value stated in (8.7). A check on the expression for $\text{var}_0(K)$ is also easily performed, using the well-known fact that

$$\text{var}_0(K) = E_0(K^2) - \{E_0(K)\}^2.$$

The value of $E_0(K^2)$, the second moment of the null distribution of K, is again obtained by multiplying possible values (in this case of K^2) by the corresponding probabilities under H_0 and summing. We find

$$E_0(K^2) = \left[(36+36)\left(\frac{1}{24}\right) + (16+16)\left(\frac{3}{24}\right) + (4+4)\left(\frac{5}{24}\right) + 0\left(\frac{6}{24}\right)\right]$$
$$= \frac{26}{3}.$$

Thus,

$$\text{var}_0(K) = \frac{26}{3} - (0)^2 = \frac{26}{3},$$

which agrees with what we obtain using (8.12) directly, namely,

$$\text{var}_0(K) = \frac{4(4-1)(2(4)+5)}{18} = \frac{26}{3}.$$

8. *Symmetry of the Distribution of K under the Null Hypothesis.* When H_0 is true, the distribution of K is symmetric about its mean 0 (see Comment 6 for verification of this when $n = 4$). This implies that

$$P_0(K \leq -x) = P_0(K \geq x), \tag{8.20}$$

for all x. Equation (8.20) is used directly to convert upper-tail probabilities to lower-tail probabilities. In particular, it follows from (8.20) that the lower α percentile for the null distribution of $\overline{K}$ is $-k_\alpha$, thus the use of $-k_\alpha$ as the critical value in procedure (8.9).

9. *Possible Values for K.* If $n = 4j$ or $n = 4j + 1, j = 0, 1, \ldots$, the statistic K (8.6) is always an even integer. Similarly, if $n = 4j + 2$ or $n = 4j + 3$, $j = 0, 1, \ldots,$ K is always an odd integer. The fact that K can assume only every other integer follows from the counting procedure used to define K (see (8.5) and (8.6)). The even or odd property of K for specific sample sizes can be deduced from the relation $K = 2K' - \{n(n-1)/2\}$ and the fact that $n(n-1)/2$ is an even integer (the product of an odd and an even integer) when $n = 4j$ or $n - 4j + 1$, $j - 0, 1, \ldots$, and is an odd integer (the product of two odd integers) when $n = 4j + 2$ or $n = 4j + 3$, $j = 0, 1, \ldots$.

10. *Large-Sample Approximation.* From the counting representation for K in (8.5) and (8.6), we see immediately that

$$E(K) = E\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Q((X_i, Y_i), (X_j, Y_j))\right]$$
$$= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E[Q((X_i, Y_i), (X_j, Y_j))].$$

$$= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} [P\{(Y_2 - Y_1)(X_2 - X_1) > 0\} - P\{(Y_2 - Y_1)(X_2 - X_1) < 0\}],$$

which, because the X and Y variables are continuous, yields

$$E(K) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} [2P\{(Y_2 - Y_1)(X_2 - X_1) > 0\} - 1]$$

$$= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \tau = \binom{n}{2} \tau, \tag{8.21}$$

from expression (8.2) for τ. The value of τ is 0 if X and Y are independent, so it follows that the expected value of K under H_0 is 0, as noted in (8.11). For the variance of K, we can use a well-known expression for the variance of a sum of random variables to obtain

$$\text{var}(K) = \left[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{var}(Q_{ij}) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{\substack{s=1 \\ (i,j) \neq (s,t)}}^{n-1} \sum_{t=s+1}^{n} \text{cov}(Q_{ij}, Q_{st}) \right], \tag{8.22}$$

where $Q_{uv} = Q((X_u, Y_u), (X_v, Y_v))$, for $1 \leq u < v \leq n$.

After considerable tedious calculation, we can show that (8.22) simplifies to

$$\text{var}(K) = [n(n-1)] \left[\frac{1}{2}(1 - \tau^2) + 4(n-2) \left\{ \delta - \left(\frac{\tau + 1}{2} \right)^2 \right\} \right], \tag{8.23}$$

where τ is given in (8.2) and

$$\delta = P\{(Y_2 - Y_1)(X_2 - X_1) > 0 \quad \text{and} \quad (Y_3 - Y_1)(X_3 - X_1) > 0\}. \tag{8.24}$$

Using a mutually exclusive breakdown of the event in δ (8.24) similar to that in (8.2), we see that

$$\begin{aligned} \delta = [&P\{Y_2 > Y_1, X_2 > X_1, Y_3 > Y_1, X_3 > X_1\} \\ &+ P\{Y_2 > Y_1, X_2 > X_1, Y_3 < Y_1, X_3 < X_1\} \\ &+ P\{Y_2 < Y_1, X_2 < X_1, Y_3 > Y_1, X_3 > X_1\} \\ &+ P\{Y_2 < Y_1, X_2 < X_1, Y_3 < Y_1, X_3 < X_1\}] \\ = [&P\{Y_1 < \min(Y_2, Y_3), X_1 < \min(X_2, X_3)\} \\ &+ P\{Y_3 < Y_1 < Y_2, X_3 < X_1 < X_2\} \\ &+ P\{Y_2 < Y_1 < Y_3, X_2 < X_1 < X_3\} \\ &+ P\{Y_1 > \max(Y_2, Y_3), X_1 > \max(X_2, X_3)\}]. \end{aligned} \tag{8.25}$$

When X and Y are independent variables (under H_0), (8.25) simplifies to

$$\begin{aligned}\delta_0 = [&P_0\{Y_1 < \min(Y_2, Y_3)\}P_0\{X_1 < \min(X_2, X_3)\} \\ &+ P_0(Y_3 < Y_1 < Y_2)P_0(X_3 < X_1 < X_2) \\ &+ P_0(Y_2 < Y_1 < Y_3)P_0(X_2 < X_1 < X_3) \\ &+ P_0\{Y_1 > \max(Y_2, Y_3)\}P_0\{X_1 > \max(X_2, X_3)\}].\end{aligned} \tag{8.26}$$

However, X_1, X_2, and X_3 are mutually independent, identically distributed random variables, as are Y_1, Y_2, and Y_3. Thus, we know that

$$P_0\{Y_1 < \min(Y_2, Y_3)\} = P_0\{X_1 < \min(X_2, X_3)\} = \frac{1}{3}, \tag{8.27}$$

$$P_0\{Y_1 > \max(Y_2, Y_3)\} = P_0\{X_1 > \max(X_2, X_3)\} = \frac{1}{3}, \tag{8.28}$$

and

$$\begin{aligned}P_0(X_3 < X_1 < X_2) = P_0(X_2 < X_1 < X_3) = P_0(Y_3 < Y_1 < Y_2) \\ = P_0(Y_2 < Y_1 < Y_3) = \frac{1}{6}.\end{aligned} \tag{8.29}$$

Combining (8.26), (8.27), (8.28), and (8.29), we obtain

$$\delta_0 = \left[\frac{1}{3}\left(\frac{1}{3}\right) + \frac{1}{6}\left(\frac{1}{6}\right) + \frac{1}{6}\left(\frac{1}{6}\right) + \frac{1}{3}\left(\frac{1}{3}\right)\right] = \frac{10}{36}. \tag{8.30}$$

From (8.23) and (8.30) and the fact that $\tau = 0$ when X and Y are independent, it follows that the null variance of K is given by

$$\begin{aligned}\text{var}_0(K) &= [n(n-1)]\left[\frac{1}{2}(1-0)^2 + 4(n-2)\left\{\frac{10}{36} - \left(\frac{0+1}{2}\right)^2\right\}\right] \\ &= [n(n-1)]\left[\frac{1}{2} + \frac{1}{9}(n-2)\right] = \frac{n(n-1)(2n+5)}{18},\end{aligned}$$

as previously noted in (8.12).

The asymptotic normality under both H_0 and general alternatives of the standardized form

$$K^* = \frac{K - E_0(K)}{\{\text{var}_0(K)\}^{1/2}} = \frac{K}{\left\{\dfrac{n(n-1)(2n+5)}{18}\right\}^{1/2}}$$

follows from Hoeffding's (1948a) U-statistic theorem applied to the bivariate setting. (For additional details, see Example 3.6.12 in Randles and Wolfe (1979).)

11. *Ties within the X-Values and/or Y-Values.* We have recommended dealing with tied X observations and/or tied Y observations by counting a zero in the Q^* (8.17) counts leading to the computation of K (8.6). This approach is satisfactory as long as the number of (X, Y) pairs containing a tied X and/or tied Y observation does not represent a sizable percentage of the total number (n) of sample pairs.

    We should, however, point out that methods other than this zero assignment to the Q^* (8.17) counts have been considered for dealing with tied X and/or tied Y observations. One could use individual randomization (e.g., flipping a fair coin) to decide whether each of the tied pairs (X or Y) is to be counted as a $+1$ (i.e., as a concordant pair—see Comment 3) or as a -1 (i.e., as a discordant pair—again, see Comment 3) in the computation of K (8.6). (Although this approach maintains many of the nice properties of K that hold when there are no tied X and/or tied Y observations, it introduces extraneous randomness that could quite easily have a direct effect on the outcome of any subsequent inferences based on such a modified value of K.) A second alternative approach in the case of the one-sided test procedures in (8.8), (8.9), (8.14), and (8.15) is to be conservative about rejecting the null hypothesis H_0; that is, we could count all the tied X and/or tied Y observations as if they were in favor of not rejecting H_0. Thus, for example, in applying either procedure (8.8) or (8.14) to test H_0 against the alternative $\tau > 0$, we would treat *all* the pairs of pairs involving tied X and/or tied γ observations as if they were discordant pairs (in favor of not rejecting H_0) leading to Q (8.5) counts of -1 in the calculation of K. (In the case of procedures (8.9) and (8.15), all the pairs of pairs involving tied X and/or tied Y observations would be considered as concordant pairs—again in favor of not rejecting H_0—leading to Q (8.5) counts of $+1$ in the calculation of K.) Any rejection of H_0 with this conservative approach to dealing with tied X and/or tied Y observations could then be viewed as providing strong evidence in favor of the appropriate alternative. For more detailed discussion of methods for handling tied X and/or tied Y observations, see Sillitto (1947), Smid (1956), Burr (1960), and Kendall (1962).

12. *Some Power Results for the Kendall Test for Independence.* We consider the upper-tail α-level test of H_0 (8.1) versus $H_1 : \tau > 0$ given by procedure (8.8). The power, or probability of correctly rejecting H_0, for τ (8.2) values "near" the null hypothesis value of 0 can be approximated by

$$\text{Power} \doteq \Phi(A_F), \tag{8.31}$$

where $\Phi(A_F)$ is the area under a standard normal density to the left of the point

$$A_F = \{[9n(n-1)/(4n+10)]^{1/2}\tau - z_\alpha\}. \tag{8.32}$$

When $F_{X,Y}$ is the bivariate normal distribution with correlation coefficient ρ, it follows that $\tau = \frac{2}{\pi}\sin^{-1}(\rho)$ (see, for example, Gibbons and Chakraborti (2010)). Thus, when $F_{X,Y}$ is bivariate normal the approximate power depends only on the value of ρ. For purposes of illustration, suppose that the common underlying distribution is bivariate normal with $\rho = .4$. For the case of $n = 9$ and $\alpha = .060$, the test rejects H_0 if and only if $K \geq 16$, or, equivalently, $\overline{K} \geq 16/36$.

Substituting $\tau = (2/\pi)\sin^{-1}(.4) = (2/\pi)(.4115) = .2620$ in (8.32), we obtain

$$\begin{aligned} A_{\text{BIV NOR}} &= \{[9(9)(8)/2(2(9)+5)]^{1/2}.2620 - 1.555\} \\ &= \{[14.09]^{1/2}(.2620) - 1.555\} = \{.9835 - 1.555\} = -.57. \end{aligned}$$

Thus, the approximate power of this test for a bivariate normal distribution with $\rho = .4$ (and *any* means and variances) is

$$\text{Power} \doteq \Phi(-.57) = 1 - \Phi(.57) = 1 - (1 - .28) = .28.$$

This compares with the simulation estimated exact power of .35 for $n = 9$, $\rho = .4$, and $\alpha = .05$, as given in Table 8.3 of Bhattacharyya, Johnson, and Neave (1970). Additional simulation estimated exact power values for the one-sided Kendall test and sample sizes $n = 5,7,9$ and significance levels $\alpha = .01$ and .05 can be found in Bhattacharyya, Johnson, and Neave (1970) for bivariate normal and bivariate exponential distributions.

13. *Sample Size Determination.* Noether (1987) shows how to determine an approximate sample size n so that the α-level one-sided test given by procedure (8.8) will have approximate power $1 - \beta$ against an alternative value of τ (8.2) greater than zero. This approximate value of n is

$$n \doteq \frac{4(z_\alpha + z_\beta)^2}{9\tau^2}. \tag{8.33}$$

As an illustration of the use of (8.33), suppose we are testing H_0 and we desire to have an upper-tail level $\alpha = .010$ test with power $1 - \beta$ at least .90 against an alternative bivariate distribution for which $\tau = .4$. Using $z_\alpha = z_{.01} = 2.326$ and $z_\beta = z_{.10} = 1.282$, we find that the approximate required sample size for the alternative $\tau = .4$ is

$$n \doteq \frac{4(2.326 + 1.282)^2}{9(.4)^2} = 36.2.$$

To be conservative, we would take $n = 37$.

14. *Trend Test.* If we take $X_i = i,\ i = 1, \ldots, n$ and consider

$$\begin{aligned} K &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q((i, Y_i), (j, Y_j)) \\ &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c(Y_j - Y_i), \end{aligned}$$

where

$$c(a) = \begin{cases} 1 & \text{if } a > 0, \\ 0 & \text{if } a = 0, \\ -1 & \text{if } a < 0, \end{cases}$$

then K can be used as a test for a time trend in the univariate random sample $Y_1, \ldots, Y_n$. This use of K to test for a time trend was suggested by Mann (1945).

15. *Other Uses for the K Statistic.* The Wilcoxon rank sum test (Section 4.1) and the Jonckheere–Terpstra test (Section 6.2) can be viewed as tests based on K (8.6) (or, equivalently, $\hat{\tau}$ (8.34)). For this interpretation see Jonckheere (1954a) and Kendall (1962, Sections 3.12 and 13.9). Also, Wolfe (1977) has used the K statistic to compare the correlation between variables X_2 and X_1 with that between the variables X_3 and X_1, when both X_2 and X_3 are potential predictors for X_1.
16. *Consistency of the K Test.* Under the assumption that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from a continuous bivariate population with joint distribution function $F_{X,Y}(x, y)$, the consistency of the tests based on K depends on the parameter τ (8.2). The test procedures defined by (8.8), (8.9), and (8.10) are consistent against the class of alternatives corresponding to $\tau >, <$, and $\neq 0$, respectively.
17. *Multivariate Concordance.* Joe (1990) has generalized Kendall's measure of association τ from the bivariate case where τ measures the strength of association between two variables X, Y to the multivariate case where $\mathbf{X} = (X_1, \ldots, X_m)$ is an m-dimensional random variable and one is interested in a measure of the strength of the association between the components $X_1, \ldots, X_m$ of $\mathbf{X}$. Let F denote the joint distribution function of $\mathbf{X}$,

$$F(x_1, \ldots, x_m) = P(X_1 \leq x_1 \text{ and } X_2 \leq x_2 \text{ and } \ldots \text{ and } X_m \leq x_m)$$

and denote the marginal distribution functions as $F_j(x_j) = P(X_j \leq x_j), j = 1, \ldots, m$. The null hypothesis of mutual independence of $X_1, \ldots, X_m$ is

$$H_0 : F(x_1, \ldots, x_m) = \prod_{j=1}^{m} F_j(x_j), \quad \text{for all}(x_1, \ldots, x_m).$$

That is, the joint distribution is equal to the product of the marginals.

Joe has defined a class of measures of the strength of association between $X_1, X_2, \ldots, X_m$. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{im}), i = 1, 2$ be two independent m-dimensional random variables each with joint distribution function F. One member of Joe's class reduces to $\overline{\tau}$, the average of all pairwise τ's. The measure $\overline{\tau}$ was introduced by Hays (1960) and is given by

$$\overline{\tau} = \sum_{u=1}^{m-1} \sum_{v=1}^{m} \frac{\tau_{uv}}{\left\{ \dfrac{m(m-1)}{2} \right\}},$$

where

$$\begin{aligned} \tau_{uv} &= P\{(X_{1u} - X_{1v})(X_{2u} - X_{2v}) > 0\} - P\{(X_{1u} - X_{1v})(X_{2u} - X_{2v}) < 0\} \\ &= 2P\{X_{1u} - X_{1v})(X_{2u} - X_{2v}) > 0\} - 1. \end{aligned}$$

Joe has also generalized Spearman's measure (see Section 8.5) and a measure due to Blomqvist (1950).

Properties

1. *Consistency*. The tests defined by (8.8), (8.9), and (8.10) are consistent against the alternatives $\tau >, <$, and $\neq 0$, respectively.
2. *Asymptotic Normality*. See Hoeffding (1948a) or Randles and Wolfe (1979, pp. 108–109).
3. *Efficiency*. See Section 8.7.

Problems

1. The data in Table 8.3 are a subset of the data obtained by Featherston (1971). Among other things, he was interested in the relationship between the weight of tapeworms *(Taenia hydatigena)* fed to dogs and the weight of the scoleces recovered from the dogs after 20 days. (A scolex is the attachment end of a tapeworm, consisting of the head and neck.) The cysticerci used in the experiment were collected from sheep carcasses and force-fed to 10 dogs via gelatine capsules. The scoleces were recovered from each dog at autopsy, 20 days after the introduction of the tapeworms. Table 8.3 gives the mean weight of the initial cysticerci and the mean weight of the recovered worms for each of the 10 dogs in the study.

Test the hypothesis of independence versus the alternative that the mean weight of introduced cysticerci is positively correlated with the mean weight of worms recovered.

2. Let R_i be the rank of X_i in the joint ranking of $X_1, \ldots, X_n$ and let S_i be the rank of Y_i in the joint ranking of $Y_1, \ldots, Y_n$. Show that knowledge of $R_1, \ldots, R_n$ and $S_1, \ldots, S_n$ is sufficient to calculate K (8.6).

3. The data in Table 8.4 are a subset of the data obtained by Sylvester (1969) in a study concerned with the anatomical and pathological status of the corticospinal and somatosensory tracts and parietal lobes of patients who had had cerebral palsy. Among other things, he was interested in the relationship between brain weights and large fiber (>7.5 μ in diameter) counts in the medullary pyramid. Table 8.4 gives the mean brain weights (g) and medullary pyramid large fiber counts for 11 cerebral palsy subjects. Test the hypothesis of independence versus the general alternative that brain weight and large fiber count in the medullary pyramid are correlated in subjects who have had cerebral palsy.

Table 8.3 Relation Between Weight of the Cysticerci of *Taenia hydatigena* Fed to Dogs and Weight of Worms Recovered at 20 Days

	Mean weight, mg	
Dog	Cysticerci	Worms recovered
1	28.9	1.0
2	32.8	7.7
3	12.0	7.3
4	9.9	7.9
5	15.0	1.1
6	38.0	3.5
7	12.5	18.9
8	36.5	33.9
9	8.6	28.6
10	26.8	25.0

Source: D. W. Featherston (1971).

Table 8.4 Mean Brain Weights and Medullary Pyramid Large Fiber Counts for Cerebral Palsy Subjects

Subject number	Brain weight, g	Pyramidal large fiber count
1	515	32,500
2	286	26,800
3	469	11,410
4	410	14,850
5	461	23,640
6	436	23,820
7	479	29,840
8	198	21,830
9	389	24,650
10	262	22,500
11	536	26,000

Source: P. E. Sylvester (1969).

4. Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed continuous bivariate random variables with joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} e^{-y}, & 0 < x < y < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

Calculate the value of τ for this bivariate distribution.

5. Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed discrete bivariate random variables with joint probability function

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{x+y}{21}, & x = 1,2,3; \quad y = 1,2, \\ 0, & \text{elsewhere.} \end{cases}$$

Calculate the value of τ for this bivariate distribution.

6. Let (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed continuous bivariate random variables with joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{2}y^2 e^{-x-y}, & 0 < x < \infty, \quad 0 < y < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

Calculate the value of τ for this bivariate distribution.

7. The data in Table 8.5 are a subset of the data considered by Clark, Vandenberg, and Proctor (1961) in a study concerned with the relationship of scores on various psychological tests and certain physical characteristics of twins. Table 8.5 gives the test scores (totals of a number of different psychological tests) of 13 dizygous (i.e., nonidentical) male twins. Test the hypothesis of independence versus the alternative that the twins' test scores are positively correlated.

8. Previously, it was shown that if X and Y are independent random variables, then τ (8.2) has a value of 0. Show that the converse is not true by constructing a joint probability distribution for the pair of random variables X and Y such that $\tau = 0$ but X and Y are not independent.

9. If we have 25 bivariate observations and $F_{X,Y}$ is bivariate normal with correlation coefficient .3, what is the approximate power of the level $\alpha = .045$ test of H_0 (8.1) versus the alternative $\tau > 0$?

Table 8.5 Psychological Test Scores of Dizygous Male Twins

Pair i	Twin X_i	Twin Y_i
1	277	256
2	169	118
3	157	137
4	139	144
5	108	146
6	213	221
7	232	184
8	229	188
9	114	97
10	232	231
11	161	114
12	149	187
13	128	230

Source: P. J. Clark, S. G. Vandenberg, and C. H. Proctor (1961).

10. For an arbitrary number, n, of bivariate observations, what are the smallest and largest values of K? Give examples of data sets where these extremes are achieved.

11. Give an example of a data set of $n \geq 10$ bivariate observations for which K has value 0. (Consider Comment 9.)

12. Suppose $n = 20$. Compare the critical region for the exact level $\alpha = .05$ test of H_0 (8.1) versus $H_2 : \tau < 0$ based on K with the critical region for the corresponding nominal level $\alpha = .05$ based on the large-sample approximation. What is the exact significance level of this .05 nominal level test based on the large-sample approximation?

13. Consider a level $\alpha = .10$ test of H_0 (8.1) versus the alternative $\tau > 0$ based on K. How many bivariate observations (n) will we need to collect in order to have approximate power at least .95 against an alternative for which $\tau = .6$?

14. A question of significance to state legislators working with tight budgets is the spending for secondary education. The data in Table 8.6 are from the Department of Education, National Center for Education Statistics and were considered by Merline (1991) in assessing the relationship between the amount of money spent on secondary education and various performance criteria for high-school seniors. Table 8.6 gives the spending ($) per high-school senior and the percentage of those seniors who graduated for each of the 50 states in the 1987–1988 school year.

Use the large-sample approximation to test the hypothesis of independence versus the alternative that spending per high-school senior and the percentage of seniors graduating are positively correlated. (Discuss any other social or economic factors that might impact on these data and, thereby, on the conclusion from this statistical analysis.)

15. For the case of $n = 5$ untied bivariate (X, Y) observations, obtain the form of the exact null (H_0) distribution of K. (See Comment 6.)

16. Johnson (1973) studied several different managerial aspects of university associated schools of nursing. Among the data she collected were the "extent of agreement (between the dean and the faculty) on the responsibilities for decision making" and "faculty satisfaction." The ranks on the two variables for the 12 institutions that were involved in Johnson's study are presented in Table 8.7.

Test the hypothesis of independence versus the alternative that faculty/dean decision-making agreement and faculty satisfaction are negatively correlated in university schools of nursing. (*Note:* Low ranks are associated with poor faculty satisfaction and little faculty/dean decision-making agreement, respectively.)

17. Consider the test of H_0 (8.1) versus $H_1 : \tau > 0$ based on K for the following $n = 10$ (X, Y) observations: (1.5, 6), (1.9, 4), (2.3, 6), (2.7, 12), (1.5, 13), (1.8, 16), (3.6, 16), (4.2, 9), (4.7,

Table 8.6 Spending per High-School Senior and the Percentage of Those Seniors Who Graduated during the 1987–1988 School Year

State	$ per Pupil	% Graduated	State	$ per Pupil	% Graduated
Alaska	7971	65.5	Ohio	3998	79.6
New York	7151	62.3	Nebraska	3943	85.4
New Jersey	6564	77.4	Hawaii	3919	69.1
Connecticut	6230	84.9	West Virginia	3858	77.3
Massachusetts	5471	74.4	California	3840	65.9
Rhode Island	5329	69.8	Indiana	3794	76.3
Vermont	5207	78.7	Missouri	3786	74.0
Maryland	5201	74.1	Arizona	3744	61.1
Wyoming	5051	88.3	New Mexico	3691	71.9
Delaware	5017	71.7	Nevada	3623	75.8
Pennsylvania	4989	78.4	Texas	3608	65.3
Oregon	4789	73.0	North Dakota	3519	88.3
Wisconsin	4747	84.9	Georgia	3434	61.0
Michigan	4692	73.6	South Carolina	3408	64.6
Colorado	4462	74.7	North Carolina	3368	66.7
New Hampshire	4457	74.1	South Dakota	3249	79.6
Minnesota	4386	90.9	Louisiana	3138	61.4
Illinois	4369	75.6	Oklahoma	3093	71.7
Maine	4246	74.4	Tennessee	3068	69.3
Montana	4246	87.3	Kentucky	3011	69.0
Washington	4164	77.1	Arkansas	2989	77.2
Virginia	4149	71.6	Alabama	2718	74.9
Iowa	4124	85.8	Idaho	2667	75.4
Florida	4092	58.0	Mississippi	2548	66.9
Kansas	4076	80.2	Utah	2454	79.4

Source: J. W. Merline (1991).

Table 8.7 Rankings for Faculty/Dean Decision-Making Agreement and Faculty Satisfaction for Participating Schools of Nursing

School	Rank for faculty/dean decision-making agreement	Rank for faculty satisfaction
1	8	8
2	9	2
3	6	10
4	12	5
5	1	12
6	11	4
7	10	6
8	2	9
9	4	7
10	5	3
11	7	11
12	3	1

Source: B. M. Johnson (1973).

0), and (4.0, 3). Compute the P-values for the competing K-procedures based on either (a) using Q^* (8.17) counts of zero, as recommended in the Ties portion of this section, or (b) dealing with the tied X and tied Y observations in a conservative manner, as presented in Comment 11. Discuss the results.

18. In Comment 4, we noted that we have $K = 2K' - \{n(n-1)/2\}$, where $K' =$ (number of concordant pairs), when there are neither tied X nor tied Y observations. Obtain the corresponding expression for the relationship between K and K' when there are no tied X pairs and a total of t ($\neq 0$) tied Y pairs (among the $\binom{n}{2}$ total Y pairs), and we use the Q^* (8.17) counts of zero to deal with the tied Y pairs. How does this expression change if there are no tied Y pairs and t tied X pairs? Discuss the necessary changes in the expression relating K and K' when there are s ($\neq 0$) tied X pairs and t ($\neq 0$) tied Y pairs.

19. Gerstein (1965) studied the long-term pollution of Lake Michigan and its effect on the water supply for the city of Chicago. One of the measurements considered by Gerstein was the annual number of "odor periods" over the period of years 1950–1964. Table 8.8 contains this information for Lake Michigan for each of these 15 years.

Test the hypothesis that the degree of pollution (as measured by the number of odor periods) had not changed with time against the alternative that there was a general increasing trend in the pollution of Lake Michigan over the period 1950–1964. (See Comment 14.)

8.2 AN ESTIMATOR ASSOCIATED WITH THE KENDALL STATISTIC (KENDALL)

Procedure

The estimator of the Kendall population correlation coefficient τ (8.2), based on the statistic K (8.6),is

$$\hat{\tau} = \frac{2K}{n(n-1)} = \overline{K}. \tag{8.34}$$

The statistic $\hat{\tau}$ is known as *Kendall's sample rank correlation coefficient* and appropriately assumes values between -1 and 1 inclusive.

Table 8.8 Annual Number of Odor Periods for Lake Michigan for the Period 1950–1964

Year	Number of odor periods
1950	10
1951	20
1952	17
1953	16
1954	12
1955	15
1956	13
1957	18
1958	17
1959	19
1960	21
1961	23
1962	23
1963	28
1964	28

Source: H. H. Gerstein (1965).

EXAMPLE 8.2 ***(Continuation of Example 8.1).***

For the canned tuna data of Table 8.1, we see from (8.34) that the sample estimate of τ is

$$\hat{\tau} = \frac{2(16)}{9(8)} = \frac{4}{9}. \tag{8.35}$$

This estimate is also found in the R output in Example 8.1.

Comments

18. *Ties.* In the presence of ties, use $\hat{\tau} = 2K/n(n-1)$, where

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q^*((X_i, Y_i), (X_j, Y_j)) \tag{8.36}$$

and $Q^*((X_i, Y_i), (X_j, Y_j))$ is defined by (8.17).

19. *Probability Estimation.* For many problems, distribution-free test statistics are used directly to estimate basic probability parameters other than the usual distributional parameters associated with the corresponding normal theory problems. In particular, note that $\hat{\tau} = 2K/[n(n-1)]$ estimates the probability parameter τ (8.2) rather than the usual correlation coefficient for the underlying bivariate population. Estimators of such readily interpretable parameters are very helpful in data analysis. (See Crouse (1966), Wolfe and Hogg (1971), and Comment 4.18.)

Properties

1. *Standard Deviation of* $\hat{\tau}$. For the asymptotic standard deviation of $\hat{\tau}$ (8.34), see Noether (1967a, p. 78), Fligner and Rust (1983), Samara and Randles (1988), and Comment 25.
2. *Asymptotic Normality.* See Hoeffding (1948a) and Randles and Wolfe (1979, pp. 108–109).

Problems

20. Estimate τ for the tapeworm data of Table 8.3.

21. What is the maximum possible value of $\hat{\tau}$ when there are no tied X and/or tied Y observations? What is the minimum possible value of $\hat{\tau}$ when there are no tied X and/or tied Y observations? Construct three examples with $n \geq 10$ with no tied X and/or tied Y observations: one in which $\hat{\tau}$ achieves its maximum value, one in which it achieves its minimum value, and one for which $\hat{\tau} = 0$.

22. Estimate τ for the cerebral palsy data of Table 8.4.

23. Estimate τ for the twin data of Table 8.5.

24. Estimate τ for the secondary education data of Table 8.6.

25. Use Comment 17 to redo Problem 21 for the case when there are t ($\neq 0$) tied X pairs (among the total of $\binom{n}{2}$ X pairs) and no tied Y observations. How is the answer affected if there are no tied X observations and $t(\neq 0)$ tied Y pairs? if there are s ($\neq 0$) tied X pairs and t ($\neq 0$) tied Y pairs?

26. Estimate τ for the nursing faculty data of Table 8.7.

8.3 AN ASYMPTOTICALLY DISTRIBUTION–FREE CONFIDENCE INTERVAL BASED ON THE KENDALL STATISTIC (SAMARA–RANDLES, FLIGNER–RUST, NOETHER)

Procedure

For an asymptotically distribution-free symmetric two-sided confidence interval for τ, with the approximate confidence coefficient $1-\alpha$, we first compute

$$C_i = \sum_{\substack{t=1 \\ t\neq i}}^{n} Q((X_i, Y_i), (X_t, Y_t)), \quad \text{for } i = 1, \ldots, n, \tag{8.37}$$

where $Q((a,b),(c,d))$ is given by (8.5). Let $\overline{C} = (1/n)\sum_{i=1}^{n} C_i = 2K/n$ and define

$$\hat{\sigma}^2 = \frac{2}{n(n-1)}\left[\frac{2(n-2)}{n(n-1)^2}\sum_{i=1}^{n}(C_i - \overline{C})^2 + 1 - \hat{\tau}^2\right], \tag{8.38}$$

where $\hat{\tau}$ is given by (8.34). The approximate $100(1-\alpha)\%$ confidence interval $(\tau_{\mathrm{L}}, \tau_{\mathrm{U}})$ for τ that is associated with the point estimator $\hat{\tau}$ (8.34) is then given by

$$\tau_{\mathrm{L}} = \hat{\tau} - z_{\alpha/2}\hat{\sigma}, \quad \tau_{\mathrm{U}} = \hat{\tau} + z_{\alpha/2}\hat{\sigma}. \tag{8.39}$$

With τ_{L} and τ_{U} given by display (8.39), we have

$$P_\tau\{\tau_{\mathrm{L}} < \tau < \tau_{\mathrm{U}}\} \approx 1-\alpha \text{ for all } \tau. \tag{8.40}$$

(For approximate upper or lower confidence bounds for τ associated with $\hat{\tau}$, see Comment 23.)

EXAMPLE 8.3 ***(Continuation of Examples 8.1 and 8.2).***

Consider the canned tuna data of Table 8.1. We illustrate how to obtain an approximate 90% symmetric two-sided confidence interval for τ. From (8.37), we see that

$$\begin{aligned} C_5 &= \sum_{j\neq 5} Q((X_5, Y_5), (X_j, Y_j)) \\ &= [Q((X_5, Y_5), (X_1, Y_1)) + Q((X_5, Y_5), (X_2, Y_2)) \\ &\quad + Q((X_5, Y_5), (X_3, Y_3)) + Q((X_5, Y_5), (X_4, Y_4)) \end{aligned}$$

$$+ Q((X_5, Y_5), (X_6, Y_6)) + Q((X_5, Y_5), (X_7, Y_7))$$
$$+ Q((X_5, Y_5), (X_8, Y_8)) + Q((X_5, Y_5), (X_9, Y_9)].$$

Using the fact that $Q((X_i, Y_i), (X_j, Y_j)) = Q((X_j, Y_j), (X_i, Y_i))$ for every $i \neq j$ and the Q counts for the canned tuna data in Table 8.2, it follows that

$$C_5 = 1 - 1 + 1 + 1 - 1 + 1 - 1 + 1 = 2.$$

Note that C_5 is simply equal to the sum of the Q values in the $j = 5$ row and the $i = 5$ column in Table 8.2. In the same way, we find

$$C_1 = 7 - 1 = 6, \quad C_2 = 6 - 2 = 4, \quad C_3 = 8 - 0 = 8, \quad C_4 = 6 - 2 = 4$$
$$C_6 = 3 - 5 = -2, \quad C_7 = 6 - 2 = 4, \quad C_8 = 6 - 2 = 4, \quad C_9 = 5 - 3 = 2.$$

Thus, we have

$$\overline{C} = \frac{1}{9}\sum_{i=1}^{9} C_i = \frac{1}{9}[6 + 4 + 8 + 4 + 2 - 2 + 4 + 4 + 2] = \frac{32}{9}.$$

Thus,

$$\sum_{i=1}^{9}(C_i - \overline{C})^2 = \sum_{i=1}^{9}\left(C_i - \frac{32}{9}\right)^2$$
$$= \left[\left(\frac{22}{9}\right)^2 + \left(\frac{4}{9}\right)^2 + \left(\frac{40}{9}\right)^2 + \left(\frac{4}{9}\right)^2 + \left(-\frac{14}{9}\right)^2\right.$$
$$\left. + \left(-\frac{50}{9}\right)^2 + \left(\frac{4}{9}\right)^2 + \left(\frac{4}{9}\right)^2 + \left(-\frac{14}{9}\right)^2\right]$$
$$= \frac{484 + 4(16) + 1600 + 2(196) + 2500}{81} = \frac{560}{9}. \tag{8.41}$$

Using the values for $\hat{\tau}$ and $\Sigma_{i=1}^{9}(C_i - \overline{C})^2$ given in (8.35) and (8.41), respectively, we see from (8.38) that

$$\hat{\sigma}^2 = \frac{2}{9(8)}\left[\frac{2(7)}{9(8)^2}\left(\frac{560}{9}\right) + 1 - \left(\frac{4}{9}\right)^2\right]$$
$$= \frac{1}{36}[1.512 + 1 - .198] = .064.$$

With $1 - \alpha = .90$ (so that $\alpha = .10$), $z_{.05} = 1.65$. Hence, from (8.39), we obtain

$$\tau_L = \frac{4}{9} - 1.65(.064)^{1/2} = .444 - .417 = .027$$

and

$$\tau_U = \frac{4}{9} + 1.65(.064)^{1/2} = .444 + .417 = .861.$$

Our approximate 90% symmetric confidence interval for τ is thus $(\tau_L, \tau_U) = (.027, .861)$.

The above results may be duplicated with the command `kendall.ci` in package `NSM3`. The arguments needed are the samples X and Y, the confidence level α, and whether the interval should be the two-sided symmetric interval described above or a one-sided interval described in Comment 24. In particular,

```
kendall.ci (x, y, alpha=.1, type="t")
```

reproduces the above bounds for this example.

Comments

20. *Interpretation as a Confidence Interval for a Probability.* The confidence interval given by (8.39) is an approximate $1 - \alpha$ confidence interval for a parameter that is a linear function of the probability $P\{(X_1 - X_2)(Y_1 - Y_2) > 0\}$. This is common practice in the field of nonparametric statistics, where probabilities are often natural and easily interpretable parameters. Recall the relation of the Wilcoxon two-sample test of Section 4.1 to the parameter $P(X < Y)$. (See Comments 4.7, 4.10, 4.14, and 4.18.)
21. *Ties.* In the presence of ties, use $Q^*((X_i, Y_i), (X_j, Y_j))$ defined by (8.17) instead of $Q((X_i, Y_i), (X_j, Y_j))$ given by (8.5) in the computation of $C_1, \ldots, C_n$ and $\hat{\tau}$.
22. *Alternative Method of Calculation.* The following equivalent formula for the term $\Sigma_{i=1}^n (C_i - \overline{C})^2$ in the definition of $\hat{\sigma}^2$ (8.38), namely,

$$\sum_{i=1}^{n} (C_i - \overline{C})^2 = \sum_{i=1}^{n} C_i^2 - \frac{4K^2}{n}, \tag{8.42}$$

    is often computationally more convenient.
23. *Concordant/Discordant Pairs Representation for the C_i's.* Let C_i' and C_i'' be the numbers of pairs (X_j, Y_j), $j \neq i$, that are concordant and discordant, respectively, with (X_i, Y_i), for $i = 1, \ldots, n$. Then, the C_i (8.37) counts can be expressed as $C_i = C_i' - C_i''$, for $i = 1, \ldots, n$.
24. *Confidence Bounds.* In many settings, we are interested only in making one-sided confidence statements about the parameter τ; that is, we wish to assert with specified confidence that τ is no larger (or, in other settings, no smaller) than some upper (lower) confidence bound based on the sample data. To obtain such one-sided confidence bounds for τ, we proceed as follows. For a specified confidence coefficient $1 - \alpha$, find z_a (not $z_{\alpha/2}$, as for the confidence interval). An approximate $100(1 - \alpha)\%$ *upper* confidence bound τ_U^* for τ is then given by

$$[-1, \tau_U^*) = [-1, \hat{\tau} + z_\alpha \hat{\sigma}), \tag{8.43}$$

    where $\hat{\tau}$ and $\hat{\sigma}^2$ are given by (8.34) and (8.38), respectively. With τ_U^* given by display (8.43), we have

$$P_\tau\{-1 \leq \tau < \tau_U^*\} \approx 1 - \alpha \text{ for all } \tau. \tag{8.44}$$

The corresponding approximate $100(1-\alpha)\%$ *lower* confidence bound τ_{L}^* for τ is given by

$$(\tau_{\mathrm{L}}^*, 1) = (\hat{\tau} - z_\alpha \hat{\sigma}, 1), \tag{8.45}$$

with

$$P_\tau\{\tau_{\mathrm{L}}^* < \tau \leq 1\} \approx 1 - \alpha \text{ for all } \tau. \tag{8.46}$$

25. *Alternative Approximate Confidence Limits.* Samara and Randles (1988) showed that the statistic $\hat{\tau}/\hat{\sigma}$ is itself distribution-free under the null hypothesis (H_0) of independence, and they tabled the upper αth percentile of its null distribution, k_α^*, for $\alpha = .005, .01, .025, .05$, and .10 and sample sizes $n = 6(1)20$. Slightly improved confidence intervals and confidence bounds can be obtained by replacing the normal percentiles $z_{\alpha/2}$ and z_α by $k_{\alpha/2}^*$ and k_α^*, respectively, in (8.39), (8.43), and (8.45).
26. *Estimating the Asymptotic Standard Deviation of* $\hat{\tau}$. The statistic $\hat{\sigma}$ (8.38) is chosen to be a consistent estimator for the asymptotic standard deviation of the point estimator $\hat{\tau}$ (8.34). It is not necessary to use all the sample observations in calculating $\hat{\sigma}$. In fact, any fixed percentage subset of the n sample observations can be employed to find the C_i values used in (8.38). For example, 25% of a random sample of n paired observations (namely, $n/4$ observations) could be used to obtain $\hat{\sigma}$.
27. *Asymptotic Coverage Probability.* Asymptotically, the true coverage probability of the interval defined by (8.39) and the bounds in (8.43) and (8.45) will agree with the nominal confidence coefficient $1 - \alpha$. Subject to Assumption A, this asymptotic (n infinitely large) result does not depend on the distribution of the underlying bivariate population. Thus, the interval given by (8.39) and the bounds in (8.43) and (8.45) have been constructed to have the asymptotically distribution-free property.
28. *Historical Development.* The initial effort at constructing asymptotically distribution-free confidence intervals and bounds for τ was due to Noether (1967a). The approximate $100(1-\alpha)\%$ confidence interval proposed by Noether is $\hat{\tau} \pm z_{\alpha/2}\hat{\sigma}_N$, where $\hat{\sigma}_N^2$ is a consistent estimator (based on U-statistics theory) of the variance of $\hat{\tau}$. However, it was later pointed out that $\hat{\sigma}_N^2$ can assume negative values, even though it is estimating the nonnegative quantity $\mathrm{var}(\hat{\tau})$. Although this distressing possibility is more likely to occur in small samples, it can be negative for sample sizes as large as $n = 30$, To avoid this problem, Fligner and Rust (1983) proposed the use of $\hat{\tau} \pm z_{\alpha/2}\hat{\sigma}_{\mathrm{FR}}$ as an asymptotically distribution-free $100(1-\alpha)\%$ confidence interval for $\hat{\tau}$, where $\hat{\sigma}_{\mathrm{FR}}^2$ is a jackknife estimator (different from $\hat{\sigma}_{\mathrm{N}}^2$) of $\mathrm{var}(\hat{\tau})$ that is consistent and cannot assume negative values. A few years later, Samara and Randles (1988) noted that although the Fligner–Rust jackknife estimator $\hat{\sigma}_{\mathrm{FR}}^2$ can never be negative, it can be zero for a variety of rank configurations, including some nonextreme cases. They suggested a final modification leading to the asymptotically distribution-free $100(1-\alpha)\%$ confidence interval in display (8.39), where $\hat{\sigma}^2$ (8.38) $= \hat{\sigma}_{\mathrm{SR}}^2$ is a third consistent estimator of $\mathrm{var}(\hat{\tau})$. The estimator $\hat{\sigma}_{\mathrm{SR}}^2 = \hat{\sigma}^2$ (8.38) is also based on U-statistics methodology (as is $\hat{\sigma}_{\mathrm{N}}^2$), but it can never be negative and is zero only in the two extreme cases where $\hat{\tau} = \pm 1$. For the approximate confidence

interval $\hat{\tau} \pm z_{\alpha/2}\hat{\sigma}$ to be simply the singleton point $\hat{\tau}(= +1 \text{ or } -1)$ in such extreme cases is not ideal, but it is also not unreasonable.

29. *Competitor Tests for Independence.* In Section 8.1, we discussed tests of independence (8.1) based on Kendall's statistic K (8.6). Noether (1967a), Fligner and Rust (1983), and Samara and Randles (1988) also proposed distribution-free tests of H_0 (8.1) based on the statistics $\hat{\tau}/\hat{\sigma}_{\mathrm{N}}$, $\hat{\tau}/\hat{\sigma}_{\mathrm{FR}}$, and $\hat{\tau}/\hat{\sigma}_{\mathrm{SR}}$, respectively, where $\hat{\tau}$ is given by (8.34) and $\hat{\sigma}^2_{\mathrm{N}}$, $\hat{\sigma}^2_{\mathrm{FR}}$, and $\hat{\sigma}^2_{SR}$ are the various consistent estimators of $\mathrm{var}(\hat{\tau})$ discussed in Comment 28. Although not generally as powerful as the procedures based on K for testing H_0 (8.1), the tests based on $\hat{\tau}/\hat{\sigma}_{\mathrm{N}}$, $\hat{\tau}/\hat{\sigma}_{\mathrm{FR}}$, and $\hat{\tau}/\hat{\sigma}_{\mathrm{SR}}$ all have the advantage (not possessed by the tests based on K) that they are also asymptotically distribution-free procedures for testing the more general null hypothesis H_0^*: $\tau = 0$.

30. *Partial Correlation Coefficients.* Let $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$ be a random sample from a continuous trivariate distribution. It is often of interest to assess the association between the X and Y variables, controlled for the third variable Z. Gripenberg (1992) proposed measuring this "partial correlation" by the parameter

$$\begin{aligned} \tau_{XY/Z} &= 2P\{(Y_2 - Y_1)(X_2 - X_1) > 0 | Z_1 = Z_2\} - 1 \\ &= E[Q((X_1, Y_1), (X_2, Y_2)) | Z_1 = Z_2], \end{aligned} \tag{8.47}$$

where $Q((X_1, Y_1), (X_2, Y_2))$ is defined by (8.5). To estimate $\tau_{XY/Z}$, Gripenberg arranged the (X_i, Y_i, Z_i) triples in an increasing order with respect to the values of the Z variable. Letting $Z_{N(1)} \leq \cdots \leq Z_{N(n)}$ denote the order statistics for $Z_1, \ldots, Z_n$, the ordered triples correspond to $(X_{N(1)}, Y_{N(1)}, Z_{N(1)}), \ldots, (X_{N(n)}, Y_{N(n)}, Z_{N(n)})$ (with respect to increasing Z values). Gripenberg's estimator for $\tau_{XY/z}$ is then given by

$$T_{XY/Z} = \frac{1}{n-1} \sum_{i=1}^{n-1} Q((X_{N(i)}, Y_{N(i)}), (X_{N(i+1)}, Y_{N(i+1)})), \tag{8.48}$$

where, once again, $Q((X_{N(i)}, Y_{N(i)}), (X_{N(i+1)}, Y_{N(i+1)})$ is given by (8.5). The approximate $100(1-\alpha)\%$ confidence interval for $\tau_{XY/Z}$ (8.47) proposed by Gripenberg is then

$$\frac{T_{XY/Z}}{\left[1 + \dfrac{bz^2_{\alpha/2}}{n}\right]} \pm \frac{\left\{\dfrac{bz^2_{\alpha/2}}{n}\left(1 - T^2_{XY/Z} + \dfrac{bz^2_{\alpha/2}}{n}\right)\right\}^{1/2}}{\left[1 + \dfrac{bz^2_{\alpha/2}}{n}\right]}, \tag{8.49}$$

where b is an arbitrary consistent estimator of $\beta = \frac{\sigma^{*2}}{1-\tau^2_{XY/Z}}$, with σ^{*2} representing the asymptotic variance of $n^{1/2}T_{XY/Z}$. Two competing estimators b are considered by Gripenberg.

Properties

1. *Asymptotic Distribution-Freeness.* For populations satisfying Assumption A, (8.40) holds. Hence, we can control the coverage probability to be approximately $1-\alpha$ for large sample size n without having more specific knowledge about the form of the underlying bivariate (X, Y) distribution. Thus, (τ_L, τ_U) is an asymptotically distribution-free confidence interval for τ over the class of all continuous bivariate distributions.

Problems

27. For the tapeworm data of Table 8.3, find a confidence interval for τ with the approximate confidence coefficient .95.

28. For the cerebral palsy data of Table 8.4, find a confidence interval for τ with the approximate confidence coefficient .90.

29. Use only six (X, Y) pairs (those corresponding to the first six lot numbers) in Table 8.1 and compute a new estimator $\hat{\sigma}^2$ (8.38) for the asymptotic variance of $\hat{\tau}$ (see Comment 26). Compare it with the estimator based on all nine observations obtained in Example 8.3.

30. For the twins data in Table 8.5, find a lower confidence bound for τ with the approximate confidence coefficient .95. (See Comment 24.)

31. For the educational expense data in Table 8.6, find a lower confidence bound for τ with the approximate confidence coefficient .95. (See Comment 24.)

32. For the nursing data of Table 8.7, find an upper confidence bound for τ with the approximate confidence coefficient .95.

33. Suppose that $(X_1, Y_1, Z_1) = (7.0, 2.5, 1.9), (X_2, Y_2, Z_2) = (6.3, 9.6, 4.1), \quad (X_3, Y_3, Z_3) = (6.9, 3.7, 12.4), \quad (X_4, Y_4, Z_4) = (3.6, 12.1, 6.5), \quad (X_5, Y_5, Z_5) = (9.0, 6.4, 11.2), \quad (X_6, Y_6, Z_6) = (3.0, 6.2, 7.7)$, and $(X_7, Y_7, Z_7) = (4.2, 0.4, 8.2)$ represent a random sample of size $n = 7$ from a trivariate probability distribution. Estimate the partial correlation $\tau_{XY/Z}$ (8.47) between X and Y conditional on Z (See Comment 30).

8.4 AN ASYMPTOTICALLY DISTRIBUTION-FREE CONFIDENCE INTERVAL BASED ON EFRON'S BOOTSTRAP

The asymptotically distribution-free confidence for the parameter τ described in Section 8.3 is based on obtaining a mathematical expression for σ^2, the variance of $\hat{\tau}$. Such an expression depends on the unknown underlying bivariate distribution and so σ^2 must be estimated from the data. The estimate given by (8.38) is consistent, and it is used to form the confidence interval of Section 8.3. In many problems, however, it will be difficult or impossible to obtain a tractable mathematical expression for the variance of the statistic of interest. Efron's bootstrap is a general method for obtaining estimated standard deviations of estimators $\hat{\theta}$ and confidence intervals for parameters θ without requiring a tractable mathematical expression for the asymptotic variance of $\hat{\theta}$. Efron's technique eliminates the mathematical intractability obstacle by relying on computing power and is known as a *computer-intensive method.* It is applicable in a great variety of problems (see Efron (1979), Efron and Gong (1983), Efron and Tibshirani (1993), Davison and Hinkley (1997), DiCiccio and Efron (1996), and Manly (2007)). In this section, we apply Efron's bootstrap method to obtain an asymptotically distribution-free

confidence interval for the parameter τ (the population measure of association defined by (8.2)) using the estimator $\hat{\tau}$ given by (8.34), where Q (8.5) is replaced by Q^* (8.17) in the definition of K.

Procedure

Denote the observed bivariate sample values as

$$Z_1 = (X_1, Y_1), \quad Z_2 = (X_2, Y_2), \ldots, Z_n = (X_n, Y_n).$$

1. Make n random draws with replacement from the bivariate sample $Z_1, Z_2, \ldots, Z_n$. This is equivalent to doing independent random sampling from the bivariate empirical distribution function $\hat{F}$, which puts probability $1/n$ on each of the data points $Z_i, i = 1, \ldots, n$.

   For the canned tuna data of Table 8.1, $n = 9$ and

   $$Z_1 = (44.4, 2.6),\ Z_2 = (45.9, 3.1),\ Z_3 = (41.9, 2.5),\ Z_4 = (53.3, 5.0),$$
   $$Z_5 = (44.7, 3.6),\ Z_6 = (44.1, 4.0),\ Z_7 = (50.7, 5.2),\ Z_8 = (45.2, 2.8),$$
   $$Z_9 = (60.1, 3.8).$$

   A possible bootstrap sample of these data is, for example, 1 copy of Z_1, 2 copies of Z_2, 0 copies of Z_3, 0 copies of Z_4, 1 copy of Z_5, 3 copies of Z_6, 0 copies of Z_7, 1 copy of Z_8, and 1 copy of Z_9.

2. Perform step 1 a large number, say, B, of times. For each draw, compute $\hat{\tau}$. Note that in computing $\hat{\tau}$, it will be necessary to use Q^* (8.17) rather than Q (8.5) in the definition of K. This is because ties will occur in most bootstrap samples because we sample with replacement. Denote the B values of $\hat{\tau}$ as $\hat{\tau}^{*1}, \hat{\tau}^{*2}, \ldots, \hat{\tau}^{*B}$. These are called the *bootstrap replications* of $\hat{\tau}$. Let $\hat{\tau}^{*(1)} \le \hat{\tau}^{*(2)} \le \cdots \le \hat{\tau}^{*(B)}$ denote the ordered values of the bootstrap replications.

   An asymptotically distribution-free confidence interval for τ, with the approximate confidence coefficient $100(1-\alpha)\%$, is $(\tau'_{\mathrm{L}}, \tau'_{\mathrm{U}})$, where

   $$\tau'_{\mathrm{L}} = \hat{\tau}^{*(k)}, \quad \tau'_{\mathrm{U}} = \hat{\tau}^{*(B+1-k)} \tag{8.50}$$

   and

   $$k = B\left(\frac{\alpha}{2}\right). \tag{8.51}$$

   If $k = B(\alpha/2)$ is an integer, then τ'_{L} is the kth-largest bootstrap replication and τ'_{U} is the $(B+1-k)$th-largest replication. For example, if $\alpha = .10$ and $B = 1,000$, $k = 1,000(.05) = 50$, τ'_{L} is the bootstrap replication occupying position 50 in the ordered list, and τ'_{U} is the bootstrap replication occupying position 951 in the ordered list. If $B(\alpha/2)$ is not an integer, we follow the convention of Effon and Tibshirani (1993, p. 160) and set $k = \langle (B+1)(\alpha/2) \rangle$, the largest integer that is less than or equal to $(B+1)(\alpha/2)$. With this value of k, τ'_{L} is the bootstrap replication occupying position k in the ordered list and τ'_{U} is the bootstrap replication occupying position $B+1-k$ in the ordered list.

EXAMPLE 8.4 ***(Continuation of Examples 8.1, 8.2, and 8.3).***

We obtained 1000 bootstrap replications of $\hat{\tau}$. Figure 8.1 is a histogram of the 1000 bootstrap replications. The 1000 bootstrap replications are as follows:

$$\hat{\tau}^{*(1)} = -.556, \hat{\tau}^{*(2)} = -.500, \ \hat{\tau}^{*(3)} = \cdots = \hat{\tau}^{*(5)} = -.444,$$

$$\hat{\tau}^{*(6)} = -.417, \ \hat{\tau}^{*(7)} = \cdots \hat{\tau}^{*(9)} = -.361,$$

$$\hat{\tau}^{*(10)} = \hat{\tau}^{*(11)} = -.306, \hat{\tau}^{*(12)} = -.250,$$

$$\hat{\tau}^{*(13)} = \hat{\tau}^{*(14)} = -.222, \ \hat{\tau}^{*(15)} = \hat{\tau}^{*(16)} = -.194,$$

$$\hat{\tau}^{*(17)} = \hat{\tau}^{*(18)} = -.167, \ \hat{\tau}^{*(19)} = -.139,$$

$$\hat{\tau}^{*(20)} = \hat{\tau}^{*(21)} = -.111, \hat{\tau}^{*(22)} = \cdots = \hat{\tau}^{*(32)} = -.083,$$

$$\hat{\tau}^{*(33)} = \cdots = \hat{\tau}^{*(40)} = -.056, \ \hat{\tau}^{*(41)} = \cdots = \hat{\tau}^{*(50)} = -.028,$$

$$\hat{\tau}^{*(51)} = \cdots = \hat{\tau}^{*(70)} = .000, \ \hat{\tau}^{*(71)} = \cdots = \hat{\tau}^{*(85)} = .028,$$

$$\hat{\tau}^{*(86)} = \cdots = \hat{\tau}^{*(91)} = .056, \ \hat{\tau}^{*(92)} = \cdots = \hat{\tau}^{*(113)} = .083,$$

$$\hat{\tau}^{*(114)} = \cdots = \hat{\tau}^{*(133)} = -.111, \ \hat{\tau}^{*(134)} = \cdots = \hat{\tau}^{*(148)} = .139,$$

$$\hat{\tau}^{*(149)} = \cdots = \hat{\tau}^{*(173)} = .167, \ \hat{\tau}^{*(174)} = \cdots = \hat{\tau}^{*(194)} = .194,$$

$$\hat{\tau}^{*(195)} = \cdots = \hat{\tau}^{*(227)} = .222, \ \hat{\tau}^{*(228)} = \cdots = \hat{\tau}^{*(264)} = .250,$$

$$\hat{\tau}^{*(265)} = \cdots = \hat{\tau}^{*(316)} = .278, \ \hat{\tau}^{*(317)} = \cdots = \hat{\tau}^{*(351)} = .306,$$

$$\hat{\tau}^{*(352)} = \cdots = \hat{\tau}^{*(393)} = .333, \ \hat{\tau}^{*(394)} = \cdots = \hat{\tau}^{*(433)} = .361,$$

$$\hat{\tau}^{*(434)} = \cdots = \hat{\tau}^{*(478)} = .389, \ \hat{\tau}^{*(479)} = \cdots = \hat{\tau}^{*(530)} = .417,$$

$$\hat{\tau}^{*(531)} = \cdots = \hat{\tau}^{*(592)} = .444, \ \hat{\tau}^{*(593)} = \cdots = \hat{\tau}^{*(640)} = .472,$$

$$\hat{\tau}^{*(641)} = \cdots = \hat{\tau}^{*(687)} = .500, \ \hat{\tau}^{*(688)} = \cdots = \hat{\tau}^{*(730)} = .528,$$

$$\hat{\tau}^{*(731)} = \cdots = \hat{\tau}^{*(769)} = .556, \ \hat{\tau}^{*(770)} = \cdots = \hat{\tau}^{*(808)} = .583,$$

$$\hat{\tau}^{*(809)} = \cdots = \hat{\tau}^{*(847)} = .611, \ \hat{\tau}^{*(848)} = \cdots = \hat{\tau}^{*(880)} = .639,$$

$$\hat{\tau}^{*(881)} = \cdots = \hat{\tau}^{*(912)} = .667, \ \hat{\tau}^{*(913)} = \cdots = \hat{\tau}^{*(935)} = .694,$$

$$\hat{\tau}^{*(936)} = \cdots = \hat{\tau}^{*(955)} = .722, \ \hat{\tau}^{*(956)} = \cdots = \hat{\tau}^{*(969)} = .750,$$

$$\hat{\tau}^{*(970)} = \cdots = \hat{\tau}^{*(986)} = .778, \ \hat{\tau}^{*(987)} = \cdots = \hat{\tau}^{*(992)} = .806,$$

$$\hat{\tau}^{*(993)} = \hat{\tau}^{*(994)} = .833, \ \hat{\tau}^{*(995)} = \cdots = \hat{\tau}^{*(999)} = .861, \hat{\tau}^{*(1000)} = .889.$$

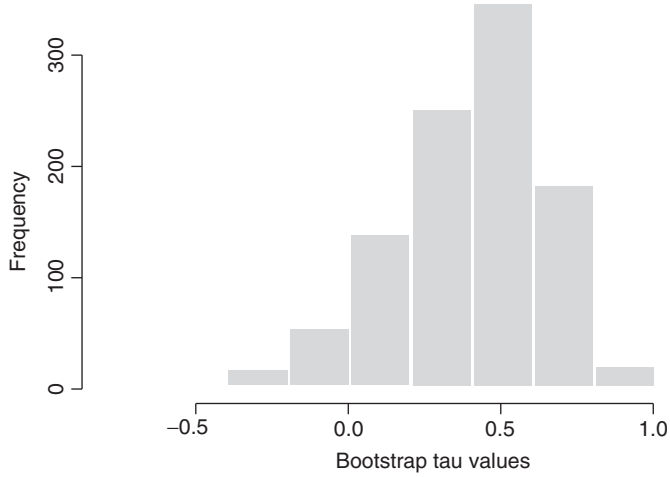

Figure 8.1 Histogram of 1000 bootstrap replications of Kendall's sample correlation coefficient for the canned tuna data of Table 8.1.

For an approximate 90% confidence interval $\alpha = .1, (\alpha/2) = 0.5$, and from (8.51), we find $k = 1000(.05) = 50$. Then from (8.50) and the ordered list of 1000 bootstrap replications,

$$\tau'_{\mathrm{L}} = \hat{\tau}^{*(50)} = -.028, \quad \tau'_{\mathrm{U}} = \hat{\tau}^{*(951)} = .722.$$

The command `kendall.ci` will provide a bootstrap confidence interval. In addition to the arguments specified in Example 8.3, one here must set `bootstrap=T` and a value for the number of replicates B. For example,

```
kendall.ci(x, y, alpha=.1, type="t", bootstrap=T, B=1000)
```

will find an interval similar, but almost certainly not identical, to the interval above or the interval found in Example 8.3. Running the above command five times results in the following intervals: $(-.063, .818)$, $(-.067, .857)$, $(-.063, .824)$. $(-.030, .824)$, and $(-.030, .871)$. Recall that the method from Section 8.3 gave the interval (.027, .861).

Comments

31. *The Bootstrap Estimated Standard Error.* For Kendall's sample correlation coefficient $\hat{\tau}$, the standard deviation of $\hat{\tau}$, which we have denoted thus far as σ, depends on the bivariate distribution function $F_{X,Y}$. We now denote $F_{X,Y}$ as F, dropping the subscripts. We could exhibit the dependence of σ on F by writing σ as $\sigma(F)$. Although F is unknown, it can be estimated by the bivariate empirical distribution function $\hat{F}$, which puts probability $1/n$ on each of the observed data points $Z_i = (X_i, Y_i), i = 1, \ldots, n$. The bootstrap estimate of $\sigma(F)$ is $\sigma(\hat{F})$, where $\sigma(\hat{F})$ is the standard deviation of $\hat{\tau}$ when the true underlying distribution is $\hat{F}$ rather than F. A tractable mathematical expression for $\sigma(\hat{F})$ is very difficult to obtain. However, $\sigma(\hat{F})$ can be estimated using the B bootstrap replications, by

$$\hat{\sigma}_B = \left\{ \frac{\sum_{i=1}^{B} (\hat{\tau}^{*i} - \hat{\tau}^{*\cdot})^2}{(B-1)} \right\}^{1/2}, \tag{8.52}$$

where

$$\hat{\tau}^{*\cdot} = \frac{\sum_{i=1}^{B} \hat{\tau}^{*i}}{B}. \tag{8.53}$$

As B tends to ∞, $\hat{\sigma}_B$ tends to $\sigma(\hat{F})$. As n tends to ∞, $\sigma(\hat{F})$ tends to $\sigma(F)$. Thus, $\hat{\sigma}_B$ can be used as an estimate of the standard deviation of $\hat{\tau}$.

32. *The Bootstrap in the One-Sample Nonparametric Framework.* In this section, we applied the bootstrap in a bivariate situation, where the data are bivariate observations $Z_i = (X_i, Y_i), i = 1, \ldots, n$, and the parameter of interest is τ. The bootstrap can be used in a wide variety of situations, including the one-sample problem, the k-sample problem, censored data problems, and complicated multivariate frameworks. (See Efron and Gong (1983), Efron and Tibshirani (1993), Davison and Hinkley (1997), and DiCiccio and Efron (1996).) In this comment, we describe the approach in the context of the one-sample nonparametric framework.

Suppose we are interested in estimating a parameter $\theta = \theta(F)$, when $X_1, \ldots, X_n$ are a random sample from an unknown distribution F. The nonparametric maximum likelihood estimate of F is the statistic $\hat{\theta} = \theta(F_n)$, where F_n is the sample distribution function. For example, if we are interested in estimating the rth moment of the F distribution, $\theta(F) = E(X^r)$, then $\theta(F_n) = (\sum_{i=1}^{n} X_i^r)/n$.

The bootstrap procedure in the one-sample problem is analogous to the procedure we described for the bivariate situation. The steps are as follows:

1. Make n random draws with replacement from the sample $X_1, \ldots, X_n$.
2. Perform step 1 a large number, say B, of times. For each draw, compute $\hat{\theta}$. Denote the B values of $\hat{\theta}$ as $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$. These are the bootstrap replications of $\hat{\theta}$.

The bootstrap estimate of the standard deviation of $\hat{\theta}$ is

$$\hat{\sigma}_B = \left\{ \frac{\sum_{i=1}^{B} (\hat{\theta}^{*i} - \hat{\theta}^{*\cdot})^2}{(B-1)} \right\}^{1/2}, \tag{8.54}$$

where

$$\hat{\theta}^{*\cdot} = \frac{\sum_{i=1}^{B} \hat{\theta}^{*i}}{B}. \tag{8.55}$$

An asymptotically distribution-free confidence interval for θ, with approximate confidence coefficient $100(1-\alpha)\%$, is $(\theta'_{\mathrm{U}}, \theta'_{\mathrm{L}})$, where

$$\theta'_{\mathrm{L}} = \hat{\theta}^{*(k)}, \quad \theta'_{\mathrm{U}} = \hat{\theta}^{*(B+1-k)}, \tag{8.56}$$

where $\hat{\theta}^{*(1)} \le \hat{\theta}^{*(2)} \le \cdots \le \hat{\theta}^{*(B)}$ are the ordered values for the bootstrap replications and $k = \langle (B+1)(\alpha/2) \rangle$, the largest integer that is less than or equal to $(B+1)(\alpha/2)$.

The confidence interval defined by (8.56) is called the *percentile interval*. Let $\hat{G}$ denote the cumulative distribution function of $\hat{\theta}^*$:

$$\hat{G}(t) = \frac{\#\{\hat{\theta}^{*i} < t\}}{B}. \tag{8.57}$$

The end points $\theta'_{\mathrm{L}}, \theta'_{\mathrm{U}}$ are, respectively, the α and $1-\alpha$ percentiles of $\hat{G}$.

The percentile confidence interval is *transformation-respecting*. If $\eta = m(\theta)$ is a monotone transformation, then a confidence interval (η_L, η_U) for the parameter η is obtained directly from the confidence interval (8.56) for θ via $\eta_L = m(\theta'_L), \eta_U = m(\theta'_U)$. For example, a confidence interval for θ^2 is obtained by squaring the end points of the confidence interval for θ.

The percentile confidence interval is *range-preserving*. For example, consider the percentile interval based on bootstrapping $\hat{\tau}$. The values of the parameter τ, Kendall's population correlation coefficient, are always between -1 and 1. The possible values of the estimator $\hat{\tau}$ also are in the interval $[-1, 1]$. Thus, the bootstrap replications of $\hat{\tau}$ must be in the interval $[-1, 1]$, as must the confidence interval end points, because the end points are particular bootstrap replications. More generally, the estimator $\hat{\theta}$ of the form $\hat{\theta} = \theta(F_n)$ satisfies the same range restrictions as $\theta = \theta(F)$, and thus, the percentile interval based on bootstrapping $\hat{\theta}$ also satisfies the same range restrictions as θ.

33. *The BC_a Confidence Interval.* Efron and Tibshirani (1993, Chapter 14) (see also DiCiccio and Efron, 1996) present a method, called the BC_a method, that gives more accurate confidence limits than does the percentile method of Comment 32. The acronym BC_a means "bias-corrected and accelerated." The BC_a method depends on a bias-correction z_0 and an acceleration a. In the one-sample nonparametric framework, z_0 can be estimated by

$$\hat{z}_0 = \Phi^{-1}\left\{\frac{\#\{\hat{\theta}^{*i} < \hat{\theta}\}}{B}\right\}, \tag{8.58}$$

where Φ denotes the standard normal cumulative distribution function. Thus, $\hat{z}_0$ is Φ^{-1} of the proportion of the bootstrap replications less than $\hat{\theta}$.

The estimate of a is

$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\theta}_\cdot - \hat{\theta}_{-i})^3}{6\left\{\sum_{i=1}^{n}(\hat{\theta}_\cdot - \hat{\theta}_{-i})^2\right\}^{3/2}}, \tag{8.59}$$

where $\hat{\theta}_{-i}$ is the estimate of θ obtained by deleting X_i, and computing $\hat{\theta}$ for the reduced sample $X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ and

$$\hat{\theta}_\cdot = \frac{\sum_{i=1}^{n}\hat{\theta}_{-i}}{n}. \tag{8.60}$$

The lower and upper end points of the $100(1-\alpha)\%$ confidence interval are

$$\theta''_L = \hat{G}^{-1}\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha/2)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right), \tag{8.61}$$

$$\theta''_U = \hat{G}^{-1}\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha/2)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha/2)})}\right), \tag{8.62}$$

where in (8.61) and (8.62), Φ is the standard normal distribution function, $z^{(\alpha/2)} = \Phi^{-1}(\alpha/2), z^{(1-\alpha/2)} = \Phi^{-1}(1-\alpha/2)$ (if, for example, $a = .10$, then

$z^{(\alpha/2)} = -1.65$, and $z^{(1-\alpha/2)} = 1.65$), $\hat{G}$ is given by (8.57), $\hat{z}_0$ is given by (8.58), and $\hat{a}$ is given by (8.59).

The end points θ''_{L} and θ''_{U} of the BC_a interval are also percentiles of the bootstrap distribution $\hat{G}$ but not necessarily the same ones as given by the percentile interval. If $\hat{a} = \hat{z}_0 = 0$, the BC_a and percentile intervals are the same.

The BC_a interval also enjoys the transformation-respecting and range-preserving properties that hold for the percentile interval. The BC_a interval, however, has an accuracy advantage. The BC_a interval has a second-order accuracy property, whereas the percentile interval is only first-order accurate. See Section 14.3 of Efron and Tibshirani (1993).

The appendix of Efron and Tibshirani (1993) describes some available bootstrap software and contains some programs in the S language, including a program for computing BC_{a} intervals.

34. *The Choice of B, the Number of Bootstrap Replications.* The choice of B depends to some extent on the particular statistic that is being bootstrapped and the complexity of the situation. Efron and Tibshirani (1993, p. 52) give some rules of thumb based on their extensive experience with the bootstrap. Roughly speaking, $B = 200$ replications are usually sufficient for estimating a standard error but much larger values of B, such as 1000 or 2000, are required for bootstrap confidence intervals.

35. *An Example Where the Bootstrap Fails.* Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution on $(0, \theta)$. The maximum likelihood estimator of θ, the upper end point of the interval, is $\hat{\theta} = \text{maximum}(X_1, \ldots, X_n) = X_{(n)}$. Efron and Tibshirani (1993, p. 81) point out that the bootstrap does not do well in this situation. Miller (1964) showed that the jackknife estimator of θ also fails in this situation, because it depends not only on $X_{(n)}$ but also on $X_{(n-1)}$, the second largest observation, and the latter contains no additional information about θ when the value of $X_{(n)}$ is available.

36. *Jackknife versus Bootstrap.* For a linear statistic of the form $\hat{\theta} = \mu + \left\{\sum_{i=1}^{n} h(X_i)/n\right\}$, where μ is a constant and h is a function, there is no loss of information in using the jackknife rather than the bootstrap. For nonlinear statistics, there is a loss of information and the bootstrap should be preferred. See Efron and Tibshirani (1993) for a detailed discussion of the relationship between the jackknife and the bootstrap.

    One disadvantage of the bootstrap is that two different people bootstrapping the same data will not in general get the same bootstrap estimate of the standard deviation or the same confidence interval. This violates what Gleser (1996) calls "the first law of applied statistics," namely: "Two individuals using the same statistical method on the same data should arrive at the same conclusion." See Gleser (1996) for other disadvantages of the bootstrap.

37. *Development of the Bootstrap.* The bootstrap was formally introduced by Efron (1979). Efron and Tibshirani (1993, p. 56), however, credit many authors for similar ideas, and they designate as "particularly notable" the contributions of Hartigan's typical value theory (1969, 1971, 1975). Hartigan recognized the wide applicability of subsample methods (a subsample of $X_1, \ldots, X_n$ is any subset of the whole sample) as a tool for assessing variability. See Efron and Tibshirani (1993, p. 56) for references to other papers that contain ideas related to bootstrapping.

Properties

See Bickel and Freedman (1981) for asymptotic consistency. See Efron and Tibshirani (1993) and Davison and Hinkley (1997) for various properties including accuracy, transformation-respecting, range-preserving, and the relationship of the bootstrap to the jackknife. See Hall (1992) for a high-level mathematical treatment of the bootstrap. See Manly (2007) for bootstrap methods in biology.

Problems

34. For the cerebral palsy data of Table 8.4, use the bootstrap method to find a confidence interval for τ with approximate confidence coefficient .90. Compare your results with those of Problem 28.

35. For the psychological test scores data of Table 8.5, use the bootstrap method to find a confidence interval for τ with approximate confidence coefficient .95.

36. Consider the case $n = 3$ where you have three bivariate observations Z_1, Z_2, and Z_3. List the possible bootstrap samples and give the corresponding probability of each being selected on a given bootstrap replication.

37. Consider the case where you have four bivariate observations Z_1, Z_2, Z_3, and Z_4. List the possible bootstrap samples and give the corresponding probability of each being selected on a given bootstrap replication.

38. Illustrate by means of an example or show directly that with n observations the number of possible bootstrap samples is $\binom{2n-1}{n}$.

39. Show that if $\hat{a} = \hat{z}_0 = 0$, the BC_a interval given by (8.61) and (8.62) reduces to the percentile interval.

8.5 A DISTRIBUTION-FREE TEST FOR INDEPENDENCE BASED ON RANKS (SPEARMAN)

Hypothesis

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a continuous bivariate population (i.e., Assumption A is satisfied) with joint distribution function $F_{X,Y}$ and marginal distribution functions F_X and F_Y. In this section, we return to the problem of testing for independence between the X and Y variables corresponding to the null hypothesis H_0 (8.1). Here, however, alternatives to H_0 will no longer be stated in terms of the correlation coefficient τ (8.2). Instead, the alternatives of interest in this section are less specifically interpretable, corresponding to the quite general (but vague; see Comment 47) concepts of positive or negative association between the X and Y variables.

Procedure

To compute the Spearman rank correlation coefficient r_s, we first order the n X observations from least to greatest and let R_i denote the rank of $X_i, i = 1, \ldots, n$, in this ordering. Similarly, we separately order the n Y observations from least to greatest and let S_i denote the rank of $Y_i, i = 1, \ldots, n$, in this ordering. The Spearman (1904) rank correlation coefficient is defined as the Pearson product moment sample correlation of

the R_i and the S_i. (See Comment 40.) When no ties within a sample are present, this is equivalent to two computationally efficient formulae:

$$r_s = \frac{12 \sum_{i=1}^{n} \left\{ \left[R_i - \frac{n+1}{2} \right] \left[S_i - \frac{n+1}{2} \right] \right\}}{n(n^2 - 1)} \tag{8.63}$$

$$= 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}, \tag{8.64}$$

where $D_i = S_i - R_i, i = 1, \ldots, n$.

a. *One-Sided Upper-Tail Test.* To test the null hypothesis of independence, H_0 (8.1), versus the directional alternative

$$H_1 : [X \text{ and } Y \text{ are positively associated}] \tag{8.65}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } r_s \geq r_{s,\alpha}; \quad \text{otherwise do not reject,} \tag{8.66}$$

where the constant $r_{s,\alpha}$ is chosen to make the type I error probability equal to α. Values of $r_{s,\alpha}$ are found with the command `qSpearman`.

b. *One-Sided Lower-Tail Test.* To test independence, H_0 (8.1), versus the directional alternative

$$H_2 : [X \text{ and } Y \text{ are negatively associated}] \tag{8.67}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } r_s \leq -r_{s,\alpha}; \quad \text{otherwise do not reject.} \tag{8.68}$$

c. *Two-Sided Test.* To test independence, H_0 (8.1), versus the general dependency alternative

$$H_3 : [X \text{ and } Y \text{ are not independent variables}] \tag{8.69}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } |r_s| \geq r_{s,\alpha/2}; \quad \text{otherwise do not reject.} \tag{8.70}$$

This two-sided procedure is the two-sided symmetric test with $\alpha/2$ probability in each tail of the null distribution of r_s.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of r_s, suitably standardized. For this standardization, we need to know the expected value and variance of r_s when the null hypothesis of independence is true. Under H_0, the expected value and variance of r_s are

$$E_0(r_s) = 0 \tag{8.71}$$

and

$$\text{var}_0(r_s) = \frac{1}{n-1}, \tag{8.72}$$

respectively. These expressions for $E_0(r_s)$ and $\text{var}_0(r_s)$ are verified by direct calculations in Comment 42 for the special case of $n = 4$. General derivations of both expressions are discussed in Comment 45.

The standardized version of r_s is

$$r_s^* = \frac{r_s - E_0(r_s)}{\{\text{var}_0(r_s)\}^{1/2}} = (n-1)^{1/2} r_s. \tag{8.73}$$

When H_0 is true, r_s^* has, as n tends to infinity, an asymptotic $N(0,1)$ distribution. (See Comment 45 for indications of the proof.) The normal theory approximation for procedure (8.66) is

$$\text{Reject } H_0 \text{ if } r_s^* \geq z_\alpha; \quad \text{otherwise do not reject,} \tag{8.74}$$

the normal theory approximation for procedure (8.68) is

$$\text{Reject } H_0 \text{ if } r_s^* \leq -z_\alpha; \quad \text{otherwise do not reject,} \tag{8.75}$$

and the normal theory approximation for procedure (8.70) is

$$\text{Reject } H_0 \text{ if } |r_s^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{8.76}$$

Ties

If there are ties among the n X observations and/or separately among the n Y observations, assign each of the observations in a tied (either X or Y) group the average of the integer ranks that are associated with the tied group. After computing r_s with these average ranks for tied observations, use procedure (8.66), (8.68), or (8.70). Note, however, that this test associated with tied X's and/or tied Y's is only approximately, and not exactly, of significance level α. (To get an exact level α test even in this tied setting, see Comment 46.)

If there are tied X's and/or tied Y's, Spearman's rank correlation coefficient calculated with Pearson's correlation does not require modification. If using the computationally efficient version of r_s at (8.64), some changes to the statistic are necessary. The statistic r_s in this case becomes

$$r_s = \frac{n(n^2-1) - 6\sum_{s=1}^{n} D_s^2 - \frac{1}{2}\left\{\sum_{i=1}^{g}\left[t_i\left(t_i^2-1\right)\right] + \sum_{j=1}^{h}\left[u_j\left(u_j^2-1\right)\right]\right\}}{\left\{\left[n(n^2-1) - \sum_{i=1}^{g} t_i\left(t_i^2-1\right)\right]\left[n(n^2-1) - \sum_{j=1}^{h} u_j\left(u_j^2-1\right)\right]\right\}^{1/2}}, \tag{8.77}$$

where in (8.77) g denotes the number of tied X groups, t_i is the size of tied X group i, h is the number of tied Y groups, and u_j is the size of tied Y group j. We note that an untied $X(Y)$ observation is considered to be a tied $X(Y)$ group of size 1. In particular, if neither the collection of X nor the collection of Y observations contains tied values, we have $g = h = n, t_j = u_j = 1, i = 1, \ldots, n$, and $j = 1, \ldots, n$. In this case of no tied X's

and no tied Y's, each term involving either $(t_i^2 - 1)$ or $(u_j^2 - 1)$ reduces to zero and the "ties" expression for r_s in (8.77) reduces to the "no-ties" form for r_s, as given in (8.64).

As a consequence of this effect that ties have on the null distribution of r_s, in order to use the large-sample approximation when there are tied X observations and/or tied Y observations, we first compute r_s^* (8.73) using average ranks and the ties-corrected version of r_s (8.77). Approximation (8.74), (8.75), or (8.76) can then be applied, as appropriate for the problem, with this value of r_s^*.

EXAMPLE 8.5 ***Proline and Collagen in Liver Cirrhosis.***

Kershenobich, Fierro, and Rojkind (1970) have studied the relation between the free pool of proline and collagen content in human liver cirrhosis. The data in Table 8.9 are based on an analysis of cirrhotic livers from seven patients, each having a histological diagnosis of portal cirrhosis.

We are interested in assessing whether there is a positive relationship between the total collagen and the free proline in cirrhotic livers. Thus, we wish to apply procedure (8.66) to test the hypothesis of independence, H_0 (8.1), versus the alternative, H_1 (8.65), of positive association. For purposes of illustration, we consider the significance level $\alpha = .01$. The statistic r_s is symmetric about 0 (see Comment 43), so $P(r_s \geq r_{s,.01}) = P(r_s \leq -r_{s,.01})$. In R, this is found with

```
qSpearman(.01, r=7),
```

where `r` is the number of samples. The result is $-.786$, so procedure (8.66) becomes

$$\text{Reject } H_0 \text{ if } r_s \geq .786.$$

Ranking the X (total collagen) values from least to greatest, using average ranks for the tied pair, we obtain $R_1 = (1 + 2)/2 = 1.5, R_2 = (1 + 2)/2 = 1.5, R_3 = 3, R_4 = 4, R_5 = 5, R_6 = 6$, and $R_7 = 7$. Similarly, ranking the Y (free proline) values from least to greatest, again using average ranks for the tied pairs, we find $S_1 = (2 + 3)/2 = 2.5, S_2 = 4, S_3 = (2 + 3)/2 = 2.5, S_4 = 1, S_5 = 5, S_6 = 6$, and $S_7 = 7$. Taking differences, we see that

$$D_1 = 2.5 - 1.5 = 1, \quad D_2 = 4 - 1.5 = 2.5, \quad D_3 = 2.5 - 3 = -.5$$
$$D_4 = 1 - 4 = -3, \quad D_5 = 5 - 5 = 0, \quad D_6 = 6 - 6 = 0$$
$$D_7 = 7 - 7 = 0.$$

Table 8.9 Free Proline and Total Collagen Contents of Cirrhotic Patients

Patient	Total collagen, X_i (mg/g dry weight of liver)	Free proline, Y_i, (μ mole/g dry weight of liver)
1	7.1	2.8
2	7.1	2.9
3	7.2	2.8
4	8.3	2.6
5	9.4	3.5
6	10.5	4.6
7	11.4	5.0

Source: D. Kershenobich, F. J. Fierro, and M. Rojkind (1970).

There are ties, so we need to use the ties-corrected version of the statistic r_s if using (8.64). If using Pearson's correlation of ranks, no modifications are necessary.

For this purpose, we note that there are $g = 6$ tied X groups, with $t_1 = 2, t_2 = t_3 = t_4 = t_5 = t_6 = 1$, and $h = 6$ tied Y groups, with $u_2 = 2, u_1 = u_3 = u_4 = u_5 = u_6 = 1$. Thus, for these tied data, the modified value of r_s (8.77) is calculated to be

$$r_s = \frac{7(7^2-1) - 6[(1)^2 + (2.5)^2 + (-.5)^2 + (-3)^2 + 3(0)^2] - \frac{1}{2}(2)(2)(2^2-1)}{\{[7(7^2-1) - 2(2^2-1)][7(7^2-1) - 2(2^2-1)]\}^{1/2}}$$

$$= \frac{7(48) - 6(16.5) - 6}{\{[7(48) - 6][7(48) - 6]\}^{1/2}} = \frac{231}{330} = .700.$$

This value of r_s is also obtained through the R command

```
cor(x, y, method="spearman")
```

where x and y are the data from Table 8.9. This value of r_s is not greater than the critical value .786, so we do not reject the null at the $\alpha = .01$ level. Note that the critical value given by R results in a significance level of $\alpha = .024$, not $\alpha = .01$.

The one-sided P-value for these data is the smallest significance level at which we can reject H_0 in favor of a positive association between total collagen and free proline in cirrhotic patients with the observed value of the test statistic $r_s = .700$. We see that the P-value is $P_0(r_s \geq .700)$. By symmetry, this is the same as $P_0(r_s \leq -.700)$. The R command pSpearman will provide the following:

```
pSpearman(-.700, r=7)=.044.
```

Thus, there is some marginal evidence that total collagen and free proline are positively associated in subjects with liver cirrhosis.

For the large-sample approximation, we use $r_s = .700$ to compute r_s^* (8.73) and obtain

$$r_s^* = (6)^{1/2}(.700) = 1.71.$$

Thus, the smallest significance level at which we can reject H_0 in favor of positive association between total collagen and free proline in subjects with liver cirrhosis using the normal theory approximation is .0436 ($z_{.0436} = 1.71$).

The R function cor.test reproduces the above analysis.

```
cor.test (x, y, method="spearman", alternative="greater")
```

produces this output:

```
 Spearman's rank correlation rho

data: x and y
S = 16.8, p-value = 0.03996
alternative hypothesis: true rho is greater than 0
sample estimates:
rho
0.7

  Warning message:
In cor.test.default(x8.9, y8.9, method = "s", alt = "g") :
Cannot compute exact p-values with ties
```

The statistic reported as S is the sum of D_i^2 from (8.64). However, in the presence of ties, this value is not accurate. Rather, it is found using Pearson's correlation of the ranks for r_s and then solving for D_i^2 in (8.64). R provides a warning about inexact P-values in the presence of ties. In this case, one should use `pSpearman` to obtain a P-value.

Comments

38. *Motivation for the Test.* The null hypothesis of this section is that the X and Y variables are independent, which, in the case of no ties, implies that any permutation of the X ranks $(R_1, \ldots, R_n)$ is equally likely to occur with any permutation of the Y ranks $(S_1, \ldots, S_n)$. As a result, under the null hypothesis H_0 (8.1) of independence, the Spearman rank correlation coefficient r_s (8.64) will have a tendency to assume values near zero. However, when the alternative H_1 : [X and Y are positively associated] is true, the rank vectors $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ will tend to agree, resulting in small differences $D_i = S_i - R_i, i = 1, \ldots, n$. Thus, when H_1 (8.65) is true, we would expect the value of $\sum_{j=1}^{n} D_j^2$ to be small and the resulting value of r_s (8.64) to be large and positive. This suggests rejecting H_0 in favor of positive association H_1 (8.65) for large positive values of r_s and motivates procedures (8.66) and (8.74). Similar rationales apply to procedures (8.68), (8.70), (8.75), and (8.76).

39. *Computation of* r_s. The value of r_s (8.63) can also be obtained in R using

    ```
    cor(rank(x), rank(y), method=''pearson'').
    ```

    The command rank provides the ranks of a sample. The default method of dealing with ties in this command is to average the ranks within a tie group.

40. *Pearson's Product Moment Sample Correlation Coefficient.* The classical Pearson product moment sample correlation coefficient for the pair of vectors $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ is given by

$$r_p = \frac{\sum_{k=1}^{n} (X_k - \overline{X})(Y_k - \overline{Y})}{\left[\sum_{i=1}^{n} (X_i - \overline{X})^2 \ \sum_{j=1}^{n} (Y_j - \overline{Y})^2\right]^{1/2}}, \tag{8.78}$$

    where $\overline{X} = \sum_{s=1}^{n} X_s/n$ and $\overline{Y} = \sum_{t=1}^{n} Y_t/n$. We note that the Spearman rank correlation coefficient r_s is simply the classical correlation coefficient applied to the rank vectors $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ instead of the actual X and Y observations, respectively. (See Problem 49.)

41. *Derivation of the Distribution of* r_s *under* H_0 *(No-Ties Case).* Without loss of generality, we take $R_1 = 1, \ldots, R_n = n$; under H_0 (8.61) all possible $n!(S_1, S_2, \ldots, S_n)$ Y-rank configurations are equally likely, therefore each has null probability $1/n!$.

    Let us consider the case $n = 4$. In the following table, we display the $4! = 24$ possible (S_1, S_2, S_3, S_4) configurations, the associated values of r_s, and the corresponding null probabilities.

(R_1, R_2, R_3, R_4)	(S_1, S_2, S_3, S_4)	Null probability	r_s
(1, 2, 3, 4)	(1, 2, 3, 4)	$\frac{1}{24}$	1
(1, 2, 3, 4)	(1, 2, 4, 3)	$\frac{1}{24}$	.8
(1, 2, 3, 4)	(1, 3, 2, 4)	$\frac{1}{24}$	.8
(1, 2, 3, 4)	(1, 3, 4, 2)	$\frac{1}{24}$	.4
(1, 2, 3, 4)	(1, 4, 2, 3)	$\frac{1}{24}$	.4
(1, 2, 3, 4)	(1, 4, 3, 2)	$\frac{1}{24}$	.2
(1, 2, 3, 4)	(2, 1, 3, 4)	$\frac{1}{24}$	.8
(1, 2, 3, 4)	(2, 1, 4, 3)	$\frac{1}{24}$	.6
(1, 2, 3, 4)	(2, 3, 1, 4)	$\frac{1}{24}$	.4
(1, 2, 3, 4)	(2, 3, 4, 1)	$\frac{1}{24}$	−.2
(1, 2, 3, 4)	(2, 4, 1, 3)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(2, 4, 3, 1)	$\frac{1}{24}$	−.4
(1, 2, 3, 4)	(3, 1, 2, 4)	$\frac{1}{24}$	.4
(1, 2, 3, 4)	(3, 1, 4, 2)	$\frac{1}{24}$	0
(1, 2, 3, 4)	(3, 2, 1, 4)	$\frac{1}{24}$	.2
(1, 2, 3, 4)	(3, 2, 4, 1)	$\frac{1}{24}$	−.4
(1, 2, 3, 4)	(3, 4, 1, 2)	$\frac{1}{24}$	−.6
(1, 2, 3, 4)	(3, 4, 2, 1)	$\frac{1}{24}$	−.8
(1, 2, 3, 4)	(4, 1, 2, 3)	$\frac{1}{24}$	−.2
(1, 2, 3, 4)	(4, 1, 3, 2)	$\frac{1}{24}$	−.4
(1, 2, 3, 4)	(4, 2, 1, 3)	$\frac{1}{24}$	−.4
(1, 2, 3, 4)	(4, 2, 3, 1)	$\frac{1}{24}$	−.8
(1, 2, 3, 4)	(4, 3, 1, 2)	$\frac{1}{24}$	−.8
(1, 2, 3, 4)	(4, 3, 2, 1)	$\frac{1}{24}$	−1

Thus, for example, the probability is $\frac{3}{24}$ under H_0 that r_s is equal to .8, because $r_s = .8$ when any of the three outcomes $(S_1, S_2, S_3, S_4) = (1, 2, 4, 3)$, (1, 3, 2, 4), or (2, 1, 3, 4) occurs and each of these outcomes has null probability $\frac{1}{24}$. Simplifying, we obtain the null distribution

Possible value of r_s	Probability under H_0
−1.0	$\frac{1}{24}$
−0.8	$\frac{3}{24}$
−0.6	$\frac{1}{24}$
−0.4	$\frac{4}{24}$
−0.2	$\frac{2}{24}$
0.0	$\frac{2}{24}$
0.2	$\frac{2}{24}$
0.4	$\frac{4}{24}$
0.6	$\frac{1}{24}$
0.8	$\frac{3}{24}$
1.0	$\frac{1}{24}$

The probability, under H_0, that r_s is greater than or equal to .6, for example, is therefore

$$\begin{aligned} P_0(r_s \geq .6) &= P_0(r_s = 1.0) + P_0(r_s = .8) + P_0(r_s = .6) \\ &= \frac{1}{24} + \frac{3}{24} + \frac{1}{24} = \frac{5}{24}. \end{aligned}$$

Note that we have obtained the null distribution of r_s without specifying the form of the underlying independent X and Y populations under H_0, beyond the point of requiring that they be continuous. This is why the test procedures based on r_s are called *distribution-free procedures*. From the null distribution of r_s, we can determine the critical value $r_{s,\alpha}$ and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific forms of the underlying continuous and independent X and Y distributions.

42. *Calculation of the Mean and Variance of r_s under the Null Hypothesis.* In displays (8.71) and (8.72), we presented formulas for the mean and variance of r_s when the null hypothesis is true. In this comment, we illustrate a direct calculation of $E_0(r_s)$ and $\text{var}_0(r_s)$ in the particular case of $n = 4$, using the null distribution of r_s obtained in Comment 41. (Later, in Comment 45, we present general derivations of $E_0(r_s)$ and $\text{var}_0(r_s)$.) The null mean, $E_0(r_s)$, is obtained by multiplying each possible value of r_s with its probability under H_0 and summing the products. Thus,

$$\begin{aligned} E_0(r_s) = &-1\left(\frac{1}{24}\right) - .8\left(\frac{3}{24}\right) - .6\left(\frac{1}{24}\right) - .4\left(\frac{4}{24}\right) - .2\left(\frac{2}{24}\right) \\ &+ 0\left(\frac{2}{24}\right) + .2\left(\frac{2}{24}\right) + .4\left(\frac{4}{24}\right) + .6\left(\frac{1}{24}\right) + .8\left(\frac{3}{24}\right) + 1\left(\frac{1}{24}\right) \\ = &\ 0. \end{aligned}$$

This is in agreement with the value stated in (8.71). A check on the expression for $\text{var}_0(r_s)$ is also easily performed, using the well-known fact that

$$\text{var}_0(r_s) = E_0(r_s^2) - \{E_0(r_s)\}^2.$$

The value of $E_0\left(r_s^2\right)$, the second moment of the null distribution of r_s, is again obtained by multiplying possible values (in this case, of r_s^2 by the corresponding probabilities under H_0 and summing). We find

$$\begin{aligned} E_0(r_s^2) &= \left[(1+1)\left(\frac{1}{24}\right) + (.64+.64)\left(\frac{3}{24}\right) + (.36+.36)\left(\frac{1}{24}\right)\right. \\ &\quad \left. + (.16+.16)\left(\frac{4}{24}\right) + (.04+.04)\left(\frac{2}{24}\right) + 0\left(\frac{2}{24}\right)\right] \\ &= \frac{1}{3}. \end{aligned}$$

Thus,

$$\text{var}_0(r_s) = \frac{1}{3} - 0^2 = \frac{1}{3},$$

which agrees with what we obtain using (8.72) directly, namely,

$$\text{var}_0(r_s) = \frac{1}{4-1} = \frac{1}{3}.$$

43. *Symmetry of the Distribution of r_s under the Null Hypothesis.* When H_0 is true, the distribution of r_s is symmetric about its mean 0. (See Comment 41 for verification of this when $n = 4$.) This implies that

$$P_0(r_s \le -x) = P_0(r_s \ge x), \tag{8.79}$$

for all x. Equation (8.79) is used directly to convert upper-tail probabilities to lower-tail probabilities. In particular, it follows from (8.79) that the lower αth percentile for the null distribution of r_s is $-r_{s,\alpha}$; thus, the use of $-r_{s,\alpha}$ as the critical value in procedure (8.68).

44. *Equivalent Form.* Let $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ be the vectors of separate ranks for the X and Y observations, respectively. We note that

$$\begin{aligned} \sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right) &= \sum_{i=1}^{n} R_i S_i - \frac{n+1}{2}\sum_{i=1}^{n} R_i \\ &\quad - \frac{n+1}{2}\sum_{i=1}^{n} S_i + \frac{n(n+1)^2}{4}. \end{aligned}$$

However, $\sum_{i=1}^{n} R_i = \sum_{i=1}^{n} S_i = \sum_{i=1}^{n} i = n(n+1)/2$. Thus, we have

$$\sum_{i=1}^{n}\left(R_i - \frac{n+1}{2}\right)\left(S_i - \frac{n+1}{2}\right) = \sum_{i=1}^{n} R_i S_i - \frac{n(n+1)^2}{4}.$$

Combining this fact with the definition of r_s in display (8.63), we obtain an alternative computational expression for r_s, namely,

$$r_s = \frac{12\sum_{i=1}^{n} R_i S_i}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right). \tag{8.80}$$

Thus, r_s is a linear function of the statistic $\sum_{i=1}^{n} R_i S_i$. Therefore, the various tests of independence discussed in this section can be as easily based on $\sum_{i=1}^{n} R_i S_i$ as on the more complicated formula for r_s given in (8.63) (or its counterpart in (8.64)).

45. *Large-Sample Approximation.* Under the null hypothesis H_0 (8.1), the rank vectors $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ are independent and each is uniformly distributed over the set of $n!$ permutations of $(1, 2, \ldots, n)$. It follows that the random variables $\sum_{i=1}^{n} R_i S_i$ and $\sum_{j=1}^{n} jS_j$ have the same null distribution. Combining this fact with the representation for r_s given in (8.80), it follows that

$$\begin{aligned} E_0(r_s) &= E_0\left[\frac{12\sum_{j=1}^{n} jS_j}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right)\right] \\ &= \frac{12\sum_{j=1}^{n} jE_0(S_j)}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right). \end{aligned}$$

Each $S_j, j = l, \ldots, n$, has a probability distribution that is uniform over the set of integers $\{1, 2, \ldots, n\}$. It follows that $E_0(S_j) = \sum_{k=1}^{n} k(1/n) = (n+1)/2$, for $j = 1, \ldots, n$. Thus, we have that

$$\begin{aligned} E_0(r_s) &= \frac{12\sum_{j=1}^{n} j\left(\frac{n+1}{2}\right)}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right) \\ &= \frac{12\frac{n(n+1)}{2}\left(\frac{n+1}{2}\right)}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right) = 0, \end{aligned}$$

as previously noted in (8.71). For the null variance of r_s, we first note that

$$\mathrm{var}_0(r_s) = \mathrm{var}_0\left[\frac{12\sum_{j=1}^{n} jS_j}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right)\right] = \frac{144}{n^2(n^2-1)^2}\mathrm{var}_0\left(\sum_{j=1}^{n} jS_j\right). \tag{8.81}$$

Using a well-known expression for the variance of a sum of random variables, we have that

$$\begin{aligned} \mathrm{var}_0\left(\sum_{j=1}^{n} jS_j\right) &= \sum_{j=1}^{n}\mathrm{var}_0(jS_j) + \sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n}\mathrm{cov}_0(jS_j, kS_k) \\ &= \sum_{j=1}^{n} j^2\mathrm{var}_0(S_j) + \sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n} jk\ \mathrm{cov}_0(S_j, S_k). \end{aligned}$$

The joint distribution of (S_j, S_k) is the same for every $j \neq k = 1, \ldots, n$ and the marginal distribution of S_j is the same for each $j = 1, \ldots, n$. It follows that

$$\text{var}_0\left(\sum_{j=1}^{n} jS_j\right) = \text{var}_0(S_1)\sum_{j=1}^{n} j^2 + \text{cov}_0(S_1, S_2)\sum_{j=1}^{n}\sum_{k=1, k\neq j}^{n} jk.$$

Using the facts that

$$\sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{6}$$

and

$$\begin{aligned}\sum_{j=1}^{n}\sum_{k=1, k\neq j}^{n} jk &= \left(\sum_{j=1}^{n} j\right)\left(\sum_{k=1}^{n} k\right) - \sum_{j=1}^{n} j^2 \\ &= \left[\frac{n(n+1)}{2}\right]^2 - \frac{n(n+1)(2n+1)}{6} \\ &= \frac{n(n^2-1)(3n+2)}{12},\end{aligned}$$

we obtain

$$\text{var}_0\left(\sum_{j=1}^{n} jS_j\right) = \left[\frac{n(n+1)(2n+1)}{6}\text{var}_0(S_1) + \frac{n(n^2-1)(3n+2)}{12}\right.$$
$$\left.\text{cov}_0(S_1, S_2)\right].$$

Moreover, under H_0 (8.1), it can be shown (see Problems 53 and 54) that $\text{var}_0(S_1) = (n^2-1)/12$ and $\text{cov}_0(S_1, S_2) = -(n+1)/12$. Thus, we have

$$\begin{aligned}\text{var}_0\left(\sum_{j=1}^{n} jS_j\right) &= \left[\frac{n(n+1)(2n+1)(n^2-1)}{72}\right] \\ &\quad - \frac{n(n^2-1)(3n+2)(n+1)}{144} \\ &= \frac{n(n+1)(n^2-1)}{144}[2(2n+1) - (3n+2)] \\ &= \frac{n^2(n+1)(n^2-1)}{144}. \qquad (8.82)\end{aligned}$$

Combining (8.81) and (8.82) yields

$$\text{var}_0(r_s) = \frac{144}{n^2(n^2-1)^2}\left[\frac{n^2(n+1)(n^2-1)}{144}\right] = \frac{1}{n-1},$$

as noted in (8.72).

The asymptotic normality under H_0 of the standardized form

$$r_s^* = \frac{r_s - E_0(r_s)}{\{\text{var}_0(r_s)\}^{1/2}} = (n-1)^{1/2} r_s$$

follows from the fact that r_s has the same null distribution as

$$\frac{12 \sum_{j=1}^{n} jS_j}{n(n^2-1)} - 3\left(\frac{n+1}{n-1}\right)$$

and standard techniques for establishing the asymptotic nonnality of a linear combination $\left(\Sigma_{j=1}^{n} \quad jS_j\right)$ of random variables. (For additional details, see Sections 8.4 and 12.3 in Randles and Wolfe (1979).)

46. *Exact Conditional Null Distribution of r_s with Ties among the X- and/or Y-Values.* To have a test with exact significance level even in the presence of tied X and/or Y observations, we must consider all the possible values of r_s corresponding to the fixed observed rank vector $(R_1, \ldots, R_n) = (r_1, \ldots, r_n)$ and every one of the $n!$ permutations of the observed rank vector $(S_1, \ldots, S_n) = (s_1, \ldots, s_n)$, where average ranks have been used to break ties in both of the rank vectors. As in Comment 41, it still follows that under H_0 each of the $n!$ possible outcomes for the ordered configurations $(s_1, \ldots, s_n)$, in conjunction with a fixed value of $(r_1, \ldots, r_n)$, both based on using average ranks to break ties, occurs with probability $1/n!$. For each such $(s_1, \ldots, s_n)$ configuration and fixed $(r_1, \ldots, r_n)$, the value of r_s is computed and the results are tabulated. We illustrate this construction for $n = 4$ and the data $(X_1, Y_1) = (2, 3.1), (X_2, Y_2) = (3.9, 4), (X_3, Y_3) = (2, 5.1)$, and $(X_4, Y_4) = (3.6, 4)$. Using average ranks to break ties, the associated X and Y rank vectors are $(R_1, R_2, R_3, R_4) = (1.5, 4, 1.5, 3)$ and $(S_1, S_2, S_3, S_4) = (1, 2.5, 4, 2.5)$, respectively. Thus, we have $D_1 = -.5, D_2 = -1.5, D_3 = 2.5, D_4 = -.5$, and an obtained value of $r_s = .1$. To assess the significance of r_s, we obtain its conditional distribution by considering the $4! = 24$ equally likely (under H_0) possible values of r_s for the fixed rank vector $(r_1, r_2, r_3, r_4) = (1.5, 4, 1.5, 3)$ in conjunction with each of the 24 permutations of the rank vector $(s_1, s_2, s_3, s_4) = (1, 2.5, 4, 2.5)$. These 24 permutations of (1, 2.5, 4, 2.5) and associated values of r_s are as follows:

(s_1, s_2, s_3, s_4)	Probability under H_0	Value or r_s
(1, 2.5, 4, 2.5)	$\frac{1}{24}$	.1
(1, 2.5, 2.5, 4)	$\frac{1}{24}$	.55
(1, 2.5, 2.5, 4)	$\frac{1}{24}$	.55
(2.5, 1, 2.5, 4)	$\frac{1}{24}$	−.2
(1, 4, 2.5, 2.5)	$\frac{1}{24}$	.85
(1, 4, 2.5, 2.5)	$\frac{1}{24}$	.85
(1, 2.5, 4, 2.5)	$\frac{1}{24}$	.1
(2.5, 1, 4, 2.5)	$\frac{1}{24}$	−.65

(s_1, s_2, s_3, s_4)	Probability under H_0	Value or r_s
(4, 1, 2.5, 2.5)	$\frac{1}{24}$	−.65
(4, 1, 2.5, 2.5)	$\frac{1}{24}$	−.65
(4, 2.5, 1, 2.5)	$\frac{1}{24}$	.1
(2.5, 4, 1, 2.5)	$\frac{1}{24}$	.85
(4, 2.5, 1, 2.5)	$\frac{1}{24}$	.1
(4, 2.5, 2.5, 1)	$\frac{1}{24}$	−.35
(4, 2.5, 2.5, 1)	$\frac{1}{24}$	−.35
(2.5, 4, 2.5, 1)	$\frac{1}{24}$	.4
(2.5, 1, 4, 2.5)	$\frac{1}{24}$	−.65
(2.5, 1, 2.5, 4)	$\frac{1}{24}$	−.2
(2.5, 2.5, 1, 4)	$\frac{1}{24}$	.55
(2.5, 2.5, 1, 4)	$\frac{1}{24}$	.55
(2.5, 4, 1, 2.5)	$\frac{1}{24}$	.85
(2.5, 4, 2.5, 1)	$\frac{1}{24}$	.4
(2.5, 2.5, 4, 1)	$\frac{1}{24}$	−.35
(2.5, 2.5, 4, 1)	$\frac{1}{24}$	−.35

This yields the null-tail probabilities

$$P_0(r_s \geq .85) = \frac{4}{24} \quad P_0(r_s \geq -.2) = \frac{16}{24}$$
$$P_0(r_s \geq .55) = \frac{8}{24} \quad P_0(r_s \geq -.35) = \frac{20}{24}$$
$$P_0(r_s \geq .4) = \frac{10}{24} \quad P_0(r_s \geq -.65) = 1$$
$$P_0(r_s \geq .1) = \frac{14}{24}.$$

This distribution is called the *conditional null distribution* or the *permutation null distribution* of r_s, given the observed sets of tied ranks $(r_1, r_2, r_3, r_4) = (1.5, 4, 1.5, 3)$ and $(s_1, s_2, s_3, s_4) = (1, 2.5, 4, 2.5)$. For the particular observed value $r_s = .1$, we have $P_0(r_s \geq .1) = \frac{14}{24}$, so that such a value does not indicate a deviation from H_0 in the direction of positive association between the X and Y variables. (We note that *both* the null expected value and null variance for r_s are different in this case of tied ranks (see Problem 51) than the corresponding expressions given in (8.71) and (8.72) for the no ties setting.)

47. *Point Estimation and Confidence Intervals Associated with* r_s. The Kendall statistic K (8.6) is directly associated with the population correlation coefficient

τ (8.2). This leads naturally to point estimators and approximate confidence intervals for τ based on K. Such is not the case for the Spearman statistic r_s (8.63). The measure of association linked with the independence tests based on r_s is

$$\eta = \frac{3[\tau + (n-2)\phi]}{n+1},$$

where τ is given by (8.2) and

$$\phi = 2P\{(Y_3 - Y_1)\}(X_2 - X_1) > 0\} - 1.$$

This measure of association η has several undesirable properties, including the facts that it is dependent on the sample size n and it is asymmetric in the X and Y labels. (For more discussion along these lines, see Fligner and Rust (1983).) As a result, point estimators and confidence intervals for η based on r_s are of little practical interest.

48. *Trend Test*. If we take $X_i = i, i = 1, \ldots, n$, and compute r_s, then the procedures based on r_s can be used as tests for a time trend in the univariate random sample $Y_1, \ldots, Y_n$.
49. *Other Uses for the r_s Statistic*. Spearman's rank correlation coefficient r_s also finds use in other settings where association is a primary issue. One such instance is in connection with Page's test for ordered alternatives in a two-way layout (see Section 7.2). Page's L statistic (7.10) is directly related to r_s. For more details, see Comment 7.22.

Properties

1. *Asymptotic Normality*. See Randles and Wolfe (1979, pp. 405–407).
2. *Efficiency*. See Section 8.7.

Problems

40. In order to study the effects of pharmaceutical and chemical agents on mucociliary clearance, doctors often use the ciliary beat frequency (CBF) as an index of ciliary activity. One accepted way to measure CBF in a subject is through the collection and analysis of an endobronchial forceps biopsy specimen. However, this technique is a rather invasive method for measuring CBF. In a study designed to assess the effectiveness of less invasive procedures for measuring CBF, Low et al. (1984) considered the alternative technique of nasal brushing. The data in Table 8.10 are a subset of the data collected by Low et al. during their investigation.

The subjects in the study were all men undergoing bronchoscopies for diagnoses of a variety of pulmonary problems. The CBF values reported in Table 8.10 are averages of 10 consecutive measurements on each subject.

Test the hypothesis of independence versus the alternative that the CBF measurements via nasal brushing and endobronchial forceps biopsy are positively associated (and, therefore, that nasal brushing is an acceptable alternative to the more invasive endobronchial forceps biopsy technique for measuring CBF).

41. Test the hypothesis of independence versus the alternative that the mean weight of introduced cysticerci is positively correlated with the mean weight of worms recovered for the tapeworm data in Table 8.3.

Table 8.10 Relation between Ciliary Beat Frequency (CBF) Values Obtained through Nasal Brushing and Endobronchial Forceps Biopsy

	CBF (hertz)	
Subject	Nasal brushing	Endobronchial forceps biopsy
1	15.4	16.5
2	13.5	13.2
3	13.3	13.6
4	12.4	13.6
5	12.8	14.0
6	13.5	14.0
7	14.5	16.0
8	13.9	14.1
9	11.0	11.5
10	15.0	14.4
11	17.0	16.0
12	13.8	13.2
13	17.4	16.6
14	16.5	18.5
15	14.4	14.5

Source: P. P. Low, C. K. Luk, M. J. Dulfano, and P. J. P. Finch (1984).

42. Test the hypothesis of independence versus the alternative that spending per high-school senior and percentage seniors graduating are positively correlated for the secondary education data in Table 8.6.

43. Show that the two expressions for r_s in displays (8.63) and (8.64) are equivalent.

44. For arbitrary number of observations, what are the smallest and largest possible values of r_s? Justify your answers.

45. Suppose $n = 5$ and we observe the data $(X_1, Y_1) = (3.7, 9.2), (X_2, Y_2) = (4.3, 9.4), (X_3, Y_3) = (5.0, 9.2), (X_4, Y_4) = (6.2, 10.4)$, and $X_5, Y_5) = (5.3, 9.2)$. What is the conditional probability distribution of r_s under H_0 (8.1) when average ranks are used to break ties among the Y's? How extreme is the observed value of r_s in this conditional null distribution? Compare this fact with that obtained by taking the observed value of r_s to the (incorrect) unconditional null distribution of r_s. (See also Problem 48.)

46. Give an example of a data set of $n \geq 10$ bivariate observations for which r_s has value 0.

47. Suppose $n = 25$. Compare the critical region for the level $\alpha = .05$ test of H_0 (8.1) versus H_2 (8.67) based on r_s with the critical region for the corresponding nominal level $\alpha = .05$ test based on the large-sample approximation.

48. For the case of $n = 5$ untied bivariate (X, Y) observations, obtain the form of the exact null (H_0) distribution of r_s. (See Comment 41.)

49. Let r_p be the Pearson product moment correlation coefficient defined in (8.78). Show that r_s (8.63) is simply this Pearson product moment correlation coefficient applied to the rank vectors $(R_1, \ldots, R_n)$ and $(S_1, \ldots, S_n)$ instead of the original $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ vectors.

50. Use the computer software R obtain the value of r_s for the secondary education data in Table 8.6, using average ranks to break the ties in the X and Y values.

51. Obtain the values of $E_0(r_s)$ and $\text{var}_0(r_s)$ corresponding to the exact conditional null distribution of r_s for the case of $n = 5$ and the tied data considered in Comment 46. Compare

these values with the corresponding values for $E_0(r_s)$ and $\text{var}_0(r_s)$ given in expressions (8.71) and (8.72) for the no ties setting. Discuss a possible reason for the difference in these null variances.

52. Use the Lake Michigan pollution data in Table 8.8 to test the hypothesis that the degree of pollution (as measured by the number of odor periods) had not changed with time against the alternative that there was a general increasing trend in the pollution of Lake Michigan over the period of 1950–1964. (See Comment 48.)

53. Let $(S_1, \ldots, S_n)$ be a vector of ranks that is uniformly distributed over the set of all $n!$ permutations of $(1, 2, \ldots, n)$. Show that the marginal probability distribution of each S_i, for $i = 1, \ldots, n$, is uniform over the set $\{1, 2, \ldots, n\}$. Use this fact to show that $E(S_i) = (n+1)/2$ and $\text{var}(S_i) = (n^2 - 1)/12$, for $i = 1, \ldots, n$.

54. Let $(S_1, \ldots, S_n)$ be a vector of ranks that is uniformly distributed over the set of all $n!$ permutations of $(1, 2, \ldots, n)$. Show that the joint marginal probability distribution of (S_i, S_j), for $i \neq j = 1, \ldots, n$, is given by

$$P(S_i = s, S_j = t) = \begin{cases} \dfrac{1}{n(n-1)}, & s \neq t = 1, \ldots, n \\ 0, & \text{otherwise.} \end{cases}$$

Use this fact to show that $\text{cov}(S_i, S_j) = -(n+1)/12$, for $i \neq j = 1, \ldots, n$.

55. The data in Table 8.11 were considered by Gentry and Pike (1970) in their study of the relationship between the mean rate of return over the period 1956 through 1969 and the 1969 value of common stock portfolios for 32 life insurance companies.

Test the hypothesis of independence versus the general alternative that the 1956–1969 mean rate of return for a stock portfolio is correlated in some fashion with its 1969 value.

8.6 A DISTRIBUTION-FREE TEST FOR INDEPENDENCE AGAINST BROAD ALTERNATIVES (HOEFFDING)

Hoeffding (1948b) proposed a test of independence that is able to detect a much broader class of alternatives to independence than the classes of alternatives that can be detected by the tests of Sections 8.1 and 8.5 based on sample correlation coefficients.

Procedure

To test the hypothesis that the X and Y random variables are independent, namely, H_0 given by (8.1), we first rank $X_1, \ldots, X_n$ jointly and let R_i denote the rank of X_i in this joint ranking, $i = 1, \ldots, n$. Then rank $Y_1, \ldots, Y_n$ jointly, and let S_i denote the rank of Y_i in this joint ranking, $i = 1, \ldots, n$. We let c_i denote the number of sample pairs (X_α, Y_α) for which both $X_\alpha < X_i$ and $Y - \alpha < Y_i$; that is,

$$c_i = \sum_{\alpha=1}^{n} \phi(X_\alpha, X_i)\phi(Y_\alpha, Y_i), \quad i = 1, \ldots, n, \tag{8.83}$$

where $\phi(a, b) = 1$ if $a < b, = 0$, otherwise.

Table 8.11 Mean Rate of Return of Common Stock Portfolios over the Period 1956–1969 and the 1969 Value of Each Equity Portfolio for 32 Life Insurance Companies

Company	Mean rate (%) of return, 1956–1969	Value of common stock portfolio, December 31, 1969 (millions of dollars)
1	18.83	96.0
2	16.98	54.6
3	15.36	84.4
4	14.65	251.5
5	14.21	131.8
6	13.68	37.3
7	13.65	109.9
8	13.07	13.5
9	12.99	76.3
10	12.81	72.6
11	11.60	42.1
12	11.51	41.5
13	11.50	56.2
14	11.41	59.3
15	11.26	1184.0
16	10.67	144.0
17	10.44	111.9
18	10.44	179.8
19	10.33	29.2
20	10.30	279.5
21	10.22	166.6
22	10.05	194.3
23	10.04	40.8
24	9.57	428.4
25	9.50	7.0
26	9.48	485.6
27	9.29	165.3
28	9.21	343.8
29	9.04	35.4
30	8.82	24.7
31	8.78	2.7
32	7.26	8.9

Source: J. Gentry and J. Pike (1970).

We set

$$Q = \sum_{i=1}^{n} (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2), \tag{8.84}$$

$$R = \sum_{i=1}^{n} (R_i - 2)(S_i - 2)c_i, \tag{8.85}$$

and

$$S = \sum_{i=1}^{n} c_i(c_i - 1), \tag{8.86}$$

and compute

$$D = \frac{Q - 2(n-2)R + (n-2)(n-3)S}{n(n-1)(n-2)(n-3)(n-4)}. \tag{8.87}$$

For a (two-sided) test of H_0 versus the alternative that X and Y are dependent (see Comment 52), at the α level of significance,

$$\text{Reject } H_0 \text{ if } D \geq d_\alpha; \quad \text{otherwise do not reject,} \tag{8.88}$$

where the constant d_α satisfies the equation $P_0(D \geq d_\alpha) = \alpha$.

Large-Sample Approximation

For the large-sample approximation, we use a statistic B, proposed by Blum, Kiefer, and Rosenblatt (1961), that is slightly different than Hoeffding's D statistic. (The tests based on B and D are, however, asymptotically equivalent, because the statistics $nD + \left(\frac{1}{36}\right)$ and nB have the same asymptotic distribution under H_0. See Comment 53.) Let

$$B = n^{-5} \sum_{i=1}^{n} [N_1(i)N_4(i) - N_2(i)N_3(i)]^2, \tag{8.89}$$

where

$$\begin{aligned}
N_1(i) &= \text{number of sample pairs } (X_\alpha, Y_\alpha) \text{ lying in the region} \\
&\quad T_1(i) = \{(x, y) : x \leq X_i \text{ and } y \leq Y_i\}, \\
N_2(i) &= \text{number of sample pairs } (X_\alpha, Y_\alpha) \text{ lying in the region} \\
&\quad T_2(i) = \{(x, y) : x > X_i \text{ and } y \leq Y_i\}, \\
N_3(i) &= \text{number of sample pairs } (X_\alpha, Y_\alpha) \text{ lying in the region} \\
&\quad T_3(i) = \{(x, y) : x \leq X_i \text{ and } y > Y_i\}, \\
N_4(i) &= \text{number of sample pairs } (X_\alpha, Y_\alpha) \text{ lying in the region} \\
&\quad T_4(i) = \{(x, y) : x > X_i \text{ and } y > Y_i\}.
\end{aligned} \tag{8.90}$$

That is, for each i, determine the number of sample pairs (X_α, Y_α) lying in each of the regions determined by the horizontal and vertical lines through the point (X_i, Y_i).

A large-sample approximation to procedure (8.88) is

$$\text{Reject } H_0 \text{ if } \frac{1}{2}\pi^4 nB \geq b_\alpha; \text{ otherwise do not reject,} \tag{8.91}$$

where the constant b_α satisfies the equation $P_0(\frac{1}{2}\pi^4 nB \geq b_\alpha) = \alpha$. Blum, Kiefer, and Rosenblatt suggested that when n is small, the error introduced when utilizing the large-sample approximation may be reduced by substituting $(n-1)B$ for nB in the left-hand side of (8.91). For a different large-sample approximation, see Comment 53.

Ties

Use average ranks and replace (8.83) by

$$c_i = \sum_{\substack{\alpha=1 \\ a \neq i}}^{n} \phi^*(X_\alpha, X_i)\phi^*(Y_\alpha, Y_i), \quad i = 1, \ldots, n, \tag{8.92}$$

where

$$\phi^*(a,b) = \begin{cases} 1, & \text{if } a < b, \\ \frac{1}{2}, & \text{if } a = b, \\ 0, & \text{otherwise.} \end{cases} \tag{8.93}$$

EXAMPLE 8.6 ***[Continuation of Example 8.5].***

We return to the data of Table 8.9 and consider the relation between the free pool of proline and collagen content in human liver cirrhosis. We apply Hoeffding's test of independence. From (8.92), we find

$$\begin{aligned}
c_1 &= \phi^*(X_2,X_1)\phi^*(Y_2,Y_1) + \phi^*(X_3,X_1)\phi^*(Y_3,Y_1) + \phi^*(X_4,X_1)\phi^*(Y_4,Y_1) \\
&\quad + \phi^*(X_5,X_1)\phi^*(Y_5,Y_1) + \phi^*(X_6,X_1)\phi^*(Y_6,Y_1) + \phi^*(X_7,X_1)\phi^*(Y_7,Y_1) \\
&= \frac{1}{2}(0) + 0\left(\frac{1}{2}\right) + 0(1) + 0(0) + 0(0) + 0(0) = 0, \\
c_2 &= \phi^*(X_1,X_2)\phi^*(Y_1,Y_2) + \phi^*(X_3,X_2)\phi^*(Y_3,Y_2) + \phi^*(X_4,X_2)\phi^*(Y_4,Y_2) \\
&\quad + \phi^*(X_5,X_2)\phi^*(Y_5,Y_2) + \phi^*(X_6,X_2)\phi^*(Y_6,Y_2) + \phi^*(X_7,X_2)\phi^*(Y_7,Y_2) \\
&= \frac{1}{2}(1) + 0(1) + 0(1) + 0(0) + 0(0) + 0(0) = \frac{1}{2}, \\
c_3 &= \phi^*(X_1,X_3)\phi^*(Y_1,Y_3) + \phi^*(X_2,X_3)\phi^*(Y_2,Y_3) + \phi^*(X_4,X_3)\phi^*(Y_4,Y_3) \\
&\quad + \phi^*(X_5,X_3)\phi^*(Y_5,Y_3) + \phi^*(X_6,X_3)\phi^*(Y_6,Y_3) + \phi^*(X_7,X_3)\phi^*(Y_7,Y_3) \\
&= 1\left(\frac{1}{2}\right) + 1(0) + 0(1) + 0(0) + 0(0) + 0(0) = \frac{1}{2}, \\
c_4 &= \phi^*(X_1,X_4)\phi^*(Y_1,Y_4) + \phi^*(X_2,X_4)\phi^*(Y_2,Y_4) + \phi^*(X_3,X_4)\phi^*(Y_3,Y_4) \\
&\quad + \phi^*(X_5,X_4)\phi^*(Y_5,Y_4) + \phi^*(X_6,X_4)\phi^*(Y_6,Y_4) + \phi^*(X_7,X_4)\phi^*(Y_7,Y_4) \\
&= 1(0) + 1(0) + 1(0) + 0(0) + 0(0) + 0(0) = 0, \\
c_5 &= \phi^*(X_1,X_5)\phi^*(Y_1,Y_5) + \phi^*(X_2,X_5)\phi^*(Y_2,Y_5) + \phi^*(X_3,X_5)\phi^*(Y_3,Y_5) \\
&\quad + \phi^*(X_4,X_5)\phi^*(Y_4,Y_5) + \phi^*(X_6,X_5)\phi^*(Y_6,Y_5) + \phi^*(X_7,X_5)\phi^*(Y_7,Y_5) \\
&= 1(1) + 1(1) + 1(1) + 1(1) + 0(0) + 0(0) = 4, \\
c_6 &= \phi^*(X_1,X_6)\phi^*(Y_1,Y_6) + \phi^*(X_2,X_6)\phi^*(Y_2,Y_6) + \phi^*(X_3,X_6)\phi^*(Y_3,Y_6) \\
&\quad + \phi^*(X_4,X_6)\phi^*(Y_4,Y_6) + \phi^*(X_5,X_6)\phi^*(Y_5,Y_6) + \phi^*(X_7,X_6)\phi^*(Y_7,Y_6) \\
&= 1(1) + 1(1) + 1(1) + 1(1) + 1(1) + 0(0) = 5, \\
c_7 &= \phi^*(X_1,X_7)\phi^*(Y_1,Y_7) + \phi^*(X_2,X_7)\phi^*(Y_2,Y_7) + \phi^*(X_3,X_7)\phi^*(Y_3,Y_7) \\
&\quad + \phi^*(X_4,X_7)\phi^*(Y_4,Y_7) + \phi^*(X_5,X_7)\phi^*(Y_5,Y_7) + \phi^*(X_6,X_7)\phi^*(Y_6,Y_7) \\
&= 1(1) + 1(1) + 1(1) + 1(1) + 1(1) + 1(1) = 6.
\end{aligned}$$

We next compute the values of Q, R, S, and D. Using (8.84) and (8.87) and the R_i's and S_i's found in Example 8.5, we obtain

$$\begin{aligned} Q &= .5(-.5)(1.5)(.5) + .5(-.5)(3)(2) + 2(1)(1.5)(.5) \\ &\quad + 3(2)(0)(-1) + 4(3)(4)(3) + 5(4)(5)(4) + 6(5)(6)(5) \\ &= 1443.81, \\ R &= -.5(.5)(0) + (-.5)(2)\left(\frac{1}{2}\right) + 1(.5)\left(\frac{1}{2}\right) \\ &\quad + 2(-1)(0) + 3(3)(4) + 4(4)(5) + 5(5)(6) \\ &= 265.75, \\ S &= 0(-1) + \frac{1}{2}\left(-\frac{1}{2}\right) + \frac{1}{2}\left(-\frac{1}{2}\right) + 0(-1) + 4(3) + 5(4) + 6(5) \\ &= 61.5 \end{aligned}$$

and

$$\begin{aligned} D &= \frac{1443.81 - 2(5)(265.75) + 5(4)(61.5)}{7(6)(5)(4)(3)} \\ &= \frac{16.31}{2520}. \end{aligned}$$

We now use these data to illustrate the computations needed to perform the large-sample approximation. For example, for the pair $(X_4, Y_4) = (8.3, 2.6)$, dividing the plane into the four regions defined by (8.90) and counting the number of sample pairs in these regions yields the $N_1(4), N_2(4), N_3(4)$, and $N_4(4)$ values defined by (8.90), namely,

$$N_1(4) = 1, \quad N_2(4) = 0, \quad N_3(4) = 3, \quad N_4(4) = 3.$$

Performing similar subdivisions and counts corresponding to the other six sample pairs, we find

$$\begin{aligned} &N_1(1) = 1, \quad N_2(1) = 2, \quad N_3(1) = 1, \quad N_4(1) = 3, \\ &N_1(2) = 2, \quad N_2(2) = 2, \quad N_3(2) = 0, \quad N_4(2) = 3, \\ &N_1(3) = 3, \quad N_2(3) = 1, \quad N_3(3) = 1, \quad N_4(3) = 3, \\ &N_1(5) = 5, \quad N_2(5) = 0, \quad N_3(5) = 0, \quad N_4(5) = 2, \\ &N_1(6) = 6, \quad N_2(6) = 0, \quad N_3(6) = 0, \quad N_4(6) = 1, \\ &N_1(7) = 7, \quad N_2(7) = 0, \quad N_3(7) = 0, \quad N_4(7) = 0. \end{aligned}$$

From (8.89), we then obtain

$$B = (7)^{-5}\{[3-2]^2 + [6-0]^2 + [6-1]^2 + [3-0]^2 + [10-0]^2 + [6-0]^2 + [0-0]^2\}$$

$$= 7^{-5}(207).$$

The sample size $n = 7$ is relatively small, so we calculate the left-hand side of (8.91) with $(7-1)B$ replacing $7B$. We find

$$\frac{1}{2}\pi^4(n-1)B = \frac{1}{2}(3.14)^4(6)(207)(7)^{-5} = 3.60.$$

The value of D given above may be reproduced using the `R` command `HoeffD`. The arguments are the two samples X and Y. Estimates of P-values and critical values d_α may be obtained from `pHoeff`. These are approximate values based on Monte Carlo simulation. For this value of D, the P-value is approximately .077. This is approximate not only due to the simulation of the distribution, but also because ties exist in the data.

For the large-sample approximation, $nD + 1/36 = .073$ and $nB = .086$. For larger n, we would see closer agreement. The command `hoeffd` in package `Hmisc` (Harrell (2012)) will perform this test and give asymptotic P-values based on B. The following is the relevant `R` output from the call `hoeffd (x, y)` where `x` is the collagen data from Table 8.9 and `y` is the proline data:

```
D

     x     y
x   1.00  0.19
y   0.19  1.00

n= 7

P
     x   y
x        0.0215
y   0.0215
```

The test statistic D has an upper bound of 1/30 for all n (Wilding and Mudholkar (2008)). `R` reports the statistic D scaled to an upper bound of 1. So, the value of D previously calculated as 16.31/2520 is scaled to $D' = (16.31/2520) \cdot 30 = .194$. The asymptotic P-value given is based on B (using n, not $n-1$ despite the low sample size). For this data, the P-value is .0215. Thus, we would reject the null hypothesis for any specified significance level α greater or equal to .0215. Note that this test is approximate due to the presence of ties. Additionally, the sample size may be inappropriate for the use of asymptotic P-values. A combination of the two factors (ties, sample size) may explain the discrepancy between the P-value based on D (0.077 and the value found here. 0.0215).

For comparison, recall that in Section 8.5, we applied the Spearman's test to the data of Table 8.9 and found the one-sided P-value to be between .05 and .10. Thus, the two-sided P-value for the test is between .10 and .20.

Comments

50. *Motivation for Hoeffding's Test.* Define

$$D^*(x,y) = P(X \le x \text{ and } Y \le y) - P(X \le x)P(Y \le y). \tag{8.94}$$

We note that $D^*(x, y) = 0$ for all (x, y) if and only if H_0 is true. This fact was used by Hoeffding in devising the test based on D. The statistic D estimates the parameter

$$\Delta_1(F) = E_F\{D^*(X', Y')\}^2, \tag{8.95}$$

where (X', Y') is a random member from the underlying bivariate population with distribution F. In other words, we may think of $D^*(x, y)$ as a measure of the deviation from H_0 at the point (x, y), and $\Delta_1(F)$ as the average value of the square of this deviation.

51. *Null Distribution of D*. In determining the null distribution of D, we can, without loss of generality, take $R_1 = i$ and obtain the associated values of D for the $n!$ possible Y rank configurations of the form $(S_1, \ldots, S_n)$. Each of these configurations has probability $[1/(n!)]$ under H_0.

52. *Consistency of D against a Broad Class of Alternatives*. The D test was designed by Hoeffding to detect a broad class of alternatives to the hypothesis of independence, and in this sense its character differs from that of the tests of independence of Sections 8.1 and 8.5 based on sample correlation coefficients. Although Hoeffding (1948b) showed that the D test is not sensitive to all alternatives to H_0, he demonstrated that under mild restrictions on the nature of the underlying bivariate population F, the test is consistent when H_0 is false. Thus, the D test detects alternatives where the X's and Y's are positively associated and alternatives where the X's and Y's are negatively associated. Furthermore, there exist populations F where X, Y are dependent and D is consistent, but the tests based on the sample correlation coefficients are not consistent.

53. *Relationship of D and B*. The statistics $nD + (\frac{1}{36})$ and nB have the same asymptotic distribution under H_0. (See Hoeffding (1948b) and Blum, Kiefer, and Rosenblatt (1961).) Thus, another large-sample approximation to procedure (8.88) is

$$\text{Reject } H_0 \text{ if } \left(\frac{1}{2}\right)\pi^4\left\{nD + \left(\frac{1}{36}\right)\right\} \geq b_\alpha; \quad \text{otherwise do not reject.}$$

54. *Development of D Test*. The test based on D was introduced by Hoeffding (1948b). The related test based on B was considered by Blum, Kiefer, and Rosenblatt (1961), who extended the approach to testing for the independence of k $(k \geq 2)$ variables. A one-sided test, similar in character to the two-sided B test, was proposed by Crouse (1966). Skaug and Tjøstheim (1993) considered the Blum–Kiefer–Rosenblatt statistic in a time-series setting and established (under mild conditions) consistency against lag one dependent alternatives. Zheng (1997) used smoothing methods to develop a nonparametric test of independence between two variables. His test is consistent against any form of dependence.

55. *Finite Sample Size Distribution of D*. Wilding and Mudholkar (2008) proposed improved methods for estimating the distribution of Hoeffding's D under the null distribution for small sample sizes. Their approximations are based on the Weibull family of distributions. The command `pHoeff` provides an approximation of this distribution using Monte Carlo simulation. For the discrete values d

of D, it provides the probability that d occurs $P(D = d)$, the lower-tail probabilities $P(D \leq d)$, and upper-tail probabilities $P(D \geq d)$. For example, if $n = 5$ and 20,000 Monte Carlo runs are used, the output is

```
          d  P(D = d)  P(D <= d)  P(D >= d)
-0.01667    0.13280    0.13280     1.00000
0.000000    0.79995    0.93275     0.86720
0.033333    0.06725    1.00000     0.06725.
```

Properties

1. *Consistency*. The test defined by (8.88) is consistent against populations for which the parameter $\Delta_1(F)$ defined by (8.95) is positive. For conditions on F that ensure that $\Delta_1(F)$ will be positive, see Hoeffding (1948b) and Yanagimoto (1970).
2. *Asymptotic Distribution*. For the asymptotic distribution of $\{nD + (\frac{1}{36})\}$, see Hoeffding (1948b) and Blum, Kiefer, and Rosenblatt (1961).

Problems

56. The data in Table 8.12 are a subset of the data obtained by Shen et al. (1970) in an experiment concerned with the hypothesis that diabetes mellitus is not simply a function of insulin deficiency and that perhaps insulin insensitivity could play an important role in the hyperglycemia of diabetes. One of the purposes of the study was to investigate the relation between the response to a glucose tolerance test and glucose impedance, a quantity describing the body tissues' resistance to glucose and expected to be constant for a given individual throughout the experimental range of glucose uptake rate in the author's study. The seven subjects represented in Table 8.12 were volunteers recently released from a minimum security prison and characterized by low plasma glucose response to oral glucose. Table 8.12 gives the weighted glucose response to an oral glucose tolerance test (X) and the glucose impedance reading (Y) for each of the seven subjects.

Use procedure (8.88) to test for impedance of weighted glucose response and glucose impedance. (Recall that procedure (8.88) is designed to detect all alternatives to the hypothesis of independence. However, if one has prior reasons or evidence to suspect that the weighted glucose response is positively correlated with glucose impedance, it would be more appropriate to focus on alternatives of positive association by using the one-sided procedure, based on Kendall's K, given by (8.8).)

Table 8.12 Weighted Glucose Response and Glucose Impedance

Subject	Weighted glucose response, X	Glucose impedance, Y
1	130	26.1
2	116	19.7
3	122	26.8
4	117	23.7
5	108	23.4
6	115	24.4
7	107	16.5

Source: S. Shen, G. M. Reaven, J. W. Farquhar, and R. H. Nakanishi (1970).

57. Apply Kendall's two-sided test based on (8.10) to the data of Table 8.12. Compare your result with the result of Problem 56.

58. Apply the large-sample approximation given in Comment 53 to the data in Table 8.9. Compare this approximation with the approximation based on (8.91) that was used in Example 8.6.

8.7 EFFICIENCIES OF INDEPENDENCE PROCEDURES

Investigation of the asymptotic relative efficiencies of tests for independence is made more difficult by our inability to define natural classes of alternatives to the hypothesis of independence. The asymptotic relative efficiencies of the test procedure (one- or two-sided) based on Kendall's statistic K (8.6) with respect to the corresponding normal theory test based on Pearson's product moment correlation coefficient r_{p} (8.78) have been found by Stuart (1954) and Konijn (1956) for a class of dependence alternatives "close" to the hypothesis of independence. Values of this asymptotic relative efficiency $e(K, r_{\mathrm{P}})$, for selected bivariate $F_{X,Y}$, are as follows:

$F_{X,Y}$:	Normal	Uniform	Double Exponential
$e(K, r_{\mathrm{p}})$:	.912	1.000	1.266

In the normal setting, natural alternatives to independence correspond to bivariate normal distributions with nonzero correlation. In this case, the asymptotic relative efficiency of the test procedure (one- or two-sided) based on Spearman's statistic r_{s} (8.63) with respect to the corresponding test procedure based on Kendall's K is 1. Moreover, the common asymptotic relative efficiency of either the test procedure based on r_{s} or the test procedure based on K with respect to the corresponding normal theory test based on r_{p} is $(3/\pi)^2 = .912$.

The point estimator and confidence interval associated with normality assumptions for the independence problem are concerned with the underlying correlation coefficient, whereas the estimator and confidence intervals based on Kendall's K relate to the parameter τ. In view of this, the estimator $\hat{\tau}$ (8.34) and the approximate confidence intervals given by (8.39) and (8.50) are not easily compared with the normal theory procedures; hence, their asymptotic efficiencies are not presented here.

We do not know of any results for the asymptotic efficiency of Hoeffding's independence test (Section 8.6).

Chapter 9

Regression Problems

INTRODUCTION

Among the most common applications of statistical techniques are those involving some sort of regression analysis. Such procedures are designed to detect and interpret stochastic relationships between a dependent (response) variable and one or more independent (predictor) variables. These regression relationships can vary from that of a simple linear relationship between the dependent variable and a single independent variable to complex, nonlinear relationships involving a large number of predictor variables.

In Sections 9.1–9.4, we present nonparametric procedures designed for the simplest of regression relationships, namely, that of a single stochastic linear relationship between a dependent variable and one independent variable. (Such a relationship is commonly referred to as a *regression line*.) In Section 9.1, we present a distribution-free test of the hypothesis that the slope of the regression line is a specified value. Sections 9.2 and 9.3 provide, respectively, a point estimator and distribution-free confidence intervals and bounds for the slope parameter. In Section 9.4, we complete the analysis for a single regression line by discussing both an estimator of the intercept of the line and the use of the estimated linear relationship to provide predictions of dependent variable responses to additional values of the predictor variable. In Section 9.5, we consider the case of two or more regression lines and describe an asymptotically distribution-free test of the hypothesis that the regression lines have the same slope; that is, that the regression lines are parallel.

In Section 9.6, we present the reader with an introduction to the extensive field of rank-based regression analysis for more complicated regression relationships than that of a straight line. In Section 9.7, we provide short introductions to a number of recent developments in the rapidly expanding area of non-rank-based nonparametric regression, where the goal is to make statistical inferences about the relationship between a dependent variable and one or more independent variables without a priori specification of a formal model describing the regression relationship. These non-rank-based approaches to nonparametric regression are generally more complicated than the level assumed throughout the rest of this text. As a result, our approach in Section 9.7 is simply to give brief descriptions of a variety of statistical techniques that are commonly used to develop such procedures and provide appropriate references for readers interested in more detailed information about them, rather than to concentrate on specific procedures and their application to appropriate data sets.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

Finally, in Section 9.8, we consider the asymptotic relative efficiencies of the straight-line regression procedures discussed in Sections 9.1–9.3 and 9.5–9.6 with respect to their competitors based on classical least squares estimators.

ONE REGRESSION LINE

Data. At each of n fixed values, $x_1, \ldots, x_n$, of the independent (predictor) variable x, we observe the value of the response random variable Y. Thus, we obtain a set of observations $Y_1, \ldots, Y_n$, where Y_i is the value of the response variable when $x = x_i$. The x's are assumed to be distinct and, without loss of generality, we take $x_1 < x_2 < \cdots < x_n$.

Assumptions

A1. Our straight-line model is

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, \ldots, n, \tag{9.1}$$

where the x's are known constants and α (the intercept) and β (the slope) are unknown parameters.

A2. The random variables $e_1, \ldots, e_n$ are a random sample from a continuous population that has median 0.

9.1 A DISTRIBUTION-FREE TEST FOR THE SLOPE OF THE REGRESSION LINE (THEIL)

Hypothesis

The null hypothesis of interest here is that the slope, β, of the postulated regression line is some specified value β_0, namely,

$$H_0 : \beta = \beta_0. \tag{9.2}$$

Thus, the null hypothesis asserts that for every unit increase in the value of the independent (predictor) variable x, we would expect an increase (or decrease, depending on the sign of β_0) of roughly β_0 in the value of the dependent (response) variable Y.

Procedure

To compute the Theil (1950a) statistic C, we first form the n differences

$$D_i = Y_i - \beta_0 x_i, \quad i = 1, \ldots, n. \tag{9.3}$$

Let

$$C = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c(D_j - D_i), \tag{9.4}$$

where

$$c(a) = \begin{cases} -1, & \text{if } a < 0, \\ 0, & \text{if } a = 0, \\ 1, & \text{if } a > 0. \end{cases} \tag{9.5}$$

Thus, for each pair of subscripts (i, j), with $1 \le i < j \le n$, score 1 if $D_j - D_i$ is positive, and score -1 if $D_j - D_i$ is negative. The statistic C (9.4) is then just the sum of these 1's and -1s.

a. *One-Sided Upper-Tail Test*. To test the null hypothesis

$$H_0 : \beta = \beta_0$$

versus the alternative that the slope is larger than the specified β_0 corresponding to

$$H_1 : \beta > \beta_0, \tag{9.6}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } \bar{C} \ge k_\alpha; \quad \text{otherwise do not reject,} \tag{9.7}$$

where the constant k_α is chosen to make the type I error probability equal to α and $\bar{C} = C/(n(n-1)/2)$. (See Comment 2.)

b. *One-Sided Lower-Tail Test*. To test the null hypothesis

$$H_0 : \beta = \beta_0$$

versus the alternative that the slope is smaller than the specified β_0 corresponding to

$$H_2 : \beta < \beta_0, \tag{9.8}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } \bar{C} \le -k_\alpha; \quad \text{otherwise do not reject.} \tag{9.9}$$

c. *Two-Sided Test*. To test the null hypothesis

$$H_0 : \beta = \beta_0$$

versus the alternative that the slope is simply not equal to the specified β_0 corresponding to

$$H_3 : \beta \ne \beta_0, \tag{9.10}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } |\bar{C}| \ge k_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{9.11}$$

This two-sided procedure is the two-sided symmetric test with $\alpha/2$ probability in each tail of the null distribution of $\bar{C}$.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of C, suitably standardized. For this standardization, we need to know the expected value and variance of C when the null hypothesis H_0 (9.2) is true. Under H_0, the expected value and variance of C are

$$E_0(C) = 0 \tag{9.12}$$

and

$$\text{var}_0(C) = \frac{n(n-1)(2n+5)}{18}, \tag{9.13}$$

respectively. (See Comment 2.)

The standardized version of C is

$$C^* = \frac{C - E_0(C)}{\{\text{var}_0(C)\}^{1/2}} = \frac{C}{\{n(n-1)(2n+5)/18\}^{1/2}}. \tag{9.14}$$

When H_0 is true, C^* has, as n tends to infinity, an asymptotic $N(0,1)$ distribution (See Comment 2). The normal theory approximation for procedure (9.7) is

$$\text{Reject } H_0 \text{ if } C^* \geq z_\alpha; \quad \text{otherwise do not reject,} \tag{9.15}$$

the normal theory approximation for procedure (9.9) is

$$\text{Reject } H_0 \text{ if } C^* \leq -z_\alpha; \quad \text{otherwise do not reject,} \tag{9.16}$$

and the normal theory approximation for procedure (9.11) is

$$\text{Reject } H_0 \text{ if } |C^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{9.17}$$

Ties

If there are ties among the D_i (9.3) differences, C may be computed as described in (9.4), but keep in mind that procedures (9.7), (9.9), and (9.11) are then approximate rather than exact. Sen (1968) suggested a way to deal with ties among the values of the independent variable x.

EXAMPLE 9.1 ***Effect of Cloud Seeding on Rainfall.***

Smith (1967) described experiments performed in Australia to investigate the effects of a particular method of cloud seeding on the amount of rainfall. In one experiment that took place in the Snowy Mountains, two areas served as target and control, respectively, and during any one period, a random process was used to determine whether clouds over the target area should be seeded. The effect of seeding was measured by the *double ratio* [T/Q (seeded)]/[T/Q (unseeded)], where T and Q are the total rainfalls in the target and control areas, respectively. Table 9.1 provides the double ratio calculated for each year of a 5-year experiment.

The slope parameter β represents the rate of change in Y per unit change in x. We apply the one-sided lower-tail test (9.9) with β_0 equal zero. This should be viewed as a

Table 9.1 Double Ratio for 5 Years in the Snowy Mountains of Australia

Years seeded, x_i	Double ratio, Y_i
1	1.26
2	1.27
3	1.12
4	1.16
5	1.03

Source: E. J. Smith (1967).

test of the null hypothesis that the double ratio does not change with time (i.e., the effects of seeding during one year do not overlap into other years) against the alternative that there is a decrease over time, either in the rainfall increases resulting from the seeding or in the ability of the experiments to detect such increases.

From (9.3), with $\beta_0 = 0$, we see that $D_i = Y_i$. We now illustrate the computations required to obtain the value of C (9.4) for these data.

(i, j)	$D_j - D_i$	$c(D_j - D_i)$
(1, 2)	.01	1
(1, 3)	−.14	−1
(1, 4)	−.10	−1
(1, 5)	−.23	−1
(2, 3)	−.15	−1
(2, 4)	−.11	−1
(2, 5)	−.24	−1
(3, 4)	.04	1
(3, 5)	−.09	−1
(4, 5)	−.13	−1

Thus, we find the value of C and $\bar{C}$ to be

$$C = \sum_{i=1}^{4} \sum_{j=i+1}^{5} c(D_j - D_i) = -6, \quad \bar{C} = \frac{C}{n(n-1)/2} = -.6.$$

Using the fact that the null distribution of C is symmetric about zero (See Comment 2), we find that the P-value for these data is $P(\bar{C} \leq -.6) =$ `pKendall(-.6, N = 5, lower.tail = T)` = .117. Thus, there is not much evidence of a decrease over time of the rainfall increases resulting from the seeding.

To illustrate the normal theory approximation (which should not be expected to be highly accurate for a sample size as small as 5), we first find from (9.14) that

$$C^* = \frac{-6}{\{5(4)(15)/18\}^{1/2}} = -1.47.$$

Thus, the smallest significance level at which we can reject $H_0 : \beta = 0$ in favor of $\beta < 0$ using the normal theory approximation is .0708, since $z_{.0708} = -1.47$. As expected for this small sample size ($n = 5$), this is not in especially good agreement with the exact P-value of .117 found previously.

The above analysis may also be carried out in R. The command `theil` requires arguments for data vectors `x` and `y`, the null hypothesized value `beta. 0` and `type="t"`, `"1"`, or `"u"` for a two-tailed, lower-tail, or upper-tail test, respectively. If taking `x` and `y` to be the data from Table 9.1, a call of the function `theil(x, y, beta.0=0, type="1")` results in the following output, which reproduces the analysis above:

```
Null:  beta less than 0
C = -6, C.bar = -0.6, P = 0.117.
```

Comments

1. *Motivation for the Test.* From (9.4), we see that C will be large when $D_j > D_i$ for many (i,j) pairs. Now

$$D_j - D_i = [Y_j - \beta_0 x_j - (Y_i - \beta_0 x_i)] = [Y_j - Y_i + \beta_0(x_i - x_j)].$$

   Furthermore, under model (9.1), the median of $Y_j - Y_i = [\beta(x_j - x_i) + (e_j - e_i)]$ is $\beta(x_j - x_i)$. Thus, under model (9.1), the median of $D_j - D_i$ is $[\beta(x_j - x_i) + \beta_0(x_i - x_j)] = (\beta - \beta_0)(x_j - x_i)$. Hence, we tend to obtain positive $D_j - D_i$ differences when $\beta > \beta_0$, and these positive differences lead to large values of C. This serves as partial motivation for procedure (9.7).
2. *Relationship to Kendall's Correlation Statistic K.* The statistic C (9.4) is simply Kendall's correlation statistic K (8.6) computed between the x and $Y - \beta_0 x$ values. In particular, a test of $\beta_0 = 0$ can be interpreted as a test for correlation between the x and Y sequences. Moreover, the null H_0 (9.2) distribution properties of the statistic C (when there are no tied D values) are identical with the corresponding distributional properties of Kendall's statistic K under its null hypothesis of independence (See Section 8.1). This leads immediately to the use of the critical values k_α in procedures (9.7), (9.9), and (9.11). In addition, the symmetry about zero for the null distribution of C follows from Comment 8.8 and the values of $E_0(C)$ and $\text{var}_0(C)$ are direct consequences of the corresponding values of $E_0(K)$ and $\text{var}_0(K)$, respectively, developed in Comment 8.10. Finally, the asymptotic $(n \to \infty)$ normality for the standardized statistic C^* under the null hypothesis H_0 (9.2) derives from the similar property for the standardized K^*, as discussed in Comment 8.10.
3. *Testing for Trends over Time.* In the special case when the x-values are the time order (as in Example 9.1), the procedures in (9.7), (9.9), and (9.11) (with β_0 set equal to zero) can be viewed as tests against a time trend and have been suggested for this use by Mann (1945). (See also Comment 8.14.)

Properties

1. *Consistency.* The tests defined by (9.7), (9.9), and (9.11) are consistent against the alternatives 0 $\beta >, <$, and $\neq \beta_0$, respectively.
2. *Asymptotic Normality.* See Comment 2.
3. *Efficiency.* See Sen (1968) and Section 9.8.

Problems

1. Johnson et al. (1970) considered the behavior of a cenosphere-resin composite under hydrostatic pressure. The authors pointed out that most deep submersible vehicles utilize a buoyancy material, known as *syntactic foam*, that is a composite of closely packed hollow glass microspheres embedded in a resin matrix. These microspheres are relatively expensive to manufacture, and the cost of the syntactic foam is principally determined by the cost of the microspheres. The authors also noted that the ash from generating stations burning pulverized coal contains a small proportion of hollow glassy microspheres, known as *cenospheres*, and these have about the right size distribution for use in syntactic foam. The cenospheres can be readily collected from the ash-disposal method used in certain British generating stations. The authors were thus interested in whether the cenospheres would, in some applications, perform as well as the manufactured microspheres.

In attempting to assess the usefulness of cenospheres as a component of syntactic foam, Johnson et al. investigated the effects of hydrostatic pressure (such as exists in the ocean depths) on the density of a cenosphere-resin composite. The results are given in Table 9.2. What is the P-value for a test of $H_0 : \beta = 0$ against the alternative $\beta > 0$ for these data?

2. Explain why the effect of the unknown intercept parameter α (See model (9.1)) is "eliminated" in the application of procedure (9.4) to a set of data.

3. Consider the tapeworm data discussed in Problem 8.1. Using the mean weight of the initial force-fed cysticerci as the independent (predictor) variable, test the hypothesis that there was virtually no change in the mean weight of the cysticerci over the 20-day period following introduction into the dogs against the alternative that the typical tapeworm grew in size during the period of the study.

4. Stitt, Hardy, and Nadel (1971) studied the relationship between the surface area and body weight of squirrel monkeys. The data in Table 9.3 represent the total surface areas (cm^3) and

Table 9.2 The Effects of Hydrostatic Pressure on the Density of a Cenosphere-Resin Composite

Specimen	Pressure (psi)	Density (g/cm^3)
1	0	0.924
2	5,000	0.988
3	10,000	0.992
4	15,000	1.118
5	20,000	1.133
6	25,000	1.145
7	30,000	1.157
8	100,000	1.357

Source: A. A. Johnson, K. Mukherjee, S. Schlosser, and E. Raask (1970).

Table 9.3 Body Weight and Total Surface Area of Squirrel Monkeys

Monkey	Body weight, g	Total surface area, cm^3
1	660	780.6
2	705	887.6
3	994	1122.8
4	1129	1125.2
5	1005	l070.4
6	923	1039.2
7	953	1040.0
8	1018	1133.4
9	1181	1148.0

Source: J. T. Stitt, J. D. Hardy, and E. R. Nadel (1971).

body weights (g) for nine squirrel monkeys. Treating body weight as the independent variable, test for the presence of a linear relationship between these two measurements in squirrel monkeys.

5. Explain the meaning of the intercept parameter α and slope parameter β in model (9.1).
6. Consider the odor periods data of Table 8.8 in Problem 8.19. Test the conjecture that over the period 1950–1964, the number of odor periods for Lake Michigan generally increased at a rate greater than two per year.

9.2 A SLOPE ESTIMATOR ASSOCIATED WITH THE THEIL STATISTIC (THEIL)

Procedure

To estimate the slope parameter β of model (9.1), compute the $N = n(n-1)/2$ individual sample slope values $S_{ij} = (Y_j - Y_i)/(x_j - x_i), 1 \le i < j \le n$. The estimator of β (Theil (1950c)) associated with the Theil statistic, C, is

$$\hat{\beta} = \text{median } \{S_{ij}, 1 \le i < j \le n\}. \tag{9.18}$$

Let $S^{(1)} \le \cdots \le S^{(N)}$ denote the ordered values of the sample slopes S_{ij}. Then if N is odd, say $N = 2k + 1$, we have $k = (N-1)/2$ and

$$\hat{\beta} = S^{(k+1)}, \tag{9.19}$$

the value that occupies position $k + 1$ in the list of the ordered S_{ij} values. If N is even, say $N = 2k$, then $k = N/2$ and

$$\hat{\beta} = [S^{(k)} + S^{(k+1)}]/2. \tag{9.20}$$

That is, when N is even, $\hat{\beta}$ is the average of the two S_{ij} values that occupy positions k and $k + 1$ in the ordered list of all N sample slopes S_{ij}.

EXAMPLE 9.2 ***Effect of Cloud Seeding on Rainfall—Example 9.1 Continued.***

Consider the double-ratio data of Table 9.1. The ordered values of the $N = 5(4)/2 = 10$ sample slopes $S_{ij} = (Y_j - Y_i)/(x_j - x_i)$ are $S^{(1)} \le \cdots \le S^{(10)} : -.150, -.130, -.080, -.070, -.0575, -.055, -.045, -.033, .010,$ and $.040$. As $N = 10$ is even, we use (9.20) with $k = \frac{10}{2} = 5$ to obtain the slope estimate $\hat{\beta} = [S^{(5)} + S^{(6)}]/2 = [-.0575 - .055]/2 = -.0563$.

Comments

4. *Generalization for Nondistinct x-Values.* Sen (1968) generalized Theil's (1950c) estimator to the case where the x's are not distinct. Let N' denote the number of positive $x_j - x_i$ differences, for $1 \le i < j \le n$. (In the case where the x's are distinct, $N' = N$.) Sen's estimator of β is the median of the N' sample slope values that can be computed from the data. In the special case when

$x_1 = x_2 = \ldots = x_m = 0$ and $x_{m+1} = x_{m+2} = \ldots = x_{m+q} = 1$ (with $n = m + q$ and $m < n$), Sen's estimator reduces to the median of the $mq(Y_j - Y_i)$ differences, where $i = 1, \ldots, m$ and $j = m + 1, \ldots, m + q$. That is, Sen's estimator reduces to the Hodges–Lehmann two-sample estimator of Section 4.2 applied to the two samples $Y_1, \ldots, Y_m$ and $Y_{m+1}, \ldots, Y_{m+q}$.

Dietz (1989) considered various nonparametric estimators of the slope including Theil's estimator. She found that Theil's estimator is robust, easy to compute, and competitive in terms of mean squared error with alternative slope estimators. She also considered various nonparametric estimators of the intercept and of the mean response at a given x-value.

5. *Sensitivity to Gross Errors.* The estimator $\hat{\beta}$ (9.18) is less sensitive to gross errors than is the classical least squares estimator

$$\bar{\beta} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2},$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $\bar{Y} = \sum_{j=1}^{n} Y_j/n$.

6. *Median versus Weighted Average.* The estimator $\hat{\beta}$ (9.18) is the median of the N individual slope estimators $S_{ij} = (Y_j - Y_i)/(x_j - x_i)$. The least squares estimator $\bar{\beta}$ (See Comment 5) is a weighted average of the S_{ij}'s.

7. *Sample Slopes.* The command `theil` will output the $n(n-1)/2$ sample slopes if the additional argument `slopes=T` is specified. If `x` and `y` are the data from Table 9.1, then `theil(x, y, slopes=T)` will output the following table:

i	j	S_ij
1	2	0.01000000
1	3	−0.07000000
1	4	−0.03333333
1	5	−0.05750000
2	3	−0.15000000
2	4	−0.05500000
2	5	−0.08000000
3	4	0.04000000
3	5	−0.04500000
4	5	−0.13000000

A different slope estimator is due to Siegel (1982). For a fixed point, the Siegel estimator computes the $n - 1$ slopes with the remaining points and takes the median of these $n - 1$ values. This is done for each point, resulting in n medians. The median of these n medians is the estimate of β.

Properties

1. *Standard Deviation of $\hat{\beta}$ (9.18).* For the asymptotic standard deviation of $\hat{\beta}$ (9.18), see Sen (1968).

2. *Asymptotic Normality.* See Sen (1968).
3. *Efficiency.* See Sen (1968) and Section 9.8.

Problems

7. Estimate β for the cenosphere-resin data of Table 9.2.

8. Compute the least squares estimator $\bar{\beta}$ (See Comment 5) for the cenosphere-resin data of Table 9.2, and compare $\bar{\beta}$ with the $\hat{\beta}$ value obtained in Problem 7. In general, which of $\hat{\beta}$ and $\bar{\beta}$ is easier to compute?

9. Estimate β for the body-weight and surface-area data for squirrel monkeys discussed in Problem 4.

10. Obtain the set of 28 ordered individual sample slopes for the cenosphere-resin data of Table 9.2.

11. Estimate β for the tapeworm data discussed in Problems 3 and 8.1.

12. Obtain the set of 45 ordered individual sample slopes for the tapeworm data discussed in Problems 3 and 8.1.

9.3 A DISTRIBUTION-FREE CONFIDENCE INTERVAL ASSOCIATED WITH THE THEIL TEST (THEIL)

Procedure

For a symmetric two-sided confidence interval for β, with confidence coefficient $1-\alpha$, first obtain the upper $(\alpha/2)$th percentile point $k_{\alpha/2}$ of the null distribution of $\overline{C}$ (9.4). Let $C_\alpha = n(n-1)/2 \cdot k_{\alpha/2} - 2$ and set

$$M = \frac{N - C_\alpha}{2}, \tag{9.21}$$

and

$$Q = \frac{N + C_\alpha}{2} = M + C_\alpha, \tag{9.22}$$

where, once again, $N = n(n-1)/2$. The $100(1-\alpha)\%$ confidence interval (β_L, β_U) for the slope β that is associated with the two-sided Theil test (Section 9.1) is then given by

$$\beta_L = S^{(M)}, \quad \beta_U = S^{(Q+1)}, \tag{9.23}$$

where $S^{(1)} \le \cdots \le S^{(N)}$ are the ordered individual sample slopes $S_{ij} = (Y_j - Y_i)/(x_j - x_i), 1 \le i < j \le n$, used in computing the point estimator $\hat{\beta}$ (9.18). That is, β_L is the sample slope S_{ij} that occupies position M in the list of N ordered sample slopes. The upper end point β_U is the sample slope S_{ij} value that occupies position $Q = M + C_\alpha$ in this ordered list. With β_L and β_U given by display (9.23), we have

$$P_\beta(\beta_L < \beta < \beta_U) = 1 - \alpha \text{ for all } \beta. \tag{9.24}$$

For upper or lower confidence bounds for β associated with appropriate one-sided Theil's test procedures, see Comment 9.

Large-Sample Approximation

For large n, the integer C_α may be approximated by

$$C_\alpha \approx z_{\alpha/2}\left\{\frac{n(n-1)(2n+5)}{18}\right\}^{1/2}. \tag{9.25}$$

In general, the value of the right-hand side of (9.25) is not an integer. To be conservative, take C_α to be the largest integer that is less than or equal to the right-hand side of (9.25) for use in (9.21) and (9.22).

EXAMPLE 9.3 ***Effect of Cloud Seeding on Rainfall—Example 9.1 Continued.***

Consider the double-ratio data of Table 9.1. We illustrate how to obtain the 95% confidence interval for β. With $1-\alpha = .95$ (so that $\alpha = .05$), we see that $k_{\alpha/2} = k_{.025} = .8$. Thus, $C_{.025} = \frac{5.4}{2}k_{.025} - 2 = 8 - 2 = 6$. Since $N = 5(4)/2 = 10$, we see from (9.21) and (9.22) that

$$M = \frac{10-6}{2} = 2$$

and

$$Q = \frac{10+6}{2} = 8.$$

Using these values of $M = 2$ and $Q = 8$ in display (9.23), we see that

$$\beta_L = S^{(2)}, \quad \beta_U = S^{(9)}$$

provide the end points of our 95% confidence interval for β. From the ordered list of sample slope values given in Example 9.2, we obtain $S^{(2)} = -.130$ and $S^{(9)} = .010$, so that our 95% confidence interval for β is

$$(\beta_L, \beta_U) = (-.130, .010).$$

Note that the critical value given by `R` results in a confidence level of $1-\alpha = .917$, not $1-\alpha = .95$. Using the `theil` command with the arguments `alpha=1-.05` and `type="t"` results in the following output:

```
1 - alpha = 0.05 two-sided CI for beta:
-0.15, 0.04
```

The theil command produces an interval whose confidence level is at least $1-\alpha$. The actual confidence level for this interval is $1-\alpha = .983$, compared to $1-\alpha = .917$ for the interval determined by hand.

Comments

8. *Use of* `R` *to Compute the End Points of the Confidence Interval (9.3).* The $n(n-1)/2$ individual sample slope values $S_{ij} = (Y_j - Y_i)/(x_j - x_i), 1 \le i \le$

$j \leq n$, can also be obtained from the `theil` command by specifying the argument `slopes=T`. For details, See Comment 7.

9. *Confidence Bounds*. In many settings, we are interested only in making one-sided confidence statements about the parameter β; that is, we wish to assert with specified confidence that β is no larger (or, in other settings, no smaller) than some upper (lower) confidence bound based on the sample data. To obtain such one-sided confidence bounds for β, we proceed as follows. For specified confidence coefficient $1-\alpha$, find the upper αth [not $(\alpha/2)$th, as for the confidence interval] percentile point k_a of the null distribution of C (9.4). Let $C_\alpha^* = \frac{n(n-1)}{2} k_\alpha - 2$ and set

$$M^* = \frac{N - C_\alpha^*}{2} \quad \text{and} \quad Q^* = \frac{N + C_\alpha^*}{2}. \tag{9.26}$$

The $100(1-\alpha)\%$ *lower* confidence bound β_{L}^* for β is then given by

$$(\beta_{\mathrm{L}}^*, \infty) = (S^{(M^*)}, \infty), \tag{9.27}$$

where, as before, $S^{(1)} \leq \cdots \leq S^{(N)}$ are the ordered individual sample slopes. With β_{L}^* given by display (9.27), we have

$$P_\beta(\beta_{\mathrm{L}}^* < \beta < \infty) = 1 - \alpha \text{ for all } \beta. \tag{9.28}$$

The corresponding $100(1-\alpha)\%$ *upper* confidence bound β_{U}^* is given by

$$(-\infty, \beta_{\mathrm{U}}^*) = (-\infty, S^{(Q^*+1)}). \tag{9.29}$$

It follows that

$$P_\beta(-\infty < \beta < \beta_{\mathrm{U}}^*) = 1 - \alpha \text{ for all } \beta. \tag{9.30}$$

For large n, the integer C_α^* may be approximated by

$$C_\alpha^* \approx z_\alpha \left\{ \frac{n(n-1)(2n+5)}{18} \right\}^{1/2}. \tag{9.31}$$

In general, the value of the right-hand side of (9.31) is not an integer. To be conservative, take C_α^* to be the largest integer that is less than or equal to the right-hand side of (9.31) for use in display (9.26).

10. *Midpoint of the Confidence Interval as an Estimator*. The midpoint of the confidence interval given by (9.23), namely, $[S^{[M)} + S^{(Q+1)}]/2$, suggests itself as a reasonable estimator of β. (Note that this actually yields a class of estimators depending on the value of α.) In general, this midpoint does not give the same value as $\hat{\beta}$ (9.18).

Properties

1. *Distribution-Freeness.* For populations satisfying Assumptions A1 and A2, (9.24) holds (See Theil (1950b, 1950c)). Hence, we can control the coverage probability to be $1-\alpha$ without having more specific knowledge about the form of the common underlying distribution of the e_i's. Thus, (β_L, β_U) is a distribution-free confidence interval for β over a very large class of populations.
2. *Efficiency.* See Sen (1968) and Section 9.8.

Problems

13. Obtain a 90% confidence interval for β for the cenosphere-resin data in Table 9.2.

14. Obtain a 95% confidence interval for β for the body weight and surface area data for squirrel monkeys discussed in Problems 4 and 9.

15. Consider a fixed set of data. Show that for $\alpha_2 > \alpha_1$, the symmetric two-sided $(1-\alpha_1)$ confidence interval for β given by (9.23) is always as long or longer than the symmetric two-sided $(1-\alpha_2)$ confidence interval for β.

16. Obtain a 95% upper confidence bound (See Comment 9) for β for the double-ratio cloud-seeding data in Table 9.1.

17. Obtain a 95% lower confidence bound (See Comment 9) for β for the tapeworm data discussed in Problems 8.1, 3, and 11.

18. Obtain a 90% confidence interval for β for the Lake Michigan odor periods data discussed in Problems 8.19 and 6.

19. Find the midpoint of the 95% confidence interval for β obtained in Problem 14 for the squirrel monkey body-weight and surface-area data. As noted in Comment 10, this midpoint can be used to estimate the value of β. Compare this midpoint estimator with the value of $\hat{\beta}$ (9.18) obtained in Problem 9.

9.4 AN INTERCEPT ESTIMATOR ASSOCIATED WITH THE THEIL STATISTIC AND USE OF THE ESTIMATED LINEAR RELATIONSHIP FOR PREDICTION (HETTMANSPERGER–McKEAN–SHEATHER)

Procedure

To estimate the intercept parameter α of model (9.1), we define

$$A_i = Y_i - \hat{\beta}x_i, \quad i = 1, \ldots, n, \tag{9.32}$$

where $\hat{\beta}$ is the point estimator of β given in (9.18). An estimator associated with the Theil statistic C and suggested by Hettmansperger, McKean, and Sheather (1997) is

$$\hat{\alpha} = \text{median } \{A_1, \ldots, A_n\}. \tag{9.33}$$

Let $A^{(1)} \leq \cdots \leq A^{(n)}$ denote the ordered A_i values (9.32). Then if n is odd, say $n = 2k+1$, we have $k = (n-1)/2$ and

$$\hat{\alpha} = A^{(k+1)}, \tag{9.34}$$

the value that occupies position $k+1$ in the list of ordered A_i values. If n is even, say $n = 2k$, then $k = n/2$ and

$$\hat{\alpha} = \frac{A^{(k)} + A^{(k+1)}}{2}. \tag{9.35}$$

That is, when n is even, $\hat{\alpha}$ is the average of the two values that occupy positions k and $k+1$ in the ordered list of all n A_i's.

Employing both the estimator $\hat{\beta}$ (9.18) for the slope and the estimator $\hat{\alpha}$ (9.33) for the intercept, our estimated linear relationship between the x and Y variables is then given by

$$\overbrace{\text{med } Y_{x=x^*}} = [\overbrace{\text{median } Y \text{ when } x = x^*}] = \hat{\alpha} + \hat{\beta}\, x^*. \tag{9.36}$$

That is, we would predict $\overbrace{\text{med } Y_{x=x^*}}$ to be the typical value of the dependent variable Y for a future setting of the independent variable x at x^*. (See also Comment 12.)

EXAMPLE 9.4 ***Effect of Cloud Seeding on Rainfall—Example 9.1 Continued.***

Once again, consider the double-ratio data of Table 9.1. From Example 9.2, we see that the slope estimate for these data is $\hat{\beta} = -.0563$. Combining this value with the (x_i, Y_j) pairs from Table 9.1, the five ordered A (9.32) values are $A^{(1)} \le \cdots \le A^{(5)}$: 1.2889, 1.3115, 1.3163, 1.3826, and 1.3852. As $n = 5$ is odd, we use (9.34) with $k = (5-1)/2 = 2$ to obtain the intercept estimate $\hat{\alpha} = A^{(3)} = 1.3163$. This estimate is provided by the command `theil`.

Combining the slope estimate of $\hat{\beta} = -.0563$ and this intercept estimate of $\alpha = 1.3163$, our final estimated linear relationship between the x and Y variables is then given by (9.36) to be

$$\overbrace{\text{med } Y_{x=x^*}} = [\overbrace{\text{median } Y \text{ when } x = x^*}] = 1.3163 - .0563x^*. \tag{9.37}$$

Thus, for example, we would estimate the median double-ratio value after 4.5 years of the cloud-seeding study to have been

$$\begin{aligned}\overbrace{\text{med } Y_{x=4.5}} &= [\overbrace{\text{median } Y \text{ when } x = 4.5 \text{ years}}] \\ &= 1.3163 - .0563(4.5) = 1.06295.\end{aligned}$$

Using (9.37) once again, the predicted double-ratio value if the study were to continue for a sixth year would be

$$\begin{aligned}\overbrace{\text{med } Y_{x=6}} &= [\overbrace{\text{median } Y \text{ when } x = 6 \text{ years}}] \\ &= 1.3163 - .0563(6) = 0.9785.\end{aligned}$$

One must always exercise caution in using an estimated linear relationship to predict typical values of the dependent variable Y for values of the independent variable x that are too different from the range of x-values used in establishing the estimated linear relationship (See Comment 12). For example, it would make no sense whatsoever to use the relationship (9.37) to estimate the median double-ratio value for negative values of x^*,

because they are not possible. In addition, although $x^* = 20$ years would certainly be a possible value for the independent variable (corresponding to 20 consecutive years of the cloud-seeding study), in order to use (9.37) to predict the typical double-ratio value Y after 20 years of the study would require the assumption that the regression relationship (9.1) remains linear for that extended time period. Although this may be a reasonable assumption to make, it is not one that comes automatically. Careful consideration should be given to its validity before using (9.37) to predict the double-ratio value that far into the future based solely on the 5 years of available data.

Comments

11. *Competing Estimators.* The intercept estimator given by (9.33) is not the only nonparametric estimator of α that has been studied in the statistical literature. In the case of symmetry of the underlying distribution for the error random variables $e_1, \ldots, e_n$ in Assumption A2, Hettmansperger and McKean (1977) proposed the competing estimator $\tilde{\alpha}$ associated with the median of the $n(n+1)/2$ Walsh averages of the n individual A_i (9.32) differences. Adichie (1967) proposed and studied the asymptotic properties of an entire class of estimators for α associated with rank tests.
12. *Appropriate Range of Values of the Independent Variable for Purposes of Prediction.* When we choose to use the estimated linear relationship (9.36) to predict typical values of the dependent variable Y for a particular setting of the independent variable x^*, caution must always be the rule. For prediction purposes, we must be relatively confident that the linear relationship holds at least approximately when x assumes the value x^*. This is seldom of concern when x^* is well situated among the values of the independent variable at which we observed sample-dependent variables in obtaining the estimated relationship (9.36) in the first place. However, when we are interested in predicting the typical value of the dependent variable Y for a setting of the independent variable x^* that is outside the range for which sample data had been collected in obtaining the estimated relationship (9.36), we must not automatically assume that the linear relationship (9.1) is still appropriate. Careful consideration must be given to justification of the reasonableness of this relationship for the particular problem of interest prior to using (9.36) for prediction purposes when considering such extended ranges of the independent variable.

Problems

20. Estimate α for the cenosphere-resin data of Table 9.2.

21. Estimate α for the body-weight and surface-area data for squirrel monkeys discussed in Problem 4.

22. Use the linear relationship (9.36) to estimate the typical density of a cenosphere-resin composite under hydrostatic pressure of 17,500 psi.

23. Use the linear relationship (9.36) to estimate the typical total surface area (cm^3) for a squirrel monkey with a body weight of 1000 g.

24. Estimate α for the tapeworm data discussed in Problems 8.1, 3, 11, and 17. Use the linear relationship (9.36) to estimate the typical weight of worms recovered from a dog that had been force-fed 20 mg cysticerci of *Taenia hydatigena.*

25. Consider the cenosphere-resin data of Table 9.2. Discuss the reasonableness of using the linear relationship (9.36) established in Problem 20 for these data to estimate the typical density of cenosphere-resin composites under hydrostatic pressures of 35,000, 75,000, and 200,000 psi.

26. Consider the body-weight and surface-area data for squirrel monkeys presented in Problem 4. Discuss the reasonableness of using the linear relationship (9.36) established in Problem 21 for these data to estimate the typical total surface area (cm^3) for squirrel monkeys with body weights of 320, 975, and 2500 g.

k(≥2) REGRESSION LINES

9.5 AN ASYMPTOTICALLY DISTRIBUTION-FREE TEST FOR THE PARALLELISM OF SEVERAL REGRESSION LINES (SEN, ADICHIE)

In this section, we discuss an asymptotically distribution-free procedure to test for parallelism of $k \geq 2$ regression lines. Thus, we are concerned with testing equality of the k slope parameters without additional constraints on the corresponding, unspecified intercepts.

Data. For the ith line, $i = 1, \ldots, k$, we observe the value of the ith response random variable Y_i at each of n_i fixed levels, $x_{i1}, \ldots, x_{in_i}$, of the ith independent (predictor) variable x_i. Thus, for the ith line, $i = 1, \ldots, k$, we obtain a set of observations $Y_{i1}, \ldots, Y_{in_i}$, where Y_{ij} is the value of the response variable Y_i when $x_i = x_{ij}$.

Assumptions

Bl. We take as our straight-line model

$$Y_{ij} = \alpha_i + \beta_i x_{ij} + e_{ij}, \quad i = 1, \ldots, k; \quad j = 1, \ldots, n_i, \tag{9.38}$$

where the x_{ij}'s are known constants and $\alpha_1, \ldots, \alpha_k$ and $\beta_1, \ldots, \beta_k$ are the unknown intercept and slope parameters, respectively.

B2. The $N = n_1 + \cdots + n_k$ random variables $e_{11}, \ldots, e_{1n_1}, \ldots, e_{k1}, \ldots, e_{kn_k}$ are mutually independent.

B3. The random variables $\{e_{i1}, \ldots, e_{in_i}\}, i = 1, \ldots, k$, are k random samples from a common continuous population with distribution function $F(.)$.

Hypothesis

The null hypothesis of interest here is that the k regression lines in model (9.38) have a common, but unspecified, slope, β, namely,

$$H_0 : [\beta_1 = \cdots = \beta_k = \beta, \text{ with } \beta \text{ unspecified}]. \tag{9.39}$$

Note that this null hypothesis does not place any conditions whatsoever on the intercept parameters $\alpha_1, \ldots, \alpha_k$. Thus, the assertion in H_0 (9.39) is simply that the k regression lines in model (9.38) are parallel.

Procedure

To construct the Sen–Adichie statistic V, we first align each of the k regression samples. Let $\bar{\beta}$ be the pooled least squares estimator for the common slope β under the null hypothesis H_0 (9.39), as given by

$$\bar{\beta} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) Y_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}, \tag{9.40}$$

where

$$\bar{x}_i = \sum_{j=1}^{n_i} \frac{x_{ij}}{n_i}, \quad \text{for } i = 1, \ldots, k. \tag{9.41}$$

For each of the k regression samples, compute the aligned observations

$$Y_{ij}^* = (Y_{ij} - \bar{\beta}\, x_{ij}), \quad i = 1, \ldots, k; \quad j = 1, \ldots, n_i. \tag{9.42}$$

Order these aligned observations Y_{ij}^* from least to greatest separately within each of the k regression samples. Let r_{ij}^* denote the rank of Y_{ij}^* in the joint ranking of the aligned observations $Y_{i1}^*, \ldots, Y_{in_i}^*$ in the ith regression sample.

Compute

$$T_i^* = \sum_{j=1}^{n_i} [(x_{ij} - \bar{x}_i) r_{ij}^*]/(n_i + 1), \quad i = 1, \ldots, k, \tag{9.43}$$

where $\bar{x}_i$ is given by (9.41). Setting

$$C_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2, \quad i = 1, \ldots, k, \tag{9.44}$$

the Sen–Adichie statistic V is then given by

$$V = 12 \sum_{i=1}^{k} \left[\frac{T_i^*}{C_i} \right]^2. \tag{9.45}$$

To test

$$H_0 : [\beta_1 = \cdots = \beta_k = \beta, \text{ with } \beta \text{ unspecified}]$$

versus the general alternative

$$H_1 : [\beta_1, \ldots, \beta_k \text{ not all equal}] \tag{9.46}$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } V \geq \chi^2_{k-1,\alpha}; \quad \text{otherwise do not reject,} \tag{9.47}$$

where $\chi^2_{k-1,\alpha}$ is the upper α percentile of a chi-square distribution with $k - 1$ degrees of freedom. Values of $\chi^2_{k-1,\alpha}$ can be obtained from the `R` command `qchisq`.

Ties

If there are ties among the n_i aligned observations Y_{ij}^* (9.42) for the ith regression sample, use average ranks to break the ties and compute the weighted sum T_i^* (9.43) contribution to V (9.45) for that sample.

EXAMPLE 9.5 ***Ammonium Flux in Coastal Sediments.***

Coastal sediments are an important reservoir for organic nitrogen (ON). The degradation and mineralization of ON in coastal sediments is bacterially mediated and is known to involve several distinct steps. Moreover, it is possible to measure the rates of the processes at each of these steps. During the first stage of ON remineralization, ammonium is generated by heterotrophic bacteria during a process called *ammonification*. Ammonium can then be released to the environment or be microbially transformed to other nitrogenous species.

Mortazavi (1997) collected four sediment cores from Apalachicola Bay, Florida, and analyzed them at the Florida State University. The flux of ammonium (μ moles N per square meter of surface area) to the overlying water was measured for each core sample every 90 minutes during a 6-hour incubation period. These data are presented in Table 9.4 for the four core samples.

We are interested in assessing whether the rate of ammonium flux is similar across these four coastal sediments (at least over the 6-hour period of the study). Thus, if we

Table 9.4 Coastal Sediment Ammonium Flux in Apalachicola Bay, Florida

Core sample, i	Time, x_{ij} (h)	Ammonium flux, Y_{ij} (μ moles N/m^2)
Core 1	0	0
	1.5	33.019
	3	111.314
	4.5	196.205
	6	230.658
Core 2	0	0
	1.5	131.831
	3	181.603
	4.5	230.070
	6	258.119
Core 3	0	0
	1.5	33.351
	3	97.463
	4.5	196.615
	6	217.308
Core 4	0	0
	1.5	8.959
	3	105.384
	4.5	211.392
	6	255.105

Source: B. Mortazavi (1997).

let β_i correspond to the rate of ammonium flux for the ith coastal sediment core sample, $i = 1, \ldots, 4$, we are interested in testing the null hypothesis H_0 (9.39) against the general alternative (9.46) that the rates are not the same for the four coastal areas in Apalachicola from which the core samples were drawn.

First, we must obtain the pooled least squares estimator $\bar{\beta}$ (9.40). The set of x_{ij} values is the same for each of the coastal sediment samples, so

$$\bar{x}_1 = \bar{x}_2 = \bar{x}_3 = \bar{x}_4 = \frac{0 + 1.5 + 3 + 4.5 + 6}{5} = 3.$$

Hence, from (9.44), we obtain

$$\begin{aligned} C_1^2 = C_2^2 = C_3^2 = C_4^2 &= (0-3)^2 + (1.5-3)^2 + (3-3)^2 + (4.5-3)^2 + (6-3)^2 \\ &= 9 + 2.25 + 0 + 2.25 + 9 + 22.5, \end{aligned}$$

which, in turn, yields

$$\sum_{i=1}^{4} \sum_{j=1}^{5} (x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{4} C_i^2 = 4(22.5) = 90.$$

For the numerator of $\bar{\beta}$ (9.40), we see that

$$\begin{aligned} \sum_{i=1}^{4} \sum_{j=1}^{5} (x_{ij} - \bar{x}_i)(Y_{ij}) = [&(0-3)(0+0+0+0) \\ &+ (1.5-3)(33.019 + 131.831 + 33.351 + 8.959) \\ &+ (3-3)(111.314 + 181.603 + 97.463 + 105.384) \\ &+ (4.5-3)(196.205 + 230.070 + 196.615 + 211.392) \\ &+ (6-3)(230.658 + 258.119 + 217.308 + 255.105)] \\ = [&0 - 310.74 + 0 + 1251.423 + 2883.57] = 3824.253. \end{aligned}$$

Combining these two quantities, we obtain the value of the pooled least squares slope estimator (9.40) to be

$$\bar{\beta} = \frac{3824.253}{90} = 42.49.$$

Next, we create the aligned observations Y_{ij}^* (9.42) for each of the core samples:

$$\begin{aligned} \text{Core 1}: \quad Y_{11}^* &= 0 - 42.49(0) = 0 \\ Y_{12}^* &= 33.019 - 42.49(1.5) = -30.716 \\ Y_{13}^* &= 111.314 - 42.49(3) = -16.156 \\ Y_{14}^* &= 196.205 - 42.49(4.5) = 5 \\ Y_{15}^* &= 230.658 - 42.49(6) = -24.282 \end{aligned}$$

$$\begin{aligned}
\text{Core 2}: Y_{21}^* &= 0 - 42.49(0) = 0\\
Y_{22}^* &= 131.831 - 42.49(1.5) = 68.096\\
Y_{23}^* &= 181.603 - 42.49(3) = 54.133\\
Y_{24}^* &= 230.070 - 42.49(4.5) = 38.865\\
Y_{25}^* &= 258.119 - 42.49(6) = 3.179
\end{aligned}$$

$$\begin{aligned}
\text{Core 3}: Y_{31}^* &= 0 - 42.49(0) = 0\\
Y_{32}^* &= 33.351 - 42.49(1.5) = -30.384\\
Y_{33}^* &= 97.463 - 42.49(3) = -30.007\\
Y_{34}^* &= 196.615 - 42.49(4.5) = 5.41\\
Y_{35}^* &= 217.308 - 42.49(6) = -37.632
\end{aligned}$$

$$\begin{aligned}
\text{Core 4}: Y_{41}^* &= 0 - 42.49(0) = 0\\
Y_{42}^* &= 8.959 - 42.49(1.5) = -54.776\\
Y_{43}^* &= 105.384 - 42.49(3) = -22.086\\
Y_{44}^* &= 211.392 - 42.49(4.5) = 20.187\\
Y_{45}^* &= 255.105 - 42.49(6) = 0.165.
\end{aligned}$$

Ordering these aligned observations Y_{ij}^* from least to greatest separately within each of the four core samples, we obtain the following within-samples rankings:

$$\begin{aligned}
&\text{Core 1}: r_{11}^* = 4, \quad r_{12}^* = 1, \quad r_{13}^* = 3, \quad r_{14}^* = 5, \quad \text{and} \quad r_{15}^* = 2.\\
&\text{Core 2}: r_{21}^* = 1, \quad r_{22}^* = 5, \quad r_{23}^* = 4, \quad r_{24}^* = 3, \quad \text{and} \quad r_{25}^* = 2.\\
&\text{Core 3}: r_{31}^* = 4, \quad r_{32}^* = 2, \quad r_{33}^* = 3, \quad r_{34}^* = 5, \quad \text{and} \quad r_{35}^* = 1.\\
&\text{Core 4}: r_{41}^* = 3, \quad r_{42}^* = 1, \quad r_{43}^* = 2, \quad r_{44}^* = 5, \quad \text{and} \quad r_{45}^* = 4.
\end{aligned}$$

The values of $T_1^*, \ldots, T_4^*$ are then obtained from (9.43) to be

$$\begin{aligned}
T_1^* &= \frac{[(0-3)(4) + (1.5-3)(1) + (3-3)(3) + (4.5-3)(5) + (6-3)(2)]}{(5+1)} = 0,\\
T_2^* &= \frac{[(0-3)(1) + (1.5-3)(5) + (3-3)(4) + (4.5-3)(3) + (6-3)(2)]}{(5+1)} = 0,\\
T_3^* &= \frac{[(0-3)(4) + (1.5-3)(2) + (3-3)(3) + (4.5-3)(5) + (6-3)(1)]}{(5+1)} = -.75,\\
T_4^* &= \frac{[(0-3)(3) + (1.5-3)(1) + (3-3)(2) + (4.5-3)(5) + (6-3)(4)]}{(5+1)} = 1.5.
\end{aligned}$$

Combining these T_i^* values with the corresponding values of C_i^2 previously obtained, we see from (9.45) that the Sen–Adichie statistic V for these data is given by

$$V = 12\left\{\frac{(0)^2}{22.5} + \frac{(0)^2}{22.5} + \frac{(-.75)^2}{22.5} + \frac{(1.5)^2}{22.5}\right\}$$
$$= 12\{0 + 0 + 0.25 + .1\} = 1.5.$$

For the Sen–Adichie procedure (9.47), we compare the value of V to the chi-square distribution with $k - 1 = 3$ degrees of freedom. We see that the observed value of $V = 1.5$ is the .318 percentile for the chi-square distribution with 3 degrees of freedom. Thus, the P-value for these data and test procedure (9.47) is .682, indicating that there is virtually no sample evidence in support of significant differences in the rates (slopes) of ammonium flux for the four coastal areas sampled.

The `R` command `sen.adichie` replicates this analysis. The argument is a list `z`. There are k items in the list `z`, one for each set of data corresponding to a specific linear relation. Each of these k items is a matrix with the first column the x values, the second the Y values.

Comments

13. *Motivation for the Test.* The pooled least squares estimator $\bar{\beta}$ (9.40) estimates some weighted combination, say β^*, of the k individual slopes $\beta_1, \ldots, \beta_k$ (9.38). From Assumptions Bl and B3, it follows that the aligned observations Y_{ij}^* (9.42), $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, will tend to have values near

$$\begin{aligned}\text{med}(Y_{ij}^*) &= \text{med}(Y_{ij} - \bar{\beta}x_{ij}) \\ &\approx \text{med}(Y_{ij}) - \beta^* x_{ij} = \alpha_i + \beta_i x_{ij} - \beta^* x_{ij} + \text{med}(e_{ij}) \\ &= \alpha_i + (\beta_i - \beta^*)x_{ij} + \text{med}(e_{ij}). \end{aligned} \tag{9.48}$$

If the null hypothesis H_0 (9.39) is true, then $\beta_1 = \cdots = \beta_k = \beta^* = \beta$ and we would expect each of $Y_{i1}^*, \ldots, Y_{in_i}^*$ to be near $\alpha_i + \text{med}(e_{ij})$, for each of the regression samples $i = 1, \ldots, k$. As the r_{ij}^* ranks are obtained separately within each of the k samples, it follows that under H_0 (9.39) the ranks $r_{i1}^*, \ldots, r_{in_i}^*$ should behave like a random permutation of the integers $1, \ldots, n_i$ and exhibit no additional relationship with the regression constants $x_{i1}, \ldots, x_{in_i}$, for $i = 1, \ldots, k$. Thus, the null hypothesis setting should lead to values of T_i^* near zero, for $i = 1, \ldots, k$, and subsequently to small values of the Sen–Adichie test statistic V (9.45). On the other hand, if the null hypothesis H_0 (9.39) is not true, then some of the β_i's will be larger than β^* and some of them will be smaller than β^*. For those regression populations for which β_i is larger than β^*, we see from (9.48) that the aligned observations Y_{ij}^* (9.42) will be positively related to the values of the corresponding regression constants x_{ij}. This would tend to produce large positive values for the corresponding T_i^*'s (9.43). For those regression populations for which β_i is smaller than β^*, we see from (9.48) that the aligned observations Y_{ij}^* (9.42) will be negatively related to the values of the corresponding regression constants x_{ij}. This would tend to produce large negative values for the corresponding T_i^*'s (9.43). Each of these T_i^* values

is squared in the calculation of the Sen–Adichie test statistic V (9.45). Therefore, regression populations with either β_i larger or smaller than β^* will tend to produce large contributions to the test statistic V, providing partial motivation for procedure (9.47).

14. *Historical Development.* The general form of the test procedure (9.47), but using a rank estimate for the common value of the slope parameter β under H_0 (9.39) in the construction of the aligned observations Y_{ij}^* (9.42), was first proposed and studied by Sen (1969). The use of the pooled least squares estimator $\bar{\beta}$ (9.40) in the construction of the Y_{ij}^*'s was first suggested by Adichie (l984).

15. *Potthoff's Conservative Test of Parallelism.* For the case $k = 2$, Potthoff (1974) proposed a Wilcoxon-type test of $\beta_1 = \beta_2$. He compared each sample slope that can be computed from line 2 data with each sample slope that can be computed from line 1 data, scoring 1 if the sample 2 slope is larger than the sample 1 slope and 0 otherwise. His statistic was the average of the $n_1(n_1 - 1)(n_2)(n_2 - 1)/4$ such indicators. (To avoid complications, he assumed no two x_{1j}'s are equal and no two x_{2j}'s are equal.) The test associated with his statistic was neither distribution-free nor asymptotically distribution-free. Instead, he used an upper bound for the null variance of the statistic to produce a conservative test procedure.

16. *Competitor Based on Joint Rankings When the Intercept Is Common.* The Sen–Adichie procedure (9.47) is based on the individual rankings of the aligned observations Y_{ij}^* (9.42) *separately* within each of the k samples. This requires a good deal more computational time than if we could use a single simultaneous ranking of all $N = n_1 + \cdots + n_k$ aligned observations. Although such a joint ranking is not appropriate for the general model (9.38), Adichie (1974) proposed a procedure based on the joint ranking of all N of the aligned observations for settings where it is also reasonable to assume equality of the k intercepts $\alpha_1, \ldots, \alpha_k$ in model (9.38). Thus, Adichie's procedure is appropriate for testing H_0 (9.39) under Assumptions B2, B3 and the following more restrictive Assumption B1′ replacing Assumption B1:
Bl′. We take as our straight-line model

$$Y_{ij} = \alpha + \beta_i x_{ij} + e_{ij}, \quad i = 1, \ldots, k; j = 1, \ldots, n_i, \tag{9.49}$$

where the x_{ij}'s are known constants, α is the common (unknown) intercept and $\beta_1, \ldots, \beta_k$ are the unknown slope parameters, respectively.

Adichie's (1974) test statistic for this more restrictive setting is quite similar in form to the Sen–Adichie test statistic V (9.45). The major difference is the use of the single simultaneous ranking of all N of the aligned observations, rather than the k separate rankings utilized in constructing V. (We note, in passing, that the assumption of a common intercept α would be quite reasonable for the ammonium flux data considered in Example 9.5.)

17. *Test Procedures for Restricted Alternatives.* The Sen–Adichie procedure (9.47) is designed to test H_0 (9.39) against the class of general alternatives H_1 (9.46). Other authors have proposed nonparametric procedures designed to test H_0 against more restricted classes of alternatives. Adichie (1976) and Rao and Gore (1984) studied asymptotically distribution-free test procedures designed to reach a decision between H_0 and the class of ordered alternatives $H_2 : [\beta_1 \leq \cdots \leq \beta_k,$

with at least one strict inequality]. Finally, Kim and Lim (1995) considered asymptotically distribution-free procedures for testing H_0 against umbrella alternatives of the form $H_3 : [\beta_1 \leq \cdots \leq \beta_p \geq \beta_{p+1} \geq \cdots \geq \beta_k$, with at least one strict inequality].

18. *Comparing Several Regression Lines with a Control.* The Sen–Adichie procedure (9.47) is designed to test H_0 (9.39) against the class of general alternatives H_1 (9.46). In this context, the test involves a comparison of each regression line with every other regression line. For settings where one of the regression lines corresponds to a standard line for a control population, we might want to make only the $k - 1$ comparisons between the noncontrol regression lines and this control line. Lim and Wolfe (1997) proposed and studied an asymptotically distribution-free procedure for testing the null hypothesis H_0 (9.39) against the "treatments" versus control alternative $H_4 : [\beta_1 \leq \beta_i, i = 2, \ldots, k$, with at least one strict inequality], where, without loss of generality, the first regression line plays the role of the control line.

Properties

1. *Asymptotic Chi-Squareness.* See Sen (1969).
2. *Efficiency.* See Sen (1969) and Section 9.8.

Problems

27. Wells and Wells (1967) discussed Project SCUD, an attempt to study the effects of cloud seeding on cyclones. The basic hypothesis of interest was that cloud seeding in areas of cyclogenesis on the east coast of the United States had no measurable effect on the development of storms there. Table 9.5, based on a subset (Experiment 1 of Table I of Wells and Wells (1967)) of the observational data from Project SCUD, gives "*RI*" and "*M*" values for 11 seeded and 10 control units. The quantity *RI* is a measure of precipitation and the quantity *M*, the geostrophic meridional circulation index, was used in predicting cyclogenesis. Cyclones

Table 9.5 Precipitation Amounts *RI* and Circulation Index *M* for Seeded and Control Units

	Seeded		Control	
Unit, j	$x_{1j}(M)$	$Y_{1j}(RI)$	$x_{2j}(M)$	$Y_{2j}(RI)$
1	24	.180	−7	.138
2	28	.175	10	.081
3	30	.178	17	.072
4	37	.021	25	.188
5	43	.260	44	.075
6	47	.715	51	.435
7	52	.441	53	.423
8	57	.205	63	.339
9	71	.417	75	.519
10	87	.498	90	.738
11	115	.603		

Source: J. M. Wells and M. A. Wells (1967).

were expected to develop only when M was predicted positive. Test that the regression lines of RI on M for seeded and control units are parallel.

28. Consider the aligned observations Y_{ij}^* (9.42), $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. Discuss why additional knowledge about the intercept parameters $\alpha_1, \ldots, \alpha_k$ is not necessary in order to use the *separate* within-samples ranks of the Y_{ij}^*'s in the construction of the test statistic V (9.45).

29. Consider the aligned observations Y_{ij}^* (9.42), $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$.

(a) Discuss why it would *not* be appropriate, in general, to use a single simultaneous ranking of all $N = n_1 + \cdots + n_k$ aligned observations in the construction of a statistic for testing H_0 (9.39).

(b) Under what conditions on the intercept parameters $\alpha_1, \ldots, \alpha_k$ might such a single simultaneous ranking be appropriate for developing a statistic to test H_0 (9.39)? (See Comment 16.)

30. Wardlaw and van Belle (1964) discussed the mouse hemidiaphragm method for assaying insulin. This procedure depends on the ability of the hormone to stimulate glycogen synthesis by the diaphragm tissue, in vitro. Hemidiaphragms are dissected from mice of uniform weight that have been starved for 18 hours. The tissues are incubated in tubes, and after incubation, the hemidiaphragms are washed with water and analyzed for glycogen content using anthrone reagent. The content is measured in terms of optical density. The procedure makes use of the fact that increasing the concentration of insulin in the incubation medium tends to increase glycogen synthesis by the hemidiaphragms. Specifically, for levels of insulin between .1 and 1.0 μ/ml, there is an approximate linear relationship between glycogen content and log concentration of insulin (See Wardlaw and Moloney (1961)).

The data in Table 9.6 are the log concentrations of insulin and the glycogen contents for 12 observations each from two varieties of insulin, namely, standard insulin and sample 1 insulin. For both standard and sample 1 lines, there are six observations at an insulin volume of .3 ml and six observations at a volume of 1.5 ml. In this insulin assay, and in many bioassays, the question of parallelism is extremely important, because the concept of relative potency (of a test preparation with respect to a standard) depends on the assumption that the dose–response lines are parallel. Using the data in Table 9.6, test the hypothesis that the dose–response lines for standard insulin and sample 1 insulin are parallel.

31. Experimental geneticists use survival under stressful conditions to compare the relative fitness of different species. Dowdy and Wearden (1991) considered data relating to the survival of

Table 9.6 Glycogen Content of Hemidiaphragms Measured by Optical Density in the Anthrone Test × 1000

	Standard insulin		Sample I insulin	
j	x_{1j} (log dose)	Y_{1j} (glycogen)	x_{2j} (log dose)	Y_{2j} (glycogen)
1	log (0.3)	230	log (0.3)	310
2	log (0.3)	290	log (0.3)	265
3	log (0.3)	265	log (0.3)	300
4	log (0.3)	225	log (0.3)	295
5	log (0.3)	285	log (0.3)	255
6	log (0.3)	280	log (0.3)	280
7	log (1.5)	365	log (1.5)	415
8	log (1.5)	325	log (1.5)	375
9	log (1.5)	360	log (1.5)	375
10	log (1.5)	300	log (1.5)	275
11	log (1.5)	360	log (1.5)	380
12	log (1.5)	385	log (1.5)	380

Source: A. C. Wardlaw and G. van Belle (1964).

Table 9.7 Numbers of *Drosophila* Flies (Three Different Species) That Survive to Adulthood after Exposure to Various Levels (ppm) of an Organic Phosphorus Insecticide

Species	Level of insecticide (ppm)	Number survived to adulthood
Drosophila melanogaster	0.0	91
	0.3	71
	0.6	23
	0.9	5
Drosophila pseudoobscura	0.0	89
	0.3	77
	0.6	12
	0.9	2
Drosophila serrata	0.0	87
	0.3	43
	0.6	22
	0.9	8

Source: S. Dowdy and S. Wearden (1991).

three species of *Drosophila* under increasing levels of organic phosphorus insecticide. Four batches of medium, identical except for the levels of insecticide they contained, were prepared. One hundred eggs from each of three *Drosophila* species were deposited on each of the four medium preparations and the level of insecticide (x) in parts per million (ppm) and number of *Drosophila* flies that survived to adulthood (y) for each combination are recorded in Table 9.7. Test the hypothesis that the three species of *Drosophila* exhibit the same response to increasing levels of insecticide in the medium studied.

32. Among the pieces of information used to assess the age of primates are measurements of skull, muzzle, and long-bone development. Reed (1973) collected such measurements for *Papio cynocephalus* baboons over a period of 5 years and developed a regression relationship between these attributes and age. A portion of Reed's data (from African colonies existing at the Southwest Foundation for Research and Education in San Antonio, Texas) is presented in Table 9.8 for male and female *Papio cynocephalus* baboons. The recorded data are age, in months, and the sum of skull, muzzle, and long-bone measurements, in millimeters.

Use these data to decide whether there is any difference in the slopes defining the linear relationships between age and the sum of skull, muzzle, and long-bone measurements for male and female *Papio cynocephalus* baboons.

GENERAL MULTIPLE LINEAR REGRESSION

9.6 ASYMPTOTICALLY DISTRIBUTION-FREE RANK-BASED TESTS FOR GENERAL MULTIPLE LINEAR REGRESSION (JAECKEL, HETTMANSPERGER–McKEAN)

The statistical procedures discussed in Sections 9.1–9.4 are concerned with the case of a straight-line relationship between a single independent (predictor) variable x and a response random variable Y. In Section 9.5, we presented a test procedure for assessing parallelism of two such straight-line relationships, each with a single independent (predictor) variable. However, in many settings where a regression relationship is of interest there are several independent (predictor) variables that potentially influence the value of a single response random variable. In this section, we present an asymptotically

Table 9.8 Age, in Months, and Sum of Skull, Muzzle, and Long-Bone Measurements, in Millimeters, for Male and Female *Papio cynocephalus* Baboons

	Male		Female	
j	x_{1j} (sum)	Y_{1j} (age)	x_{2j} (sum)	Y_{2j} (age)
1	175.0	1.36	175.0	1.58
2	183.0	2.20	183.0	2.48
3	190.0	3.05	190.0	3.40
4	200.0	4.45	200.0	4.92
5	211.0	6.19	211.0	6.87
6	220.0	7.78	220.0	8.66
7	230.0	9.70	230.0	10.86
8	239.5	11.66	239.5	13.14
9	245.5	12.96	245.5	14.67
10	260.0	16.33	260.0	18.68
11	271.5	19.21	271.5	22.14
12	284.0	22.52	284.0	26.18
13	291.0	24.46	291.0	28.57
14	302.5	27.78	302.5	32.68
15	314.0	31.25	314.0	37.03
16	318.5	32.65	318.5	38.80
17	327.0	35.36	327.0	42.22
18	337.0	38.65		
19	345.5	41.52		
20	360.0	46.61		
21	375.0	52.10		
22	384.5	55.69		
23	397.0	60.55		
24	411.0	66.18		
25	419.5	69.68		
26	428.5	73.47		
27	440.0	78.41		
28	454.5	84.81		

Source: O. M. Reed (1973).

distribution-free rank-based procedure for testing appropriate hypotheses in such a setting, commonly known as *multiple linear regression.*

Data. Let $\mathbf{x}' = (x_1, \ldots, x_p)$ be a row vector of p independent (predictor) variables and let $\mathbf{x}_1' = (x_{11}, \ldots, x_{p1}), \ldots, \mathbf{x}_n' = (x_{1n}, \ldots, x_{pn})$ denote n fixed values of this vector. At each of these fixed vectors $\mathbf{x}_1', \cdots \mathbf{x}_n'$, we observe the value of the single response random variable Y. Thus, we obtain a set of observations $Y_1, \ldots, Y_n$, where Y_i is the value of the response variable when $\mathbf{x}' = \mathbf{x}_i'$.

Assumptions

C1. Our model for multiple linear regression is

$$Y_i = \xi + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i = \xi + \mathbf{x}_i' \boldsymbol{\beta}, \quad i = 1, \ldots, n, \quad (9.50)$$

where $\mathbf{x}_1' = (x_{11}, \ldots, x_{p1}), \ldots, \mathbf{x}_n' = (x_{1n}, \ldots, x_{pn})$ are vectors of known constants, ξ is the unknown "intercept" parameter, and $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_p)$ is a row

vector of unknown parameters, commonly referred to as the *set of regression coefficients*. For convenience later, we also write expression (9.50) in matrix notation. Let $\mathbf{Y}' = (Y_1, \ldots, Y_n), \boldsymbol{\xi}' = (\xi, \ldots, \xi)$, and set

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{21} & \ldots & x_{p1} \\ x_{12} & x_{22} & \ldots & x_{p2} \\ \vdots & \vdots & & \vdots \\ x_{1,n-1} & x_{2,n-1} & \ldots & x_{p,n-1} \\ x_{1n} & x_{2n} & \ldots & x_{pn} \end{bmatrix}. \tag{9.51}$$

Then, using matrix notation, the multiple linear regression model (9.50) can also be written as

$$\mathbf{Y} = \boldsymbol{\xi} + \mathbf{X}\boldsymbol{\beta}. \tag{9.52}$$

C2. The error random variables $e_1, \ldots, e_n$ are a random sample from a continuous distribution that is symmetric about its median 0, has cumulative distribution function $F(\cdot)$, and probability density function $f(\cdot)$ satisfying the mild condition that $\int_{-\infty}^{\infty} f^2(t)dt < \infty$.

Hypothesis

We are interested in testing the null hypothesis that a specific subset $\boldsymbol{\beta}_q$ of the regression parameters $\boldsymbol{\beta}$ are zero. Without loss of generality (because the ordering of the $(x_1, \beta_1), \ldots, (x_p, \beta_p)$ pairs in model (9.50) is arbitrary), we take this subset $\boldsymbol{\beta}_q$ to be the first q components of $\boldsymbol{\beta}$; that is, we take $\boldsymbol{\beta}'_q = (\beta_1, \ldots, \beta_q)$. Thus, we wish to test the null hypothesis

$$H_0 : [\boldsymbol{\beta}'_q = \mathbf{0}; \boldsymbol{\beta}'_{p-q} = (\beta_{q+1}, \ldots, \beta_p) \text{ and } \xi \text{ unspecified}]. \tag{9.53}$$

Thus, the null hypothesis asserts that the independent variables $x_1, \ldots, x_q$ do not play significant roles in determining the value of the dependent variable Y. (In many settings, we are interested in assessing the effect of *all* the independent variables simultaneously, which corresponds to taking $q = p$ in H_0 (9.53). Also see Problem 35.)

Procedure

To compute the Jaeckel–Hettmansperger–McKean test statistic HM, we proceed in several distinct steps. First, we obtain an unrestricted estimator for the vector of regression parameters $\boldsymbol{\beta}$. Let $R_i(\boldsymbol{\beta})$ denote the rank of $Y_i - \mathbf{x}'_i\boldsymbol{\beta}$ among $Y_1 - \mathbf{x}'_1\boldsymbol{\beta}, \ldots, Y_n - \mathbf{x}'_n\boldsymbol{\beta}$, as a function of $\boldsymbol{\beta}$, for $i = 1, \ldots, n$ (See Comment 20). The unrestricted estimator for $\boldsymbol{\beta}$, corresponding to a special case of a class of such estimators proposed by Jaeckel (1972), is then that the value of $\boldsymbol{\beta}$, say, $\hat{\boldsymbol{\beta}}$, that minimizes the measure of dispersion (once again, see Comment 20)

$$D_j(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \sum a(R_i(\boldsymbol{\beta}))(Y_i - \mathbf{x}'_i\boldsymbol{\beta}),$$

where a is a nondecreasing function on the integers $1, 2, \ldots, n$ such that $\Sigma_k a(k) = 0$. Typically, a is written as a score function ϕ on [0,1] by the relation $a(k) = \phi(k/(n+1))$.

The function ϕ is standardized so that $\int \phi = 0$ and $\int \phi^2 = 1$. One such ϕ is the Wilcoxon score function:

$$\phi(x) = \sqrt{12}\left(x - \frac{1}{2}\right), x \in [0, 1].$$

Using this ϕ, the dispersion is

$$D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = (12)^{1/2}(n+1)^{-1}\sum_{i=1}^{n}\left[R_i(\boldsymbol{\beta}) - \frac{n+1}{2}\right](Y_i - \mathbf{x}_i'\boldsymbol{\beta}). \tag{9.54}$$

The estimator $\hat{\boldsymbol{\beta}}$ does not, in general, have a closed-form expression (See Comment 21 for a special case where such a closed-form expression is available), and iterative computer methods are generally necessary to obtain numerical solutions.

The second step in the computation of the Jaeckel–Hettmansperger–McKean test statistic *HM* involves repeating the steps leading to $\hat{\boldsymbol{\beta}}$ except now the minimization of the Jaeckel dispersion measure $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is obtained under the condition imposed by the null hypothesis H_0 (9.53), namely, that $\boldsymbol{\beta}_q = \mathbf{0}$, with $\boldsymbol{\beta}_{p-q}$ unspecified. Let $\hat{\boldsymbol{\beta}}_0$ denote the value of $\boldsymbol{\beta}$ that minimizes $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ in (9.54) under the null constraint that $\boldsymbol{\beta}_q = \mathbf{0}$.

Let $D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ and $D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0)$ denote the overall minimum and the minimum under the null constraint that $\boldsymbol{\beta}_q = \mathbf{0}$, respectively, of the Jaeckel dispersion measure $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ in (9.54) and set

$$D_J^* = D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0) - D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{9.55}$$

We note that D_J^* represents the drop or reduction in Jaeckel dispersion from fitting the full model as opposed to the reduced model corresponding to the null hypothesis H_0 (9.53) constraint that $\boldsymbol{\beta}_q = \mathbf{0}$.

The third and final step in the construction of the Jaeckel–Hettmansperger–McKean test statistic *HM* is the computation of a consistent estimator (See Comment 23) of a scale parameter τ. For the Wilcoxon score,

$$\tau = [12]^{-1/2}\left[\int_{-\infty}^{\infty} f^2(t)dt\right]^{-1}. \tag{9.56}$$

Combining the results of these three construction steps, the Jaeckel–Hettmansperger–McKean test statistic *HM* is given by

$$HM = \frac{2D_J^*}{q\hat{\tau}}. \tag{9.57}$$

When H_0 (9.53) is true, the statistic *HM* has, as n tends to infinity, an asymptotic F distribution with degrees of freedom q and $n - p - 1$, corresponding to the q constraints placed on β under H_0 and the total number p of predictors.

To test

$$H_0 : [\boldsymbol{\beta}_q' = \mathbf{0}; \boldsymbol{\beta}_{p-q}' = (\boldsymbol{\beta}_{q+1}, \ldots, \beta_p) \text{ and } \xi \text{ unspecified}]$$

against the general alternative

$$H_0 : [\boldsymbol{\beta}_q' \neq \mathbf{0}; \boldsymbol{\beta}_{p-q}' = (\beta_{q+1}, \ldots, \beta_p) \text{ and } \xi \text{ unspecified}] \tag{9.58}$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } HM \geq F_{q,n-p-1,\alpha}; \quad \text{otherwise do not reject,} \tag{9.59}$$

where $F_{q,n-p-1,\alpha}$ is the upper α percentile of an F distribution with q and $n-p-1$ degrees of freedom. The values of $F_{q,n-p-1,\alpha}$ may be obtained from the `R` command `qf`.

Instead of an $F_{q,n-p-1,\alpha}$ critical value, one may remove the q from the denominator of the statistics HM and use a $\chi^2_{q,\alpha}$, the upper α percentile of a chi-square distribution. However, Hettmansperger and McKean (1977) and McKean and Sheather (1991) pointed out that the chi-square distribution is often too light tailed for use with small or moderate size samples.

Ties

If there are ties among $Y_1 - \mathbf{x}_1'\boldsymbol{\beta}, \ldots, Y_n - \mathbf{x}_n'\boldsymbol{\beta}$, use average ranks to break the ties in the computation of the minimum $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Similarly, if there are ties among $Y_1 - \mathbf{x}_1'\boldsymbol{\beta}_0 \ldots, Y_n - \mathbf{x}_n'\boldsymbol{\beta}_0$, use average ranks to break the ties in the computation of the minimum $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0)$.

EXAMPLE 9.6 ***Snow Goose Departure Times.***

Wildlife science involves the study of how environmental conditions affect wildlife habits. Freund et al. (2010) report data on such a study to assess how a variety of environmental conditions affect the time that lesser snow geese leave their overnight roost sites to fly to their feeding areas. The data in Table 9.9 represent the following observations collected at a refuge near the Texas coast for 36 days of the 1987–1988 winter season:

TIME(Y) :	minutes before ($-$) or after ($+$) sunrise,
TEMP(x_1) :	air temperature in degrees Celsius,
HUM(x_2) :	relative humidity,
LIGHT(x_3) :	light intensity,
CLOUD(x_4) :	percent cloud cover.

Here, we consider a multiple regression analysis to assess the influence that the environmental conditions temperature (TEMP), relative humidity (HUM), light intensity (LIGHT), and percent cloud cover (CLOUD) have on the departure times (TIME) of lesser snow geese in this region of the country. For illustrative purposes, we consider the following three distinct null hypotheses:

$$H_{01} : [\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0; \xi \text{ unspecified}], \tag{9.60}$$

$$H_{02} : [\beta_1 = \beta_2 = 0; \beta_3, \beta_4, \text{ and } \xi \text{ unspecified}], \tag{9.61}$$

Table 9.9 Environmental Conditions Related to Snow Goose Departure Times

DATE	TIME	TEMP	HUM	LIGHT	CLOUD
11/10/87	11	11	78	12.6	100
11/13/87	2	11	88	10.8	80
11/14/87	−2	11	100	9.7	30
11/15/87	−11	20	83	12.2	50
11/17/87	−5	8	100	14.2	0
11/18/87	2	12	90	10.5	90
11/21/87	−6	6	87	12.5	30
11/22/87	22	18	82	12.9	20
11/23/87	22	19	91	12.3	80
11/25/87	21	21	92	9.4	100
11/30/87	8	10	90	11.7	60
12/05/87	25	18	85	11.8	40
12/14/87	9	20	93	11.1	95
12/18/87	7	14	92	8.3	90
12/24/87	8	19	96	12.0	40
12/26/87	18	13	100	11.3	100
12/27/87	−14	3	96	4.8	100
12/28/87	−21	4	86	6.9	100
12/30/87	−26	3	89	7.1	40
12/31/87	−7	15	93	8.1	95
01/02/88	−15	15	43	6.9	100
01/03/88	−6	6	60	7.6	100
01/05/88	−14	2	92	9.0	60
01/07/88	−8	2	96	7.1	100
01/08/88	−19	0	83	3.9	100
01/10/88	−23	−4	88	8.1	20
01/11/88	−11	−2	80	10.3	10
01/12/88	5	5	80	9.0	95
01/14/88	−23	5	61	5.1	95
01/15/88	−7	8	81	7.4	100
01/16/88	9	15	100	7.9	100
01/20/88	−27	5	51	3.8	0
01/21/88	−24	−1	74	6.3	0
01/22/88	−29	−2	69	6.3	0
01/23/88	−19	3	65	7.8	30
01/24/88	−9	6	73	9.5	30

Source: R. J. Freund, W. J. Wilson and D. Mohr (2010).

and

$$H_{03} : [\beta_2 = 0; \beta_1, \beta_3, \beta_4, \text{ and } \xi \text{ unspecified}]. \tag{9.62}$$

Consider the null hypothesis H_{01} (9.60). To test if all four parameters are 0, the function `rfit` from package `Rfit` (Kloke and McKean, 2011) is used. The data from Table 9.9 is in the `R` data frame `goose`. This data set includes a column of l's labeled `INT` to represent the intercept term ξ if desired. The calls to perform the rank regression are

```
rfit(TIME ~ TEMP + HUM + LIGHT + CLOUD, data=goose)
```

or

```
rfit(TIME ~ INT + TEMP + HUM + LIGHT + CLOUD, data=goose,
    intercept=F)
```

These produce identical results. The default score function is the Wilcoxon score. The estimates for the intercepts ξ and β_1 through β_4 are output as a result of either of the above calls:

```
Coefficients :

                    TEMP         HUM       LIGHT       CLOUD
-51.41229030 1.03912341 0.12628642 2.53480480 0.08951666
```

The estimates for the parameters β_i and ξ may be used to predict the time the geese leave for their feeding area given values x_i of the environmental variables `TEMP`, `HUM`, `LIGHT`, and `CLOUD`. This estimated regression relation is

$$\hat{Y} = -51.41 + 1.04x_1 + 0.13x_2 + 2.53x_3 + 0.09x_4.$$

More details on these parameter estimates may be obtained by the summary command on the rank fit object generated by `rfit`.

```
Coefficients;
       Estimate Std.Error t.value    p.value
     -51.412290  9.159212  -5.6132  4.128e-06
TEMP   1.039123  0.271468   3.8278  0.0006116
HUM    0.126286  0.116753   1.0817  0.2880251
LIGHT  2.534805  0.770792   3.2886  0.0025748
CLOUD  0.089517  0.045066   1.9863  0.0561974.
```

This provides the same parameter estimates for ξ and β_i. In addition, there are individual tests on whether a parameter is zero or not based on asymptotic normality.

It is possible that not all four of the predictor variables, x_i, are needed to model the response variable Y well in this example. Hypotheses tests on subsets of the β_i parameters can be used to choose a suitable model.

To test the hypotheses (9.60), (9.61), and (9.62), we will compare two models to each other: a full model and a reduced model. For hypothesis H_{01} (9.60), we are interested in comparing a full model with all $q = 4$ parameters β_i and the intercept ξ in it to a reduced model with no β_i parameters, only the intercept. We use D_J^*, the difference of the Jaeckel dispersions D_J for each of these models, to perform the comparison. In (9.55) the first term on the right is the Jaeckel dispersion for the reduced model and the second term is the dispersion for the full model. This difference in dispersions is standardized to become the test statistic *HM*.

The reduced model goes with the null hypothesis, the full model is paired with the alternative. For H_{01}, a reduced model with none of the four β_i is fit with a call of

```
r.01 <- rfit(TIME ~ INT, data=goose, intercept=F)
```

and the alternative hypothesis full model with all parameters is modeled by

```
f.01 <- rfit(TIME ~ TEMP + HUM + LIGHT + CLOUD,
        data=goose)
```

The `R` command `drop.test` will perform the test of H_{01} versus the alternative using the statistic *HM*. This command has two arguments: the first is the rank fit for the full model and the second is the rank fit for the reduced model. We compare the rank fits `f.01` and `r.01` with

```
drop.test(f.01, r.01)
```

The output of this call is

```
Drop in Dispersion Test
 F-Statistic     p-value
  1.7708e+01  1.1619e-07
```

The F-statistic is the value of *HM*. The individual components of this statistic given in (9.57) may be viewed directly. If the full versus reduced model comparison is saved to an R object by using the command `h.01 <- drop.test(f.01, r.01)`, then displaying the components of this analysis with `names(h.01)` shows that there are six pieces of information available: `F`, `p.value`, `RD` ("reduction in dispersion"), `tauhat`, `df1`, and `df2`. These values refer, respectively, to the statistic *HM*, the associated upper-tail *P*-value, $D_J^*, \hat{\tau}$, and the numerator and denominator degrees of freedom for the statistic. The *P*-value and *HM* are automatically displayed with `drop.test`. The others may be printed with the `$` indexing convention of R:

```
h.01$RD=294.0261
h.01$tauhat=8.30223
h.01$df1=4
h.01$df2=31
```

Note that the numerator degrees of freedom is $q = 4$ and the denominator degrees of freedom is $n - p - 1 = 36 - 4 - 1$, as expected. The estimate of $\hat{\tau}$ is found using the method of Koul et al. (1987). From (9.57),

$$HM = \frac{2D_J^*}{q\hat{\tau}} = \frac{2 \cdot 294.0261}{4 \cdot 8.30223} = 17.70766$$

agreeing with the output of `drop.test` above. For a particular α, a critical value could be obtained using the R command `qf(alpha, df 1=4, df 2=31, lower.tail=F)`. The *P*-value is obtained with `pf(17.70766, df 1=4, df2=31, lower.tail=F)` or taken from the output of `drop.test`. Given the low *P*-value for this data and hypothesis, we reject H_{01} in favor of the alternative hypothesis that not all of β_1 through β_4 are 0. Due to complexity of minimizing the Jaeckel dispersion measure, the values of the statistics found above may differ slightly when running R under various hardware and software configurations.

To get additional information about potential contributions of some of the individual independent (predictor) variables x_i, we make use of additional hypotheses. First, consider H_{02} (9.61). This null tests if $\beta_1 = \beta_2 = 0$ versus not both are 0. Under both H_{02} and the alternative H_{12}, ξ, β_3, and β_4 are unspecified. The full model is the same in this case as when testing H_{01} so the correct rank fit is again `f.01`. The reduced model is fit with

```
r.02 <- rfit(TIME ~ LIGHT + CLOUD, data=goose)
```

and the test is performed by

```
drop.test(f.01, r.02)
```

The output of this call is

```
Drop in Dispersion Test
 F-Statistic   p-value
   6.6681478 0.0039024
```

This null is rejected, implying that the variables `TEMP` and/or `HUM` contribute significantly (over and above the contributions of `LIGHT` and `CLOUD`) to the determination of

the time that lesser snow geese leave their overnight roost sites to fly to their feeding areas.

Finally, consider the third null hypothesis H_{03} (9.62). This hypothesis states that $\beta_2 = 0$, so the appropriate reduced model is fit by

```
r.03 <- rfit(TIME ~ TEMP + LIGHT + CLOUD, data=goose)
```

and the full model rank fit is again `f.01`. Using `drop. test (f.01, r.03)`, the result is

```
Drop in Dispersion Test
  F-Statistic    p-value
      1.72126   0.19916
```

Based on the large P-value, there is no evidence that the relative humidity `HUM` contributes significantly (over and above the contributions of `TEMP. LIGHT` and `CLOUD`) to the determination of the time that lesser snow geese leave their overnight roost sites to fly to their feeding areas. This does not conflict with the test of the hypothesis H_{02} because the alternative in that case is that not *both* β_1 `(TEMP)` and β_2 `(HUM)` are 0.

Comments

19. *Motivation for the Test.* Use of a measure of dispersion to assess the effectiveness of a model fit to a set of data is common in regression analysis. The estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ are chosen to minimize the Jaeckel dispersion associated with the differences $\mathbf{Y}_i - \mathbf{x}_j'\boldsymbol{\beta}, i = 1, \ldots, n$, under no restrictions on $\boldsymbol{\beta}$ and under the null hypothesis restriction that $\boldsymbol{\beta} = (\boldsymbol{\beta}_q, \boldsymbol{\beta}_{p-q}) = (\mathbf{0}_q, \boldsymbol{\beta}_{p-q})$, respectively. Thus, the numerator of the Jaeckel–Hettmansperger–McKean test statistic $2D_J^* = 2[D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{0}}) - D_J(\mathbf{Y} - \mathbf{X}\hat{\beta})]$ is twice the drop or reduction in the Jaeckel dispersion from fitting the full model as opposed to the reduced null hypothesis model (9.53) with $\boldsymbol{\beta}_q = \mathbf{0}_q$. Large values of this drop in dispersion will lead to large values of *HM* (9.57) and are indicative of lack of agreement between the collected data and the null hypothesis. This serves as partial motivation for procedure (9.59).

20. *Translation lnvariance—"Effect" of the "Intercept" Parameter* ξ. The Jaeckel dispersion measure $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is translation invariant in the sense that it is not affected by the unknown value of the "intercept" parameter, ξ. We note that the rank $R_i(\boldsymbol{\beta})$ of $Y_i - \mathbf{x}_i'\boldsymbol{\beta}$ among $Y_1 - \mathbf{x}_1'\boldsymbol{\beta}$, as a function of $\boldsymbol{\beta}$, is exactly the same as the rank of $Y_i - \xi - \mathbf{x}_i'\boldsymbol{\beta}$ among $Y_1 - \xi - \mathbf{x}_1'\boldsymbol{\beta}, \ldots Y_n - \xi - \mathbf{x}_n'\boldsymbol{\beta}$, for $i = 1, \ldots, n$. It follows that the Jaeckel measure of dispersion is independent of the value of the intercept parameter ξ, because

$$D_J(\mathbf{Y} - \xi - \mathbf{X}\boldsymbol{\beta}) = (12)^{1/2}(n+1)^{-1}\sum_{i=1}^{n}\left[R_i(\boldsymbol{\beta}) - \frac{n+1}{2}\right](Y_i - \xi - \mathbf{x}_i'\boldsymbol{\beta})$$

$$= D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \xi[(12)^{1/2}(n+1)^{-1}]\sum_{i=1}^{n}\left[R_i(\boldsymbol{\beta}) - \frac{n+1}{2}\right]$$

$$= D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \text{because} \sum_{i=1}^{n}\left[R_i(\boldsymbol{\beta}) - \frac{n+1}{2}\right] = 0.$$

21. *Closed-Form Expression for $\hat{\boldsymbol{\beta}}$-Special Case of Straight-Line Regression.* One situation where the unrestricted estimator $\hat{\boldsymbol{\beta}}$ minimizing $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ in (9.54) has a closed-form expression is when we have only a single independent (predictor) variable so that model (9.50) corresponds to a straight-line regression $Y_i = \xi + \beta_1 x_i + e_i, i = 1, \ldots, n$. For this setting, the Jaeckel (1972) estimator of the slope β_1 is a weighted median of the set of all pairwise slopes $S_{ij} = (Y_j - Y_i)/(x_j - x_i)$, for (i, j) such that $x_i \neq x_j$. This particular estimator was first derived using different criteria and studied by Adichie (1967).

22. *Estimation of the "Intercept" ξ.* Because the Jaeckel dispersion measure $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is independent of the unknown value of the "intercept" ξ (See Comment 20), the estimator $\hat{\boldsymbol{\beta}}$ obtained by minimizing $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ does not provide any information relative to ξ. Hettmansperger and McKean (1977) suggested using the full-model residuals $e_i^* = Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$, for $i = 1, \ldots, n$, to estimate ξ. In particular, they proposed the estimator

$$\hat{\xi} = \text{median} \left\{ \frac{e_i^* + e_j^*}{2}, 1 \leq i \leq j \leq n \right\}. \tag{9.63}$$

(We note, in passing, that $\hat{\xi}$ is simply the Hodges–Lehmann one-sample estimator $\hat{\theta}$ (3.23) applied to the full-model residuals $e_1^*, \ldots, e_n^*$.)

23. *Estimation of the Parameter τ (9.56).* Part of the construction of the Jaeckel–Hettmansperger–McKean test statistic, *HM* (9.57), is the computation of a consistent estimator of the parameter τ (9.56). A variety of approaches leading to a number of competing consistent estimators have been considered in the literature. Hettmansperger and McKean (1977) suggested a consistent estimator for τ based on the length of a Wilcoxon signed rank confidence interval (See Section 3.3) applied to the full-model residuals $e_1^*, \ldots, e_n^*$ discussed in Comment 22. Koul, Sievers, and McKean (1987) recommended an estimator of τ based on the empirical distribution function of the absolute differences of the full-model residuals $e_1^*, \ldots, e_n^*$. An approach to the estimation of τ based on window- or kernel-type estimation of the probability density function $f(\cdot)$ has been considered by Schuster (1974) and Schweder (1975, 1981).

24. *Generalized Score Functions.* The rank regression procedure discussed in this section is based on the use of the Wilcoxon-type scoring function of the ranks in the construction of the Jaeckel dispersion function $D_J(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ in (9.54). Other scoring functions, such as $\boldsymbol{\Phi}^{-1}(\cdot)$ associated with the van der Waerden test discussed in Comment 4.12, were also considered by Jaeckel (1972) in the construction and study of an entire class of rank-based dispersion measures for the multiple linear regression setting.

25. *Test for a More General Null Hypothesis.* Hettmansperger, McKean, and Sheather (1997) described a generalization of the null hypothesis presented in H_0 (9.53). They discussed statistical procedures for the more inclusive problem of testing $H_0^* : \mathbf{M}\boldsymbol{\beta} = 0$ versus the general alternative $H_A^* : \mathbf{M}\boldsymbol{\beta} \neq \mathbf{0}$, where $\mathbf{M}$ is an arbitrary full row rank $q \times p$ matrix, for some $q \leq p$.

    As an example, consider the null hypothesis $H_{04} : [\beta_1 = \beta_2; \beta_3, \beta_4,$ and ξ unspecified] against the alternative $H_{14} : [\beta_1 \neq \beta_2; \beta_3, \beta_4,$ and ξ unspecified] for the snow geese data from Example 9.6. This is equivalent to setting $\boldsymbol{M}$ to the

single row matrix $[1 \ -1 \ 0 \ 0]$ so that $\mathbf{M}\boldsymbol{\beta} = \mathbf{0}$ is the same as H_{04}. In R, we again use `drop.test` with the full model `f.01` from Example 9.6 and the reduced model from

```
r.04 <- rfit(TIME ~ I(TEMP + HUM) + LIGHT + CLOUD,
        data=goose)
```

where the `I` command uses the + symbol to force `TEMP` and `HUM` to be considered as a sum, rather than using the + symbol to represent additional terms in the rank regression model. The estimated rank regression equation for the reduced model is

```
Coefficients :
                I (TEMP + HUM)     LIGHT      CLOUD
  -65.6185750       0.2366911 3.3860983 0.1346024
```

and the test of the hypotheses H_{04} against H_{14} is given by

```
Drop in Dispersion Test
 F-Statistic   p-value
   8.4534141 0.0066766
```

Thus, there is good evidence that the independent variables temperature and relative humidity do not contribute in the same degree to the determination of the time that lesser snow geese leave their overnight roost sites to fly to their feeding areas. This finding is in good agreement with what had been established previously in Example 9.6.

26. *Extension to the General Linear Model.* Our discussion of rank-based regression in this section has only touched upon a small portion of a much more extensive rank-based approach to the large class of linear models. Although a discussion of this more general setting is beyond the scope of this text, we recommend that the interested reader take advantage of two excellent survey articles on this topic by Draper (1988) and Hettmansperger, McKean, and Sheather (1997).

Properties

1. *Consistency.* Under certain regularity conditions (see, for example, Hettmansperger, McKean, and Sheather (1997)), the test defined by (9.59) is consistent against the alternatives H_1 (9.58).
2. *Asymptotic Chi-Squareness.* See McKean and Hettmansperger (1976).
3. *Efficiency.* See McKean and Hettmansperger (1976) and Section 9.8.

Problems

33. In heart catheterization, a 3-mm-diameter Teflon catheter (tube) is inserted into a major vein or artery at the femoral region and maneuvered up into the heart itself to assess the heart's physiology and functional ability. Heart catheterizations are sometimes performed on children with congenital heart defects. In such cases, the length of the catheter is often determined by a physician's educated guess. Rice (2007) considered a data set obtained by Weindling (1977) in a preliminary study involving 12 children. For each child, the exact catheter length required was determined by using a fluoroscope to check that the tip of the catheter had reached the pulmonary artery. The 12 catheter lengths (cm) and the heights (in) and weights (lb) for the 12 children in the study are given in Table 9.10.

Table 9.10 Required Length of Heart Catheter as a Function of Height and Weight

Child	Height, in	Weight, lb	Length of heart catheter, cm
1	42.8	40.0	37.0
2	63.5	93.5	49.5
3	37.5	35.5	34.5
4	39.5	30.0	36.0
5	45.5	52.0	43.0
6	38.5	17.0	28.0
7	43.0	38.5	37.0
8	22.5	8.5	20.0
9	37.0	33.0	33.5
10	23.5	9.5	30.5
11	33.0	21.0	38.5
12	58.0	79.0	47.0

Source: J. A. Rice (2007).

Treating length of heart catheter as the independent variable, test for the importance of height and weight in determining the required catheter length.

34. Iman (1994) considered data obtained by Leaf et al. (1989) in a study of options for reducing concentrations of total plasma triglycerides. Leaf et al. obtained measurements of the following variables on each of 13 patients:

Y: Total triglyceride level, mmol/l,
x_1: Sex of the patient (coded as female = 0, male = 1),
x_2: Whether patient is obese (coded as no = 0, yes = 1),
x_3: Chylo-microns,
x_4: Very low density lipoprotein (VLDL),
x_5: Low density lipoprotein (LDL),
x_6: High density lipoprotein (HDL),
x_7: Age of the patient.

These data are presented in Table 9.11.

(a) Including all the measured variables, find the approximate P-value for a test of the null hypothesis that obesity does not play a significant role in determination of the total triglyceride level.
(b) Including all the measured variables, find the approximate P-value for a test of the null hypothesis that none of the lipoproteins play significant roles in determination of the total triglyceride level.
(c) Does age play a significant role in the determination of the total triglyceride level, when all the measured variables are taken into account? Justify your answer.
(d) Find an approximate P-value for a test of the null hypothesis that none of the measured variables contribute significantly to the determination of the total triglyceride level.

35. Consider the multiple linear regression model in (9.50). Often it is the case that we are interested in testing whether *any* of the independent variables $x_1, \ldots, x_p$ have significant effects on the determination of the value of the dependent random variable Y. This corresponds to taking $q = p$ in the statement of the null hypothesis (9.53). For this setting, what would be the form of the null constraint estimator $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$? Provide a closed-form expression for $D_J(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{0}})$ for this setting in terms of the ordered Y observations, $Y^{(1)} \leq \cdots \leq Y^{(n)}$.

36. Freund et al. (2010) presented a set of data relating survival times (TIME) of liver transplant patients to the following information collected from the patients prior to their transplant operations:

Table 9.11 Blood Plasma Measurements Related to Total Triglyceride Level

Patient	Total triglyceride level	Sex/Obese	Chylo-microns	VLDL	LDL	HDL	Age
1	20.19	1/1	3.11	4.51	2.05	0.67	53
2	27.00	0/1	4.90	6.03	0.67	0.65	51
3	51.75	0/0	5.72	7.98	0.96	0.60	54
4	51.36	0/1	7.82	9.58	1.06	0.42	56
5	28.98	1/1	2.62	7.54	1.42	0.36	66
6	21.70	0/1	1.48	3.96	1.09	0.23	37
7	14.40	1/1	0.57	8.60	2.16	0.83	41
8	15.14	1/1	0.60	5.46	1.58	0.85	55
9	50.00	1/1	6.29	13.03	1.48	0.28	43
10	23.73	1/1	1.94	7.12	0.91	0.57	58
11	29.33	0/1	0.52	8.94	1.58	0.88	39
12	19.98	0/1	1.11	5.85	1.19	0.62	41
13	13.28	1/0	1.61	3.73	1.58	0.62	54

Source: D. A. Leaf, W. E. Connor, R. Illingworth, S. P. Bacon, and G. Sexton (1989) and R. L. Iman (1994).

CLOT: a measure of the clotting potential of the patient's blood

PROG: a subjective index of the patient's prospect of recovery

ENZ: a measure of a protein present in the body

LIV: a measure relating to white blood cell count.

These data for 54 liver transplant patients are presented in Table 9.12. Examine the relationship of survival time (TIME) to the four measured preoperation variables. Which of them provide significant input into the determination of survival time for liver transplant patients?

37. In Section 9.1, we discussed a procedure designed to test the effect of a single independent (predictor) variable x on a dependent random variable Y when the anticipated relationship between x and Y is linear. Sometimes, the anticipated relationship between x and Y is better described by a higher order polynomial in x, rather than a simple linear relationship. Discuss how the general procedure presented in this section can be used to test for a relationship between x and Y that is best described by a polynomial of degree $p > 1$.

38. Consider the cenosphere-resin composite data of Problem 1. In that problem, you were asked to assess the significance of a possible linear relationship between hydrostatic pressure, x, and the density of the cenosphere-resin composite, Y. Suppose that someone suggested that the relationship between x and Y might be better represented by a cubic polynomial through the expression

$$E[Y|x] = \xi + \beta_1 x + \beta_2 x^2 + \beta_3 x^3,$$

where ξ, β_1, β_2, and β_3 are unknown parameters. (See Problem 37.)

(a) Find the approximate P-value for an appropriate test of the null hypothesis that there is no (cubic, quadratic, or linear) significant relationship between x and Y.

(b) Find the approximate P-value for an appropriate test of the null hypothesis that the relationship between x and Y is actually quadratic, as opposed to cubic.

(c) Find the approximate P-value for an appropriate test of the null hypothesis that the relationship between x and Y is actually linear, as opposed to either quadratic or cubic.

39. Hettmansperger, McKean, and Sheather (1997) described the following generalization of the null hypothesis presented in H_0 (9.53). They discussed statistical procedures for the more general problem of testing $H_0^* : \mathbf{M}\boldsymbol{\beta} = \mathbf{0}$ versus the general alternative $H_A^* : \mathbf{M}\boldsymbol{\beta} \neq \mathbf{0}$, where

Table 9.12 Survival Times of Liver Transplant Patients and Related Biological Measurements

Patient	TIME	CLOT	PROG	ENZ	LIV
1	34	3.7	51	41	1.55
2	58	8.7	45	23	2.52
3	65	6.7	51	43	1.86
4	70	6.7	26	68	2.10
5	71	3.2	64	65	0.74
6	72	5.2	54	56	2.71
7	75	3.6	28	99	1.30
8	80	5.8	38	72	1.42
9	80	5.7	46	63	1.91
10	87	6.0	85	28	2.98
11	95	5.2	49	72	1.84
12	101	5.1	59	66	1.70
13	101	6.5	73	41	2.01
14	109	5.2	52	76	2.85
15	115	5.4	58	70	2.64
16	116	5.0	59	73	3.50
17	118	2.6	74	86	2.05
18	120	4.3	8	119	2.85
19	123	6.5	40	84	3.00
20	124	6.6	77	46	1.95
21	125	6.4	85	40	1.21
22	127	3.7	68	81	2.57
23	136	3.4	83	53	1.l2
24	144	5.8	61	73	3.50
25	148	5.4	52	88	1.81
26	151	4.8	61	76	2.45
27	153	6.5	56	77	2.85
28	158	5.1	67	77	2.86
29	168	7.7	62	67	3.40
30	172	5.6	57	87	3.02
31	178	5.8	76	59	2.58
32	181	5.2	52	86	2.45
33	184	5.3	51	99	2.60
34	191	3.4	77	93	1.48
35	198	6.4	59	85	2.33
36	200	6.7	62	81	2.59
37	202	6.0	67	93	2.50
38	203	3.7	76	94	2.40
39	204	7.4	57	83	2.16
40	215	7.3	68	74	3.56
41	217	7.4	74	68	2.40
42	220	5.8	67	86	3.40
43	276	6.3	59	100	2.95
44	295	5.8	72	93	3.30
45	310	3.9	82	103	4.55
46	311	4.5	73	106	3.05
47	313	8.8	78	72	3.20
48	329	6.3	84	83	4.13
49	330	5.8	83	88	3.95
50	398	4.8	86	101	4.10
51	483	8.8	86	88	6.40
52	509	7.8	65	115	4.30
53	574	11.2	76	90	5.59
54	830	5.8	96	114	3.95

Source: R. J. Freund, W. J. Wilson and D. Mohr (2010).

M is an arbitrary full row rank $q \times p$ matrix, for some $q \leq p$ (See Comment 25). Within this more general setting, what matrix **M** corresponds to the special case of the null hypothesis H_0 in (9.53)?

40. In an attempt to gain a better understanding of the complexities of air pollution in general and to predict pollutant levels in particular, the Los Angeles Pollution Control District routinely records the levels of pollutants and several meteorological conditions at various sites around the city. As reported by Rice (2007), the data in Table 9.13 represent the maximum level of an oxidant (a photochemical pollutant) and the morning averages of the four meteorological variables: wind speed, temperature, humidity, and insolation (measure of amount of sunlight) over a 30-day period in a single summer.

Ignoring the distinct possibility that there is some degree of correlation between maximum oxidant levels collected on adjacent days (which would violate Assumption C3 regarding the independence of the observations of the dependent variable), examine the relationship of oxidant level to the four meteorological variables. Which of them contribute significantly to the maximum oxidant level on a given day for the Los Angeles Pollution Control District?

41. For the data discussed in Problem 40, consider a multiple linear regression of the maximum oxidant level on the four meteorological measurements. Find the approximate P-value for

Table 9.13 Maximum Oxidant Level, Wind Speed, Temperature, Humidity, and Insolation for a 30-Day Summer Period in the Los Angeles Pollution Control District

Day	Oxidant level	Wind speed	Temperature	Humidity	Insolation
1	15	50	77	67	78
2	20	47	80	66	77
3	13	57	75	77	73
4	21	38	72	73	69
5	12	52	71	75	78
6	12	57	74	75	80
7	12	53	78	64	75
8	11	62	82	59	78
9	12	52	82	60	75
10	20	42	82	62	58
11	11	47	82	59	76
12	17	40	80	66	76
13	20	42	81	68	71
14	23	40	85	62	74
15	17	48	82	70	73
16	16	50	79	66	72
17	10	55	72	63	69
18	11	52	72	61	57
19	11	48	76	60	74
20	9	52	77	59	72
21	5	52	73	58	67
22	5	48	68	63	30
23	4	65	67	65	23
24	7	53	71	53	72
25	18	36	75	54	78
26	17	45	81	44	81
27	23	43	84	46	78
28	23	42	83	43	78
29	24	35	87	44	77
30	25	43	92	35	79

Source: J. A. Rice (2007).

Table 9.14 Number of *Chaoborus* Larvae and Water Quality of Samples

Sample	Number of larvae	Depth	Brackishness	Dissolved oxygen
1	35	8.4	8.0	1.0
2	10	2.0	6.5	8.5
3	9	3.5	6.2	6.5
4	30	10.4	5.0	1.5
5	20	6.5	6.5	7.5
6	23	6.2	7.3	4.5
7	28	12.4	6.4	4.0
8	8	7.0	6.0	10.0
9	29	5.8	6.1	3.0
10	4	3.0	5.4	11.0
11	18	6.0	7.3	4.5
12	14	5.5	6.6	5.5
13	32	9.0	6.5	2.5
14	6	1.1	5.8	7.0

Source: S. Dowdy and S. Wearden (1991).

an appropriate test of the null hypothesis that the regression coefficients for wind speed and humidity are the same, as are the regression coefficients for temperature and insolation. (See Comment 25.)

42. Dowdy and Wearden (1991) considered the relationship between several environmental factors and the number of larvae of the phantom midge, genus *Chaoborus*, which is similar to a mosquito in appearance, but is not bloodsucking. The larva burrows into the sediment at the bottom of a body of water and remains there during the daylight hours. At night, it migrates to the surface of the water to feed. The larva is itself eaten by larger animals and therefore plays an important role in the food chain for freshwater fish. A team of biologists studied a recreational lake created by damming a small stream and recorded the following measurements at each of 14 sampling points in the lake:

Y: number of larvae of *Chaoborus* collected in a grab sample of the sediment from an area of approximately 225 cm^2 of lake bottom.

X_1: depth (meters) of the lake at the sampling point.

x_2: brackishness (conductivity) of the water at the lake bottom (recorded in mhos per decimeter).

x_3: dissolved oxygen (milligrams per liter) in the water sampled from the lake bottom.

The data from these 14 sampling points are presented in Table 9.14. Examine the relationship of the number of *Chaoborus* larvae to the three measured water quality variables. Which of them provide significant input into the determination of the number of *Chaoborus* larvae in a lake environment?

NONPARAMETRIC REGRESSION ANALYSIS

9.7 AN INTRODUCTION TO NON-RANK-BASED APPROACHES TO NONPARAMETRIC REGRESSION ANALYSIS

In all the previous sections of this chapter, the *modus operandi* has been to consider a specific regression model with associated parameters (e.g., straight line, two straight

lines, multiple linear regression) and then to discuss appropriate rank-based procedures for making statistical inferences about the unknown parameters. They have all been nonparametric in nature, in that the inferential procedures were not dependent upon the assumption of a particular underlying distribution for the error terms. Recently, however, there has been considerable research activity in the literature in an arena that has become known generally as *nonparametric regression.* Although it maintains an indifference to the form of the underlying distribution for the error terms, the distinction in this new area of endeavor is that even a specific regression model is no longer stipulated a priori. The data are asked to provide not only the eventual statistical inference but also aid with the development of an appropriate regression relationship between the dependent random variable and the independent predictor variables(s). Thus, the intent of these nonparametric regression procedures is to permit the data to aid in both the selection of an appropriate model for the regression relationship and the inferences eventually drawn from this model.

All the procedures previously discussed in this chapter have also been rank-based, in the sense that some form of ranking was used in arriving at the appropriate inferences. When the model itself is open for data input, however, ranks are no longer sufficient to provide both model selection and inferential procedures. Hence, the procedures associated with this field known as *nonparametric regression* do not generally utilize rankings in reaching their conclusions. As a result, they are often more complicated and computationally intensive than the level assumed throughout the rest of this text. Consequently, our approach in this section will be to discuss briefly some of the statistical techniques that are commonly used in developing such nonparametric regression procedures rather than to provide details of specific procedures and their applications to appropriate data sets. More detailed summaries of various aspects of this general area of nonparametric regression are provided in Chapters 13 and 14 and, for example, in Cleveland (1994), Eubank (1999), Hastie and Tibshirani (1990), Wasserman (2006), and Ryan (2009).

We concentrate here on the setting where we are interested in obtaining information about the relationship between a single dependent random variable Y and a single independent (predictor) variable x. For available procedures in the area of nonparametric regression when there are multiple independent variables, the reader is referred to work by Friedman (1991) and Stone (1994), for example.

Data. At each of n fixed values, $x_1, \ldots, x_n$, of the independent (predictor) variable x, we observe the value of the response random variable Y. Thus, we obtain a set of observations $Y_1, \ldots, Y_n$, where Y_i is the value of the response variable when $x = x_i$.

Assumptions

The most general nonparametric regression relationship between Y_i and x_i is given by

$$Y_i = \mu(x_i) + e_i, \quad \text{for} \quad i = 1, \ldots, n, \tag{9.64}$$

where the random variables $e_1, \ldots, e_n$ are a random sample from a continuous population that has median 0.

The goal, of course, is to make valid statistical inferences about the form of the regression function $\mu(\cdot)$. Depending on the specific approach to nonparametric regression under consideration, a variety of additional *regularity* conditions are often imposed on the form of $\mu(\cdot)$ to enable development of appropriate statistical methodology.

As there is likely to be a good deal of fuzziness (variability) in the response data $Y_1, \ldots, Y_n$, it is often difficult to describe the relationship between x and Y, as expressed in the median $\mu(x)$, without the aid of a more formal model. Therefore, we search for ways to dampen, or "smooth," the fluctuations present in the Y observations as we move along the various x values. In this section, we discuss a variety of ways to approach this smoothing of the data. Ryan (2009) referred to each of these smoothing techniques as a "smoother" and to the associated estimates $\hat{\mu}(x_i)$ at the nx_i values as a "smooth." The first four smoothers discussed in this section are linear smoothers, in the sense that the estimates $\hat{\mu}(x_1), \ldots, \hat{\mu}(x_n)$ in a particular smooth are always linear combinations of the observations $Y_1, \ldots, Y_n$. The fifth smoother is nonlinear.

Running Line Smoother. One of the earliest attempts at nonparametric regression is associated with the running line smoother proposed and studied by Cleveland (1979). For this smoother, a moving window of points is utilized and a simple least squares linear regression line is computed each time a point is deleted and another added as the window moves along the x values. The plot of these running lines as a function of the independent variable x is referred to as the *running line smoother estimator* for $\mu(x)$.

A number of issues are important relative to the construction of running line smoothers. First, one must decide on how many points are to be used in each window (i.e., the window *size* or *size of the neighborhood)* for which the least squares regression line is to be computed. Clearly, a window size that is too small will result in very little smoothing of the data, whereas a window size that is too large will virtually force a single straight-line relationship on the data, regardless of its validity. This choice of window size is discussed in Hastie and Tibshirani (1987), where they indicate that a window size of roughly 10–15% of the data is reasonable. Another matter of concern with running line smoothers is how to deal with the extremes of the data, where symmetric windows are not possible. Statistical inference about $\hat{\mu}(x)$ associated with running line smoothers is addressed in Hastie and Tibshirani (1990).

Kernel Regression Smoother. As with the running line smoother, the kernel regression smoother utilizes *neighborhood data* to provide its estimate of the regression function $\mu(x)$. In this setting, the neighborhoods are often referred to as *strips* and the size of a strip is called the *bandwidth*. One of the clear distinctions between running line smoothers and kernel regression smoothers is in how they weight the observations in a given window. For a running line smoother, the points in a neighborhood are equally weighted, although, of course, they could have differing influences on the estimation process. On the other hand, for a kernel regression smoother, the distance of the points in a neighborhood from the center of a neighborhood, say, x_0, is used to differentially weight their contributions. Basically, in the process of estimating $\mu(x_0)$, no weight is given to those observations outside of the neighborhood centered at x_0 and the greatest weight in the neighborhood is given to those observations Y_i for which the corresponding x_i are closest to x_0. A *kernel function* is utilized to assign these differential weights to the observations across the various neighborhoods.

Altman (1992) addressed the question of how many strips to use in constructing a kernel regression smoother, as well as some related procedures for statistical inference. The selection of a kernel function and its relationship to the stipulation of both the number of strips and the bandwidth is discussed in Hastie and Tibshirani (1990) and Härdle (1992). One particular shortcoming of kernel regression smoothers is that their performance at the boundaries of the predictor region can be rather poor, as documented by Hastie and Loader (1993) and Fan and Marron (1993).

Local Regression Smoother. Local regression smoothers were first introduced by Cleveland (1979), where he referred to the process as *locally weighted regression*. Local regression smoothers once again use overlapping neighborhoods and, as with the kernel regression smoothers, weight the contributions of points to the estimation of $\mu(x_0)$ in an inverse relationship to their distances from x_0. The estimation in a particular neighborhood is thus like a local weighted least squares fit.

Robust versions of local regression smoothers, which downweight large residuals, have also been proposed (see, e.g., Cleveland (1994) and Cleveland, Grosse, and Shyu (1992)) for the setting where the random errors have a symmetric distribution. Computational methods for local regression smoothers are presented in Cleveland and Grosse (1991). Approaches to statistical inferences for $\mu(x_0)$, as well as diagnostic checks associated with local regression smoothers, are discussed in Cleveland, Grosse, and Shyu (1992).

Running line, kernel regression, and local regression smoothers are discussed in more detail in Chapter 14.

Spline Regression Smoothers. A *spline* is a curve pieced together from a number of individually constructed curve/line segments; that is, a spline is simply a piecewise polynomial. (Smith (1979) and Eubank (1999) provided nice discussions of this general concept.) When each segment of a spline contains only linear terms, it is called a *linear spline*.

The application of splines to regression problems in a general sense is discussed in Wegman and Wright (1983), where they refer to splines associated with a regression model as *regression splines*. The simplest of these regression spline smoothers are those associated with linear splines. The junctures where these lines are pieced together are known as *knots.* When the positions of these knots are known a priori, the use of linear regression spline smoothers is relatively straightforward. However, when the positions of the knots are also unknown, the problem becomes a good deal more complicated. Applications of higher order polynomial splines (in particular, quadratic and cubic splines) are discussed in Eubank (1999).

A different approach to the use of splines in regression problems is associated with the development of *smoothing splines.* For these procedures, the regression smooth results from minimization of a sum of squares augmented by a smoothing term related to the order of the desired smoothing spline. For further information on smoothing splines, the reader is referred to Eubank (1999) and Wahba (1990).

Wavelet Smoother. This smoother represents the observed data with a set of basis functions and their corresponding coefficients. Wavelet functions are commonly used as a basis because they are able to model data sampled from very general relations between Y and x. The wavelet estimate is not linear. The coefficients are modified nonlinearly by thresholding rules which generally set some of the coefficients to 0 and shrink the remaining toward 0. The estimate of μ is found with these modified coefficients, not the original values.

Donoho and Johnstone (1994, 1995) created popular threshold methods for wavelets called *VisuShrink and SureShrink*. Under the assumptions of normal errors, they showed that the asymptotic rates of convergence for these wavelet smoothers are optimal or near optimal. New methods of thresholding with improved estimation properties have been derived. For example, Cai (1999) collapsed groups of coefficients together in order to attain optimal convergence rates with improved visual smoothness of the estimate. Others have extended wavelet smoothers to variable designs for the x_i (see, e.g., Kovac

and Silverman (2000)) and non-normality of errors (e.g., Nason (1996)). Chapter 13 discusses wavelet smoothing in greater detail.

General Discussion. As mentioned previously, all the nonparametric regression procedures discussed in this section are considerably more computationally intensive than the material presented elsewhere in the text. As a result, computer software is essential for the implementation of these nonparametric regression smoothers. Such software is available in R.

Finally, we note that in determining which of these approaches to nonparametric regression is most appropriate for a given problem, the decision invariably comes down to the relative importance of minimizing bias versus minimizing variance. All these smoothers produce biased estimators for the regression function $\mu(x)$ so that the desired trade-off between the sizes of the variance and bias (along with computational capabilities, of course) often strongly influences the choice of a particular nonparametric regression smoother.

9.8 EFFICIENCIES OF REGRESSION PROCEDURES

The asymptotic relative efficiencies of the Theil procedures of Sections 9.1–9.3 with respect to their normal theory counterparts based on the least squares estimator of β have been found by Sen (1968) to be given by the expression

$$e_F = \varepsilon^2 \left[12\sigma_F^2 \left\{ \int_{-\infty}^{\infty} f^2(u)du \right\}^2 \right] = \varepsilon^2 e_F^*, \tag{9.65}$$

where σ_F^2 is the variance of the common underlying (continuous) distribution $F(\cdot)$ for the random variables $e_1, \ldots, e_n$ in (9.1), $f(\cdot)$ is the probability density function corresponding to F, and ϵ^2 is the limiting value (n tending to infinity) of ϵ_n^2, where ϵ_n is the product moment correlation coefficient between $(x_1, \ldots, x_n)$ and $(1, \ldots, n)$ as given in expression (6.2) of Sen (1968). The parameter $\int_{-\infty}^{\infty} f^2(u)du$ is the area under the curve associated with $f^2(\cdot)$, the square of the common probability density function. We note that the expression e_F (9.65) is simply ϵ^2 times the corresponding Pitman efficiencies (e_F^*) in the one-sample, two-sample, and k-sample location settings (See Sections 3.11, 4.5, and 6.10).

We note that ϵ_n clearly depends on the design configuration for the values $(x_1, \ldots, x_n)$ of the independent variable. An important special case where $\epsilon_n = 1$ is the equally spaced, no-replications design, where $x_i = x_1 + (i-1)a$, for some $a > 0$ and $i = 1, \ldots, n$. When $\epsilon^2 = 1$, values of e_F (9.65) correspond to e_F^* and can be obtained from display (3.116).

The asymptotic relative efficiency under a sequence of near alternatives of the Sen–Adichie parallelism test based on V (9.45) with respect to the corresponding normal theory procedure based on least squares estimators was found to be e_F^* (9.65) by Sen (1969). The asymptotic relative efficiency under a sequence of contiguous alternatives of the Jaeckel-Hettmansperger-McKean rank-based multiple linear regression test based on HM (9.57) with respect to the corresponding least squares competitor test procedure was found by McKean and Hettmansperger (1976) to be e_F^* (9.65) as well. Once again, the values of e_F^* can be found in display (3.116).

Chapter 10

Comparing Two Success Probabilities

INTRODUCTION

In Chapter 2, we described inferential procedures for a single success probability. These procedures are based on the proportion of successes observed in n independent Bernoulli trials. Recall that each observation could be classified a success or failure (depending on whether or not a specified attribute was present). In this chapter, the object is to compare two unknown success probabilities, p_1, p_2, on the basis of the corresponding rates of success in independent samples.

Section 10.1 describes approximate tests and confidence intervals for $p_1 - p_2$. The tests and confidence intervals are based on the large-sample approximations. Section 10.2 presents Fisher's exact (conditional) test. In Section 10.3, we introduce the odds ratio and present inferential procedures (tests, estimators, and confidence intervals) for the odds ratio. Section 10.4 describes tests, estimators, and confidence intervals for analyzing k strata of 2×2 tables. Section 10.5 considers the efficiency properties.

Data. We observe the outcomes of $n_{1\cdot}$ independent repeated Bernoulli trials, each with success probability p_1. We also observe the outcomes of $n_{2\cdot}$ independent repeated Bernoulli trials, each with success probability p_2. The data are represented in Table 10.1.

Assumptions

In Section 10.1, we make the following assumptions.

- **A1.** $\mathcal{O}_{11}$ is the number of successes observed in $n_{1\cdot}$ independent Bernoulli trials, each with success probability p_1.
- **A2.** $\mathcal{O}_{21}$ is the number of successes observed in $n_{2\cdot}$ independent Bernoulli trials, each with success probability p_2.
- **A3.** The Bernoulli trials corresponding to sample 1 are independent of the Bernoulli trials corresponding to sample 2.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

Table 10.1 2×2 Table of Outcomes

	Successes	Failures	Totals
Sample 1	$\mathcal{O}_{11}$	$\mathcal{O}_{12}$	$n_{1.}$
Sample 2	$\mathcal{O}_{21}$	$\mathcal{O}_{22}$	$n_{2.}$
Totals:	$n_{.1}$	$n_{.2}$	$n_{..}$

(10.1)

The hypothesis of interest in Section 10.1 is

$$H_0 : p_1 = p_2 = p \tag{10.2}$$

with the common value p being unspecified.

An exact conditional test of H_0 is achievable but computationally difficult, as follows. The random quantity

$$D = \frac{\mathcal{O}_{11}}{n_{1.}} - \frac{\mathcal{O}_{21}}{n_{2.}} \tag{10.3}$$

is an estimator of $p_1 - p_2$. Suppose, for our data, the observed value of D is D_{obs}. Using the independent binomial distributions of $\mathcal{O}_{11}$ and $\mathcal{O}_{21}$, one can, for a specified value of p, compute $P_p(D \geq D_{\text{obs}})$. Then define the P-value to be

$$P = \max_{0 \leq p \leq 1} P_p(D \geq D_{\text{obs}}).$$

This computationally intensive unconditional test is due to Barnard (1945), but later its originator and others have preferred Fisher's exact conditional test. See Barnard (1945, 1947), Suissa and Shuster (1985), and Haber (1986, 1987) for more on Barnard's unconditional test and Fisher's exact test. See Mehta and Hilton (1993) for a discussion of conditional versus unconditional tests and the computational difficulties incurred by unconditional tests, especially for contingency tables beyond the 2×2 case.

In this chapter, we defer Fisher's exact test to Section 10.2. In Section 10.1, we present approximate tests and confidence intervals for $p_1 - p_2$. These approximate procedures are based on the large-sample approximations.

10.1 APPROXIMATE TESTS AND CONFIDENCE INTERVALS FOR THE DIFFERENCE BETWEEN TWO SUCCESS PROBABILITIES (PEARSON)

To test H_0, given by (10.2), we use the following large-sample tests.

Large-Sample Test Procedures

The test statistic is a suitably standardized version of $\hat{p}_1 - \hat{p}_2$, where

$$\hat{p}_1 = \frac{\mathcal{O}_{11}}{n_{1.}}, \quad \hat{p}_2 = \frac{\mathcal{O}_{21}}{n_{2.}}. \tag{10.4}$$

The standard deviation of $D = \hat{p}_1 - \hat{p}_2$ can be estimated by

$$\widehat{\text{SD}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_{1.}} + \frac{\hat{p}(1-\hat{p})}{n_{2.}}}, \tag{10.5}$$

where

$$\hat{p} = \frac{\mathcal{O}_{11} + \mathcal{O}_{21}}{n_{1.} + n_{2.}} \tag{10.6}$$

is an estimator of the hypothesized common success rate p. Recall that $\mathcal{O}_{11}$ is the number of successes in sample 1, $\mathcal{O}_{21}$ is the number of successes in sample 2, and $n_{1.} + n_{2.}$ is the total number of trials for both samples. Thus, $\hat{p}$ may be viewed as being obtained by pooling the data in the two samples. The standardized version of $\hat{p}_1 - \hat{p}_2$ is

$$A = \frac{\hat{p}_1 - \hat{p}_2}{\widehat{\mathrm{SD}}(\hat{p}_1 - \hat{p}_2)}. \tag{10.7}$$

When H_0 is true, the asymptotic distribution of A is $N(0, 1)$.

a. *Approximate One-Sided Upper-Tail Test.* We denote the difference in success rates by

$$p_d = p_1 - p_2. \tag{10.8}$$

To test

$$H_0 : p_d = 0$$

versus

$$H_1 : p_d > 0,$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } A \geq z_\alpha; \text{ otherwise do not reject.} \tag{10.9}$$

b. *Approximate One-Sided Lower-Tail Test.* To test

$$H_0 : p_d = 0$$

versus

$$H_2 : p_d < 0,$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } A \leq -z_\alpha; \text{ otherwise do not reject.} \tag{10.10}$$

c. *Approximate Two-Sided Test.* To test

$$H_0 : p_d = 0$$

versus

$$H_3 : p_d \neq 0,$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } |A| \geq z_{a/2}; \text{ otherwise do not reject.} \tag{10.11}$$

Approximate confidence intervals for p_d are obtained as follows. For confidence intervals, a different estimator is utilized to estimate the standard deviation of $\widehat{p}_1 - \widehat{p}_2$ than was used in the testing procedure given earlier. Let

$$\widetilde{\mathrm{SD}}(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_{1.}} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_{2.}}}, \tag{10.12}$$

where

$$\widehat{p}_1 = \frac{\mathcal{O}_{11}}{n_{1.}}, \quad \widehat{p}_2 = \frac{\mathcal{O}_{21}}{n_{2.}}.$$

In estimating the standard deviation of $\widehat{p}_1 - \widehat{p}_2$ for use in the confidence interval procedure, we no longer assume $p_1 = p_2$, as in the case when H_0 is true. That is why $\widetilde{\mathrm{SD}}(\widehat{p}_1 - \widehat{p}_2)$ given by (10.12) differs from $\widehat{\mathrm{SD}}(\widehat{p}_1 - \widehat{p}_2)$ given by (10.5).

For a symmetric two-sided confidence interval for p_d, with the approximate confidence coefficient $1 - \alpha$, set

$$p_{d,\mathrm{L}} = \widehat{p}_1 - \widehat{p}_2 - z_{\alpha/2} \cdot \widetilde{\mathrm{SD}}(\widehat{p}_1 - \widehat{p}_2) \tag{10.13}$$

and

$$p_{d,\mathrm{U}} = \widehat{p}_1 - \widehat{p}_2 + z_{\alpha/2} \cdot \widetilde{\mathrm{SD}}(\widehat{p}_1 - \widehat{p}_2). \tag{10.14}$$

With $p_{d,\mathrm{L}}$ and $p_{d,\mathrm{U}}$ given by displays (10.13) and (10.14), respectively, we have, for all (p_1, p_2) pairs,

$$P_{(p_1,p_2)}(p_{d,\mathrm{L}} < p_d < p_{d,\mathrm{U}}) \approx 1 - \alpha. \tag{10.15}$$

EXAMPLE 10.1 ***Care Patterns for Black and White Patients with Breast Cancer.***

Diehr et al. (1989) pointed out that it is well known that survival of women with breast cancer tends to be lower in blacks than whites. See the references of their paper for documentation of this fact. Diehr and her colleagues were interested in whether differences seen in survival could be accounted for by differences in diagnostic methods and treatment. Their study sought to determine if there are statistically significant patterns of care and, if so, whether these differences can be attributed to differences between black and white patients in age, stage, type of insurance, type of hospital, or type of physician.

The information used in the Diehr et al. (1989) study concerning management and treatment during the first 4 months of diagnosis was abstracted from a systematic sample of inpatient and outpatient records of female patients in 107 participating hospitals. Diehr and her colleagues reported on a subset of 10 breast cancer patterns. The 10 were chosen because they were applicable to most patients and thus could be assessed for enough black patients to make the study possible. One pattern of interest was *liver scan.* Did patients with local or regional disease have a liver scan or CT scan of the liver? Such scans are not routinely required for a patient with local or regional disease because the likelihood of finding an abnormality with the scan is low in the absence of abnormal liver chemistry or hepatomegaly. In the Diehr et al. study, it was found that black patients with local disease were more likely to have a liver scan or a CT scan than were white patients. The

Table 10.2 Patients with Local or Regional Disease Receiving Liver Scan in Hospital 8

	Liver scan		
Patients	Yes	No	Totals
Black	4	8	12
White	1	20	21
Totals:	5	28	33

Source: P. Diehr, J. Yergan, J. Chu, P. Feigl, G. Glaefke, R. Moe, M. Bergner, and J. Rodenbaugh (1989).

percentage of black patients receiving appropriate care was about 10 percentage points lower than that of white patients, even after controlling for other factors. In particular, considering the data for patients in the 19 hospitals that had enough black patients for individual analysis, Diehr and her colleagues found that black patients were more likely than white patients to receive liver scan, and this tendency could not be attributed simply to chance. To see how this conclusion was reached, we first focus on the liver scan data of hospital 8 given in Table 10.2. We will illustrate how, for that hospital, the approximate test based on A can be used to see if there was a significant difference between the chance of a white patient receiving a scan and the chance of a black patient receiving a scan. (We return to this problem in Section 10.2 and apply Fisher's exact test. Later, in Section 10.4, we apply the Mantel–Haenszel test to the data from the 19 hospitals with the most black patients to get an overall conclusion.)

Let p_1 be the unknown probability that a black patient in hospital 8 with local or regional disease will receive a liver scan and let p_2 be the unknown probability that a white patient in hospital 8 with local or regional disease will receive a liver scan. We write the null hypothesis as

$$H_0 : p_1 = p_2,$$

or equivalently, $p_d = 0$, where $p_d = p_1 - p_1$.

Diehr and her colleagues suspected that a deviation from H_0 would be in the direction of the one-sided alternative $p_1 > p_2$. To test against this one-sided alternative, we use procedure (10.8). We find, from (10.3) to (10.6),

$$\widehat{p}_1 = \frac{4}{12} = .3333, \quad \widehat{p}_2 = \frac{1}{21} = .0476, \quad \widehat{p} = \frac{4+1}{12+21} = \frac{5}{33} = .1515,$$

and

$$\widehat{\text{SD}} = \sqrt{\frac{(.1515)(.8485)}{12} + \frac{(.1515)(.8485)}{21}} = .1296.$$

Then, from (10.7),

$$A = \frac{.3333 - .0476}{.1296} = 2.20.$$

The approximate $\alpha = .05$ test given by procedure (10.8) is reject H_0 if $A \geq 1.65$, accept H_0 otherwise. As $A = 2.20$, we reject H_0 at that level. The P-value is the probability that a standard normal is greater than $A = 2.20$. Using `pnorm(2.20, lower.tail = F)`,

we find $P = .014$. This constitutes strong evidence that in hospital 8 the chance that a black patient with local or regional breast cancer receives a liver scan is higher than the corresponding chance that a white patient with local or regional breast cancer receives a liver scan.

To find an approximate 95% confidence interval for $p_1 - p_2$, we first compute, via (10.12),

$$\widetilde{\mathrm{SD}}(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\frac{.3333(.6667)}{12} + \frac{(.0476)(.9524)}{21}} = .1439.$$

Then, the lower and upper confidence limits, given by (10.13) and (10.14), respectively, are

$$p_{d,\mathrm{L}} = .3333 - .0476 - 1.96(.1439) = .004,$$

$$p_{d,\mathrm{U}} = .3333 - .0476 + 1.96(.1439) = .568.$$

The 2 × 2 Chi-Squared Test of Homogeneity

The large-sample two-sided test based on A can also be presented via Karl Pearson's famous chi-squared statistic. It is motivated as follows. Suppose the null hypothesis H_0 is true. Then, the best estimator of the common success probability is $\widehat{p}$, given by (10.6). In the notation of Table 10.1, this can be written as

$$\widehat{p} = \frac{n_{.1}}{n_{..}}. \tag{10.16}$$

Using this estimator, the expected values of the random quantities $\mathcal{O}_{11}, \mathcal{O}_{12}, \mathcal{O}_{21}$, and $\mathcal{O}_{22}$ in Table 10.1 can be estimated, respectively, by E_{11}, E_{12}, E_{21}, and E_{22}, where

$$\begin{aligned}
E_{11} &= n_{1.} \times \widehat{p} = \frac{n_{1.} \times n_{.1}}{n_{..}}, \\
E_{12} &= n_{1.} \times (1 - \widehat{p}) = \frac{n_{1.} \times n_{.2}}{n_{..}}, \\
E_{21} &= n_{2.} \times \widehat{p} = \frac{n_{2.} \times n_{.1}}{n_{..}}, \\
E_{22} &= n_{2.} \times (1 - \widehat{p}) = \frac{n_{2.} \times n_{.2}}{n_{..}}.
\end{aligned} \tag{10.17}$$

A measure of the discrepancy between the observed frequen cies, the $\mathcal{O}$'s, and the estimated expected frequencies under the hypothesis, the E's, is the chi-squared statistic given by

$$\chi^2 = \frac{(\mathcal{O}_{11} - E_{11})^2}{E_{11}} + \frac{(\mathcal{O}_{12} - E_{12})^2}{E_{12}} + \frac{(\mathcal{O}_{21} - E_{21})^2}{E_{21}} + \frac{(\mathcal{O}_{22} - E_{22})^2}{E_{22}}. \tag{10.18}$$

The chi-squared statistic can be written in abbreviated notation as

$$\chi^2 = \sum \frac{(\mathcal{O} - E)^2}{E}, \tag{10.19}$$

where we have omitted the subscripts on the $\mathcal{O}$'s and the E's, but it is to be understood that the summation $\sum$ is over the four cells of Table 10.1. Note that the differences "observed minus expected," that is, $\mathcal{O} - E$, are squared, eliminating the balancing out of positive and negative discrepancies. Also, each squared difference is weighted by the inverse of the corresponding E, so that the differences involving small E's assume the greatest importance.

It can be shown that

$$A^2 = \chi^2, \tag{10.20}$$

where A is given by (10.7). This implies that the two-sided approximate α-level test of $p_d = 0$ versus $p_d \neq 0$ given by procedure (10.11) is equivalent to the test

$$\text{Reject } H_0 \text{ if } \chi^2 \geq \chi^2_{\alpha,1}; \text{ otherwise do not reject,} \tag{10.21}$$

where $\chi^2_{\alpha,1}$ is the upper α percentile point of the chi-squared distribution with 1 degree of freedom. There is a shortcut formula for the calculation of the chi-squared test statistic, namely,

$$\chi^2 = \frac{n_{..}(\mathcal{O}_{11}\mathcal{O}_{22} - \mathcal{O}_{21}\mathcal{O}_{12})^2}{n_{.1} \times n_{.2} \times n_{1.} \times n_{2.}}. \tag{10.22}$$

For the liver scan data of Table 10.2, we can use (10.22) to find

$$\chi^2 = \frac{33(4 \times 20 - 1 \times 8)^2}{(5)(28)(12)(21)} = 4.85,$$

agreeing (allowing for round-off error) with the value obtained by squaring $A = 2.20$.

R implements the two tests in this section with the command `prop.test`. One may enter a matrix where each row contains the number of successes and failures for a particular population or one can enter two vectors: one for the number of successes (`x`) and one for the corresponding number of observations (`n`). For the data in Table 10.2, the analysis may be performed with the success `x` and number `n` vectors by calling

```
prop.test(x=c(4, 1), n=c(12, 21), correct=F, alternative=
    "greater")
```

Setting the argument correct to false prevents a continuity correction from being applied to the data. Alternatively, we could have constructed a matrix with the successes and failures in rows:

```
table10.2 <- matrix(c(4, 8, 1, 20), byrow=T, nrow=2)
```

then use this matrix in `prop.test`:

```
prop.test(table10.2, correct=F, alternative="greater")
```

The output from each of these calls is identical:

```
X-squared = 4.849, df = 1, p-value = 0.01383
alternative hypothesis: greater
95 percent confidence interval:
0.04918621 1.00000000
sample estimates:
     prop1      prop2
0.33333333 0.04761905
```

The inference is based on A^2 (`X-squared`), not A. Note that the confidence interval given coincides with the direction of the alternative. To obtain the two-sided confidence interval calculated in Example 10.1, change the argument `alternative` to "`two.sided`".

The chi-squared test, as developed here, is called a *chi-squared test of homogeneity*. This is because, for Table 10.1, we have considered the data to be based on a sample of size $n_{1.}$ from one population and a separate independent sample of size $n_{2.}$ from a second population. Thus, for the liver scan data of Table 10.2, $n_{1.} = 12, n_{2.} = 21$, and the null hypothesis specifies that $p_1 = p_2$, where p_1 denotes the probability that a black patient with local or regional disease will receive a liver scan and p_2 denotes the probability that a white patient with local or regional disease will receive a liver scan. The null hypothesis is called the *homogeneity hypothesis* because it specifies that the chance of success is the same for both populations.

The 2 × 2 Chi-Squared Test of Independence

In contrast to the homogeneity framework, 2×2 tables also arise when none of $n_{.1}, n_{.2}, n_{1.}$ or $n_{2.}$ is fixed, but instead when each observation from a general population is cross-classified on the basis of two characteristics (having characteristic C, say, not having characteristic C; having characteristic D, say, not having characteristic D). The question is whether the occurrences of the characteristics are *independent*. We now describe how the chi-squared statistic defined by (10.18) is also appropriate for a test of independence.

We rewrite Table 10.1 using slightly different notation (see Table 10.3).

Table 10.3 2×2 Table of Outcomes

	C	Not C	Totals
D	$\mathcal{O}_{11}$	$\mathcal{O}_{11}$	$n_{1.}$
Not D	$\mathcal{O}_{21}$	$\mathcal{O}_{22}$	$n_{2.}$
Totals:	$n_{.1}$	$n_{.2}$	$n_{..}$

(10.23)

Let $p_{ij}, i = 1, 2, j = 1, 2$ denote the true unknown joint probability of falling into cell (i, j) of Table 10.3. Thus

$$\begin{aligned} p_{11} &= P(C \text{ and } D), & p_{12} &= P(\text{not } C \text{ and } D), \\ p_{21} &= P(C \text{ and not } D), & p_{22} &= P(\text{not } C \text{ and not } D). \end{aligned} \tag{10.24}$$

The marginal probabilities are

$$\begin{aligned} p_{1.} &= p_{11} + p_{12}, & p_{2.} &= p_{21} + p_{22}, \\ p_{.1} &= p_{11} + p_{21}, & p_{.2} &= p_{12} + p_{22}. \end{aligned} \tag{10.25}$$

The hypothesis H_I of independence asserts that all joint probabilities are equal to the product of their marginal probabilities, namely,

$$H_I : p_{ij} = p_{i.} \times p_{.j}, \quad i = 1, 2, \quad j = 1, 2. \tag{10.26}$$

If, for example, E_{11} denotes the expected number of observations that fall in cell (1, 1) of Table 10.3, we have

$$E_{11} = n_{..} \times P(C \text{ and } D) = n_{..} \times p_{11},$$

and under H_I,

$$E_{11} = n_{..} \times P(C) \times P(D) = n_{..} \times p_{.1} \times p_{1.}$$

It is natural to estimate $P(C) = p_{1.}$ by $(\mathcal{O}_{11} + \mathcal{O}_{21})/n_{..}$ and $P(D) = p_{1.}$ by $(\mathcal{O}_{11} + \mathcal{O}_{12})/n_{..}$. That is, *P(C)* is estimated by the relative frequency of event *C,* and *P(D)* is estimated by the relative frequency of event *D.* Thus, under the hypothesis of independence, the E's are estimated as (we are abusing notation and using the same symbol E_{ij} for the expected number of observations falling into the (i,j) cell and an estimator of that expected number)

$$E_{11} = n_{..} \times \left(\frac{\mathcal{O}_{11} + \mathcal{O}_{21}}{n_{..}}\right) \times \left(\frac{\mathcal{O}_{11} + \mathcal{O}_{12}}{n_{..}}\right) = \frac{n_{.1} \times n_{1.}}{n_{..}},$$

and

$$E_{12} = \frac{n_{.2} \times n_{1.}}{n_{..}},$$

$$E_{21} = \frac{n_{.1} \times n_{2.}}{n_{..}},$$

$$E_{22} = \frac{n_{.2} \times n_{2.}}{n_{..}}. \tag{10.27}$$

Note that the E's given by display (10.27) agree with the E's given by display (10.17). It follows that the χ^2 statistic given by (10.20) can also be used to test independence. Specifically, an approximate α-level test of H_I, versus alternatives where association holds between the two characteristics, is

$$\text{Reject } H_I \text{ if } \chi^2 \geq {\chi^2}_{\alpha,1}; \text{ otherwise do not reject.} \tag{10.28}$$

We illustrate this test in Example 10.2.

EXAMPLE 10.2 ***Death Penalty and Gun Registration.***

The data in Table 10.4 were reported by Clogg and Shockey (1988), whose source was the 1982 General Social Survey.

For these data, we can calculate χ^2 via (10.18) or (10.22). In R, we use the command `chisq.test`. The argument is a matrix containing the characteristic data:

```
table10.4 <- matrix(c(784, 236, 311, 66), byrow=T, nrow=2)
```

Table 10.4 Gun Registration and Death Penalty Cross-Classification

Gun registration	Death Penalty		Totals
	Favor	Oppose	
Favor	784	236	1020
Oppose	311	66	377
Totals:	1095	302	1397

Source: 1982 General Social Survey; see C. C. Clogg and J. W. Shockey (1988).

Using this matrix in `chisq.test(table10.4, correct=F)` results in

```
Pearson's Chi-squared test

data:  table10.4
X-squared = 5.1503, df = 1, p-value = 0.02324
```

We find $\chi^2 = 5.15$ with a P-value of .023, indicating that there is an association between the two characteristics, namely, attitude toward gun registration and attitude toward the death penalty.

Although χ^2 measures, via the formal hypothesis test, the significance of association between two characteristics, it does *not* measure the *degree* of association. In Section 10.3, we will discuss a measure of the degree of association based on the odds ratio.

The R command `chisq.test` produces the expected and observed values for the data. These may be accessed with the `$` notation followed by `expected` or `observed`. Thus, for the data in Table 10.2, the expected values in the cells are obtained with `chisq.test(table10.2, correct=F) $expected`. This results in

```
          [,1]     [,2]
[1,] 1.818182 10.18182
[2,] 3.181818 17.81818
```

Comments

1. *Sample-Size Determination.* Suppose we want to determine sample sizes so that our estimate $\widehat{p}_1 - \widehat{p}_2$ of the true difference $p_1 - p_2$ will be within D of the true value, with probability equal to $1 - \alpha$. Equating the desired precision, D, to the actual precision, $(z_{\alpha/2}) \cdot \mathrm{SD}(\widehat{p}_1 - \widehat{p}_2)$, yields the equation

$$D = z_{\alpha/2} \cdot \sqrt{\frac{p_1(1-p_1)}{n_{1.}} + \frac{p_2(1-p_2)}{n_{2.}}}.$$

Taking $n_{1.} = n_{2.} = m$ and solving for m yields

$$m = \frac{(z_{\alpha/2})^2 \cdot [p_1(1-p_1) + p_2(1-p_2)]}{D^2}. \tag{10.29}$$

We cannot use (10.29) as it stands, because p_1 and p_2 are not known. (The purpose of the experiment is to obtain information about the unknown values of

p_1, p_2, and $p_1 - p_2$.) However, the term in square brackets in the numerator of the right-hand side of (10.29) is largest when $p_1 = p_2 = \frac{1}{2}$. Thus, a sufficiently large-sample would be

$$m = \frac{(z_{\alpha/2})^2 \cdot \left[\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\right]}{D^2} = \frac{(z_{\alpha/2})^2}{2D^2}.$$

This sample size assures the desired reliability regardless of the values of p_1 and p_2. In situations in which it is known that p_1 and p_2 are definitely less than some maximum value p^* (say), which is less than $\frac{1}{2}$, p^* can be substituted for p_1 and p_2 in (10.29), yielding

$$m = \frac{(z_{\alpha/2})^2[2(p^*)(1-p^*)]}{D^2}. \tag{10.30}$$

The same is true if p_1 and p_2 are known to be greater than some value p^*, which is greater than $\frac{1}{2}$.

2. *Testing $p_1 - p_2$ Equals Some Specific Nonzero Value.* In this section, the tests based on A were formulated for the null hypothesis $p_1 = p_2$, or, equivalently, $p_1 - p_2 = 0$. To test $p_1 - p_2 = \delta_0$ (say), where δ_0 is any specified nonzero value between -1 and 1, use the statistic A', defined as

$$A' = \frac{(\widehat{p}_1 - \widehat{p}_2) - \delta_0}{\widehat{\mathrm{SD}}(\widehat{p}_1 - \widehat{p}_2)}. \tag{10.31}$$

Note that the denominator of A' uses $\widehat{\mathrm{SD}}(\widehat{p}_1 - \widehat{p}_2)$, as given by (10.5), rather than the estimator $\widetilde{\mathrm{SD}}(\widehat{p}_1 - \widehat{p}_2)$, given by (10.12), based on pooling the two sample proportions together. The statistic A' should be referred to percentiles of the normal distribution. Significantly large values of A' indicate $p_1 - p_2 > \delta_0$; significantly small values of A' indicate $p_1 - p_2 < \delta_0$.

3. *Different Sampling Schemes.* In the framework of the test of homogeneity, we are testing $H_0 : p_1 - p_2$, a comparison of success probabilities using two binomial samples. In the notation of Table 10.1, we observe $\mathcal{O}_{11}$ successes out of $n_{1.}$ observations in sample 1, a sample from an underlying population 1 (say), and we observe $\mathcal{O}_{21}$ successes out of $n_{2.}$ observations in sample 2, a sample from an underlying population 2. Here, $n_{1.}$ and $n_{2.}$ are fixed, but $n_{.1}$ and $n_{.2}$ are not fixed but random (although $n_{.1}$ and $n_{.2}$ are constrained to sum to $n_{..}$). In the framework of testing for independence, the sampling scheme is different. It is known as *cross-sectional sampling*. A total of $n_{..}$ subjects are obtained from an underlying population, and then each subject falls, in the notation of Table 10.3, into the 4 cells of the 2×2 table according to whether or not the subject possesses characteristic C and whether or not he has characteristic D. Here none of the row and column totals $n_{.1}, n_{.2}, n_{1.}, n_{2.}$ are fixed; only $n_{..}$ is fixed.

4. *Determining If the Sample Sizes Are Large Enough for the Large-Sample Approximation.* The tests and confidence interval of this section depend on an approximation to exact probabilities. The approximation is close if the samples are large. They should be large enough so that the E's, defined by display (10.17), each are no smaller than five.

5. *Yates' Correction for Continuity.* Yates (1934) proposed a continuity correction for the chi-squared test. The corrected χ^2 statistic is

$$\chi_c^2 = \frac{n_{..}(|\mathcal{O}_{11}\mathcal{O}_{22} - \mathcal{O}_{21}\mathcal{O}_{12}| - n_{..}/2)^2}{n_{.1} \times n_{.2} \times n_{1.} \times n_{2.}}. \tag{10.32}$$

This correction can be used in both the test for independence and the test for homogeneity. To get a P-value, refer χ_c^2 to the chi-squared distribution with 1 degree of freedom. To perform a formal test of the null hypothesis, reject H_0 at the (approximate) type I error probability α if $\chi_c^2 \geq \chi_{\alpha,1}^2$ and accept H_0 if $\chi_c^2 < \chi_{\alpha,1}^2$; otherwise do not reject. This correction is implemented in the `R` commands `prop.test` and `chisq.test` by the argument `correct=T`.

There is disagreement in the statistical literature about the virtue of the continuity correction based on χ_c^2. See Storer and Kim (1990) and the references therein. On the basis of their study, Storer and Kim hold the view that in the context of comparing two binomial samples (i.e., testing homogeneity with two marginals $n_{1.}$ and $n_{2.}$ fixed), Yates' continuity correction should not be used. They also state "... our results suggest that for any reasonable sample size one will not be led far astray by the simple uncorrected χ^2 statistic." Storer and Kim actually compare seven tests of the homogeneity hypothesis $H_0 : p_1 = p_2$, including the approximate test based on χ^2, the approximate test based on χ_c^2, and Fisher's exact test. Fisher's exact test is conditional on all marginal totals $n_{.1}, n_{.2}, n_{1.}, n_{2.}$ being fixed. We present this test in Section 10.2.

6. *McNemar's Test.* Instead of having two independent samples to form the 2×2 table, as is the setup for the homogeneity test, there will be experiments in which the categorical data are based on dependent samples. Dependent samples can occur in matched-pair studies. For example, a pair may consist of a sibling and a parent. This is the situation for the Hodgkins tonsillectomy example in this comment. Dependent samples can also occur when the same subject is measured at two different times.

Johnson and Johnson (1972), in a study that was interested in testing the theory that the tonsils protect the body against invasion of the lymph nodes by a Hodgkin's disease virus (also see Problem 3), obtained tonsillectomy data on 85 Hodgkin's cases and a sibling of each case. The data showed 41 tonsillectomies among the Hodgkin's cases and 33 tonsillectomies among the siblings. The pairing of a case with sibling means that the rates for the two groups are not independent. The pairing should be taken into account in the analysis in order to achieve the best chance of detecting a departure from the null hypothesis. The null hypothesis asserts that Hodgkin's cases and their siblings have the same rates of tonsillectomy. A proper way to test the null hypothesis is to apply the one-sample binomial test (Section 2.1) to Table 10.5, obtained by Johnson and Johnson.

If there is no association between tonsillectomy and Hodgkin's disease, then the probability is $\frac{1}{2}$ that a patient–sibling pair falls in the upper-right cell and $\frac{1}{2}$ that it falls in the lower-left cell, given that the pair falls off the main diagonal. Since the pairs are independent, the ratio $\frac{15}{22}$ can be compared with $\frac{1}{2}$ by a binomial

Table 10.5 Tonsillectomy Rates for Hodgkin's Disease Patients and Siblings

		Sibling		
		Tonsillectomy	No tonsillectomy	Totals
Hodgkin's Patients	Tonsillectomy	26	15	41
	No tonsillectomy	7	37	44
	Totals:	33	52	85

Source: S. K. Johnson and R. E. Johnson (1972).

test, as described in Chapter 2. Specifically, let

$$\widehat{p}_{.1} = \text{proportion of silbings with tonsillectomy} = \frac{26+7}{85},$$

$$\widehat{p}_{1.} = \text{proportion of Hodgkin's patients with tonsillectomy} = \frac{26+15}{85}.$$

The statistics $\widehat{p}_{.1}$ and $\widehat{p}_{1.}$ are estimates of $p_{.1}$ and $p_{1.}$, respectively. The difference between $\widehat{p}_{1.}$ and $\widehat{p}_{.1}$ is

$$d = \widehat{p}_{1.} - \widehat{p}_{.1} = \frac{26+15}{85} - \frac{26+7}{85} = \frac{15-7}{85}$$

and the estimated standard deviation of this difference is

$$\widehat{\text{SD}}(d) = \frac{\sqrt{15+7}}{85}.$$

An approximate large-sample test can be applied by referring

$$\frac{d}{\widehat{\text{SD}}(d)} = \frac{\frac{15-7}{85}}{\frac{\sqrt{15+7}}{85}} = \frac{15-7}{\sqrt{15+7}} = 1.71$$

to a $N(0,1)$ distribution. The approximate one-sided P-value is .044, indicating that there is evidence that the rate of tonsillectomy is higher for Hodgkin's cases than for their siblings.

The exact one-sided P-value is $\Pr(B \geq 15)$, where B is a binomial random variable with $p = \frac{1}{2}$ and $n = 22$. This is easily found using the `R` command `pbinom`.

To find $\Pr(B \geq 15)$, we use

$$\Pr(B \geq 15) = \Pr(B > 14).$$

Using `pbinom(14, size=22, prob=1/2 lower.tail=F)`, the P-value is .0669. The one-sided approximate P-value of .044 found by the normal approximation is in reasonable agreement with this exact one-sided P-value of .067. Note, however, that if you were using an $\alpha = .05$ level, you would accept $H_0' : p_{1.} = p_{.1}$ in favor of the alternative $p_{1.} > p_{.1}$ with the one-sided exact McNemar's test, but you would reject H_0 with the normal approximation.

Table 10.6 Data on Matched Pairs

		Controls		
		Factor present	Factor absent	Totals
Cases	Factor present	$\mathcal{O}_{11}$	$\mathcal{O}_{12}$	$n_{1.}$
	Factor absent	$\mathcal{O}_{21}$	$\mathcal{O}_{22}$	$n_{2.}$
	Totals:	$n_{.1}$	$n_{.2}$	$n_{..}$

The R command `mcnemar.test` will implement this procedure. The data is specified as a 2×2 matrix. This command provides two-sided P-values. Appropriate one-sided P-values may be derived using this two-sided value.

More generally, suppose we are dealing with a retrospective study where each case has been matched with a control. We wish to compare the frequency of an antecedent factor (in the preceding example, tonsillectomy) among the cases with the frequency of the antecedent factor among the controls. The data can be summarized as in Table 10.6.

The null hypothesis is $H_0' : p_{1.} = p_{.1}$, which is equivalent to $p_{12} = p_{21}$.

To test $H_0' : p_{1.} = p_{.1}$, the hypothesis that asserts the cases and controls have the same population proportions of the antecedent factor, refer

$$d = \frac{\widehat{p}_{1.} - \widehat{p}_{.1}}{\widehat{\mathrm{SD}}(d)} = \frac{\mathcal{O}_{12} - \mathcal{O}_{21}}{\sqrt{\mathcal{O}_{12} + \mathcal{O}_{21}}} \tag{10.33}$$

to a $N(0, 1)$ distribution.

The two-sided test of $H_0' : p_{1.} = p_{.1}$ versus the alternative $p_{1.} \neq p_{.1}$ at the (approximate) α level is reject H_0' if $|d| \geq z_{\alpha/2}$; otherwise do not reject.

The one-sided (approximate) α-level test of H_0' versus the alternative $p_{1.} > p_{.1}$ is reject H_0' if $d \geq z_\alpha$; otherwise do not reject. Similarly, the one-sided (approximate) α-level test of H_0' versus the alternative $p_{1.} < p_{.1}$ is reject H_0' if $d \leq -z_\alpha$; otherwise do not reject.

See McNemar (1947), Mosteller (1952), and Agresti (2013) for further details.

7. *Edwards' Correction for Continuity.* Recall McNemar's test (Comment 6) in the matched-pairs situation. The test is based on the statistic

$$d = \frac{\mathcal{O}_{12} - \mathcal{O}_{21}}{\sqrt{\mathcal{O}_{12} + \mathcal{O}_{21}}}.$$

The approximate two-sided α-level test of $H_0' : p_{1.} = p_{.1}$ versus the alternative $p_{1.} \neq p_{.1}$ is reject H_0' if $|d| \geq z_{\alpha/2}$; otherwise do not reject. An equivalent test is to compute

$$d^2 = \frac{(\mathcal{O}_{12} - \mathcal{O}_{21})^2}{\mathcal{O}_{12} + \mathcal{O}_{21}} \tag{10.34}$$

and reject H_0' if $d^2 \geq \chi^2_{\alpha,1}$, do not reject H_0' if $d^2 < \chi^2_{\alpha,1}$, where $\chi^2_{\alpha,1}$ is the upper α percentile point of the chi-squared distribution with 1 degree of freedom. To

correct this test for continuity, a correction due to Edwards (1948) is based on computing

$$\chi_e^2 = \frac{(|\mathcal{O}_{12} - \mathcal{O}_{21}| - 1)^2}{\mathcal{O}_{12} + \mathcal{O}_{21}} \tag{10.35}$$

and referring χ_e^2 to the chi-squared distribution with one degree of freedom.

Properties

1. *Asymptotic Distribution of Pearson's Chi-Squared Statistic.* See Agresti (2013, Sections 3.2.1 and 16.3.3).
2. *Asymptotic Equivalence of Pearson's Chi-Squared Statistic and the Likelihood Ratio Statistic.* See Agresti (2013, Section 16.3.4).
3. *Power of the Chi-Squared Test.* See Agresti (2013, Section 6.6.4).

Problems

1. Andrews (1995) investigated bodily shame as a possible mediating factor between abusive experiences (sexual and physical) and later depression in a community sample of adult women. A total of 101 women, who ranged in age from 32 to 56 years, were selected from an original longitudinal study of 289 women performed between 1980 and 1983 in Islington, an inner-city area of London, England (Brown et al. 1986). The 3-year study was designed to investigate the onset and course of depressive disorder. The investigators concentrated on working-class women with at least one child at home in order to get a group of women who were at high risk of developing clinical depression. Table 10.7 is a 2×2 table adapted from Andrews (1995) for the purpose of investigating association between childhood abuse and depression. (Physical and sexual abuse were combined into one category, abuse.)

From Table 10.7, we see that 17 of 31 women with childhood abuse had been depressed, whereas 22 of 70 women who had not expressed childhood abuse had been depressed. Test for independence of childhood abuse and depression against alternatives of association.

2. In the study by Andrews (1995) described in Problem 1, there was also an investigation into the possible association between abuse in adulthood and depression. Table 10.8, adapted from Andrews, gives the results.

Test for independence of adulthood abuse and depression against alternatives of association.

3. Vianna, Greenwald, and Davies (1971) considered a series of 101 Hodgkin's disease patients, with the purpose of testing the theory that the tonsils protect the body against invasion of the lymph nodes by a Hodgkin's disease virus. (The existence of such a virus has not been established.) Among the 101 Hodgkin's cases, they found 67 had had a tonsillectomy, whereas in a control group of 107 patients with other complaints, 43 had had a tonsillectomy. Compute 99% confidence limits for the difference in true rates.

Table 10.7 Abuse in Childhood and Depression in the 8-Year Study Period

Abuse in childhood	Depression	No depression	Totals
Yes	17	14	31
No	22	48	70
Totals:	39	62	101

Source: B. Andrews (1995).

Table 10.8 Abuse in Adulthood and Depression in the 8-Year Study Period

Abuse in adulthood	Depression	No depression	Totals
Yes	23	15	38
No	16	47	63
Totals:	39	62	101

Source: B. Andrews (1995).

Table 10.9 ABC-ELISA and Standard ELISA

Standard ELISA		+	−	Totals
ABC-ELISA	+	82	13	95
	−	6	0	6
	Totals:	88	13	101

Source: D. F. Cruess (1989).

4. R. Goode and D. Coursey, in a study of the theory that the tonsils serve as a reservoir harboring the virus that causes mononucleosis, obtained data on Stanford students seeking treatment for mononucleosis at the Stanford University Student Health Service. The data are given in Miller (1980). Among 46 students 21 years old diagnosed as having mononucleosis, they found that only 8 had had a tonsillectomy, whereas among 139 students of the same age, in the health center for other complaints, 48 had had a tonsillectomy. Compute 95% confidence limits for the difference in rates.

5. Verify directly the equivalence of (10.18) and (10.22).

6. Cruess (1989) points out that the error of applying the ordinary chi-squared statistic to paired data occurs frequently in the medical literature. Cruess cites in particular the study of Shen et al. (1988). They compared the results of two tests, ABC-ELISA and standard ELISA, on 101 hydatidosis patients. (Hydatidosis, or hydatid disease, is infestation with echinococcus, a genus of tapeworms.) Shen et al. used the ordinary unpaired chi-squared test and reported a P-value < 0.005. This was inappropriate because each case was tested using both laboratory procedures and thus the data were paired. Instead of the ordinary unpaired chi-squared test, McNemar's test should have been performed. Table 10.9 gives the information on the 101 pairs.

Perform McNemar's test. What is the P-value? What do you conclude concerning the hypothesis of equal proportions positive for both ELISA tests?

7. Suppose you are planning an experiment to investigate two success rates p_1, p_2. Determine the value of the sample size m for each sample (in the equal sample-size case) so that your estimate, $\hat{p}_1 - \hat{p}_2$, of the true difference, $p_1 - p_2$, will be within .2 of the true difference with probability .95.

8. Verify (10.20). That is, show χ^2 given by (10.18) is equal to the square of A, where A is given by (10.7).

9. Astin et al. (1995) studied posttraumatic stress disorder (PTSD) and childhood abuse in battered women. PTSD prevalence rates were compared among 50 battered women and 37 maritally distressed women who had not experienced battering. The results are given in Table 10.10.

Is there a significant difference in the PTSD rates for battered women versus maritally distressed women (who had not experienced battering)?

Table 10.10 PTSD Rates

PTSD	Battered women	Maritally distressed women who had not experienced battering	totals
Yes	29	7	36
No	21	30	51
Totals:	50	37	87

Source: M. C. Astin, S. M. Ogland-Hand, E. M. Coleman, and D. S. Foy (1995).

10. Recall Table 10.3 and the definition of independence given by H_I (10.26). We define the conditional probabilities $p_{j|i} = p_{ij}/p_{i.}, i = 1, 2, j = 1, 2$. Thus, for example, $p_{1|2}$ is the conditional probability of the observation landing in column 1 (i.e., has characteristic C) given that the observation has landed in row 2 (i.e., does not have characteristic D). Show that H_I is equivalent to the equalities $p_{1|1} = p_{1|2}$ and $p_{2|1} = p_{2|2}$ being satisfied.

11. Show that the four equalities of H_I (given by (10.26)) are satisfied if and only if $p_{11} = p_{1.} \times p_{.1}$.

10.2 AN EXACT TEST FOR THE DIFFERENCE BETWEEN TWO SUCCESS PROBABILITIES (FISHER)

Recall the basic 2×2 table given in display (10.1). Fisher's (1934) exact test is based on the conditional distribution of $\mathcal{O}_{11}$ given the row and column sums $n_{1.}, n_{2.}, n_{.1}, n_{.2}$. The conditional distribution of $\mathcal{O}_{11}$ is

$$\Pr(\mathcal{O}_{11} = x | n_{1.}, n_{2.}, n_{.1}, n_{.2}) = \frac{\binom{n_{1.}}{x}\binom{n_{2.}}{n_{.1} - x}}{\binom{n_{..}}{n_{.1}}}. \tag{10.36}$$

The range of possible values for x is $n_{\mathrm{L}} \le x \le n_{\mathrm{U}}$, where $n_{\mathrm{L}} = \max(0, n_{1.} + n_{.1} - n_{..})$ and $n_{\mathrm{U}} = \min(n_{1.}, n_{.1})$. The conditional probability distribution defined by (10.36) is a member of a family of distributions known as *hypergeometric distributions*. Expression (10.36) can be put in a more readily usable form by simplifying the binomial coefficients appearing in the numerator and denominator. Such simplification allows us to rewrite that equation as

$$\Pr(\mathcal{O}_{11} = x | n_{1.}, n_{2.}, n_{.1}, n_{.2}) = \frac{n_{.1}! n_{.2}! n_{1.}! n_{2.}!}{n_{..}! x! \mathcal{O}_{12}! \mathcal{O}_{21}! \mathcal{O}_{22}!}. \tag{10.37}$$

Fisher's exact test judges whether $\mathcal{O}_{11}$ is significantly small or significantly large with respect to the conditional distribution defined by (10.36). Specifically, to test $H_0 : p_1 = p_2$ versus the alternative $p_1 < p_2$, Fisher's exact test is reject H_0 if $\mathcal{O}_{11} \le q_\alpha$, otherwise do not reject, where q_α is chosen from the conditional distribution so that $Pr(\mathcal{O}_{11} \le q_\alpha | n_{1.}, n_{2.}, n_{.1}, n_{.2}) = \alpha$. Similarly, to test $H_0 : p_1 = p_2$ versus the alternative $p_1 > p_2$, Fisher's exact test is reject H_0 if $\mathcal{O}_{11} \ge r_\alpha$, otherwise do not reject, where r_α satisfies $Pr(\mathcal{O}_{11} \ge r_\alpha | n_{1.}, n_{2.}, n_{.1}, n_{.2}) = \alpha$. A two-sided α-level test of $H_0 : p_1 = p_2$ versus the alternative $p_1 \ne p_2$ is reject H_0 if $\mathcal{O}_{11} \le q_{\alpha_1}$ or if $\mathcal{O}_{11} \ge r_{\alpha_2}$, otherwise do not reject, where $q_{\alpha_1}, r_{\alpha_2}$ are chosen to give α_1 probability in the lower tail and α_2 probability in the upper tail, where $\alpha_1 + \alpha_2 = \alpha$. Critical values for these tests can be obtained from the tables of Finney et al. (1963), and P-values for the test can be obtained from R.

EXAMPLE 10.3 ***Example 10.1 Continued.***

Recall the liver scan data of hospital 8. We now illustrate how, for that hospital, Fisher's exact test can be used to see if there was a significant difference between the chance of a white patient receiving a scan and the chance of a black patient receiving a scan. (Later, in Section 10.4, we use a test due to Mantel and Haenszel (1959) to get an overall conclusion based on the data from the 19 hospitals with the most black patients.)

Let p_1 be the unknown probability that a black patient receives a liver scan and let p_2 be the unknown probability that a white patient receives a liver scan. The null hypothesis is $H_0 : p_1 = p_2$. Diehr and her colleagues suspected that a deviation from H_0 would be in the direction of the one-sided alternative $p_1 > p_2$, and thus, they reported the one-sided P-value corresponding to large values of $\mathcal{O}_{11}$. To find the P-value corresponding to their observed value $\mathcal{O}_{11} = 4$ (see Table 10.2) for hospital 8, we need to evaluate the probabilities of the tables giving a value as large or larger than the observed value of $\mathcal{O}_{11}$. These tables are

$$\begin{array}{ccc} 4 & 8 & 12 \\ 1 & 20 & 21 \\ \hline 5 & 28 & 33 \end{array} \qquad \begin{array}{ccc} 5 & 7 & 12 \\ 0 & 21 & 21 \\ \hline 5 & 28 & 33 \end{array}$$

with the table on the left corresponding to Table 10.2 and the table on the right being the only more extreme response in the direction $p_1 > p_2$. Next, we use (10.37) to calculate the probabilities associated with each of these two tables, corresponding to $x = 4$ and $x = 5$.

x	Table	Probability	
4	$\begin{array}{ccc} 4 & 8 & 12 \\ 1 & 20 & 21 \\ \hline 5 & 28 & 33 \end{array}$	$\dfrac{5!28!12!21!}{33!4!8!1!20!} = .0438$	(10.38)
5	$\begin{array}{ccc} 5 & 7 & 12 \\ 0 & 21 & 21 \\ \hline 5 & 28 & 33 \end{array}$	$\dfrac{5!28!12!21!}{33!5!7!0!21!} = .0033$	(10.39)

Thus, the P-value for hospital 8 is

$$P = .0033 + .0438 = .047. \tag{10.40}$$

This constitutes strong evidence that in hospital 8 the chance that a black patient with local or regional breast cancer receives a liver scan is higher than the corresponding chance that a white patient with local or regional breast cancer receives a liver scan.

The `R` command `fisher.test` will implement this procedure. The data is specified as a 2×2 matrix. This command provides two-sided and one-sided P-values by specifying the appropriate value to the argument `alternative`. One may also use the command `oddsratio` from package `epitools` (Aragon (2012)) to obtain the P-value for the two-sided test.

Comments

8. *Justification for Fisher's Exact Test.* Result (10.36) is justified as follows. Suppose that in a group of $n_{..}$ individuals, $n_{1.}$ possess a certain attribute and $n_{..} - n_{1.}$ do not. If a sample of $n_{.1}$ individuals is drawn randomly from the $n_{..}$ individuals, without replacement, the chance that x of these individuals possess the attribute is given by the right-hand side of (10.36).
9. *Use of Fisher's Exact Test as a Test of Independence.* In this section, we have introduced Fisher's exact test in the context of a test of homogeneity. It can, however, also be used to test independence (just as the approximate chi-squared test of Section 10.1 can be used as a test of homogeneity and also as a test of independence). If, using the notation of Table 10.3, the events C and D are independent, then the conditional distribution of $\mathcal{O}_{11}$ within the restricted set of samples having fixed row and column tables is again given by (10.36).
10. *Limited Choice of α Values.* For small sample sizes, the probability distribution of $\mathcal{O}_{11}$, given by expression (10.36), is highly discrete (i.e., has a small number of possible values with corresponding probabilities). Thus, the user's choices for α, when performing the formal test, are limited. Equivalently, there will, in small-sample-size cases, be a small number of possible P-values. This is illustrated in Problem 12, where, for the data in Table 10.11, there are only three possible values for $\mathcal{O}_{11}$ with three corresponding P-values.
11. *Use and Misuse of Statistics.* Cruess (1989) reviewed the statistics of the 201 scientific articles published during the calendar year 1988 in *The American Journal of Tropical Medicine and Hygiene.* He determined that 148 of the articles had at least one detectable statistical error; most of the errors involved improper documentation or application of statistical hypothesis testing. Among others, Cruess cites the Mendis, Ihalamulla, and David (1988) article (considered in Problem 12) as one in which the sample sizes were not large enough to justify the large-sample approximation used to compute P-values. The errors cited by Cruess are typical of the uses and misuses of statistics in journals in other areas of medical research. Ironically, the papers containing errors are often the most clearly written. Indeed, it is possible to ascertain errors only if the authors supply sufficient data and detail so that other researchers can check their results. Thus, it may be unfair to be overly critical of the papers that contain errors.

Properties

1. *Uniformly Most Powerful Unbiased (UMPU) Property of Fisher's Exact Test.* See Tocher (1950) and Agresti (2013, Section 3.5.5).
2. *Minimal Sample Sizes Required to Achieve a Certain Power for Specified Significance Levels.* See Gail and Gart (1973) and Suissa and Shuster (1985).

Problems

12. The data in Table 10.11 are from a study by Mendis, Ihalamulla, and David (1988). The researchers compared the reactivity of heterologous human immune sera from patients with multiple malaria attacks to sera from primary-attack patients.

Table 10.11 Reactivity of Multiple Malaria Attack Patients and Primary-Attack Patients

Patients	Reactivity		Totals
	Low	High	
Multiple attack	0	5	5
Primary attack	2	1	3
Totals:	2	6	8

Source: K. N. Mendis, R. I. Ihalamulla, and P. H. David (1988).

We write the null hypothesis as $H_0 : p_1 = p_2$, where p_1 is the unknown probability that a multiple-attack patient will have low reactivity and p_2 is the unknown probability that a primary-attack patient will have low reactivity. What is the P-value achieved by these data if we use Fisher's exact text of H_0 against the alternative $p_1 < p_2$? What is your conclusion?

13. Consider Problem 12 and the reactivity data of Table 10.11. Mendis, Ihalamulla, and David (1988) claimed a "significantly higher incidence of reactivity" in the multiple-attack patients. (In our notation of Problem 12, their conclusion corresponds to the alternative $p_1 < p_2$, or equivalently, $1 - p_1 > 1 - p_2$.) Mendis, Ihalamulla, and David (1988) based their conclusion on the χ^2 statistic defined by (10.18). Apply the approximate test based on χ^2. Compare their conclusion with your conclusion obtained in Problem 12, recalling that the P-value obtained in Problem 12 is one-sided and the P-value based on χ^2 is two-sided.

14. Return to the reactivity data of Table 10.11. Find the P-value based on Yate's continuity correction to χ^2 (see Comment 5). Compare your result with those of Problems 12 and 13.

15. Show that expressions (10.36) and (10.37), for the conditional distribution of $\mathcal{O}_{11}$ given the row and column sums, are equivalent.

16. For the Diehr et al. (1989) study, the liver scan data for hospital 16 are given in Table 10.12. Let p_1 be the probability that a black patient in hospital 16 with local or regional disease will receive a liver scan and let p_2 be the probability that a white patient in hospital 16 with local or regional disease will receive a liver scan. Test $H_0 : p_1 = p_2$ versus the alternative $p_1 > p_2$ using Fisher's exact test. What is the P-value?

17. For the data of Table 10.12, perform two chi-squared approximations, one without a continuity correction and with Yates's continuity correction. Compare the approximate P-values with the exact P-value obtained in Problem 16.

Table 10.12 Patients with Local or Regional Disease Receiving Liver Scan in Hospital 16

Patients	Liver scan		Totals
	Yes	No	
Black	2	3	5
White	3	12	15
Totals:	5	15	20

Source: P. Diehr, J. Yergan, J. Chu, P. Feigl, G. Glaefke, R. Moe, M. Bergner, and J. Rodenbaugh (1989).

10.3 INFERENCE FOR THE ODDS RATIO (FISHER, CORNFIELD)

Although the χ^2 statistic, given by (10.18) of Section 10.1, measures, via the formal hypothesis test, the significance of association between two characteristics, it does *not* measure the *degree* of association. This is because χ^2 has the defect of depending not only on the true probabilities of landing in the four cells of the 2×2 table but also on the total number of subjects. One commonly used measure for association that does not have this defect and also is readily interpretable is the sample odds ratio. To introduce the sample odds ratio, $\widehat{\theta}$, we first define the corresponding population odds ratio parameter θ.

Recall the notation of Tables 10.1 and 10.3 and the joint probabilities given by display (10.24). The odds, given that the subject is in row 1 of the 2×2 table, that the subject will be in column 1 (instead of column 2) are

$$\theta^{(1)} = \frac{\dfrac{p_{11}}{p_{11}+p_{12}}}{\dfrac{p_{12}}{p_{11}+p_{12}}} = \frac{p_{11}}{p_{12}}. \tag{10.41}$$

In terms of the notation of Table 10.3 with characteristics C and D, (10.41) can be written as

$$\theta^{(1)} = \frac{P(C|D)}{P(\text{not } C|D)} = \frac{p_{11}}{p_{12}}. \tag{10.42}$$

Similarly, the odds, given that the subject is in row 2 of the 2×2 table, that the subject will be in column 1 (instead of column 2) are

$$\theta^{(2)} = \frac{P(C|\text{ not } D)}{P(\text{not } C|\text{ not } D)} = \frac{p_{21}}{p_{22}}. \tag{10.43}$$

The odds ratio is the parameter

$$\theta = \frac{\theta^{(1)}}{\theta^{(2)}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}. \tag{10.44}$$

The odds ratio can be any number between 0 and ∞. If the cell probabilities p_{11}, p_{12}, p_{21}, p_{22} are all positive, then independence of the characteristics C and D implies $\theta = 1$ and, conversely, $\theta = 1$ implies C and D are independent. If p_{11} or p_{22} is 0 (and p_{12} and p_{21} are not 0), then $\theta = 0$. If p_{12} or p_{21} is 0 (and p_{11} and p_{22} are not 0), then $\theta = \infty$. θ is undefined in each of the four cases (i) $p_{11} = 0$ and $p_{12} = 0$, (ii) $p_{21} = 0$ and $p_{22} = 0$, (iii) $p_{11} = 0$ and $p_{21} = 0$, and (iv) $p_{12} = 0$ and $p_{22} = 0$.

θ measures the strength of the association. A table having $1 < \theta < \infty$ is such that the probability of the subject landing in column 1, given that the subject is known to be in row 1, is higher than the probability that the subject will land in column 1, given that the subject is known to be in row 2. That is, in the notation of the conditional probabilities (see Problem 10), $p_{1|1}$ is greater than $p_{1|2}$. Correspondingly, a table for which $0 < \theta < 1$ has $p_{1|1} < p_{1|2}$.

Unconditional Procedures

Estimator of θ. The unconditional maximum likelihood estimator of θ is the sample odds ratio

$$\widehat{\theta} = \frac{\mathcal{O}_{11}\mathcal{O}_{22}}{\mathcal{O}_{12}\mathcal{O}_{21}}. \tag{10.45}$$

If $\mathcal{O}_{11}$ or $\mathcal{O}_{22}$ is 0 (and $\mathcal{O}_{12}$ and $\mathcal{O}_{21}$ are not zero), then $\widehat{\theta} = 0$. If $\mathcal{O}_{12}$ or $\mathcal{O}_{21}$ is 0 (and $\mathcal{O}_{11}$ and $\mathcal{O}_{22}$ are not zero), then $\widehat{\theta} = \infty$. $\widehat{\theta}$ is undefined in each of the four cases (i) $\mathcal{O}_{11} = 0$ and $\mathcal{O}_{12} = 0$, (ii) $\mathcal{O}_{21}$ and $\mathcal{O}_{22} = 0$, (iii) $\mathcal{O}_{11} = 0$ and $\mathcal{O}_{21} = 0$, and (iv) $\mathcal{O}_{l2} = 0$ and $\mathcal{O}_{22} = 0$. To eliminate these difficulties, we may use the adjusted version,

$$\widehat{\theta}_a = \frac{(\mathcal{O}_{11} + .5)(\mathcal{O}_{22} + .5)}{(\mathcal{O}_{12} + .5)(\mathcal{O}_{21} + .5)}. \tag{10.46}$$

(See Comment 20).

Confidence Intervals

We define

$$\nu = \ln(\theta) \tag{10.47}$$

and

$$\widehat{\nu} = \ln(\widehat{\theta}). \tag{10.48}$$

The distribution of $\widehat{\nu}$ converges more rapidly to its asymptotic distribution than does the distribution of $\widehat{\theta}$ to its asymptotic distribution. Thus, it is preferable to base tests and confidence intervals for the odds ratio on the asymptotic distribution of $\widehat{\nu}$.

The standard deviation of $\widehat{\nu}$ can be estimated by

$$\widehat{\mathrm{SD}}(\widehat{\nu}) = \sqrt{\frac{1}{\mathcal{O}_{11}} + \frac{1}{\mathcal{O}_{12}} + \frac{1}{\mathcal{O}_{21}} + \frac{1}{\mathcal{O}_{22}}}, \tag{10.49}$$

or the adjusted version

$$\widehat{\mathrm{SD}}(\widehat{\nu}) = \sqrt{\frac{1}{\mathcal{O}_{11} + .5} + \frac{1}{\mathcal{O}_{12} + .5} + \frac{1}{\mathcal{O}_{21} + .5} + \frac{1}{\mathcal{O}_{22} + .5}}. \tag{10.50}$$

The null hypothesis $\theta = 1$ is equivalent to the null hypothesis $\nu = 0$.

a. *Approximate One-Sided Upper-Tail Test.* To test

$$H_0 : \nu = 0$$

versus

$$H_1 : \nu > 0,$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } \frac{\widehat{\nu}}{\widehat{\mathrm{SD}}(\widehat{\nu})} \geq z_\alpha; \quad \text{otherwise do not reject.} \tag{10.51}$$

b. *Approximate One-Sided Lower-Tail Test.* To test

$$H_0 : \nu = 0$$

versus

$$H_2 : \nu < 0,$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } \frac{\widehat{\nu}}{\widehat{\text{SD}}(\widehat{\nu})} \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{10.52}$$

c. *Approximate Two-Sided Test.* To test

$$H_0 : \nu = 0$$

versus

$$H_3 : \upsilon \neq 0,$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } \left| \frac{\widehat{\nu}}{\widehat{\text{SD}}(\widehat{\nu})} \right| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{10.53}$$

For a symmetric two-sided confidence interval for ν, with the approximate confidence coefficient $1 - \alpha$, set

$$\nu_{\text{L}} = \widehat{\nu} - z_{\alpha/2}\widehat{\text{SD}}(\widehat{\nu}) \tag{10.54}$$

and

$$\nu_{\text{U}} = \widehat{\nu} + z_{\alpha/2}\widehat{\text{SD}}(\widehat{\nu}). \tag{10.55}$$

A confidence interval for the odds ratio θ is obtained by exponentiating ν_{L} and ν_{U} given by (10.54) and (10.55). That is, a symmetric two-sided confidence interval for θ, with the approximate confidence coefficient $1 - \alpha$, is $(\theta_{\text{L}}, \theta_{\text{U}})$, where

$$\theta_{\text{L}} = e^{\nu_{\text{L}}}, \tag{10.56}$$

$$\theta_{\text{U}} = e^{\nu_{\text{U}}}. \tag{10.57}$$

Exact conditional tests and confidence intervals for θ, which are more computationally tedious than the unconditional procedures, are presented in Comments 14 and 15, respectively.

EXAMPLE 10.4 ***Example 10.1 Continued.***

We return to the liver scan data of Table 10.2. From (10.45), we find

$$\widehat{\theta} = \frac{4(20)}{8(1)} = 10.$$

That is, we estimate the odds to be 10 times higher for black patients (with local or regional disease) to receive a liver scan than for white patients (with local or regional disease) to receive a liver scan.

From (10.47) and (10.48), we find, respectively,

$$\widehat{\nu} = \ln(\widehat{\theta}) = 2.30,$$

and

$$\widehat{\mathrm{SD}}(\widehat{\nu}) = \sqrt{\frac{1}{4} + \frac{1}{8} + \frac{1}{1} + \frac{1}{20}} = 1.19.$$

Computing $\widehat{\nu}/\widehat{\mathrm{SD}}(\widehat{\nu})$, we obtain

$$\frac{\widehat{\nu}}{\widehat{\mathrm{SD}}(\widehat{\nu})} = 1.93.$$

The corresponding approximate one-sided P-value, for testing $H_0 : \nu = 0$ versus $H_1 : \nu > 0$, is .027 (obtained from `pnorm (1.93, lower.tail = F)`). This is strong evidence that $\nu > 0$ or, equivalently, that the odds ratio θ is greater than 1.

A symmetric two-sided confidence interval for ν, with the approximate confidence coefficient .95, is found from (10.54) and (10.55) to be

$$\nu_{\mathrm{L}} = 2.30 - (1.96)(1.19) = -.03,$$

$$\nu_{\mathrm{U}} = 2.30 + (1.96)(1.19) = 4.63.$$

Exponentiating, we find (see (10.56) and (10.57))

$$\theta_{\mathrm{L}} = e^{-.03} = .97,$$

$$\theta_{\mathrm{U}} = e^{4.63} = 102.5.$$

The `R` command `oddsratio` from package `epitools` will implement this procedure. The data is specified as a 2×2 matrix. The argument `method` should be set to "`wald`". Some of the `R` output for this data is given below:

```
$data
      Yes No Total
Black   4  8    12
White   1 20    21
Total   5 28    33

$measure
  odds ratio with 95% C.I. estimate      lower     upper
                     White       10  0.9635896  103.7786

$p.value
 two-sided  midp.exact  fisher.exact  chi.square
     White  0.05047275    0.04713571  0.02766249
```

The appropriate P-value is found under the column labeled `chi.square`. The odds ratio estimate ($\hat{\theta} = 10$) and associated confidence interval are found in the `$measure` section. These results match (within rounding error) those obtained earlier. The P-value under the column heading `fisher.exact` is associated with the two-sided test from Section 10.2.

Comments

12. *Odds Ratio and Estimated Odds Ratio under Reversal of Rows and Columns.* If the roles of the rows and columns are reversed, then the odds ratio does not change. Reversing the roles of the rows and columns produces a new 2×2 table with underlying cell probabilities p'_{11}, p'_{12}, p'_{21}, p'_{22}, where $p'_{11} = p_{11}$, $p'_{12} = p_{21}$, $p'_{21} = p_{12}$, $p'_{22} = p_{22}$. Thus θ', the odds ratio for the new table, is

$$\theta' = \frac{p'_{11}p'_{22}}{p'_{12}p'_{21}} = \frac{p_{11}p_{22}}{p_{21}p_{12}} = \theta.$$

Similarly, the estimated odds ratio $\widehat{\theta}$ does not change when the roles of the rows and columns are reversed.

13. *Effect of Interchanging the Order of the Rows.* If rows 1 and 2 are interchanged, then the new table has underlying cell probabilities p'_{11}, p'_{12}, p'_{21}, p'_{22}, where $p'_{11} = p_{21}$, $p'_{12} = p_{22}$, $p'_{21} = p_{11}$, $p'_{22} = p_{12}$ and the odds ratio for the new table is

$$\theta' = \frac{p'_{11}p'_{22}}{p'_{12}p'_{21}} = \frac{p_{21}p_{12}}{p_{22}p_{11}} = \frac{1}{\theta},$$

where θ is the odds ratio for the original table. Similarly, if columns 1 and 2 of the original table are interchanged, the new odds ratio is the inverse of the original odds ratio. The same results also hold for the estimated odds ratio. That is, if the rows are interchanged (or if the columns are interchanged), $\widehat{\theta'}$, the estimated odds ratio for the new table, is the inverse of $\widehat{\theta}$, the estimated odds ratio for the original table.

14. *Exact Conditional Test for the Odds Ratio.* The conditional distribution of $\mathcal{O}_{11}$ given $n_{1.}$, $n_{.1}$, and θ (due to Fisher, 1935) is

$$P(\mathcal{O}_{11} = x | n_{1.}, n_{.1}, \theta) = \frac{\binom{n_{1.}}{x}\binom{n_{..} - n_{1.}}{n_{.1} - x}\theta^x}{\sum_{y=n_{\mathrm{L}}}^{n_{\mathrm{U}}}\binom{n_{1.}}{y}\binom{n_{..} - n_{1.}}{n_{.1} - y}\theta^y} \tag{10.58}$$

where (recall that) $n_{\mathrm{L}} = \max(0, n_{1.} + n_{.1} - n_{..})$, $n_{\mathrm{U}} = \min(n_{1.}, n_{.1})$. If $\mathcal{O}_{11}^{\mathrm{obs}}$ denotes the observed value of $\mathcal{O}_{11}$ for your 2×2 table, the P-value for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ is

$$P = \sum_{\{\text{all } x\text{-values} \geq \mathcal{O}_{11}^{\mathrm{obs}}\}} P(\mathcal{O}_{11} = x | n_{1.}, n_{.1}, \theta).$$

For testing $H_0 : \theta = \theta_0$ versus $H_2 : \theta < \theta_0$, the P-value is

$$P = \sum_{\{\text{all } x\text{-values} \leq \mathcal{O}_{11}^{\mathrm{obs}}\}} P(\mathcal{O}_{11} = x | n_{1.}, n_{.1}, \theta).$$

With the choice $\theta_0 = 1$, these procedures reduce to Fisher's exact test.

The exact conditional test's P-value can be obtained using the `R` command `fisher.test` (see Comment 17).

15. *Exact Conditional Confidence Intervals for the Odds Ratio.* Exact confidence intervals for θ can be obtained by inverting the tests of Comment 14. For a confidence interval (θ_L, θ_U), with confidence coefficient $\geq 1 - \alpha$, θ_L is the value of θ_0 for which the P-value is $\alpha/2$ when testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$.

    Similarly, θ_U is the value of θ_0 for which the P-value is $\alpha/2$ when testing $H_0 : \theta = \theta_0$ versus $H_2 : \theta < \theta_0$. The `R` command `fisher.test` can be used to obtain the exact conditional confidence interval. See Comment 17.

16. *Conditional Estimator of the Odds Ratio.* The conditional maximum likelihood estimator of θ is the value ($\tilde{\theta}$, say) of θ that maximizes the right-hand side of (10.38). This yields a different estimator than the unconditional maximum likelihood estimator $\widehat{\theta}$ given by (10.45). The value $\tilde{\theta}$ can be found by solving the equation $\mathcal{O}_{11} = E_c(\mathcal{O}_{11})$, where E_c denotes expectation with respect to the conditional distribution given by (10.58). This equation has a unique solution (Cornfield, 1956), which can be obtained by using iterative methods. $\tilde{\theta}$ can be obtained using the `R` command `fisher.test`. See Comment 17.

17. *Use of* `R` *to Perform the Conditional Procedures.* The conditional methods of Comments 14, 15, and 16 can be performed using the `fisher.test` command. For the data of Table 10.2, this gives the one-sided P-value .0471 for testing $\theta = 1$, agreeing with what we obtained in Example 10.3 using Fisher's exact test. (Recall that the tests of Comment 14 with $\theta_0 = 1$ reduce to Fisher's exact tests.) The `fisher.test` command also yields exact confidence intervals and the conditional maximum likelihood $\tilde{\theta}$. `R` gives the 95% confidence interval for θ as (`0.77,514.55`). The approximate unconditional 95% confidence interval for θ found using the results of Example 10.4 is (0.96, 103.78). The conditional maximum likelihood estimator of θ is 9.25. Recall from Example 10.4 that the unconditional maximum likelihood estimator of θ is 10. The `R` procedures are based on the algorithms due to Mehta and Patel (1986a, 1986b) and Clarkson et al. (1993). For a survey on exact methods for contingency tables, see Agresti (2013).

18. *Yule's Measure of Association.* The odds ratio θ measures association, but it takes values from 0 to ∞. If there is a preference for a measure that lies between -1 and 1, we can use Yule's (1900, 1912) Q, which is defined as

$$\begin{aligned} Q &= \frac{\theta - 1}{\theta + 1} \\ &= \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}}. \end{aligned} \tag{10.59}$$

    Its sample analog for an observed 2×2 table is

$$\begin{aligned} \widehat{Q} &= \frac{\widehat{\theta} - 1}{\widehat{\theta} + 1} \\ &= \frac{\mathcal{O}_{11}\mathcal{O}_{22} - \mathcal{O}_{12}\mathcal{O}_{21}}{\mathcal{O}_{11}\mathcal{O}_{22} + \mathcal{O}_{12}\mathcal{O}_{21}}. \end{aligned} \tag{10.60}$$

Since Q is an increasing function of θ, confidence limits for Q can be obtained from confidence limits for θ. In particular, an approximate $1-\alpha$ confidence interval for Q is (Q_L, Q_U), where

$$Q_L = \frac{\theta_L - 1}{\theta_L + 1}, \tag{10.61}$$

$$Q_U = \frac{\theta_U - 1}{\theta_U + 1}, \tag{10.62}$$

where θ_L and θ_U are given by (10.56) and (10.57), respectively.

For the liver scan data of Table 10.2, from (10.60), we find

$$\widehat{Q} = \frac{(10-1)}{10+1} = .82.$$

From (10.61) and (10.62) and Example 10.4, an approximate 90% confidence interval for Q is (Q_L, Q_U), where

$$Q_L = \frac{(1.40-1)}{(1.40+1)} = .17,$$

$$Q_U = \frac{(70.8-1)}{70.8+1} = .97.$$

19. *Odds Ratio under Binomial Sampling.* Recall the binomial model of Section 10.1. There $\mathcal{O}_{11}$ is the number of successes in $n_{1\cdot}$ independent Bernoulli trials, each with success probability p_1, and $\mathcal{O}_{21}$ is the number of successes in $n_{2\cdot}$ independent Bernoulli trials, each with success probability $p_{2\cdot}$ (Also, the sample 1 trials are assumed independent of the sample 2 trials.) Then, the odds ratio is defined to be

$$\theta = \frac{\dfrac{p_1}{1-p_1}}{\dfrac{p_2}{1-p_2}}. \tag{10.63}$$

20. *Adjusted Unconditional Estimator* $\widehat{\theta}_a$. The denominator of $\widehat{\theta}$ can be 0 with positive probability, and thus in particular, the mean and variance of $\widehat{\theta}$ do not exist. The adjusted estimator $\widehat{\theta}_a$ removes this difficulty. See Haldane (1955) and Gart and Zweiful (1967).

21. *Asymptotic Theory for* $\ln(\widehat{\theta})$. The delta method provides asymptotic normality results for functions g (that satisfy mild regularity conditions) of random variables, which themselves have an asymptotic multivariate normal distribution. The method produces an explicit expression for the variance of g, which in turn can be used to obtain a consistent estimator of that variance. The delta method is frequently used in categorical data analysis, where, under multinomial sampling, the cell entries, suitably standardized, have an asymptotic (singular) multivariate normal distribution (cf. Goodman and Kruskal (1963) and Agresti (2013, Section 3.1.7)). Applying the delta method to the function

$$g(\mathcal{O}_{11}, \mathcal{O}_{12}, \mathcal{O}_{21}, \mathcal{O}_{22}) = \ln\left(\frac{\mathcal{O}_{11}\mathcal{O}_{22}}{\mathcal{O}_{12}\mathcal{O}_{21}}\right) = \ln(\widehat{\theta}), \tag{10.64}$$

provides justification for procedures (10.51)–(10.55). The expression for the variance will depend on the p_{ij}'s that are then replaced by their sample counterparts $\mathcal{O}_{ij}/n_{..}$ to obtain a consistent estimator of the variance. Agresti (2013, Section 3.1.7) provides the elementary details for the case where g is given by (10.64).

Properties

1. *Bias of Odds Ratio Estimators.* See Haldane (1955), Gart and Zweiful (1967), and Agresti (2013, Section 3.1.1).
2. *Invariance Properties of Odds Ratio and Estimated Odds Ratio.* See Edwards (1963) and Comment 12.

Problems

18. For the data of Table 10.4, estimate θ and compute a 95% confidence interval for θ. Interpret your results.

19. For the data of Table 10.7, estimate θ and compute a 95% confidence interval for θ. Interpret your results.

20. For the data of Table 10.8, estimate θ and compute a 95% confidence interval for θ. Interpret your results.

21. For the data of Table 10.12, compute θ and obtain an approximate 95% confidence interval for θ.

22. The relative risk is defined to be

$$r = \frac{p_{1|1}}{p_{1|2}},$$

where (recall) $p_{1|1} = p_{11}/(p_{11} + p_{12})$ and $p_{1|2} = p_{21}/(p_{21} + p_{22})$. Show that if $r = 1$, then the hypothesis of independence holds and, conversely, independence implies $r = 1$.

23. Show that the odds ratio θ and the relative risk r satisfy

$$\theta = r \times \frac{1 - p_{1|2}}{1 - p_{1|1}}.$$

24. For the data of Table 10.4, compute $\widehat{Q}$ and obtain an approximate 95% confidence interval for Q.

25. For the data of Table 10.7, compute $\widehat{Q}$ and obtain an approximate 95% confidence interval for Q.

26. For the data of Table 10.8, compute $\widehat{Q}$ and obtain an approximate 95% confidence interval for Q.

27. For the data of Table 10.11, compute $\tilde{\theta}$ and compare it to $\widehat{\theta}$.

28. For the data of Table 10.12, get an exact 95% (conditional) confidence interval for θ.

10.4 INFERENCE FOR *k* STRATA OF 2 × 2 TABLES (MANTEL AND HAENSZEL)

Recall the Diehr et al. (1989) study discussed in Example 10.1 and Section 10.1. In that example, we computed the respective probabilities, for hospital 8, of white and black patients receiving liver scans. Now we wish to apply an overall test that assesses the

respective probabilities across the 19 hospitals (with the most black patients) in the Diehr et al. study.

More formally, suppose our data consist of k strata, and within each stratum we have a 2×2 table. The two rows of the 2×2 table in the ith stratum are viewed as data from two independent binomial distributions with respective success probabilities $(p_1^{(i)}, p_2^{(i)})$, $i = 1, 2, \ldots, k$. In the Diehr et al. study, $k = 19$ and the strata correspond to hospitals. Let

$$\begin{aligned} p_1^{(i)} &= \text{probability that a black patient in hospital } i \text{ (with local or regional disease) will receive a liver scan,} \\ p_2^{(i)} &= \text{probability that a white patient in hospital } i \text{ (with local or regional disease) will receive a liver scan.} \end{aligned}$$

The data in the ith 2×2 table can be represented in the notation of Table 10.13.

Mantel and Haenszel (1959) proposed an approximate test of the null hypothesis H_0, which specifies that within each stratum, the success probabilities are equal. That is

$$H_0 : p_1^{(1)} = p_2^{(1)},\ p_1^{(2)} = p_2^{(2)}, \ldots,\ p_1^{(k)} = p_2^{(k)}. \tag{10.65}$$

We let θ_i denote the odds ratio for the ith table. Recall that θ_i is defined as

$$\theta_i = \frac{p_1^{(i)}}{1 - p_1^{(i)}} \Big/ \frac{p_2^{(i)}}{1 - p_2^{(i)}}. \tag{10.66}$$

H_0 can be rewritten as

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k = 1. \tag{10.67}$$

That is, we are testing that there is a common odds ratio and it is equal to 1. Note that H_0 allows for the common success probabilities to differ from hospital to hospital. The alternatives of interest are that $p_1^{(i)} \geq p_2^{(i)}$ for all $i = 1, \ldots, k$ or $p_1^{(i)} \leq p_2^{(i)}$ for all $i = 1, \ldots, k$. In terms of the liver scan problem, the alternatives specify that across the 19 hospitals, the probabilities that black patients receive liver scans are higher than the respective probabilities that white patients receive liver scans or across the 19 hospitals the black patients have lower probabilities than those of the white patients.

Table 10.13 2×2 Table for ith Stratum

	Successes	Failures	Totals
Sample 1	$\mathcal{O}_{11i}$	$\mathcal{O}_{12i}$	$n_{1.i}$
Sample 2	$\mathcal{O}_{21i}$	$\mathcal{O}_{22i}$	$n_{2.i}$
Totals:	$n_{.1i}$	$n_{.2i}$	$n_{..i}$

Approximate Conditional Procedure

In the ith table, given that the marginal totals $n_{1.i}, n_{2.i}, n_{.1i}, n_{.2i}$, are fixed, the random variable $\mathcal{O}_{11i}$ has a hypergeometric distribution

$$P(\mathcal{O}_{11i} = x) = \frac{\binom{n_{1.i}}{x}\binom{n_{2.i}}{n_{.1i} - x}}{\binom{n_{..i}}{n_{.1i}}}. \tag{10.68}$$

The null mean $E_0(\mathcal{O}_{11i})$ is given by

$$E_0(\mathcal{O}_{11i}) = \frac{(n_{1.i})(n_{.1i})}{n_{..i}}, \tag{10.69}$$

and the null variance $\text{var}_0(\mathcal{O}_{11i})$ is given by

$$\text{var}_0(\mathcal{O}_{11i}) = \frac{(n_{1.i})(n_{2.i})(n_{.1i})(n_{.2i})}{n_{..i}^2(n_{..i} - 1)}. \tag{10.70}$$

The Mantel and Haenszel (1959) statistic is

$$\text{MH} = \frac{\sum_{i=1}^{k}\{\mathcal{O}_{11i} - E_0(\mathcal{O}_{11i})\}}{\sqrt{\sum_{i=1}^{k}\text{var}_0(\mathcal{O}_{11i})}}. \tag{10.71}$$

An approximate α-level one-sided test of H_0 given by (10.65) against the alternatives

$$H_1 : p_1^{(i)} \geq p_2^{(i)}, \quad i = 1, \ldots, k \tag{10.72}$$

(with at least one inequality strict) is

$$\text{Reject } H_0 \text{ if MH} \geq z_\alpha; \quad \text{otherwise do not reject.} \tag{10.73}$$

An approximate α-level one-sided test of H_0 against the alternatives

$$H_2 : p_1^{(i)} \leq p_2^{(i)}, \quad i = 1, \ldots, k \tag{10.74}$$

(with at least one inequality strict) is

$$\text{Reject } H_0 \text{ if MH} \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{10.75}$$

An approximate α-level two-sided test of H_0 against the alternatives

$$H_3 : p_1^{(i)} \geq p_2^{(i)} \quad \text{for all } k \text{ or } p_1^{(i)} \leq p_2^{(i)} \text{ for all } k \tag{10.76}$$

(with at least one inequality strict) is

$$\text{Reject } H_0 \text{ if } (\text{MH})^2 \geq \chi^2_{\alpha,1}; \quad \text{otherwise do not reject.} \tag{10.77}$$

EXAMPLE 10.5 ***Liver Scan Data for 19 Hospitals.***

Table 10.14 gives, for the Diehr et al. (1989) study, the percent of patients with local or regional disease receiving liver scan. The data are for the 19 hospitals with the most black patients.

In Table 10.15, we give the values of $E_0(\mathcal{O}_{11i})$ and $\text{var}_0(\mathcal{O}_{11i})$, obtained using (10.69) and (10.70), respectively.

The 19 2×2 tables formed from Table 10.14 are as follows:

Hospital 1

4	9	13
12	34	46
16	43	59

Hospital 2

4	6	10
34	33	67
38	39	77

Hospital 3

7	2	9
6	7	13
13	9	22

Hospital 4

5	5	10
59	56	115
64	61	125

Hospital 5

7	7	14
22	69	91
29	76	105

Hospital 6

5	6	11
41	80	121
46	86	132

Hospital 7

3	6	9
8	72	80
11	78	89

Hospital 8

4	8	12
1	20	21
5	28	33

Hospital 9

7	2	9
77	38	115
84	40	124

Hospital 10

4	6	10
20	70	90
24	76	100

Hospital 11

1	8	9
16	76	92
17	84	101

Hospital 12

4	10	14
10	91	101
14	101	115

Hospital 13

9	18	27
27	118	145
36	136	172

Hospital 14

3	5	8
35	45	80
38	50	88

Hospital 15

9	5	14
69	20	89
78	25	103

Hospital 16

2	3	5
3	12	15
5	15	20

Hospital 17

6	1	7
45	31	76
51	32	83

Hospital 18

14	10	24
12	70	82
26	80	106

Hospital 19

15	15	30
43	129	172
58	144	202

From (10.71) and Table 10.15, we obtain

$$\text{MH} = \frac{113 - 81.004}{\sqrt{39.383}} = 5.10.$$

Table 10.14 Percent of Patients with Local or Regional Disease Receiving Liver Scans in 19 Hospitals with the Most Black Patients

Hospital	% with scan (number of patients eligible)			
	White		Black	
1	26.1	(46)	30.8	(13)
2	50.8	(67)	40.0	(10)
3	46.2	(13)	77.8	(9)
4	51.3	(115)	50.0	(10)
5	24.2	(91)	50.0	(14)
6	33.9	(121)	45.5	(11)
7	10.0	(80)	33.3	(9)
8	4.8	(21)	33.3	(12)
9	67.0	(115)	77.8	(9)
10	22.2	(90)	40.0	(10)
11	17.4	(92)	11.1	(9)
12	9.9	(101)	28.6	(14)
13	18.6	(145)	33.3	(27)
14	43.8	(80)	37.5	(8)
15	77.5	(89)	64.3	(14)
16	20.0	(15)	40.0	(5)
17	59.2	(76)	85.7	(7)
18	14.6	(82)	58.3	(24)
19	25.0	(172)	50.0	(30)

Table 10.15 Null Means and Variances of $\mathcal{O}_{11i}$ for Liver Scan Data

i	$\mathcal{O}_{11i}$	$E_0(\mathcal{O}_{11i})$	$\text{var}_0(\mathcal{O}_{11i})$
1	4	3.525	2.038
2	4	4.935	2.204
3	7	5.318	1.347
4	5	5.120	2.317
5	7	3.867	2.449
6	5	3.833	2.307
7	3	1.112	0.886
8	4	1.818	1.012
9	7	6.097	1.839
10	4	2.400	1.658
11	1	1.515	1.159
12	4	1.704	1.326
13	9	5.651	3.789
14	3	3.455	1.805
15	9	10.602	2.245
16	2	1.250	0.740
17	6	4.301	1.537
18	14	5.887	3.470
19	15	8.614	5.255
	$\sum_{i=1}^{19} \mathcal{O}_{11i} = 113$	$\sum_{i=1}^{19} E(\mathcal{O}_{11i}) = 81.004$	$\sum_{i=1}^{19} \text{var}_0(\mathcal{O}_{11i}) = 39.383$

For MH $= 5.10$, the two-sided P-value associated with the test at (10.73) is found to be approximately 1.7×10^{-7} using `pnorm(5.1, lower.tail=F)`. Thus, there is very strong evidence that the hospitals do not have a common odds ratio that is 1. Since MH is significantly positive, there is strong evidence that across hospitals the odds that black patients get liver scans are higher than the odds that white patients get liver scans.

The Mantel–Haenszel test is implemented in `R` with the command `mantelhaen.test`. The data is input as a collection of 2×2 matrices in a three-dimensional array format. One- and two-sided tests are available through the argument `alternative`. The test statistic is given as $(\text{MH})^2$ for all tests, rather than MH. Using the data from Example 10.5, the following output is produced by `mantelhaen.test(x, correct=F, alternative="g")`:

```
   Mantel-Haenszel chi-squared test without continuity
      correction

data:    x
Mantel-Haenszel X-squared = 25.9938, df = 1, p-value = 1.713e-07
alternative hypothesis:  true common odds ratio is greater
    than 1
```

Comments

22. *Types of Alternatives the Mantel–Haenszel Test Detects.* Consider the one-sided procedure given by (10.73). This procedure rejects H_0 for significantly large values of MH. This test will be consistent against H_1 (given by (10.72)) if at least one of the inequalities is strict. However, if for some strata $p_1^{(i)} > p_2^{(i)}$ and for others the inequality goes in the other direction, the MH test will have less power because the statistic would then tend to add positive and negative deviations of the form $\{\mathcal{O}_{11i} - E_0(\mathcal{O}_{11i})\}$. Roughly speaking, the one-sided procedure based on (10.73) is consistent against alternatives A for which $E_A\left(\sum_{i=1}^{k}\{\mathcal{O}_{11i} - E_0(\mathcal{O}_{11i})\}\right) > 0$, where E_A indicates that the expectation is computed under alternative A. Similar considerations apply to the one-sided procedure given by (10.75) and the two-sided procedure given by (10.77).
23. *The Mantel–Haenszel Test Viewed as a Test of Conditional Independence.* Suppose the k 2×2 tables are tables of cross-classified data and in each stratum we are interested in whether factors C and D are (conditionally) independent. The alternative of positive association would correspond to C and D being positively associated across strata. The one-sided procedure given by (10.73) tests conditional independence against alternatives that C and D are positively associated across strata. As in Comment 22, the MH test will have less power if for some strata C and D are positively associated and, for others, C and D are negatively associated. Similar considerations apply to the one-sided procedure given by (10.75) and the two-sided procedure defined by (10.77).
24. *Zelen's Exact Test for a Common Odds Ratio.* The null hypothesis H_0, given by (10.67), specifies that the strata have a common odds ratio equal to 1. Zelen (1971) developed an exact conditional test of

$$H_0^* : \theta_1 = \theta_2 = \cdots = \theta_n = \theta \text{ (say)}, \tag{10.78}$$

where the common value θ is unspecified. The hypothesis H_0^* is called the *hypothesis of homogeneity*.

Recall Table 10.13 and let τ denote any collection of k 2×2 tables. Furthermore, define

$$T = \{\tau : \text{table } i \text{ in } \tau \text{ has marginal totals } n_{.1i}, n_{.2i}, n_{1.i}, n_{2.i}\}. \tag{10.79}$$

The $\mathcal{O}_{11i}$ values can vary, but the marginal totals are fixed. Zelen's test is based on the restricted reference set

$$T(s) = \left\{ \tau : \tau \in T \text{ and } \sum_{i=1}^{k} \mathcal{O}_{11i} = s \right\}. \tag{10.80}$$

Note that $T(s)$ is contained in T. The conditional probability of obtaining a specific set of k 2×2 tables that is a member of $T(s)$ is

$$P(\tau|s) = \frac{\prod_{i=1}^{k} \binom{n_{1.i}}{\mathcal{O}_{11i}} \binom{n_{2.i}}{\mathcal{O}_{21i}} \Big/ \binom{n_{..i}}{n_{.1i}}}{\sum_{\tau \in T} \prod_{i=1}^{k} \binom{n_{1.i}}{\mathcal{O}_{11i}} \binom{n_{2.i}}{\mathcal{O}_{21i}} \Big/ \binom{n_{..i}}{n_{..i}}}. \tag{10.81}$$

Zelen's test is conditional on $\sum_{i=1}^{k} \mathcal{O}_{11i}$ as well as the marginal totals $n_{.1i}, n_{.2i}$, $n_{1.i}, n_{2.i}$ for each table. For all k 2×2 tables having the same marginals as the observed marginals, the probability given by (10.81) is calculated. The tables are then ordered according to those probabilities. The P-value is the sum of those probabilities for those tables whose probabilities are less than or equal to the probability of the observed table. That is, suppose τ_0 is our observed set of k 2×2 tables with marginal totals $n_{.1i}, n_{.2i}, n_{1.i}, n_{2.i}$ for the ith table and suppose that for τ_0, we have $\sum_{i=1}^{k} \mathcal{O}_{11i} = s$. Let

$$T^*(s) = \{\tau : \tau \in T(s) \text{ and } P(\tau|s) \leq P(\tau_0|s)\}.$$

Then, the P-value for Zelen's test is

$$P = \sum_{\tau \in T^*(s)} P(\tau|s).$$

This P-value is two-sided; the test does not indicate the direction of a deviation from H_0^*.

For example, suppose $k = 2$ and our observed set of two tables is

$$\tau_0 = \begin{Bmatrix} 2 & 2 & 1 & 4 \\ 1 & 5, & 5 & 1 \end{Bmatrix}.$$

Thus, our observed value is $\mathcal{O}_{111} + \mathcal{O}_{112} = 3$. The four tables in the reference set $T(3)$ and their conditional probabilities given by (10.81) are as follows.

	Table	Conditional probability
1.	$\begin{Bmatrix} 0 & 4 & 3 & 2 \\ 3 & 3, & 3 & 3 \end{Bmatrix}$	$\dfrac{\left[\left\{\binom{4}{0}\binom{6}{3}\Big/\binom{10}{3}\right\}\cdot\left\{\binom{5}{3}\binom{6}{3}\Big/\binom{11}{6}\right\}\right]}{\Sigma} = .2841$
2.	$\begin{Bmatrix} 1 & 3 & 2 & 3 \\ 2 & 4, & 4 & 2 \end{Bmatrix}$	$\dfrac{\left[\left\{\binom{4}{1}\binom{6}{2}\Big/\binom{10}{3}\right\}\cdot\left\{\binom{5}{2}\binom{6}{4}\Big/\binom{11}{6}\right\}\right]}{\Sigma} = .6387$
3.	$\begin{Bmatrix} 2 & 2 & 1 & 4 \\ 1 & 5, & 5 & 1 \end{Bmatrix}$	$\dfrac{\left[\left\{\binom{4}{2}\binom{6}{1}\Big/\binom{10}{3}\right\}\cdot\left\{\binom{5}{1}\binom{6}{5}\Big/\binom{11}{6}\right\}\right]}{\Sigma} = .0767$
4.	$\begin{Bmatrix} 3 & 1 & 0 & 5 \\ 0 & 6, & 6 & 0 \end{Bmatrix}$	$\dfrac{\left[\left\{\binom{4}{3}\binom{6}{0}\Big/\binom{10}{3}\right\}\cdot\left\{\binom{5}{0}\binom{6}{6}\Big/\binom{11}{6}\right\}\right]}{\Sigma} = .0004,$

where

$$\begin{aligned}\Sigma = &\left\{\binom{4}{0}\binom{6}{3}\Big/\binom{10}{3}\right\}\cdot\left\{\binom{5}{3}\binom{6}{3}\Big/\binom{11}{6}\right\} \\ &+\left\{\binom{4}{1}\binom{6}{2}\Big/\binom{10}{3}\right\}\cdot\left\{\binom{5}{2}\binom{6}{4}\Big/\binom{11}{6}\right\} \\ &+\left\{\binom{4}{2}\binom{6}{1}\Big/\binom{10}{3}\right\}\cdot\left\{\binom{5}{1}\binom{6}{5}\Big/\binom{11}{6}\right\} \\ &+\left\{\binom{4}{3}\binom{6}{0}\Big/\binom{10}{3}\right\}\cdot\left\{\binom{5}{0}\binom{6}{6}\Big/\binom{11}{6}\right\}\end{aligned}$$

$$= .0722 + .1623 + .0195 + .0001 = .2541.$$

Thus, the four tables have conditional probabilities $.0722/.2541 = .2841$, $.1623/.2541 = .6387$, $.0193/.2541 = .0767$, and $.0001/.2541 = .0004$. Of these four tables, Table 4 has the lowest probability and Table 3 (which is τ_0) has the second lowest. The P-value is

$$P = .0004 + .0767 = .077.$$

Thus, the observed data do not support H_0^*. For small data sets, Zelen's test is performed with the command `zelen.test`. This test requires an input array of k 2×2 matrices. For the example data, the P-value found is again .077.

Mehta, Patel, and Gray (1985) developed an algorithm for performing Zelen's test and inverting the test to get a confidence interval for the common odds ratio.

An approximate P-value based on a large-sample test of H_0^* proposed by Breslow and Day (1980) is available through the `R` command `rma.mhs`. See Comment 25. The Breslow–Day statistic has, under H_0^*, an asymptotic chi-squared distribution with $k - 1$ degrees of freedom. See Breslow and Day (1980) and Jones et al. (1989). Jones et al. compared seven tests of H_0^*, including the Breslow–Day test. Jones et al. found generally low power for all the tests of H_0^* in the study, especially when the data are sparse. In sparse-data situations, a test due to Liang and Self (1985) performed best.

25. *Monte Carlo Sampling to Estimate the Exact P-Value for Zelen's Test.* For some large data sets, computation of the exact P-value for Zelen's test is not feasible. Senchaudhuri, Mehta, and Patel (1995) presented a Monte Carlo method of control variates that, with a little extra computational effort, can estimate the exact P-value with greater accuracy than can be obtained by crude Monte Carlo sampling. For precise descriptions of the method of control variates and the method of crude Monte Carlo sampling, see Senchaudhuri, Mehta, and Patel (1995). Both Monte Carlo methods sample a large number N (say) of tables τ_i from the reference set given by (10.80) with respective probabilities $P(\tau_i|s)$ given by (10.81). Senchaudhuri, Mehta, and Patel (1995) presented the data set (Dixon v. Margolis 1991) given in Table 10.16. The data may be viewed in the format of 12 2×2 tables. The data related to promotions, in 1985, 1987, and 1988, of black and white police officers in various ranks.

The software StatXact (2010) can be used to obtain a confidence interval for Zelen's P-value based on the method of control variates. It produces an interval based on a sample of a user-specified number of tables. For a sample

Table 10.16 Promotions of Black and White Police Officers

		Black		White	
Year	Rank	Promoted	Total	Promoted	Total
1985	Sgt.	10	84	66	414
	SA Sgt.	3	28	40	110
	Master Sgt.	3	12	37	162
	SA Master Sgt.	0	2	16	62
1987	Sgt.	1	98	32	487
	SA Sgt.	0	29	28	120
	Master Sgt.	3	17	28	176
	SA Master Sgt.	2	5	16	65
1988	Sgt.	4	107	43	591
	SA Sgt.	1	36	4	113
	Master Sgt.	2	20	43	198
	SA Master Sgt.	1	5	18	112

Source: P. Senchaudhuri, C R. Mehta, and N. R. Patel (1995).

of 10,000 tables, the corresponding 99% confidence interval for the P-value of Zelen's test was (.042, .055). Thus, there is strong evidence against H_0^*. That is, the data indicate that there is not a common odds ratio. (Senchaurdhuri et al. sampled 100,000 tables and their 99% confidence interval for the P-value was (.0464, .0467).)

The Breslow and Day (1980) statistic for this data is obtained using the `R` command `rma.mh` from `package metafor` (Viechtbauer (2010)). The data may be entered by providing four input vectors, one for each cell of the 2×2 tables. Assuming that these vectors are labeled `O.11, O.12, O.21` and `O.22`, the command is performed with `rma.mh(O.11, O.12, O.21, O.22)`. The Breslow–Day statistic and P-value are accessed with `$BD` and `$BDp`, respectively. The numeric values for these statistics and the data in Table 10.16 are 16.02 and 0.14.

26. *Point and Interval Estimation of a Common Odds Ratio.* If we assume (perhaps on the basis of Zelen's exact test of H_0^*) that there is a common odds ratio θ, it is natural to obtain a point estimate and confidence interval for θ. Mantel and Haenszel (1959) suggested

$$\widehat{\theta}_{\text{MH}} = \frac{\sum_{i=1}^{k} (\mathcal{O}_{11i}\mathcal{O}_{22i}/n_{..i})}{\sum_{i=1}^{k} (\mathcal{O}_{12i}\mathcal{O}_{21i}/n_{..i})}. \tag{10.82}$$

Robins, Breslow, and Greenland (1986) estimated the variance of $\log(\widehat{\theta}_{\text{MH}})$ by

$$\begin{aligned}\widehat{\sigma}^2_{\text{RGB}} = {} & \frac{\sum_{i=1}^{k} (\mathcal{O}_{11i} + \mathcal{O}_{22i})(\mathcal{O}_{11i}\mathcal{O}_{22i})/n_{..i}^2}{2\left(\sum_{i=1}^{k} \mathcal{O}_{11i}\mathcal{O}_{22i}/n_{..i}\right)^2} \\ & + \frac{\sum_{i=1}^{k} \left\{(\mathcal{O}_{11i} + \mathcal{O}_{22i})(\mathcal{O}_{12i}\mathcal{O}_{21i}) + (\mathcal{O}_{12i} + \mathcal{O}_{21i})(\mathcal{O}_{11i}\mathcal{O}_{22i})\right\}/n_{..i}^2}{2\left(\sum_{i=1}^{k} \mathcal{O}_{11i}\mathcal{O}_{22i}/n_{..i}\right)\left(\sum_{i=1}^{k} \mathcal{O}_{12i}\mathcal{O}_{21i}/n_{..i}\right)} \\ & + \frac{\sum_{i=1}^{k} (\mathcal{O}_{12i} + \mathcal{O}_{21i})(\mathcal{O}_{12i}\mathcal{O}_{21i})/n_{..i}^2}{2\left(\sum_{i=1}^{k} \mathcal{O}_{12i}\mathcal{O}_{21i}/n_{..i}\right)^2}. \end{aligned} \tag{10.83}$$

An approximate $1 - \alpha$ confidence interval for $\ln(\theta)$ is $(\psi_{\text{L}}, \psi_{\text{U}})$, where

$$\psi_{\text{L}} = \ln(\widehat{\theta}_{\text{MH}}) - z_{\alpha/2}\widehat{\sigma}_{\text{RGB}} \tag{10.84}$$

$$\psi_{\text{U}} = \ln(\widehat{\theta}_{\text{MH}}) + z_{\alpha/2}\widehat{\sigma}_{\text{RGB}} \tag{10.85}$$

and where $\widehat{\theta}_{\text{MH}}$ is given by (10.82) and $\widehat{\sigma}^2_{\text{RGB}}$ is given by (10.83). An approximate $1 - \alpha$ confidence interval for the common odds ratio θ is $(\theta_{\text{L}}, \theta_{\text{U}})$, where

$$\theta_{\text{L}} = e^{\psi_{\text{L}}}, \tag{10.86}$$

$$\theta_{\text{U}} = e^{\psi_{\text{U}}}, \tag{10.87}$$

where $\psi_{\text{L}}, \psi_{\text{U}}$ are given by (10.84) and (10.85), respectively.

The command `rma.mh` will compute $\widehat{\theta}_{\text{MH}}$ and the confidence interval. Partial output of the procedure using the data in Table 10.16 is

```
  Model Results (OR scale):

estimate     ci.lb     ci.ub
  0.4785    0.3232    0.7084
```

Thus, the estimate $\widehat{\theta}_{\text{MH}}$ of θ is .48 and the approximate 95% confidence interval for θ is (.32, .71).

Properties

1. *Relationship of Mantel–Haenzel Statistic to Cochran's (1954) Statistic.* See Fleiss (2003) and Agresti (2013, Section 6.4.2).
2. *Efficiency of Mantel–Haenzel Estimator of a Common Odds Ratio.* See Hauck (1979), Breslow (1981), Donner and Hauck (1986), Hauck and Donner (1988), and Section 10.5.

Problems

29. Mittal (1991) considered pooling tables and the paradoxes that could arise from such pooling. For example, with the two 2×2 tables

Table 1		Table 2	
$\mathcal{O}_{111}$	$\mathcal{O}_{121}$	$\mathcal{O}_{112}$	$\mathcal{O}_{122}$
$\mathcal{O}_{211}$	$\mathcal{O}_{221}$	$\mathcal{O}_{212}$	$\mathcal{O}_{222}$

the pooled 2×2 table is

Table 3

$\mathcal{O}_{111} + \mathcal{O}_{112}$	$\mathcal{O}_{121} + \mathcal{O}_{122}$
$\mathcal{O}_{211} + \mathcal{O}_{212}$	$\mathcal{O}_{221} + \mathcal{O}_{222}$.

It may happen that the indication obtained by analyzing Table 3 could be different than the individual indications obtained by analyzing Tables 1 and 2. For example, it is possible that the estimated odds ratios $\widehat{\theta}_1$ and $\widehat{\theta}_2$ for Tables 1 and 2, respectively, satisfy

$$\widehat{\theta}_1 = \frac{\mathcal{O}_{111}\mathcal{O}_{221}}{\mathcal{O}_{121}\mathcal{O}_{221}} \geq 1, \quad \widehat{\theta}_2 = \frac{\mathcal{O}_{112}\mathcal{O}_{222}}{\mathcal{O}_{122}\mathcal{O}_{212}} \geq 1$$

but the estimated odds ratio $\widehat{\theta}_3$ for Table 3 satisfies

$$\widehat{\theta}_3 = \frac{(\mathcal{O}_{111} + \mathcal{O}_{112}) \times (\mathcal{O}_{221} + \mathcal{O}_{222})}{(\mathcal{O}_{121} + \mathcal{O}_{122}) \times (\mathcal{O}_{211} + \mathcal{O}_{212})} \leq 1.$$

Or it may be that $\widehat{\theta}_1 \leq 1$ and $\widehat{\theta}_2 \leq 1$ but $\widehat{\theta}_3 \geq 1$. Such anomalies fall into a category known as *Simpson's Paradox*. Create an example satisfying Simpson's Paradox.

Table 10.17 Fractions of Students with Tonsillectomies

Age	18	19	20	21	22	23	24
IM	$\frac{6}{23}$	$\frac{3}{42}$	$\frac{12}{41}$	$\frac{8}{46}$	$\frac{5}{15}$	$\frac{2}{9}$	$\frac{4}{9}$
C	$\frac{17}{49}$	$\frac{26}{96}$	$\frac{34}{112}$	$\frac{48}{139}$	$\frac{45}{118}$	$\frac{29}{66}$	$\frac{36}{75}$

Source: R. G. Miller (1980).

30. Apply the Mantel–Haenszel test to the data of Table 10.16. What is your conclusion? Is the conclusion surprising in view of the test results obtained in Comment 25?

31. Recall the study of R. Goode and D. Coursey reported in Miller (1980) and considered in Problem 4. They surveyed students seen at Stanford University's Student Health Center between January 1968 and May 1973. They checked the charts of students treated for infectious mononucleosis (IM) for confirmation of the disease and to determine any history of tonsillectomy. The control group consisted of students seen at the Health Center between April and September 1973, who came in for any ailment and were willing to check on a survey sheet whether or not they had undergone a tonsillectomy. Within the 18–24 age groups, the data are given in Table 10.17. Do the data indicate that tonsillectomy reduces the risk of contracting infectious mononucleosis?

32. Perform Zelen's test (see Comments 24 and 25) to the data of Table 10.17. What is your conclusion?

33. Suppose we have 3 2×2 tables of the form

	Successes	Failures
Treatment A	$\mathcal{O}_{11i}$	$\mathcal{O}_{12i}$
Treatment B	$\mathcal{O}_{21i}$	$\mathcal{O}_{22i}$

and the data are as follows:

Table 1		Table 2		Table 3	
2	5	4	11	3	11
1	6	1	7	2	9

Perform Zelen's test of H_0^*. What do you conclude?

34. Refer to the three 2×2 tables of Problem 33. Assuming a common odds ratio θ, compute $\widehat{\theta}_{\text{MH}}$ (see Comment 26) and obtain an approximate 95% confidence interval for θ.

35. Even if the evidence in a given data set indicates that H_0^* is not true, explain the potential usefulness of computing $\widehat{\theta}_{\text{MH}}$ for that data set.

36. Suppose $k = 2$ and our observed set of two 2×2 tables is $\tau_0 = \begin{Bmatrix} 3 & 2 & 1 & 4 \\ 1 & 5, & 5 & 1 \end{Bmatrix}$

(a) List the tables in the reference set $T(4)$ defined by (10.80).

(b) Compute $P(\tau|4)$ for each $\tau \in T(4)$.

(c) Find the P-value for Zelen's test corresponding to the observed table τ_0.

10.5 EFFICIENCIES

Storer and Kim (1990) studied the size and power of seven tests for the two-sample binomial problem considered in Sections 10.1 and 10.2. The tests compared included the uncorrected chi-squared statistic given by (10.19) in Section 10.1, the chi-squared statistic with Yates' continuity correction (Comment 5), and Fisher's exact test (Section 10.2). In the equal sample-sizes case, an exact unconditional test due to Suissa and Shuster (1985) does quite well in terms of power. In the general case, the authors find the sample sizes required by Fisher's exact test are 10–20% higher than those required by some of the more powerful procedures considered. One of their conclusions is that for any reasonable sample size, one will not be led far astray by using the uncorrected chi-squared statistic.

For estimating the odds ratio (Sections 10.3 and 10.4), Hauck and Donner (1988) studied the asymptotic relative efficiency (ARE) of the Mantel–Haenszel estimator $\widehat{\theta}_{\mathrm{MH}}$ (Comment 26) relative to the conditional maximum likelihood estimator proposed by Birch (1964). They assumed a common odds ratio $\theta_i = \theta$, $i = 1, \ldots, k$, and $0 < \theta < \infty$. In this situation, $p_1^{(i)}$ can be expressed in terms of θ and $p_2^{(i)}, p_1^{(i)} = (\theta p_2^{(i)})/\{1 - p_2^{(i)} + \theta p_2^{(i)}\}$ and the ARE can be studied by varying the $p_1^{(i)}, i = 1, \ldots, k$, and θ. Hauck and Donner (1988) considered the case where the number of strata increases indefinitely for fixed within-stratum sample sizes. For the situations they considered, the ARE does not drop below .9. They pointed out the ARE decreases monotonically as θ moves away from 1 in either direction. For the less extreme values of θ, $.2 \leq \theta \leq .5$, the smallest ARE is .931. Hauck and Donner (1988) pointed out that the results they obtained in their 1988 study are similar to ones they obtained (Donner and Hauck (1986)) for the fixed-number-of-strata asymptotic case where the ARE was found to be high over a wide range of designs likely to arise in practice. For other efficiency results concerning $\widehat{\theta}_{\mathrm{MH}}$, see Breslow (1981) and Tarone, Gart, and Hauck (1983).

Chapter 11

Life Distributions and Survival Analysis

INTRODUCTION

In Sections 11.1–11.4 we consider a sample of lifelengths from an underlying life distribution (i.e., a distribution that puts all of its probability on nonnegative values). There are many nonparametric classes of life distributions that are used to describe aging. We focus on six natural classes that have easily understood physical interpretations. The classes are the increasing failure rate (IFR) class, the increasing failure rate average (IFRA) class, the new better than used (NBU) class, the new better than used in expectation (NBUE) class, the decreasing mean residual life (DMRL) class, and the initially increasing then decreasing mean residual life (IDMRL) class. In Sections 11.1–11.4 we describe tests of the null hypothesis of exponentiality. Section 11.1 considers IFR and IFRA alternatives. Section 11.2 considers NBU and NBUE alternatives. Section 11.3 considers DMRL alternatives and also presents confidence bands for the mean residual life function. Section 11.4 considers IDMRL alternatives, and the tests presented are designed to detect a trend change in the mean residual life. Section 11.5 presents a nonparametric confidence band for the distribution function. Sections 11.6 and 11.7 are devoted to censored data. Section 11.6 contains, for censored data, an estimator of the distribution function, confidence bands for the distribution function, and confidence bands for the quantile function. Section 11.7 presents a two-sample test for censored data. Section 11.8 considers asymptotic relative efficiencies.

Data. We obtain n observations, $X_1, \ldots, X_n$.

Assumptions for Sections 11.1–11.4

A1. The observations are a random sample from the underlying continuous population; that is, the X's are independent and identically distributed according to a continuous distribution F.

A2. F is a life distribution; that is, $F(a) = 0$ for $a < 0$. Equivalently, the X's are nonnegative.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

11.1 A TEST OF EXPONENTIALITY VERSUS IFR ALTERNATIVES (EPSTEIN)

Hypothesis

We first define the failure rate function, denoted by $r(x)$. If F has a corresponding density f, $r(x)$ is defined as

$$r(x) = \frac{f(x)}{\overline{F}(x)} \tag{11.1}$$

for those x such that $\overline{F}(x) > 0$. ($\overline{F}(x) = 1 - F(x)$ and $\overline{F}(x)$ is known as the survival function.) The failure rate is also called the hazard rate.

The failure rate has the following physical interpretation. The product $r(x)\delta_x$ is the probability that an item (unit, person, part) alive at age x will fail in the interval $(x, x + \delta_x)$, where δ_x is small. An IFR corresponds to deleterious aging; that is, the failure rate increases as age increases. A decreasing failure rate (DFR) corresponds to beneficial aging; that is, the failure rate decreases as age increases. A constant failure rate corresponds to a model where the failure rate neither increases nor decreases with age but is, in fact, independent of age.

The null hypothesis is

$$H_0 : r(x) = \lambda, \text{ for some } \lambda > 0, \text{ and all } x > 0. \tag{11.2}$$

The null hypothesis asserts that the failure rate is a constant λ (λ is unspecified); that is, the failure rate does not depend on x. The null hypothesis specifies that F is an exponential distribution. (One characterization of exponential distributions is that F is an exponential distribution if and only if its failure rate is constant.) Thus H_0 can be rewritten as

$$H_0 : F(x) = \begin{cases} 1 - e^{-\lambda x} \ (\lambda \text{ unspecified}), & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{11.3}$$

Equivalently, H_0 can be expressed as

$$H_0 : \overline{F}(x) = \begin{cases} e^{-\lambda x} \ (\lambda \text{ unspecified}), & x \geq 0, \\ 1, & x < 0. \end{cases} \tag{11.4}$$

A life distribution F is said to be in the *increasing failure rate* (IFR) class if its failure rate is nondecreasing. Similarly, a life distribution F is said to be in the *decreasing failure rate* (DFR) class if its failure rate is nonincreasing. For mathematically formal definitions of these classes, see, for example, Barlow and Proschan (1981). Whereas the IFR class is used to model deleterious aging, the DFR class is used to model beneficial aging. When the failure rate r exists, r is IFR if

$$r(x) \leq r(y) \quad \text{for all } x < y. \tag{11.5}$$

Similarly, when the failure rate r exists, r is DFR if

$$r(x) \geq r(y) \quad \text{for all } x < y. \tag{11.6}$$

Procedure

Let $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ denote the order statistics and define $X_{(0)} = 0$. The normalized spacings are $D_1, D_2, \ldots, D_n$, where

$$D_i = (n - i + 1)(X_{(i)} - X_{(i-1)}), \quad i = 1, \ldots, n. \tag{11.7}$$

Let

$$S_i = \sum_{u=1}^{i} D_u, \quad i = 1, \ldots, n \tag{11.8}$$

and define $S_0 = 0$. S_i is called the *total time on test* at $X_{(i)}$. The total-time-on-test statistic is

$$\mathcal{E} = \frac{\sum_{i=1}^{n-1} S_i}{S_n}. \tag{11.9}$$

a. *One-Sided Test against IFR Alternatives.* To test

$$H_0 : F \text{ is exponential}$$

versus

$$H_1 : F \text{ is IFR (and not exponential)},$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } \mathcal{E} \ge e_\alpha; \quad \text{otherwise do not reject,} \tag{11.10}$$

where the constant e_α is chosen to make the type I error probability equal to α; that is, $P_0\{\mathcal{E} \ge e_\alpha\} = \alpha$. The *R* function `epstein` returns the statistic $\mathcal{E}$ and the corresponding probability. The arguments are x(the data), alternative (which takes the option IFR, DFR, and two sided) and exact (where exact = False uses the large-sample approximation if $n > 9$).

b. *One-Sided Test against DFR Alternatives.* To test

$$H_0 : F \text{ is exponential}$$

versus

$$H_2 : F \text{ is DFR (and not exponential)},$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } \mathcal{E} \le \frac{n-1}{2} - e_\alpha; \quad \text{otherwise do not reject.} \tag{11.11}$$

See Comment 6.

c. *Two-Sided Test against IFR and DFR Alternatives.* To test

$$H_0 : F \text{ is exponential}$$

versus

$$H_3 : F \text{ is IFR or DFR (and not exponential)},$$

at the α level of significance (with equal probabilities $\alpha/2$ in the tails),

$$\text{Reject } H_0 \text{ if } \mathcal{E} \geq e_{\alpha/2} \text{ or if } \mathcal{E} \leq \frac{n-1}{2} - e_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{11.12}$$

Large-Sample Approximation

Let

$$\mathcal{E}^* = \frac{\mathcal{E} - E_0(\mathcal{E})}{\sqrt{\text{var}_0(\mathcal{E})}} = \frac{\mathcal{E} - \frac{(n-1)}{2}}{\sqrt{\frac{n-1}{12}}}. \tag{11.13}$$

Then, as $n \to \infty$, the distribution of $\mathcal{E}^*$ tends to the $N(0, 1)$ distribution.

The large-sample approximation to procedure (11.10) is

$$\text{Reject } H_0 \text{ if } \mathcal{E}^* \geq z_\alpha; \quad \text{otherwise do not reject.} \tag{11.14}$$

The large-sample approximation to procedure (11.11) is

$$\text{Reject } H_0 \text{ if } \mathcal{E}^* \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{11.15}$$

The large-sample approximation to procedure (11.12) is

$$\text{Reject } H_0 \text{ if } |\mathcal{E}^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{11.16}$$

EXAMPLE 11.1 ***Methylmercury Poisoning.***

Van Belle (1972) consulted on an experiment at the Florida State University designed to study the effect of methylmercury poisoning on the lifelengths of fish. In the experiment, goldfish were subjected to various dosages of methylmercury. At one dosage level, the ordered times to death (in days) were 42, 43, 51, 61, 66, 69, 71, 81, 82, and 82. We will apply the Epstein test of H_0 versus IFR alternatives. The calculations are summarized in Table 11.1.

From the fourth column of Table 11.1, we obtain

$$\sum_{i=1}^{9} S_i = 420 + 429 + 493 + 563 + 593 + 608 + 616 + 646 + 648 = 5016.$$

Then from (11.9)

$$\mathcal{E} = \frac{\sum_{i=1}^{9} S_i}{S_{10}} = \frac{5016}{648} = 7.74.$$

To apply the large-sample approximation, we note, using (11.13),

$$\mathcal{E}^* = \frac{7.74 - 4.5}{\sqrt{\frac{9}{12}}} = 3.74.$$

Table 11.1 Calculation of $\mathcal{E}$ for the Methylmercury Data

i	$X_{(i)}$	$D_{(i)}$	$S_{(i)}$
1	42	420	420
2	43	9	429
3	51	64	493
4	61	70	563
5	66	30	593
6	69	15	608
7	71	8	616
8	81	30	646
9	82	2	648
10	82	0	648

Let `methyl<-c(42, 43, 51, 61, 66, 69, 71, 81, 82, 82)`. Then `epstein(methyl, alt="ifr", exact=T)` yields $\mathcal{E} = 7.74$ with a P-value .00002. For the large-sample approximation, `epstein(methyl, alt="ifr")` returns $\mathcal{E}^* = 3.74$ with a two-sided P-value .00009.

Comments

1. *The Total-Time-on-Test Statistic.* In a life-testing situation, n independent items may be put on test to study their survival. Let $X_1, \ldots, X_n$ denote the observed lifelengths and let $X_{(1)} \le \cdots \le X_{(n)}$ denote their ordered values, and let $X_{(0)} = 0$. At time $X_{(i)}$, the total time spent on test thus far by the n items is

$$\begin{aligned} &nX_{(1)} + (n-1)(X_{(2)} - X_{(1)}) + \cdots + (n-i+1)(X_{(i)} - X_{(i-1)}) \\ &\quad = \sum_{u=1}^{i} (n-u+1)(X_{(u)} - X_{(u-1)}) \\ &\quad = \sum_{u=1}^{i} D_u = S_i, \end{aligned}$$

where (recall) S_i is given by (11.8). The total-time-on-test transformation transforms $X_{(1)}, \ldots, X_{(n)}$ into $T_1, \ldots, T_n$, where

$$T_i = \frac{\sum_{u=1}^{i} D_u}{\sum_{u=1}^{n} D_u} = \frac{S_i}{S_n}. \tag{11.17}$$

The quantities $T_1, \ldots, T_n$ are called the *total-time-on-test transforms* and

$$\mathcal{E} = \sum_{i=1}^{n-1} T_i \tag{11.18}$$

is known as the *total-time-on-test statistic*. When H_0 is true, $T_1, \ldots, T_{n-1}$ have the same distribution as the order statistics in a sample of size $n-1$ from the uniform $(0, 1)$ distribution (see Epstein (1960)). Since the sum of $n-1$

ordered values equals the sum of $n-1$ unordered values, it follows that under $H_0, \mathcal{E}$ has the same distribution as $U_1 + U_2 + \cdots + U_{n-1}$, where $U_1, \ldots, U_{n-1}$ are independent and identically distributed $U(0,1)$ random variables. Since $E(U_1) = \frac{1}{2}$ and $\text{var}(U_1) = \frac{1}{12}$, it follows that

$$E_0(\mathcal{E}) = \frac{n-1}{2} \tag{11.19}$$

and

$$\text{var}_0(\mathcal{E}) = \frac{n-1}{12}. \tag{11.20}$$

From the central limit theorem, we obtain the large-sample approximation based on $\mathcal{E}^*$ (see (11.13)). The approach to normality is fast and the approximation is very good when $n \geq 9$.

2. *The Barlow–Doksum Class of Monotonic Tests.* Barlow and Proschan (1966) showed that the T's given by (11.17) tend to be larger for an IFR distribution than for an exponential distribution. This led Barlow and Doksum (1972) to consider tests based on statistics that are monotonic in the following sense. A statistic $\mathcal{T}(T_1, \ldots, T_n)$ is monotonic in the T's if $\mathcal{T}(T_1, \ldots, T_n) \geq \mathcal{T}(T_1', \ldots, T_n')$ whenever $T_i \geq T_i'$, $i = 1, \ldots, n-1$. Barlow and Doksum studied the particular class of monotonic statistics given by $\mathcal{T}_J = \sum_{i=1}^{n-1} J(T_i)$, where J is a nondecreasing function on $(0,1)$. The function J is chosen to make the test based on $\mathcal{T}_J$ asymptotically most powerful for a given parametric alternative. The choice $J(u) = u$ yields the total-time-on-test statistic $\mathcal{E}$ that is shown to be asymptotically most powerful for Makeham alternatives. The Makeham parametric family has failure rate function

$$r_m(x) = \lambda\{1 + \theta(1 - \exp(-\lambda x))\}, \quad \theta \geq 0, \quad x > 0, \quad \lambda > 0. \tag{11.21}$$

The failure rate, r_m, is increasing when $\theta > 0$. When $\theta = 0$, $r_m(x) = \lambda$, the failure rate of an exponential distribution with parameter λ.

For other IFR tests, see Bickel and Doksum (1969), Problem 3, Klefsjö (1983), Comment 5, and the survey papers by Doksum and Yandell (1984) and Hollander and Proschan (1984).

3. *The Increasing Failure Rate Average (IFRA) Class.* The failure rate, $r(x)$, may have an increasing trend, but it may not be strictly nondecreasing, as is required to be a member of the IFR class. The failure rate may fluctuate due perhaps to seasonal variations. In a medical setting, an early increasing failure rate may decrease for a period due to medical intervention. A distribution F is in the IFRA class if its average failure rate increases. A distribution F is in the DFRA class if its average failure rate decreases. For more formal definitions, see Barlow and Proschan (1981).

The IFR class is contained in the IFRA class and the DFR class is contained in the DFRA class. If F is IFR, then it is IFRA, but the converse is not true. That is, there are IFRA distributions that are not IFR. Similarly, if F is DFR then it is DFRA, but there are DFRA distributions that are not DFR. We may write these containment relations as

$$\text{IFR} \subset \text{IFRA}$$

$$\text{DFR} \subset \text{DFRA}.$$

The exponential distributions form the boundary of the IFR class and the DFR class; that is, the exponential distributions are in the IFR class and in the DFR class, and they are the only distributions that are both IFR and DFR. Similarly, the exponential distributions form the boundary of the IFRA and DFRA classes; that is, the exponential distributions are in the IFRA class and in the DFRA class and are the only distributions that are both IFRA and DFRA.

Barlow and Scheuer (1971) and Wang (1987) estimate F when it is known to be in the IFRA class.

4. *Use of $\mathcal{E}$ for Testing IFRA.* Barlow and Proschan (1966) proved that if $X_1, \ldots, X_n$ is a sample from an IFRA distribution and $Y_1, \ldots, Y_n$ is a sample from an exponential distribution, then

$$\mathcal{E}(X_1, \ldots, X_n) \overset{\text{st}}{\geq} \mathcal{E}(Y_1, \ldots, Y_n). \tag{11.22}$$

Here, $\mathcal{E}(X_1, \ldots, X_n)$ denotes $\mathcal{E}$ computed for the X-sample and $\mathcal{E}(Y_1, \ldots, Y_n)$ denotes $\mathcal{E}$ computed for the Y-sample. The notation $\overset{\text{st}}{\geq}$ means *stochastically greater than*. The random variable Z is said to be stochastically greater than the random variable Z' if $P(Z \leq x) \leq P(Z' \leq x)$ for every x.

Motivated by result (11.22), Barlow (1968) suggested rejecting H_0 in favor of IFRA alternatives if $\mathcal{E}$ is large.

Hollander and Proschan (1975) showed that the total-time-on-test statistic $\mathcal{E}$ arises in a natural way for testing exponentiality against NBUE alternatives (see Section 11.2 and Comments 13 and 14). Hollander and Proschan (1975) showed that the consistency class of the test that rejects for large (small) values of $\mathcal{E}$ contains the NBUE (NWUE) distributions. Thus the statistic $\mathcal{E}$ can be more suitably viewed as a test for determining the larger NBUE (NWUE) class.

Other tests of H_0 versus IFRA alternatives have been proposed by Barlow (1968), Barlow and Campo (1975), Bergman (1977), Deshpande (1983), and Klefsjö (1983). Klefsjö's test is presented in Comment 5. IFR and IFRA tests for incomplete data, where some of the items are not observed up to their failure times, have been proposed by Barlow and Proschan (1969). The maximum likelihood estimator ($\widehat{G}_n$, say) of F, when it is known that F is in the IFR class, was obtained by Grenander (1956) and Marshall and Proschan (1965) (also see Barlow et al. (1972) and Robertson, Wright, and Dykstra (1988)). Hollander and Proschan (1984) illustrated the calculation of $\widehat{G}_n$ using the methylmercury poisoning data of Table 11.1.

5. *Klefsjö's IFR and IFRA Tests.* Klefsjö's (1983) proposed IFR and IFRA tests based on the normalized spacings $D_1, \ldots, D_n$ defined by (11.7). Klefsjö's tests are motivated by a graphical procedure known as the total-time-on-test transform (see Barlow and Campo (1975) and Klefsjö (1983)). Klefsjö rejects H_0 in favor of H_1 for large values of

$$A = \frac{\sum_{j=1}^{n} \alpha_j D_j}{S_n},$$

where S_n is defined by (11.8) and

$$\alpha_j = 6^{-1}\{(n+1)^3 j - 3(n+1)^2 j^2 + 2(n+1) j^3\}.$$

The null distribution of A is determined by using the result that under $H_0, D_1, D_2, \ldots, D_n$ are independent and identically distributed according to the exponential distribution given by (11.3). Klefsjö provides null distribution tables of

$$A^* = A\left(\frac{7560}{n^7}\right)^{1/2}$$

for $n = 5(5)75$. He gives the upper and lower .01, .05, and .10 percentiles. Significantly large values of A^* indicate IFR alternatives; significantly small values of A^* indicate DFR alternatives. Klefsjö shows that under H_0, A^* can be treated asymptotically as a $N(0,1)$ random variable. Klefsjö also shows that the test that rejects H_0 for large values of A is consistent against the class of continuous IFR distributions.

Klefsjö's (1983) test of H_0 versus F is IFRA (and not exponential) is based on the statistic B, where

$$B = \frac{\sum_{j=1}^{n} \beta_j D_j}{S_n},$$

where

$$\beta_j = 6^{-1}\left\{2j^3 - 3j^2 + j(1 - 3n - 3n^2) + 2n + 3n^2 + n^3\right\}.$$

Klefsjö provides null distribution tables of

$$B^* = B\left(\frac{210}{n^5}\right)^{1/2}$$

for $n = 5(5)75$, giving the upper and lower .01, .05, and .10 percentiles. Significantly large values of B indicate IFRA alternatives; significantly small values of B indicate DFRA alternatives. Klefsjö shows that under H_0, B^* can be treated asymptotically as a $N(0,1)$ random variable. Klefsjö also shows that the test that rejects H_0 for large values of B is consistent against the class of continuous IFRA distributions.

For Klefsjö's IFR test use the `R` command `klefsjo.ifr.mc(x, alternative="two.sided", exact=FALSE, min.reps=100, max.reps=1000, delta= 10^-3)`. Here, x is a vector of data of length n, alternative choices are two.sided, `ifr`, and `dfr` with the default being two.sided, exact is TRUE/FALSE and determines whether the exact test or the large-sample approximation is used if $n \geq 9$. If $n < 9$, the exact test is used. The default value is FALSE, so the large-sample approximation is used unless otherwise specified, min.reps is the minimum number of replications for the Monte Carlo approximation with the default = 100, max.rep is the maximum number of replications for the Monte Carlo approximation, and delta is the measure of accuracy for the convergence. If the probability converges to within delta, a stop occurs before reaching the maximum number of replications.

For Klefsjö's IFRA test, use `klefsjo.ifra.mc(x, alternative= "two.sided", exact=FALSE, min.reps=100, max.reps=1000, delta=10^-3)` where x, exact, min.reps max.reps are defined as in Klefsjö's IFR test above. For alternative, the choices are two.sided, ifra, dfra with the default value being two.sided.

6. *Symmetry of the Null Distribution of $\mathcal{E}$.* Under H_0, the distribution of $\mathcal{E}$ is symmetric about its mean $\frac{(n-1)}{2}$; that is,

$$P_0\Big(\mathcal{E} - \frac{n-1}{2} \geq x\Big) = P_0\Big(\mathcal{E} - \frac{n-1}{2} \leq -x\Big) \tag{11.23}$$

It follows that the lower α percentile point ($e_\alpha^{(2)}$, say) of the null distribution of $\mathcal{E}$ can be obtained from the upper α percentile point $e_\alpha^{(1)}$ via

$$e_\alpha^{(2)} = (n-1) - e_\alpha^{(1)}. \tag{11.24}$$

7. *Some IFR Distributions, Some DFR Distributions.* The exponential distribution has density function

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0, \quad \lambda > 0.$$

Its failure rate is constant, that is,

$$r(x) = \lambda, \quad x > 0.$$

Thus, the exponential distribution is both IFR and DFR.

One commonly used generalization of the exponential distribution is the Weibull distribution with density function

$$f(x) = \lambda\alpha(\lambda x)^{\alpha-1}e^{-(\lambda x)^\alpha}, \quad x > 0, \quad \alpha > 0 \quad \lambda > 0.$$

Its failure rate is

$$r(x) = \lambda\alpha(\lambda x)^{\alpha-1}.$$

The failure rate is increasing for $\alpha > 1$ and decreasing for $\alpha < 1$. Thus the Weibull distributions for which $\alpha > 1$ are IFR distributions, and the Weibull distributions with $\alpha < 1$ are DFR distributions. For $\alpha = 1$, the Weibull distribution reduces to the exponential distribution.

Another frequently used family of distributions, which is a generalization of the exponential, is the gamma family. The gamma density function is

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}, \quad x > 0, \quad \alpha > 0, \quad \lambda > 0,$$

where $\Gamma(\alpha)$ is the gamma function (cf. Kalbfleisch and Prentice (1980, p. 23)). The gamma density does not have a closed-form expression for its failure rate. Its failure rate is increasing when $\alpha > 1$ and is decreasing when $\alpha < 1$. Thus those members of the gamma family for which the parameter α is greater than 1 are IFR distributions, those members corresponding to $\alpha < 1$ are DFR distributions. When $\alpha = 1$, the gamma distribution reduces to the exponential distribution.

For other IFR and DFR distributions, see Barlow and Proschan (1981) and Kalbfleisch and Prentice (1980). Often investigators pool data from different IFR distributions to increase the effective sample size. Gurland and Sethuraman (1995) show that such a pooling may actually reverse the IFR property of the individual samples to a DFR property for the mixture.

Table 11.2 Intervals in Hours between Failures of the Air-Conditioning System of Plane 8044

i	1	2	3	4	5	6	7	8	9	10	11	12
X_i	487	18	100	7	98	5	85	91	43	230	3	130

Source: F. Proschan (1963).

Table 11.3 Ordered Survival Times in Days of Guinea Pigs under Regimen 4.3

i	1	2	3	4	5	6	7	8	9
$X_{(i)}$	10	33	44	56	59	72	74	77	92
i	10	11	12	13	14	15	16	17	18
$X_{(i)}$	93	96	100	100	102	105	107	107	108
i	19	20	21	22	23	24	25	26	27
$X_{(i)}$	108	108	109	112	113	115	116	120	121
i	28	29	30	31	32	33	34	35	36
$X_{(i)}$	122	122	124	130	134	136	139	144	146
i	37	38	39	40	41	42	43	44	45
$X_{(i)}$	153	159	160	163	163	168	171	172	176
i	46	47	48	49	50	51	52	53	54
$X_{(i)}$	183	195	196	197	202	213	215	216	222
i	55	56	57	58	59	60	61	62	63
$X_{(i)}$	230	231	240	245	251	253	254	254	278
i	64	65	66	67	68	69	70	71	72
$X_{(i)}$	293	327	342	347	361	402	432	458	555

Source: T. Bjerkedal (1960).

Properties

1. *Consistency.* See Hollander and Proschan (1975) for the consistency class of the test defined by (11.10). In particular, if F is continuous and IFR (and not exponential), the test is consistent.
2. *Asymptotic Normality.* See Barlow (1968), Bickel and Doksum (1969), and Doksum and Yandell (1984).
3. *Efficiency.* See Bickel and Doksum (1969), Klefsjö (1983), and Section 11.8.

Problems

1. Table 11.2 is based on a subset of data considered by Proschan (1963). Proschan investigated the life distribution of the air-conditioning system of a fleet of Boeing 720 jet airplanes. Table 11.2 presents intervals (in hours) between failures of the air-conditioning system of plane 8044. Using the data of Table 11.2, test H_0 versus IFR alternatives.

2. Bjerkedal (1960) studied the lifelengths of guinea pigs injected with different amounts of tubercle bacilli. One reason for choosing this species is that guinea pigs are known to have a high susceptibility to human tuberculosis. The data in Table 11.3 are a subset of the survival data considered by Bjerkedal.

The data in Table 11.3 correspond to study M (in Bjerkedal's terminology), in which animals in a single cage are under the same regimen. The regimen number is the common logarithm of the number of bacillary units in .5 ml of the challenge solution. The data in

Table 11.3 are for regimen 4.3. Regimen 4.3 corresponds to 2.2×10^4 bacillary units per .5 ml ($\log_{10}(2.2 \times 10^4) = 4.342$). There are 72 observations, and Table 11.3 gives the ordered values $X_{(1)} \leq \cdots \leq X_{(72)}$. Using the data of Table 11.3, test H_0 versus IFR alternatives.

3. Bickel and Doksum (1969) considered a class of test statistics based on the ranks of the D's. Let R_i denote the rank of D_i in the joint ranking of $D_1, \ldots, D_n$. One member of the Bickel–Doksum class is

$$W_1 = \sum_{i=1}^{n} i \log\Big(1 - \frac{R_i}{n+1}\Big). \tag{11.25}$$

When F is IFR (DFR) the spacings tend to show a downward (upward) trend (cf. Doksum and Yandell (1984). Thus, H_0 is rejected in favor of $H_1(H_2)$ for significantly large (small) values of W_1. Bickel and Doksum showed that tests based on W_1 are asymptotically equivalent to those based on $\mathcal{E}$. They found, however, that for finite sample sizes, $\mathcal{E}$ does better than W_1 in terms of power. For finite n, the null distribution of W_1 can be obtained using the fact that, under H_0, all $n!$ possible outcomes of $(R_1, \ldots, R_n)$ are equally likely (and thus each has probability $1/n!$). Determine the null distribution of W_1 for the case $n = 4$. That is, give the possible values of W_1 and the corresponding probabilities. What is the critical region of the $\alpha = \frac{1}{24}$ test of H_0 versus H_1 based on W_1?

4. Bickel and Doksum (1969) showed that, under H_0,

$$W_1^* = \frac{\sqrt{n}(W_1' + \frac{1}{2})}{s_1} \tag{11.26}$$

tends to a $N(0,1)$ distribution as $n \to \infty$. In (11.26),

$$W_1' = \frac{W_1}{n(n+1)} \tag{11.27}$$

and

$$s_1^2 = \frac{n-1}{12(n+1)}. \tag{11.28}$$

Thus an approximate α level test of H_0 versus H_1 rejects H_0 if $W_1^* \geq z_\alpha$ and accepts H_0 if $W_1^* < z_\alpha$. Similarly, an approximate α level test of H_0 versus H_2 rejects H_0 if $W_1^* \leq -z_\alpha$ and accepts H_0 if $W_1^* > -z_\alpha$. Apply the large-sample test of H_0 versus H_1, based on W_1^* to the methylmercury data of Example 11.1.

5. Apply the large-sample test based on W_1^* to the data of Table 11.2.

6. Verify the symmetry of the null distributions of $\mathcal{E}$ as expressed in (11.23) (*Hint*: Recall that, under H_0, $\mathcal{E}$ has the same distribution as $U_1 + \ldots + U_{n-1}$, where $U_1, \ldots, U_{n-1}$ are independent $U(0,1)$ random variables.)

7. Describe a situation in which DFR alternatives (i.e., beneficial aging) might occur.

8. Describe a situation where it is natural to expect the underlying distribution to satisfy the IFRA property but not satisfy the IFR property (i.e., where F might be expected to be a member of the IFRA class but not a member of the IFR class).

11.2 A TEST OF EXPONENTIALITY VERSUS NBU ALTERNATIVES (HOLLANDER–PROSCHAN)

Hypothesis

The hypothesis of interest is

$$H_0 : P(X \geq x + y | X \geq x) = P(X \geq y), \quad \text{for all } x, y \geq 0. \tag{11.29}$$

The vertical bar (|) in the probability statement on the left-hand side of (11.29) is to be read as "given that." (11.29) asserts that the probability of surviving an additional time period y, given that the item has survived to time x, is equal to the probability that a new item will survive an initial period y. Insisting that the equality holds for all x, y is equivalent to asserting that used items of all ages are no better and no worse than new items. This property is equivalent to the underlying population being an exponential population (see Comment 8). Thus (11.29) is another way of expressing H_0 given by display (11.3). Using the survival function $\overline{F}$, (11.29) can be written as

$$H_0 : \frac{\overline{F}(x+y)}{\overline{F}(x)} = \overline{F}(y), \quad \text{all } x, y \geq 0, \tag{11.30}$$

and again as

$$H_0 : \overline{F}(x+y) = \overline{F}(x)\overline{F}(y), \quad \text{all } x, y \geq 0. \tag{11.31}$$

We now turn to the new better than used (NBU) and new worse than used (NWU) alternatives. The distribution F is said to be in the NBU class if

$$\overline{F}(x+y) \leq \overline{F}(x)\overline{F}(y), \quad \text{all } x, y \geq 0. \tag{11.32}$$

Similarly, the distribution F is said to be in the NWU class if

$$\overline{F}(x+y) \geq \overline{F}(x)\overline{F}(y), \quad \text{all } x, y \geq 0. \tag{11.33}$$

If F is NBU, then new items are better than used items of any age. Similarly, if F is NWU, then new items are worse than used items of any age. The boundary members of the classes are the exponential distributions. An exponential distribution is both NBU and NWU.

Procedure

Let $X_{(1)} \leq \cdots \leq X_{(n)}$ denote the ordered X's. Compute

$$T = \sum_{i>j>k} \psi(X_{(i)}, X_{(j)} + X_{(k)}), \tag{11.34}$$

where

$$\psi(a,b) = \begin{cases} 1, & \text{if } a > b, \\ 0, & \text{if } a < b. \end{cases} \tag{11.35}$$

Note that the summation in (11.34) is over all $n(n-1)(n-2)/6$ ordered triples (i, j, k) with $i > j > k$.

a. *One-Sided Test against NBU Alternatives.* To test

$$H_0 : F \text{ is exponential}$$

versus

$$H_4 : F \text{ is NBU} \qquad \text{(and not exponential)}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } T \le t_{1,\alpha}; \quad \text{otherwise do not reject,} \tag{11.36}$$

where the constant $t_{1,\alpha}$ satisfies $P_0\{T \le t_{1,\alpha}\} = \alpha$. The *R* function `nb.mc` gives a *P*-value corresponding to T based on Monte Carlo sampling. If the value of T, say t, is less than the null mean $n(n-1)(n-2/8))$, `nb.mc` returns a value that approximates $P_0(T \le t)$. If t is greater than the null mean, `nb.mc` approximates $P_0(T \ge t)$. The *R* function `newbet` gives an approximate *P*-value corresponding to the standardized value T^* defined by (11.39).

b. *One-Sided Test against NWU Alternatives.* To test

$$H_0 : F \text{ is exponential}$$

versus

$$H_5 : F \text{ is NBU} \qquad \text{(and not exponential)}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } T \ge t_{2,\alpha}; \quad \text{otherwise do not reject,} \tag{11.37}$$

where the constant $t_{2,\alpha}$ satisfies $P_0\{T \ge t_{2,\alpha}\} = \alpha$.

c. *Two-Sided Test against NBU and NWU Alternatives.* To test

$$H_0 : F \text{ is exponential}$$

versus

$$H_6 : F \text{ is NBU or NWU} \qquad \text{(and not exponential)}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } T \le t_{1,\alpha_1} \text{ or if } T \ge t_{2,\alpha_2}; \quad \text{otherwise do not reject,} \tag{11.38}$$

where $\alpha_1 + \alpha_2 = \alpha$.

Large-Sample Approximation

Define

$$T^* = \frac{T - E_0(T)}{[\text{var}_0(T)]^{1/2}} = \frac{T - \left\{\frac{n(n-1)(n-2)}{8}\right\}}{\left\{\left(\frac{3}{2}\right)n(n-1)(n-2)\left[\left(\frac{5}{2592}\right)(n-3)(n-4) + (n-3)\left(\frac{7}{432}\right) + \left(\frac{1}{48}\right)\right]\right\}^{1/2}}. \tag{11.39}$$

When H_0 is true, the statistic T^* has an asymptotic (n tending to infinity) $N(0,1)$ distribution.

The normal theory approximation to procedure (11.36) is

$$\text{Reject } H_0 \text{ if } T^* \le -z_\alpha; \quad \text{otherwise do not reject.} \tag{11.40}$$

The normal theory approximation to procedure (11.37) is

$$\text{Reject } H_0 \text{ if } T^* \ge z_\alpha; \quad \text{otherwise do not reject.} \tag{11.41}$$

The normal theory approximation to procedure (11.38) is

$$\text{Reject } H_0 \text{ if } T^* \le -z_{\alpha_1} \text{ or if } T^* \ge z_{\alpha_2}; \quad \text{otherwise do not reject,} \tag{11.42}$$

where $\alpha_1 + \alpha_2 = \alpha$.

The *R* function `newbet` gives the *P*-value based on the normal approximation.

Ties

If $X_{(i)} = X_{(j)} + X_{(k)}$, compute T by replacing $\psi(X_{(i)}, X_{(j)} + X_{(k)})$ with $\psi^*(X_{(i)}, X_{(j)} + X_{(k)})$, where

$$\psi^*(a,b) = \begin{cases} 1, & \text{if } a > b, \\ \frac{1}{2}, & \text{if } a = b, \\ 0, & \text{if } a < b. \end{cases} \tag{11.43}$$

EXAMPLE 11.2 ***Example 11.1 Continued.***

We return to the methylmercury poisoning data of Example 11.1. Recall that the ordered lifelengths $X_{(1)} \le \ldots \le X_{(10)}$ are 42, 43, 51, 61, 66, 69, 71, 81, 82, and 82. Although, in general, the ψ functions of (11.34) need to be computed for the $n(n-1)(n-2)/6$ $(X_{(i)}, X_{(j)}, X_{(k)})$ triples with $i > j > k$, we note that $X_{(10)} < X_{(1)} + X_{(2)}$; thus for this data set, all 120 ψ functions must be zero. (Since $X_{(10)} < X_{(1)} + X_{(2)}$, there is no (i,j,k) triple with $i > j > k$ satisfying $X_{(i)} > X_{(j)} + X_{(k)}$.) Thus $T = 0$. Hollander and Proschan (1972) showed that for $n \ge 3$,

$$P_0\{T = 0\} = \frac{1}{\binom{2n-2}{n}}. \tag{11.44}$$

Thus, with $n = 10$, we find

$$P_0\{T = 0\} = \frac{1}{\binom{18}{10}} = \frac{1}{43758} = .00002.$$

Thus the *P*-value is .00002, and this is strong evidence against exponentiality in favor of deleterious aging. (Result (11.44) was derived under the assumption that F is continuous. We have a tie in the methylmercury poisoning data set, and thus the *P*-value of .00002 is approximate.)

Table 11.4 Intervals in Hours between Failures of the Air-Conditioning System of Plane 7907

i	X_i	$X_{(i)}$
1	194	15
2	15	29
3	41	33
4	29	41
5	33	181
6	181	194

Source: F. Proschan (1963).

EXAMPLE 11.3 ***Intervals between Failures for Air-Conditioning System.***

Table 11.4 is based on a subset of data considered by Proschan (1963). Proschan investigated the life distribution of the air-conditioning system of a fleet of Boeing 720 jet airplanes. The table presents intervals between failures of the air-conditioning system of plane 7907.

Before testing exponentiality, Proschan considered the following question: Can we view the X's as a random sample from a common population? We are considering a particular plane, and increased operation of the plane might lead to shorter (or longer) intervals between failures. Proschan states, "A trend toward longer intervals, if established, could be the result of greater experience, debugging, or elimination of faulty parts, whereas a trend toward shorter intervals could be the result of wearout, aging, or poor maintenance."

To see whether it is appropriate to consider the successive intervals from airplane 7907 as values from a common population, we followed Proschan and applied Mann's test for trend (see Comment 8.14) and found that there is no significant evidence of a trend; thus we proceed as if the X's are a random sample from a common population. We now apply procedure (11.36) to test the hypothesis of exponentiality against NBU alternatives. In this particular setting, we may interpret the hypothesis as a statement that the air-conditioning system, upon repair, is as good as new. Table 11.5 illustrates the calculations required to compute T.

Thus, we have

$$T = \sum_{i>j>k} \psi(X_{(i)}, X_{(j)} + X_{(k)}) = 12.$$

Let `ac<-c(194, 15, 41, 29, 33, 181)`. The *R* function `newbet(ac)` returns the value $T = 12$, $T^* = -.786$, and a corresponding P-value of .22. The *R* function `nb.mc(ac, alt="nbu")` yields the Monte Carlo sampling approximation $P_0(T \leq 12) = .15$. Thus there is not sufficient evidence to reject H_0 and we accept that the underlying distribution is some exponential distribution.

Comments

8. *Characterization of the Exponential Distribution.* If (11.31) holds for all $x, y \geq 0$, then it can be shown (cf. Barlow and Proschan (1981, p. 57)) that $P(X \geq y) = e^{-\lambda y}$ for some $\lambda > 0$. That is, the underlying population is exponential. Even

Table 11.5 $\psi(X_{(i)}, X_{(j)} + X_{(k)})$ Values Corresponding to the 20 Ordered (i, j, k) Triples with $i > j > k$

(i, j, k)	$(X_{(i)}, X_{(j)}, X_{(k)})$	$X_{(j)} + X_{(k)}$	$\psi(X_{(i)}, X_{(j)} + X_{(k)})$
(3, 2, 1)	(33, 29, 15)	44	0
(4, 2, 1)	(41, 29, 15)	44	0
(5, 2, 1)	(181, 29, 15)	44	1
(6, 2, 1)	(194, 29, 15)	44	1
(4, 3, 1)	(41, 33, 15)	48	0
(5, 3, 1)	(181, 33, 15)	48	1
(6, 3, 1)	(194, 33, 15)	48	1
(5, 4, 1)	(181, 41, 15)	56	1
(6, 4, 1)	(194, 41, 15)	56	1
(6, 5, 1)	(194, 181, 15)	196	0
(4, 3, 2)	(41, 33, 29)	62	0
(5, 3, 2)	(181, 33, 29)	62	1
(6, 3, 2)	(194, 33, 29)	62	1
(5, 4, 2)	(181, 41, 29)	70	1
(6, 4, 2)	(194, 41, 29)	70	1
(6, 5, 2)	(194, 181, 29)	210	0
(5, 4, 3)	(181, 41, 33)	74	1
(6, 4, 3)	(194, 41, 33)	74	1
(6, 5, 3)	(194, 181, 33)	214	0
(6, 5, 4)	(194, 181, 41)	222	0

though under H_0 the population distributions are restricted to be exponential, we have retained the term *nonparametric* for tests based on T and $\mathcal{E}$, because those tests are designed to detect large nonparametric classes of distributions.

9. *Relationship of NBU to IFR and IFRA.* The NBU class is larger than the IFR and IFRA classes and properly contains those classes. Symbolically,

$$\text{IFR} \subset \text{IFRA} \subset \text{NBU}.$$

The corresponding containment relations for the NWU class in relation to the DFR and DFRA classes are

$$\text{DFR} \subset \text{DFRA} \subset \text{NWU}.$$

Thus, for example, an IFR distribution is also an NBU distribution, but there are NBU distributions that are not IFR or IFRA. For example, the underlying population may be NBU, but its failure rate $r(x)$ may fluctuate (and in particular not be increasing or increasing on average), due perhaps to seasonal variations. The NBU test is designed to detect this larger class. Hollander and Proschan show that the consistency class of the NBU (NWU) test that rejects for small (large) values of T includes the continuous NBU (NWU) distributions.

10. *The NBU Test Employs New Items.* The reader should note that we can test whether new is better (worse) than used employing only new items. That is, we need a sample only of lifelengths of new items to perform the NBU (NWU) test.

11. *Motivation for the NBU Test.* Define

$$T^*(x,y) = \overline{F}(x)\overline{F}(y) - \overline{F}(x+y). \tag{11.45}$$

Note that $T^*(x,y) = 0$ for all (x,y) if and only if H_0 is true. This fact was used by Hollander and Proschan in devising the test based on T. The statistic $\frac{1}{4} - \{2T/[n(n-1)(n-2)]\}$ estimates the parameter

$$\Delta_{\text{NBU}}(F) = E_F\{T^*(X',Y')\}, \tag{11.46}$$

where X', Y' are independent and each is from the underlying life population with distribution F. We may view $T^*(x,y)$ as a measure of the deviation from H_0 at the point (x,y) and $\Delta_{\text{NBU}}(F)$ as the average value of this deviation. When F is NBU and continuous, the parameter $\Delta_{\text{NBU}}(F)$ is positive. When sampling from such a population the value of $\frac{1}{4} - \{2T/[n(n-1)(n-2)]\}$ tends to be large or, equivalently, T tends to be small. This partially motivates procedure (11.36). Asymptotic normality of T is directly obtained from Hoeffding's U-statistic theory because $2T/[n(n-1)(n-2)]\}$ is a U-statistic. (See Hollander and Proschan, 1972.)

12. *NBU Test for Censored Data.* Chen, Hollander, and Langberg (1983a) extended the NBU test to censored data by estimating the parameter $\Delta_{\text{NBU}}(F)$ using the Kaplan and Meier (1958) estimator of F. (See Comment 35.)

13. *The New Better Than Used in Expectation (NBUE) Class.* The NBUE class is larger than the NBU class. In order to define the NBUE class, we first introduce the mean residual life function. The mean residual life function, corresponding to a distribution F, gives, for each value of $x \geq 0$, the expected remaining life at time x. More formally, the mean residual life (mrl) function corresponding to a distribution F is defined as

$$m(x) = \begin{cases} E_F(X - x | X > x), & \text{for those } x \text{ such that } \overline{F}(x) > 0, \\ 0, & \text{for those } x \text{ such that } \overline{F}(x) = 0, \end{cases} \tag{11.47}$$

where X has the distribution F. In (11.47), note that $X - x$ is the residual life of an item with lifelength X, given that it has survived to time x. Also note that $m(0)$ is the mean μ of the distribution F, that is, $m(0) = E(X)$.

A distribution F with finite mean is said to be a member of the *new better than used in expectation* (NBUE) class if its corresponding mean residual life function m satisfies

$$m(0) \geq m(x) \quad \text{for all } x. \tag{11.48}$$

The NBUE class has the following interpretation. A used NBUE item of any fixed age has a smaller mean residual lifelength than does a new item.

The *new worse than used in exceptation* (NWUE) is similarly defined. A distribution F with finite mean is said to be a member of the NWUE class if its corresponding mean residual life function satisfies

$$m(0) \leq m(x) \quad \text{for all } x. \tag{11.49}$$

The NWUE class has the following interpretation. A used NWUE item of any fixed age has a larger mean residual lifelength than does a new item.

The boundary members of the NBUE and NWUE classes are the exponential distributions. The exponential distributions, given by (11.3), have mrl functions $m(x) = (\lambda)^{-1}$, $x \geq 0$. This is another characterization of the exponential distributions, namely, that F is an exponential distribution if and only if its mrl function is constant.

The relation of the NBUE class to the smaller classes IFR, IFRA, and NBU is given by the following containment relations:

$$\text{IFR} \subset \text{IFRA} \subset \text{NBU} \subset \text{NBUE}. \tag{11.50}$$

For the dual classes used to model beneficial aging,

$$\text{DFR} \subset \text{DFRA} \subset \text{NWU} \subset \text{NWUE}. \tag{11.51}$$

14. *Using the $\mathcal{E}$ Statistic to Test against NBUE Alternatives.* Let X be a random value from F. Hollander and Proschan (1975) considered the parameter

$$\Delta_{\text{NBUE}}(F) = E_F\{\overline{F}(X)[m(0) - m(X)]\}$$

as a measure of deviation, for a given F, from exponentiality toward "NBUE-ness." In the definition of $\Delta_{\text{NBUE}}(F)$, $m(x)$ is the mrl function defined by (11.47). Hollander and Proschan estimated $\Delta_{\text{NBUE}}(F)$ with its sample counterpart, $\Delta_{\text{NBUE}}(F_n)$, where F_n is the empirical distribution function defined by (11.93). This yields the statistic

$$K = \frac{\sum_{i=1}^{n} d_i X_{(i)}}{n^2},$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the ordered X's and

$$d_i = \frac{3n}{2} - 2i + \frac{1}{2}.$$

Dividing K by $\overline{X}$ to make it scale-invariant, Hollander and Proschan proposed $K^* = K/\overline{X}$ as a statistic for testing H_0 against NBUE alternatives or NWUE alternatives. Significantly large (small) values of K^* lead to rejection of H_0 in favor of NBUE (NWUE) alternatives. Hollander and Proschan showed that

$$nK^* = \mathcal{E} - \frac{n-1}{2}, \tag{11.52}$$

where $\mathcal{E}$ is the total-time-on-test statistic defined by (11.9). Thus tests based on K^* are equivalent to tests based on $\mathcal{E}$. Hence, the total-time-on-test statistic, originally proposed to detect IFR (DFR) alternatives and later proposed to detect IFRA (DFRA) alternatives, can be used to detect the larger class of NBUE (NWUE) alternatives. For testing against NBUE (NWUE) alternatives, use the *R* functions `newbet` and `nb.mc`.

Klefsjö (1983), by considering a graphical method known as the total-time-on-test transform (cf. Barlow and Campo, 1975) was also led to

the K^* statistic as a test statistic for exponentiality versus NBUE alternatives. Borges, Proschan, and Rodrigues (1984) developed a test of exponentiality versus NBUE alternatives based on the sample coefficient of variation $s/\overline{X}$, where $s^2 = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. The test proposed by Borges, Proschan, and Rodrigues is equivalent to a test studied by Lee, Locke, and Spurrier (1980). Under H_0, the distribution of $S' = \sqrt{n}\{(s/\overline{X}) - 1\}$ is asymptotically $N(0, 1)$. Significantly large values of S' indicate NBUE alternatives, and significantly small values (i.e., large negative values) of S' indicate NWUE alternatives.

Let $U_i = S_i/S_n$, where S_i is given by (11.8). Barlow and Doksum (1972) proposed the statistic $D^+ = \max_{1\le i\le n}\{U_i - i/n\}$ for testing H_0 against IFR alternatives. Koul (1978b) showed that the test that rejects H_0 for significantly large values of D^+ can be more appropriately viewed as a test of H_0 against the larger NBUE class. The null distribution of D^+, tabled by Birnbaum and Tingey (1951), can be used in this life-testing context. Asymptotically, under H_0,

$$P\left\{n^{1/2}D^+ \le x\right\} = 1 - e^{-2x^2}.$$

Whitaker and Samaniego (1989) derive estimates of F for the situation when it is known that F is a member of the NBUE class.

15. *Koul's NBU Tests.* Koul (1977, 1978a) suggested other statistics for testing H_0 versus NBU alternatives. Koul (1977) proposed the statistic $S = \min_{1\le k\le j\le n} T_{kj}$, where for $1 \le k \le j \le n$, $T_{kj} = nS_{kj} - (n-k)(n-j)$ and $S_{kj} = \sum_{i=1}^{n}\psi(X_{(i)}, X_{(k)} + X_{(j)})$. S/n^2 estimates the parameter $\alpha(F) = \inf_{x,y\ge 0}\{\overline{F}(x+y) - \overline{F}(x)\overline{F}(y)\}$. The parameter $\alpha(F)$ can be viewed as a measure of the deviation of F from H_0 toward H_4. When F is exponential, $\alpha(F) = 0$; it is negative when F is NBU. Thus, the test rejects H_0 in favor of H_4 when S is significantly small. Koul gives critical values of S for $\alpha = .005, .01, .025, .05, .10, .20$ and $n = 3(1)30(5)50$. Koul (1977) does not provide a dual test of H_0 versus H_5 (NWU alternatives). Koul (1978a) suggested a class of tests of H_0 versus NBU alternatives indexed by a function ψ satisfying mild conditions. The Hollander–Proschan NBU test corresponds to the choice $\psi(u) = u$. Koul advocated the choice $\psi(u) = u^{1/2}$.

16. *The Boyles–Samaniego Estimator.* Boyles and Samaniego (1984) considered the problem of estimating F when it is known that F is NBU. Their estimator is

$$\widehat{F}_{\mathrm{NBU}}(x) = \max_i\left\{\frac{\overline{F}_n(x + X_{(i)})}{\overline{F}_n(X_{(i)})}\right\},$$

where $X_{(1)} < \cdots < X_{(n)}$ are the ordered X's and $\overline{F}_n$ is the sample survival function. $\overline{F}_n(x) = 1 - F_n(x)$, where F_n is the sample distribution function defined by (11.105). Boyles and Samaniego show that $\widehat{F}_{\mathrm{NBU}}$ is in the NBU class of distributions and that it is relatively easy to compute. It is not, however, a consistent estimator of F for all F in the NBU class. They do show that it is a consistent estimator of F when F is NBU and $T = \inf\{M : F(M) = 1\}$ is well defined and finite.

17. *The NBU Class and Replacement Policy Comparisons.* The NBU class plays an important role in replacement policy comparisons that arise in the study of renewal processes and repairable systems. A renewal process is a sequence

Table 11.6 Ordered Survival Times (Days from Diagnosis)

7	429	579	968	1877
47	440	581	1077	1886
58	445	650	1109	2045
74	455	702	1314	2056
177	468	715	1334	2260
232	495	779	1367	2429
273	497	881	1534	2509
285	532	900	1712	
317	571	930	1784	

Source: M. M. Siddiqui and E. A. Gehan (1966).

of independent, identically distributed, nonnegative random variables that (with probability 1) are not all zero. An example of a renewal process is the following. Consider a system operating over an indefinite period of time. Upon failure, the system is repaired or replaced. Assuming negligible time for repair, the successive intervals between failures are independent, identically distributed random variables of a renewal process.

Under an age replacement policy, a unit is replaced upon failure or at age T, whichever comes first. Let $N(t)$ = number of renewals in $[0, t]$ for an ordinary renewal process and $N_A(t, T)$ = number of failures in $[0, t]$ under an age replacement policy with replacement interval T. It can be shown that $N(t)$ is stochastically larger than $N_A(t, T)$ for all $t \geq 0$, $T \geq 0$, if and only if F is NBU. For this result and related results, see Barlow and Proschan (1981, p. 179).

Properties

1. *Consistency.* See Hollander and Proschan (1972) for the consistency class of the test defined by (11.36). In particular, if F is continuous, NBU (and not exponential), the test is consistent.
2. *Asymptotic Normality.* See Hollander and Proschan (1972).
3. *Efficiency.* See Hollander and Proschan (1972), Koul (1978b), Klefsjö (1983), and Section 11.8.

Problems

9. The data in Table 11.6 are from a study discussed by Siddiqui and Gehan (1966) and also considered by Bryson and Siddiqui (1969). The data are survival times (measured from the date of diagnosis) of 43 patients suffering from chronic granulocytic leukemia. For these data, $T = 8327$. Apply the large-sample approximation (11.40) to test against NBU alternatives. (Note that one might be reluctant to postulate IFR alternatives here because, after the diagnosis of leukemia, medical treatment may cause the failure rate to decrease for a period of time.)

10. Either show directly or illustrate by means of an example that the maximum possible value of T (based on a sample of size n) is $n(n-1)(n-2)/6$ and the minimum possible value is 0.

11. Apply the NBU test to the data of Table 11.2. Compare your result to the result obtained using the test based on $\mathcal{E}$.

12. Let $\mathcal{F}_{a,b}$ denote the class of distributions with support $[a, b]$, where $b < 2a$. (Roughly speaking, each F in $\mathcal{F}_{a,b}$ puts all its probability in the interval $[a, b]$, where $b < 2a$.) Show that if

$X_1, \ldots, X_n$ is a random sample from a distribution F that is in the class $\mathcal{F}_{a,b}$, then $P_F(T = 0) = 1$.

13. (Problem 12 Continued) Show that every F in $\mathcal{F}_{a,b}$ is an NBU distribution. (*Hint:* Consider the four cases (i) $x < a$ and $y < a$, (ii) $x \geq a$ and $y \geq a$, (iii) $x < a$ and $y \geq a$, and (iv) $x \geq a$ and $y < a$. Show that in each of these cases, $F(x, y)$ satisfies the inequality of (11.32).)

14. (Problems 12 and 13 Continued) Consider the NBU test that rejects for small values of T. Using the fact that, for $n \geq 3$, $P_0[T = 0] = 1/\binom{2n-2}{n}$, show that when $n \geq 3$ and $\alpha \geq 1/\binom{2n-2}{n}$, the power of the NBU test equals 1 for every F in the class $\mathcal{F}_{a,b}$.

15. Verify directly (or illustrate using an example) result (11.52).

16. Apply the NBU test to the guinea pig survival data of Table 11.3.

11.3 A TEST OF EXPONENTIALITY VERSUS DMRL ALTERNATIVES (HOLLANDER–PROSCHAN)

Hypothesis

The null hypothesis is

$$H_0 : F \text{ is exponential.} \tag{11.53}$$

The alternatives are expressed in terms of the mrl function $m(x)$ defined by (11.47). The alternatives are the decreasing mean residual life (DMRL) alternatives, and the increasing mean residual life (IMRL) alternatives. DMRL distributions model situations where deterioration takes place with age; IMRL distributions model beneficial aging.

The distribution F is said to be a member of the *decreasing mean residual life* (DMRL) class if $F(0) = 0$ and

$$m(x) \geq m(y), \quad \text{for all } x < y \text{ such that } \overline{F}(x) \text{ and } \overline{F}(y) > 0. \tag{11.54}$$

Similarly, the distribution F is said to be a member of the *increasing mean residual life* (IMRL) class if $F(0) = 0$ and

$$m(x) \leq m(y), \quad \text{for all } x < y \text{ such that } \overline{F}(x) \text{ and } \overline{F}(y) > 0. \tag{11.55}$$

The distributions that are both DMRL and IMRL are the exponential distributions; that is, the exponentials are the boundary members of the classes.

Procedure

Let $X_{(1)} \leq \cdots \leq X_{(n)}$ denote the ordered X's. Compute

$$V^* = \frac{V}{\overline{X}}, \tag{11.56}$$

where

$$V = \frac{\sum_{i=1}^{n} c_{(i)} X_{(i)}}{n^4} \tag{11.57}$$

and

$$c_i = \left(\frac{4}{3}\right)i^3 - 4ni^2 + 3n^2 i - \left(\frac{1}{2}\right)n^3 + \left(\frac{1}{2}\right)n^2 - \left(\frac{1}{2}\right)i^2 + \left(\frac{1}{6}\right)i.$$

Let

$$V' = \left\{\sqrt{(210)n}\right\}V^*. \tag{11.58}$$

a. *One-Sided Test against DMRL Alternatives.* To test

$$H_0 : F \text{ is exponential}$$

versus

$$H_7 : F \text{ is DMRL (and not exponential)},$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } V' \geq v_{1,\alpha}; \quad \text{otherwise do not reject,} \tag{11.59}$$

where the constant $v_{1,\alpha}$ is chosen to make the type I error probability equal to α; that is, $P_0\{V' \geq v_{1,\alpha}\} = \alpha$. The *R* function `dmrl.mc` returns the Monte Carlo test and the large-sample approximation. By specifying Monte Carlo is TRUE/FALSE determines whether the Monte Carlo test or the large-sample approximation is used if $n \geq 9$. If $n < 9$ the Monte Carlo test is used. The default value is FALSE, so the large-sample approximation will be used unless specified not to (see Example 11.4).

b. *One-Sided Test against IMRL Alternatives.* To test

$$H_0 : F \text{ is exponential}$$

versus

$$H_8 : F \text{ is IMRL (and not exponential)},$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } V' \leq v_{2,\alpha}; \quad \text{otherwise do not reject,} \tag{11.60}$$

where the constant $v_{2,\alpha}$ is chosen to make the type I error probability equal to α; that is, $P_0\{V' \leq v_{2,\alpha}\} = \alpha$.

c. *Two-Sided Test against DMRL and IMRL Alternatives.* To test

$$H_0 : F \text{ is exponential}$$

versus

$$H_9 : F \text{ is DMRL or IMRL (and not exponential)}$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } V' \geq v_{1,\alpha_1} \text{ or if } V' \leq v_{2,\alpha_2}; \quad \text{otherwise do not reject,} \tag{11.61}$$

where $\alpha = \alpha_1 + \alpha_2$.

Large-Sample Approximation

Under H_0, the asymptotic distribution of V' tends to the $N(0,1)$ distribution. Thus, the large-sample approximation to procedure (11.59) is

$$\text{Reject } H_0 \text{ if } V' \geq z_\alpha; \quad \text{otherwise do not reject.} \tag{11.62}$$

The large-sample approximation to procedure (11.60) is

$$\text{Reject } H_0 \text{ if } V' \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{11.63}$$

The (equal-tailed) large-sample approximation to procedure (11.61), with $\alpha_1 = \alpha_2 = \alpha/2$ is

$$\text{Reject } H_0 \text{ if } |V'| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{11.64}$$

EXAMPLE 11.4 ***Methylmercury Poisoning.***

We return to the methyl-mercury poisoning data of Example 11.1 and illustrate the calculation of the DMRL statistic via Table 11.7.

By summing the fourth column of Table 11.7, we obtain

$$\sum_{i-1}^{10} c_i X_{(i)} = 29{,}730.$$

Thus, from (11.56), we obtain

$$V = \frac{29{,}730}{10{,}000} = 2.9730.$$

Because $\overline{X} = \left(\sum_{i=1}^{10} X_i\right)/10 = 64.8$, from (11.55) and (11.58), we obtain

$$V^* = \frac{2.9730}{64.8} = .0459$$

and

$$V' = \sqrt{2100}(.0459) = 2.10.$$

Using the *R* command `dmrl.mc(methyl, alt="dmrl", exact=T)` returns $V^* = .046$ with a P-value of .01. For the large-sample approximation, use `dmrl.mc(methyl, alt="dmrl")` to get $V' = 2.10$ with a P-value of .018. Thus the test indicates that a DMRL model is preferable to an exponential model, and there is strong evidence of age deterioration.

Table 11.7 Calculation of V for the Methylmercury Poisoning Data

i	$X_{(i)}$	$c_{(i)}$	$c_i X_{(i)}$
1	42	−189	−7,938
2	43	−1	−43
3	51	122	6,222
4	61	188	11,468
5	66	205	13,530
6	69	181	12,489
7	71	124	8,804
8	81	42	3,402
9	82	−57	−4,674
10	82	−165	−13,530

Comments

18. *Implications among Life Distribution Classes.* The relationships between the five classes we have discussed, namely, IFR, IFRA, NBU, NBUE, DMRL and their dual classes DFR, DFRA, NWU, NWUE, IMRL are given in Figure 11.1. Where no implication is shown, no implication exists, as may be demonstrated by counterexample.

    Referring to Figure 11.1, we see, for example, that if F is IFR, then F is DMRL. That is, the IFR class is contained in the DMRL class. Similarly, the DMRL class is contained in the NBUE class. That there is no arrow connecting DMRL to the IFRA and NBU classes is meant to signify that a containment relation does not hold between DMRL and IFRA or between DMRL and NBU. Thus, for example, there are IFRA distributions that are not DMRL distributions and there are DMRL distributions that are not IFRA distributions. Similarly, there are NBU distributions that are not DMRL distributions and there are DMRL distributions that are not NBU distributions (see Bryson and Siddiqui (1969)).

    The classes in Figure 11.1 consist of life distributions that can be used to model situations where the lifelengths of items tend to deteriorate with age. The dual classes in Figure 11.1 can be used to model situations where the lifelengths tend to improve with age. The boundary members of each class and its dual are the exponential distributions that are used to model situations where lifelengths neither deteriorate nor improve with age. Thus, for example, the only distributions that are both DMRL and IMRL are the exponential distributions.

19. *Motivation for the DMRL Test.* Let

$$D(x, y) = \overline{F}(x)\overline{F}(y)\{m(x) - m(y)\}, \tag{11.65}$$

    where $m(x)$ is the mean residual life function (see (11.47)). $D(x, y) = 0$ for all $x \leq y$ if and only if H_0 is true. Let X and Y be independent random variables, each with life distribution F. The parameter

$$\Delta_{\text{DMRL}}(F) = E_F\{I(X < Y)D(X, Y)\} \tag{11.66}$$

    can be considered a measure of "DMRLness." In (11.66), $I(X < Y)$ is 1 if $X < Y$, and 0 otherwise. For each $x < y$, $D(x, y)$ is a weighted measure of the

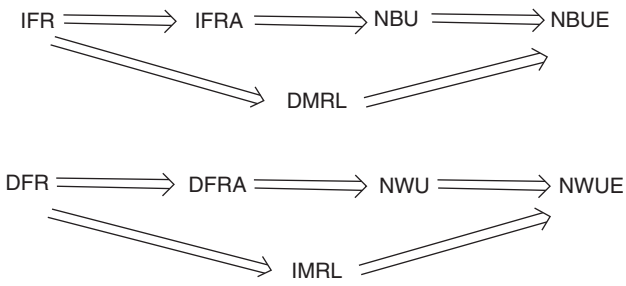

Figure 11.1 Implications among life distributions.

deviation from H_0 toward H_7 and $\Delta_{\mathrm{DMRL}}(F)$ is an average value of this deviation. The weights $\overline{F}(x)$ and $\overline{F}(y)$ represent the proportions of the population still alive at times x and y, respectively, thus furnishing comparisons concerning the mean residual lifelengths from x and y, respectively. The Hollander–Proschan statistic is obtained by substituting the empirical distribution function F_n for F in (11.66). The asymptotic normality of the statistic follows from Stigler's (1974) results on linear functions of order statistics (see Stigler (1974) and Hollander and Proschan (1975)). The exact null distribution of V' is given by Langenberg and Srinivasan (1979) for α in the upper and lower .01, .05 and .10 regions for $n = 2(1)20(5)60$. Chen, Hollander, and Langberg (1983b) extended the DMRL test to censored data by estimating $\Delta_{\mathrm{DMRL}}(F)$ using the Kaplan and Meier (1958) estimator.

Aly (1990) derives a test of H_0 versus H_7 using a different parameter to estimate "DMRLness." He finds his test outperforms V' for the three distributions he considered. His test has Pitman efficiencies with respect to V' of 1.219, 1.0714, and 1.4272 for linear failure rate, Makeham, and Weibull alternatives, respectively.

20. *The Empirical Mean Residual Life Function.* The empirical mean residual function (mrl) function, $\widehat{m}(x)$, for a sample $X_1, \ldots, X_n$, from F is obtained by replacing $\overline{F}$ by the empirical survival function, $\overline{F}_n$ in (11.47). The expression for $\widehat{m}(x)$ when this substitution is made reduces to

$$\widehat{m}(x) = \frac{\sum_{j=1}^{S(x)} (X_j^* - x)}{S(x)}, \tag{11.67}$$

where $S(x)$ denotes the number of items at time x, out of the initial sample $X_1, \ldots, X_n$, that exceed x and $X_1^*, \ldots, X_{S(x)}^*$, are those observations that exceed x. Note that $\widehat{m}(x)$ is the average, less x, of the sample values that are greater than x. Yang (1978) and Hall and Wellner (1979) showed strong consistency of $\widehat{m}(x)$ as an estimator of $m(x)$. Hall and Wellner (1979) derived nonparametric simultaneous confidence bands for $m(x)$ (see Comment 21). Guess, Hollander, and Proschan (1986) derived tests of exponentiality versus a trend change in the mrl function. They considered the situation where the turning point is known. Hawkins, Kochar, and Loader (1992) considered the situation where the turning point is unknown. The turning point procedures are discussed in Section 11.4. Mi (1994) proposed an estimator of $m(x)$ which is continuous and decreasing (increasing) when F is DMRL (IMRL).

Insurance companies are particularly interested in estimating the mrl of insurance applicants in order to set premiums. McLain and Ghosh (2011) estimate the conditional mrl function $m(x|\mathbf{z}) = E_F(X - x|X > x, \mathbf{Z} = \mathbf{z})$ in the situation where baseline covariates $\mathbf{z}$ are available and survival times are subject to censoring.

21. *Confidence Bands for the Mean Residual Life Function.* Hall and Wellner (1979) developed nonparametric simultaneous confidence bands for the mrl function. (Additional bands are presented in Csörgő and Zitikis (1996).) Assume that $E(X^r) < \infty$ for some $r > 2$. The Hall–Wellner bands are

$$\widehat{m}(x) - \frac{D_n}{\overline{F}_n(x)}, \quad \widehat{m}(x) + \frac{D_n}{\overline{F}_n(x)}, \quad 0 \le x < \infty, \tag{11.68}$$

where

$$\overline{F}_n(x) = \frac{\text{number of } X\text{-values in the sample} > x}{n} \tag{11.69}$$

is the empirical survival function, $\widehat{m}(x)$ is the mean residual life function given by (11.67), and

$$D_n = \frac{a_\alpha S_n}{n^{1/2}}, \tag{11.70}$$

where

$$S_n = \sqrt{\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}} \tag{11.71}$$

is the sample standard deviation. The value a_α is determined so that the coverage probability of the bands is approximately $1 - \alpha$. Values of a_α are given in the *R* object `b.mrl`.

The value a_α is chosen from the distribution of $\sup_{0 \le t \le 1} |\mathcal{W}_t|$, where $\mathcal{W}_t$ is the value at t of standard Brownian motion $\mathcal{W}$ (see the Weiner process, Billingsley (1968, p. 61)). The value a_α satisfies

$$P\left(\sup_{0 \le t \le 1} \mathcal{W}_t \le a_\alpha\right) = 1 - \alpha.$$

Hall and Wellner show that if F is continuous and $E(X^r) < \infty$ for some $r > 2$, then as $n \to \infty$, the limiting value of the probability given by the left-hand side of (11.72) is $1 - \alpha$. The distribution of $Y = \sup_{0 \le t \le 1} \mathcal{W}_t$ is given by (cf. Billingsley, 1968, p. 79)

$$\begin{aligned} P(Y \le a) \equiv P(a) &= \sum_{k=-\infty}^{\infty} (-1)^k \{\Phi((2k+1)a) - \Phi((2k-1)a)\} \\ &= 1 - 4\{\overline{\Phi}(a) - \overline{\Phi}(3a) + \overline{\Phi}(5a) - \cdots\}, \end{aligned}$$

where Φ denotes the standard normal cumulative distribution function and $\overline{\Phi} = 1 - \Phi$. Hall and Wellner point out the approximation $P(a) \cong 1 - 4\overline{\Phi}(a)$ gives three-place accuracy for $\alpha > 1.4$.

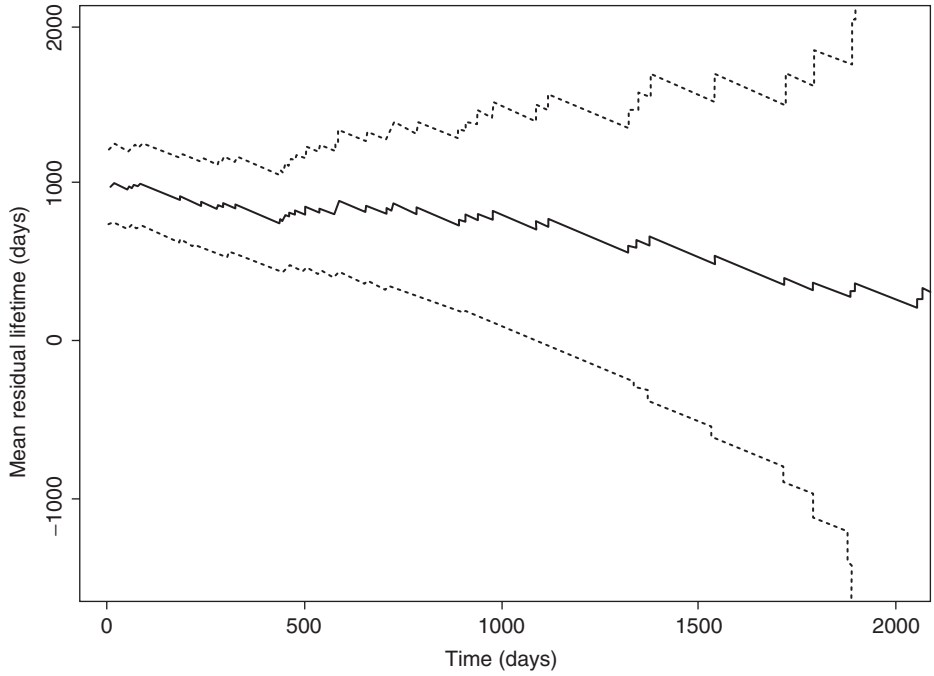

Figure 11.2 An approximate 95% confidence band for the mean residual life for the chronic granulocytic leukemia data of Table 11.6.

With the choice of a_α as given in `b.mrl`, we have

$$P\left(\widehat{m}(x) - \frac{D_n}{\overline{F}_n(x)} \le m(x) \le \widehat{m}(x) + \frac{D_n}{\overline{F}_n(x)}, \text{ for all } x \ge 0\right) \cong 1 - \alpha. \tag{11.72}$$

Thus, for example, to obtain an approximate 95% simultaneous confidence band, use a $a_{.05} = 2.2414$ when substituting into (11.70).

Figure 11.2 is a plot of $\widehat{m}(x)$ and an approximate 95% simultaneous confidence band for $m(x)$ for the chronic granulocytic leukemia data of Table 11.6. Note that the graph of $\widehat{m}(x)$ changes at each ordered X-value and is a line of slope -1 between adjacent X-values. The same is also true of the band.

The *R* program `mrl` computes the bands.

Suppose, rather than simultaneous confidence bands for all $x \ge 0$, we desire tighter bands for $m(x)$ at a specific point. Hall and Wellner (1979) show that if $\overline{F}(x) > 0$, then

$$\frac{n^{1/2}(\widehat{m}(x) - m(x))\{\overline{F}_n(x)\}^{1/2}}{s_n^*(x)}$$

tends in distribution to a $N(0, 1)$ as $n \to \infty$, where $s_n^*(x)$ is the sample standard deviation of the observations that exceed x. Using the X^* notation of (11.67),

$$s_n^*(x) = \left\{ \frac{\sum_{j=1}^{S(x)} (X_j^* - \overline{X}(x))^2}{n - 1} \right\},$$

where $\overline{X}(x) = \sum_{j=1}^{S(x)} X_j^*/S(x)$ is the mean of the observations that exceed x. It follows that

$$\left(\widehat{m}(x) - \frac{z_{\alpha/2} s_n^*(x)}{\left\{ n\overline{F}_n(x) \right\}^{1/2}},\ \widehat{m}(x) + \frac{z_{\alpha/2} s_n^*(x)}{\left\{ n\overline{F}_n(x) \right\}^{1/2}} \right)$$

is an approximate $100(1-\alpha)\%$ confidence interval for $m(x)$ at the point x.

Properties

1. *Consistency.* See Hollander and Proschan (1975) for the consistency class of the test defined by (11.59). In particular, if F is continuous, DMRL (and not exponential), the test is consistent.
2. *Asymptotic Normality.* See Hollander and Proschan (1975).
3. *Efficiency.* See Hollander and Proschan (1975), Klefsjö (1983), and Section 11.8.

Problems

17. Apply the DMRL test to the chronic granulocytic leukemia data of Table 11.6.

18. Apply the DMRL test to the air-conditioning system failure data of Table 11.2. Compare your result to the results from Problem 11.

19. (a) Calculate the estimated mean residual life function for the air-conditioning system data of Table 11.4.
(b) Calculate approximate 95% confidence bands for the true mean residual life function.

20. Apply the DMRL test to the guinea pig survival data of Table 11.3. Compare your results to the results in Problems 2 and 16.

21. (a) Calculate the estimated mean residual life function for the guinea pig survival data of Table 11.3.
(b) Calculate approximate 95% confidence bands for the true mean residual life function.

22. Table 11.8, based on data in Zacks (1992), gives the pneumatic pressure (kg/cm^2) required to break 20 concrete cubes of dimensions $10 \times 10 \times 10$ cm^3.
(a) Calculate an approximate 92% confidence band for $m(x)$.
(b) Calculate an approximate 92% confidence interval for $m(229.7)$. Compare the limits of the band at 229.7 with the limits of the interval.

23. Describe a situation in which it might be expected that the mean residual life function would be initially increasing and then later decreasing.

Table 11.8 Pneumatic Pressures Required to Break Concrete Cubes

94.9	106.9	229.7	275.7	144.5	112.8	159.3	153.1	270.6	322.0
216.4	544.6	266.2	263.6	138.5	79.0	114.6	66.1	131.2	91.1

Source: S. Zacks (1992).

11.4 A TEST OF EXPONENTIALITY VERSUS A TREND CHANGE IN MEAN RESIDUAL LIFE (GUESS–HOLLANDER–PROSCHAN)

Hypothesis

The null hypothesis is $H_0 : F$ is exponential. The alternatives are specified in terms of two nonparametric classes of distributions defined via the mean residual life function. The initially increasing, then decreasing mean residual life (IDMRL) class models aging that is initially beneficial and then is adverse. The decreasing initially, then increasing mean residual life (DIMRL) class models aging that is initially adverse and then is beneficial. A distribution with finite mean is said to be a member of the IDMRL class if there exists a turning point $\tau \geq 0$ such that

$$m(x) \leq m(y), \quad \text{for } 0 \leq x \leq y < \tau,$$
$$m(x) \geq m(y), \quad \text{for } \tau \leq x \leq y,$$

where $m(x)$ is the mrl function. Similarly, a distribution F with finite mean is said to be a member of the DIMRL class if there exists a turning point $\tau \geq 0$ such that

$$m(x) \geq m(y), \quad \text{for } 0 \leq x \leq y < \tau,$$
$$m(x) \leq m(y), \quad \text{for } \tau \leq x \leq y.$$

Procedure

We treat the case where the turning point τ is known, using a procedure due to Guess, Hollander, and Proschan (1986). The GHP statistic can be used to detect IDMRL and DIMRL alternatives. In Comment 25, we consider the case where the turning point is not known and describe a procedure due to Hawkins, Kochar, and Loader (1992). The HKL statistic can be used to detect IDMRL alternatives, but it is not designed to detect DIMRL alternatives.

To define the GHP test statistic, we set

$$T_1 = \sum_{i=1}^{i^*} B_1\left(\frac{(n-i+1)}{n}\right)\left(X_{(i)} - X_{(i-1)}\right) + B_1\left(\frac{(n-i^*)}{n}\right)(\tau - X_{(i^*)})$$
$$+ B_2\left(\frac{(n-i^*)}{n}\right)\left(X_{(i^*+1)} - \tau\right) + \sum_{i=i^*+2}^{n} B_2\left(\frac{(n-i+1)}{n}\right)(X_{(i)} - X_{(i-1)}), \tag{11.73}$$

where (letting $X_{(1)} < \cdots < X_{(n)}$ denote the ordered X-values with $X_{(0)} = 0$) the integer i^* is defined by

$$0 < X_{(1)} < \cdots < X_{(i^*)} \leq \tau < X_{(i^*+1)} < \cdots < X_{(n)}. \tag{11.74}$$

The functions B_1 and B_2 in (11.73) are defined as

$$B_1(u) = \left[\frac{2}{3} - F_n(\tau) + \frac{1}{2}F_n^2(\tau)\right]u + \left[-1 + F_n(\tau) - \frac{1}{2}F_n^2(\tau)\right]u^2 + \frac{1}{3}u^4, \tag{11.75}$$

$$B_2(u) = \left[-\frac{1}{6} + \frac{1}{2}F_n(\tau) - \frac{1}{2}F_n^2(\tau) + \frac{1}{3}F_n^3(\tau)\right]u + \left[\frac{1}{2} - F_n(\tau) + \frac{1}{2}F_n^2(\tau)\right]u^2 - \frac{1}{3}u^4. \tag{11.76}$$

where F_n is the empirical distribution function defined by (11.93). For data where there are ties, use

$$T_1 = \sum_{i=1}^{i^*} B_1\left(\frac{s_{i-1}}{n}\right)\left(\widetilde{X}_{ik} - \widetilde{X}_{(i-1)k}\right) + B_1\left(\frac{s_{i^*}}{n}\right)\left(\tau - \widetilde{X}_{i^*k}\right)$$
$$+ B_2\left(\frac{s_{i^*}}{n}\right)\left(\widetilde{X}_{(i^*+1)k} - \tau\right) + \sum_{i=i^*+2}^{k} B_2\left(\frac{s_{i-1}}{n}\right)\left(\widetilde{X}_{ik} - \widetilde{X}_{(i-1)k}\right), \tag{11.77}$$

where

$$0 = \widetilde{X}_{0k} < \widetilde{X}_{1k} < \cdots < \widetilde{X}_{i^*k} \le \tau < \widetilde{X}_{(i^*+1)k} < \cdots < \widetilde{X}_{kk} \tag{11.78}$$

are the distinct ordered observations,

$$n_i = \text{number of observed deaths at time } \widetilde{X}_{ik}, \tag{11.79}$$

$$s_i = n - \sum_{t=0}^{i} n_t, \quad \text{for } i = 0, 1, \ldots, k < n. \tag{11.80}$$

In (11.80), $n_i \neq 0, i = 1, \ldots, k$ but n_0 is allowed to be 0.

Under H_0, the distribution of $n^{1/2}T_1$ tends, as $n \to \infty$, to a normal distribution with mean 0 and variance

$$\sigma_{T_1}^2 = \mu^2\left[-\frac{1}{15}F^5(\tau) + \frac{1}{6}F^4(\tau) - \frac{1}{6}F^3(\tau) + \frac{1}{10}F^2(\tau) - \frac{1}{30}F(\tau) + \frac{1}{120}\right], \tag{11.81}$$

where μ is the mean of F. The test, for the case where τ is known, uses the statistic $n^{1/2}T_1/\hat{\sigma}_{T_1}$, where

$$\widehat{\sigma}_{T_1}^2 = \overline{X}^2\left[-\frac{1}{15}F_n^5(\tau) + \frac{1}{6}F_n^4(\tau) - \frac{1}{6}F_n^3(\tau) + \frac{1}{10}F_n^2(\tau) - \frac{1}{30}F_n(\tau) + \frac{1}{210}\right]. \tag{11.82}$$

a. *One-Sided Test against IDMRL Alternatives.* To test

$$H_0 : F \text{ is exponential}$$

versus

$$H_{10} : F \text{ is IDMRL} \quad \text{(and not exponential)},$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } \frac{n^{1/2}T_1}{\widehat{\sigma}_{T_1}} \ge z_\alpha; \quad \text{otherwise do not reject.} \tag{11.83}$$

b. *One-Sided Test against DIMRL Alternatives.* To test

$$H_0 : F \text{ is exponential}$$

versus

$$H_{11} : F \text{ is DIMRL} \quad \text{(and not exponential)}$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } \frac{n^{1/2}T_1}{\widehat{\sigma}_{T_1}} \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{11.84}$$

c. *One-Sided Test against IDMRL and DIMRL Alternatives.* To test

$$H_0 : F \text{ is exponential}$$

versus

$$H_{12} : F \text{ is IDMRL or DIMRL} \quad \text{(and not exponential)},$$

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } \frac{n^{1/2}T_1}{\widehat{\sigma}_{T_1}} \geq z_{\alpha_1} \text{ or if } \frac{n^{1/2}T_1}{\widehat{\sigma}_{T_1}} \leq -z_{\alpha_2}; \quad \text{otherwise do not reject,} \tag{11.85}$$

where $\alpha = \alpha_1 + \alpha_2$.

The *R* function `tc` returns the value of T_1, $\hat{\sigma}^2_{T_1}$ and $T_1^* = n^{1/2}T_1/\hat{\sigma}_{T_1}$ along with the corresponding P-value. (See Example 11.5.)

EXAMPLE 11.5 ***Lifelengths of Guinea Pigs Injected with Tubercle Bacilli.***

Bjerkedal (1960) studied the lifelengths of guinea pigs injected with different amounts of tubercle bacilli. Guinea pigs are known to have a high susceptibility to human tuberculosis. This is one reason experimenters choose the species. In Bjerkedal's study (M), the animals in a single cage are under the same regimen. The regimen number is the common logarithm of the number of bacillary units in .5ml of the challenge solution. In Table 11.3 (see Problem 2), we presented the data for regimen 4.3, which corresponds to 2.2×10^4 bacillary units per .5 ml (because $\log_{10}(2.2 \times 10^4) = 4.342$).

It is natural to postulate DIMRL alternatives in this situation. The motivation is that initially the injection of tubercle bacilli causes an adverse stage of aging, but after the guinea pigs have survived this initial adverse stage, their natural systems recover to yield a beneficial stage.

Hall and Wellner (1981) used regimen 4.3 and fit a parametric distribution that is in the DIMRL class. They estimated the turning point as $\widehat{\tau} = 91.9$. We apply the DIMRL test to regimen 5.5 of Table 11.9, using 91.9 as the "known" turning point. This is a reasonable a priori choice because regimens 4.3 and 5.5 are closely related.

For the DIMRL test applied to the pigs data of Table 11.9 with $\tau = 91.9$, we obtain $T_1 = -.7956$, $\hat{\sigma}^2_{T_1} = 7.1072$, and $n^{1/2}T_1/\hat{\sigma}_{T_1} = -2.53$, yielding a P-value of .006. Thus there is strong evidence to reject H_0 in favor of H_{11}.

To do the calculations using *R*, apply `tc(pigs, tau = 91.9, alt="dimrl")` to obtain $T_1 = -.7956$, and $P = .006$.

Table 11.9 Ordered Survival Times in Days of Guinea Pigs under Regimen 5.5

i	1	2	3	4	5	6	7	8	9
$X_{(i)}$	43	45	53	56	56	57	58	66	67
i	10	11	12	13	14	15	16	17	18
$X_{(i)}$	73	74	79	80	80	81	81	81	82
i	19	20	21	22	23	24	25	26	27
$X_{(i)}$	83	83	84	88	89	91	91	92	92
i	28	29	30	31	32	33	34	35	36
$X_{(i)}$	97	99	99	100	100	101	102	102	102
i	37	38	39	40	41	42	43	44	45
$X_{(i)}$	103	104	107	108	109	113	114	118	121
i	46	47	48	49	50	51	52	53	54
$X_{(i)}$	123	126	128	137	138	139	144	145	147
i	55	56	57	58	59	60	61	62	63
$X_{(i)}$	156	162	174	178	179	184	191	198	211
i	64	65	66	67	68	69	70	71	72
$X_{(i)}$	214	243	249	329	380	403	511	522	598

Source: T. Bjerkedal (1960).

Comments

22. *Motivation for the IDMRL Test.* The motivation for the IDMRL test is similar to the motivation (see Comment 19) for the DMRL test of Section 11.3. The test statistic T_1 estimates a parameter $\Delta_{\text{IDMRL}}(F)$ that is a weighted measure of the degree to which F satisfies the IDMRL property. Specifically,

$$\Delta_{\text{IDMRL}}(F) = E_F\{I(X < Y < \tau)D(X,Y) - I(\tau < X < Y)D(X,Y)\}, \quad (11.86)$$

where D is defined by (11.65) and the indicator functions in (11.86) are defined as follows. Let X, Y be independent random variables each with distribution F. Then $I(X < Y < \tau)$ is 1 if $X < Y < \tau$ and 0 otherwise. Similarly, $I(\tau < X < Y)$ is 1 if $\tau < X < Y$ and 0 otherwise.

Asymptotic normality of T_1, similarly standardized, is proved directly in Guess, Hollander, and Proschan (1986). Exact null distributions of T_1 can be obtained, but the distributions depend on τ and thus creating tables for different τ values is impractical. There are exact tables, however, for the related problem where one knows the proportion ρ of the population that "dies" at or before the turning point. (See Comment 24.)

23. *Some Situations Where Knowledge of τ May Be Available.* Knowledge of τ might be available in a situation where one is studying a biological organism in a physical model of a disease process. In such a situation, it may be the case that the first 3 months (say) constitute an incubation period. As another example, consider a training program for future doctors or a recruiting program for a military service. The value of τ may be known by the length of the intensive stage designed to eliminate the weaker students or recruits.

24. *The IDMRL Test When the Proportion ρ (of the Population that Dies before or at the Turning Point) Is Known.* In a training program, for example, past experience

Table 11.10 Estimated Large-Sample Percentiles of $T^{(2)}$

$1-\alpha$	.90	.95	.99
$(1-\alpha)$th percentile	1.41	1.59	1.93

Source: D. L. Hawkins, S. Kochar, and C. Loader (1992).

with earlier classes of recruits may provide knowledge of the proportion ρ of the population that dies at or before the turning point. Such a ρ would satisfy $F(\tau) = \rho$. Guess, Hollander, and Proschan (1986) proposed a statistic (similar to the T_1 statistic) for this situation and provided exact critical values for the case $\rho = .25$ for the sample sizes $n = 2, \ldots, 30$ in the lower and upper $\alpha = .01, .05$, and .10 regions. Their statistic can be used to detect both IDMRL and DIMRL alternatives. Tables for $\rho = 0(.1)1, .75, \frac{1}{3}$, and $\frac{2}{3}$ are given in Guess (1984).

25. *Case Where the Turning Point Is Unknown.* Hawkins, Kochar, and Loader (1992) presented two statistics, $T^{(1)}$ and $T^{(2)}$, for the case where the turning point is unknown. Their statistics can be used to test exponentiality versus IDMRL alternatives, but they are not designed to detect DIMRL alternatives. In Hawkins, Kochar, and Loader (1992), Monte Carlo power comparisons showed that $T^{(2)}$ outperformed $T^{(1)}$, and thus we present $T^{(2)}$ here. The HKL statistic is an appropriately standardized estimator of a function of F that is 0 when F is exponential and is positive when F is IDMRL (see Hawkins, Kochar, and Loader (1992) for details). The statistic can be expressed as

$$T^{(2)} = n^{1/2}(\overline{X})^{-1} \max_{0 \le k \le n} \xi_k, \tag{11.87}$$

where

$$\xi_k = A - 2\left(1 - \frac{k}{n}\right) \sum_{j=k}^{n-1} \left(1 - \frac{j}{n}\right) D_j^* + 4 \sum_{j=k}^{n-1} \left(1 - \frac{j}{n}\right)^2 D_j^*, \tag{11.88}$$

and

$$A = -X_{(1)} + \sum_{j=1}^{n-1} c_j D_j^*, \tag{11.89}$$

where

$$c_j = 1 - \frac{j}{n} - 2\left(1 - \frac{j}{n}\right)^2 \tag{11.90}$$

and

$$D_j^* = X_{(j+1)} - X_{(j)}.$$

The HKL test of H_0 versus H_{10} (F is IDMRL) rejects H_0 for significantly large values of $T^{(2)}$ and accepts H_0 otherwise. (To simplify the notation, in (11.87)–(11.90) the dependence on n of $T^{(2)}$, ξ_k, A, and c_j has been suppressed. In HKL, these quantities are called $T_n^{(2)}$, ξ_{nk}, A_n, and c_{nj}.)

HKL do not provide exact tables for the null distribution of $T^{(2)}$ but instead base their test on estimated critical values obtained from an asymptotic approximation given by (2.6) of their paper. Table 11.10 contains selected estimated percentiles of the asymptotic distribution of $T^{(2)}$.

From Table 11.10, we see that for large n,

$$P_0\,(T^{(2)} \geq 1.41) \cong .10, \quad P_0(T^{(2)} \geq 1.59) \cong .05, \quad P_0(T^{(2)} \geq 1.93) \cong .01.$$

Thus, for example, the approximate $\alpha = .05$ test of exponentiality versus IDMRL alternatives rejects H_0 if $T^{(2)} \geq 1.59$ and accepts H_0 if $T^{(2)} < 1.59$.

Aly (1990) proposed competitors of the Guess, Hollander, and Proschan (1986) tests for the case where the turning point τ (or the proportion ρ that dies before or at the turning point) is known and a competitor of the Hawkins, Kochar, and Loader (1992) tests when neither τ nor ρ is known.

Properties

1. *Consistency.* The test defined by (11.83) is consistent against those F distributions for which the parameter $T(F)$, given by (2.1) of Guess, Hollander, and Proschan (1986), is positive. In particular, if F is continuous, and IDMRL (and not exponential), the test is consistent.
2. *Asymptotic Normality.* See Guess, Hollander, and Proschan (1986).
3. *Efficiency.* See Hawkins, Kochar, and Loader (1992) and Section 11.8. Asymptotic relative efficiencies are unavailable, but HKL give some Monte Carlo power comparisons of their tests with respect to the GHP test.

Problems

24. The data for Bjerkedal's study M, regimen 6.6, are given in Table 11.11. Test for a trend change in the mrl function at $\tau = 89$.

25. Calculate the estimated mrl function for the guinea pig survival data of Table 11.9.

26. Calculate the estimated mrl function for the guinea pig survival data of Table 11.11.

27. Test for a trend change in the mrl using the chronic granulocytic leukemia data of Table 11.6.

28. Describe a situation (different from those mentioned in Comment 23) where one might expect a trend change in the mean residual life.

11.5 A CONFIDENCE BAND FOR THE DISTRIBUTION FUNCTION (KOLMOGOROV)

Assumption

A1. The observations are a random sample from the underlying continuous population. That is, the X's are independent and identically distributed according to a continuous distribution F.

Table 11.11 Ordered Survival Times in Days of Guinea Pigs under Regimen 6.6

i	1	2	3	4	5	6	7	8	9
$X_{(i)}$	12	15	22	24	24	32	32	33	34
i	10	11	12	13	14	15	16	17	18
$X_{(i)}$	38	38	43	44	48	52	53	54	54
i	19	20	21	22	23	24	25	26	27
$X_{(i)}$	55	56	57	58	58	59	60	60	60
i	28	29	30	31	32	33	34	35	36
$X_{(i)}$	60	61	62	63	65	65	67	68	70
i	37	38	39	40	41	42	43	44	45
$X_{(i)}$	70	72	73	75	76	76	81	83	84
i	46	47	48	49	50	51	52	53	54
$X_{(i)}$	85	87	91	95	96	98	99	109	110
i	55	56	57	58	59	60	61	62	63
$X_{(i)}$	121	127	129	131	143	146	146	175	175
i	64	65	66	67	68	69	70	71	72
$X_{(i)}$	211	233	258	258	263	297	341	341	376

Source: T. Bjerkedal (1960).

Note that we do not require Assumption A2 (used in Sections 11.1–11.4) that F be a life distribution. Here, the X's can also assume negative values.

We seek a simultaneous confidence band for the unknown distribution function; that is, we seek random functions (i.e., functions that depend on the observed sample values $X_1, \ldots, X_n$) $\ell(x)$ and $u(x)$ satisfying

$$P\left\{\ell(x) \le F(x) \le u(x), \text{ for all } x\right\} \ge 1 - \alpha.$$

We then say $\{\ell(x), u(x)\}$ is a simultaneous confidence band (or, more simply, a confidence band) for $F(x)$ with confidence at least $100(1-\alpha)\%$.

The bands are based on the null distribution of the Kolomogorov statistic. (See Comment 27.) They are defined as

$$\ell(x) = \begin{cases} F_n(x) - d_\alpha, & \text{if } F_n(x) - d_\alpha \ge 0, \\ 0, & \text{if } F_n(x) - d_\alpha < 0, \end{cases} \tag{11.91}$$

and

$$u(x) = \begin{cases} F_n(x) + d_\alpha, & \text{if } F_n(x) + d_\alpha \le 1, \\ 1, & \text{if } F_n(x) + d_\alpha > 1. \end{cases} \tag{11.92}$$

In (11.91) and (11.92), $F_n(x)$ is the empirical distribution function of the X's defined by

$$F_n(x) = \frac{\text{number of } X\text{'s in the sample} \le x}{n}, \tag{11.93}$$

and d_α is the upper α percentile point of the distribution of Kolmogorov's statistic D (see Comments 26 and 5.41), defined as

$$D = \sup_{-\infty < x < \infty} \{|F_n(x) - F(x)|\}; \tag{11.94}$$

Table 11.12 Calculation of the Confidence Band for $F(x)$ for the Methylmercury Poisoning Data

x	$F_{10}(x)$	$\ell(x)$	$u(x)$
$x < 42$	0	0	.409
$42 \le x < 43$	.1	0	.509
$43 \le x < 51$	.2	0	.609
$51 \le x < 61$	.3	0	.709
$61 \le x < 66$	.4	0	.809
$66 \le x < 69$	.5	.091	.909
$69 \le x < 71$	.6	.191	1
$71 \le x < 81$	.7	.291	1
$81 \le x < 82$	.8	.391	1
$x \ge 82$	1	.591	1

that is, d_α satisfies

$$P_F\left(\sup_{-\infty<x<\infty}\{|F_n(x) - F(x)|\} < d_\alpha\right) = 1 - \alpha. \tag{11.95}$$

The *R* program `kd` computes the value of D. The *R* program `kolmogorov` computes the probability under the null hypothesis that $D \ge d$.

Large-Sample Approximation

Kolmogorov (1933) and Smirnov (1939) (see also Feller (1948)) proved that as $n \to \infty$, $P_F(D \le z/\sqrt{n})$ tends to $L(z)$, where

$$L(z) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2z^2}. \tag{11.96}$$

Smirnov (1948) presents a table of values of $L(z)$. For large n, d_α can be approximated by

$$d_\alpha \doteq \frac{z_\alpha^*}{\sqrt{n}}, \tag{11.97}$$

where $L(z_\alpha^*) = 1 - \alpha$. The values of z_α^* for $\alpha = .20, .10, .05, .02, .01$ are 1.07, 1.22, 1.36, 1.52, 1.63, respectively. Thus, for example, with $\alpha = .05$, $d_{.05} \doteq 1.36/\sqrt{n}$. The large-sample approximation is reasonably good for $n \ge 38$.

EXAMPLE 11.6 ***Example 11.1 Continued.***

We use the data of Example 11.1 to determine a confidence band for the distribution of lifelengths. The ordered times to death (in days) are 42, 43, 51, 61, 66, 69, 71, 81, 82, 82. We illustrate the calculation of the 95% band. Table 11.12 illustrates the calculation of $F_n(x)$, $\ell(x)$, and $u(x)$. Figure 11.3 is a plot of $F_n(x)$ and the 95% confidence band.

The *R* function `ecdf.ks.CI(x, main=NULL, sub=NULL, xlab=deparse (substitute(x)),...)` creates the 95% confidence band. Here x is a vector of data

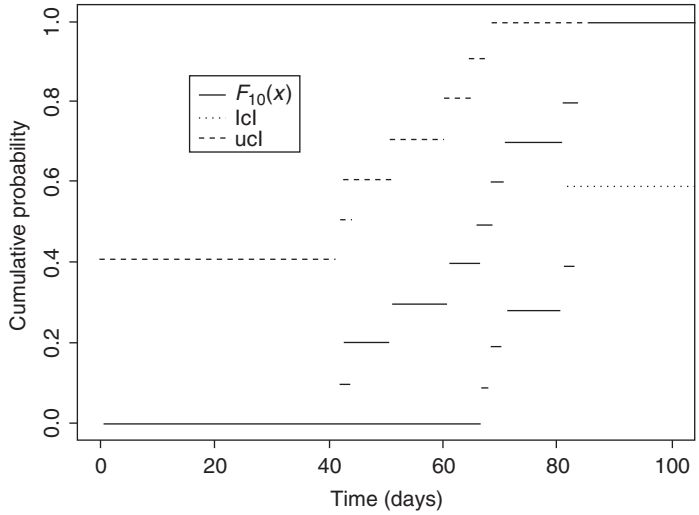

Figure 11.3 A 95% confidence band for $F(x)$ for the methylmercury poisoning data.

of length n and main is the title of the plot. The default is `ecdf(`x`)` + 95% K.S.Bands. The "..." signifies you can add any plotting options available. The function returns a list with three elements, namely, lower (the lower values of the confidence band), upper (the upper values of the confidence band), and D (the value of Kolmogorov's statistic D).

The null distribution of D is derived under the assumption that the underlying distribution is continuous. If there are tied observations (as is the case for the methylmercury poisoning data of this example), the confidence band and the goodness-of-fit test based on D are approximate, not exact.

Comments

26. *Derivation of the Confidence Band.* Using the null distribution of D, we can determine a critical value d_α such that

$$P_F(|F_n(x) - F(x)| \le d_\alpha, \text{ for all } x) = 1 - \alpha. \tag{11.98}$$

The subscript F in the notation P_F means that the probability is being computed under the assumption that $X_1, \ldots, X_n$ is a random sample from F. We can rewrite (11.98) as

$$P_F(-d_\alpha \le F_n(x) - F(x) \le d_\alpha, \text{ for all } x) = 1 - \alpha. \tag{11.99}$$

Equation (11.99) is equivalent to

$$P_F(d_\alpha \ge -F_n(x) + F(x) \ge -d_\alpha, \text{ for all } x) = 1 - \alpha. \tag{11.100}$$

Equation (11.100) is equivalent to

$$P_F(F_n(x) - d_\alpha \le F(x) \le F_n(x) + d_\alpha, \text{ for all } x) = 1 - \alpha. \tag{11.101}$$

From (11.101), we can conclude that $\{F_n(x) - d_\alpha, F_n(x) + d_\alpha\}$ is a $1 - \alpha$ confidence band for F. It is, however, possible that the upper boundary $F_n(x) + d_\alpha$ may exceed 1. We know that $F(x)$ itself cannot exceed 1, so we can lower the upper boundary to $u(x) = \text{minimum}[F_n(x) + d_\alpha, 1]$. Similarly, it is possible that the lower boundary $F_n(x) - d_\alpha$ may be less than 0. We know that $F(x)$ cannot be less than 0, so we can raise the lower boundary to $\ell(x) = \text{maximum}[F_n(x) - d_\alpha, 0]$. These adjustments yield the band given by (11.91) and (11.92).

27. *Goodness-of-Fit Test Based on D.* Suppose we have a random sample $X_1, \ldots, X_n$ from a population with distribution function $F(x)$. Suppose further there is reason to believe (perhaps based on previous experience) that F_0 is some completely specified distribution. For example, F_0 may be specified to be a normal distribution with mean 1 and standard deviation 2 or an exponential distribution (see (11.3)) with scale parameter $\lambda = \frac{1}{2}$. Kolmogorov's test of the null hypothesis,

$$H_0^* : F(x) = F_0(x) \quad \text{for all } x, \tag{11.102}$$

against the alternative,

$$H_A^* : F(x) \neq F_0(x) \text{ for at least one } x, \tag{11.103}$$

is based on

$$D = \sup_{-\infty < x < \infty} \{|F_n(x) - F_0(x)|\}, \tag{11.104}$$

where $F_n(x)$, the sample distribution function, is

$$F_n(x) = \frac{\text{number of } X\text{'s in the sample} \leq x}{n}. \tag{11.105}$$

Alternatively, $F_n(x)$ can be expressed as

$$F_n(x) = \begin{cases} 0, & x < X_{(1)}, \\ \frac{i}{n}, & X_{(i)} \leq x < X_{(i+1)}, \\ 1, & x > X_{(n)}. \end{cases} \tag{11.106}$$

In (11.104), $\sup_{-\infty < x < \infty}$ denotes the supremum over all x of the absolute value of the difference $F_n(x) - F_0(x)$. If $F_n(x)$ and $F_0(x)$ are plotted as ordinates against x as abscissa, D is the value of the largest vertical distance between F_n and F_0. The supremum occurs at one of the $X_{(i)}$'s (i.e., at one of the jump points of F_n) or just to the left of one of the $X_{(i)}$'s.

Formally, Kolmogorov's test, at the α level of significance, is

$$\text{Reject } H_0^* \text{ if } D \geq d_\alpha; \quad \text{otherwise do not reject,} \tag{11.107}$$

where d_α satisfies $P_{F_0}(D \geq d_\alpha) = \alpha$. The *R* function `find.kol(d,n)` gives the probability under the null hypothesis that $D \geq d$.

The motivation for the test is as follows. The sample distribution function $F_n(x)$ has many desirable properties as an estimator of the underlying distribution $F(x)$ from which the sample is drawn. In particular, $F_n(x)$ converges to

$F(x)$. If the hypothesized distribution F_0 is the true distribution F, then $F_n(x)$ should be "close" to $F_0(x)$. The statistic D_n is the largest vertical distance between F_n and F_0 and this largest distance should be small if H_0^* is true.

When all n observations are distinct, D can be computed as

$$D = \max_{i=1,\ldots,n} M_i, \tag{11.108}$$

where

$$M_i = \max\left\{\left|\frac{i}{n} - F_0(X_{(i)})\right|, \left|\frac{(i-1)}{n} - F_0(X_{(i)})\right|\right\}. \tag{11.109}$$

If there are tied observations, let k denote the number of distinct observations and let $Z_{(1)} < \cdots < Z_{(k)}$ denote the ordered distinct observations. Then D can be computed as

$$D = \max_{i=1,\ldots,n} M_i', \tag{11.110}$$

where

$$M_i' = \max\{|F_n(Z_{(i)}) - F_0(Z_{(i)}|, |F_n(Z_{(i-1)}) - F_0(Z_{(i)})|\}. \tag{11.111}$$

When F is continuous, the statistic D has a continuous distribution. The statistic D is distribution-free under H_0^* when F_0 is a continuous distribution. To see this, let $X_{(0)} < \cdots < X_{(n+1)}$ denote the order statistics, where $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$. When $X_1, \ldots, X_n$ are independent and identically distributed according to the continuous distribution F_0, it can be shown using the probability integral transformation (cf. Casella and Berger (2002, pp. 54–55)) that $F_0(X_1), \ldots, F_0(X_n)$ are independent and identically distributed according to the uniform distribution on [0, 1]. It follows that $F_0(X_{(1)}), \ldots, F_0(X_{(n)})$ have the same joint distribution as that of the order statistics from a sample of size n from the uniform distribution on $[0, 1]$. Therefore, in determining the distribution of D under H_0^*, without loss of generality F_0 can be taken to be the uniform distribution on $[0, 1]$. That is, $F_0(x) = 0$ for $x < 0$, $F_0(x) = x$ for $0 \le x \le 1$, $F_0(x) = 1$ for $x > 1$. This simplifies the calculations used to determine critical values d_α. For further details, see Birnbaum (1952) and Miller (1956).

To illustrate the test based on D we return to settling velocity data of Table 3.12. Suppose that we wish to test if the data are from a normal population with mean 14 and standard deviation 2. That is, suppose the hypothesized distribution is $F_0 = N(14, 2)$. The seven ordered values are 12.8, 12.9, 13.3, 13.4, 13.7, 13.8, 14.5. The values $F_0(X_{(1)}), \ldots, F_0(X_{(7)})$ for use in (11.109) are calculated as follows. Let Y denote a $N(14, 2)$ random variable. Then,

$$F_0(X_{(1)}) = P(Y \le 12.8) = P\left(\frac{Y - 14}{2} \le \frac{12.8 - 14}{2}\right) = P(Z \le -.6)$$
$$= .2743,$$

where Z has a $N(0, 1)$ distribution. Letting $\Phi(z)$ denote the area under the $N(0, 1)$ curve to the left of z (i.e., $P(Z \le z)$), we find the value .2743 from

`pnorm(−.6)`. Note that, by symmetry, $\Phi(-.6) = 1 - \Phi(.6) = P(Z \geq .6)$. Similarly, we find

$$F_0(X_{(2)}) = P(Y \leq 12.9) = P\left(\frac{Y-14}{2} \leq \frac{12.9-14}{2}\right) = P(Z \leq -.55)$$
$$= \Phi(-.55) = .2912,$$
$$F_0(X_{(3)}) = P(Y \leq 13.3) = P\left(\frac{Y-14}{2} \leq \frac{13.3-14}{2}\right) = P(Z \leq -.35)$$
$$= \Phi(-.35) = .3632,$$
$$F_0(X_{(4)}) = P(Y \leq 13.4) = P\left(\frac{Y-14}{2} \leq \frac{13.4-14}{2}\right) = P(Z \leq -.3)$$
$$= \Phi(-.3) = .3821,$$
$$F_0(X_{(5)}) = P(Y \leq 13.7) = P\left(\frac{Y-14}{2} \leq \frac{13.7-14}{2}\right) = P(Z \leq -.15)$$
$$= \Phi(-.15) = .4404,$$
$$F_0(X_{(6)}) = P(Y \leq 13.8) = P\left(\frac{Y-14}{2} \leq \frac{13.8-14}{2}\right) = P(Z \leq -.1)$$
$$= \Phi(-.1) = .4602,$$
$$F_0(X_{(7)}) = P(Y \leq 14.5) = P\left(\frac{Y-14}{2} \leq \frac{14.5-14}{2}\right) = P(Z \leq -.25)$$
$$= \Phi(.25) = .5987.$$

Table 11.13 illustrates the calculation of D using (11.108) and (11.109).

From Table 11.13, (11.108) and (11.109), we find $D = .4013$. The *R* function `kolmogorov(x, fnc,...)` computes D and gives the P-value. Here, x is a vector of length n and fnc is the functional form of the pdf of F. The first argument must be the data. The ... is for all the parameters besides the data that fnc needs to operate.

With `velocity<-c(12.8, 12.9, 13.3, 13.4, 13.7, 13.8, 14.5)` apply `kolmogorov(velocity, pnorm, mean=14, sd=2)` to obtain $D = .4013$ and $P = .157$. Therefore, there is not strong evidence to reject H_0^*. We should not, however, be surprised that H_0^* is not rejected. With the small sample size $n = 7$, Kolmogorov's test will not have good power against reasonable alternatives to H_0^*. Even when the sample sizes are large, there are many types of distributions for which the Kolmogorov statistic D will have low power (cf. Fan (1996)).

28. *Goodness-of-Fit Test for a Composite Null Hypothesis.* In Comment 26, D was used to test the simple null hypothesis H_0^*, where the underlying distribution, F, is completely specified under the null hypothesis. If, instead, the underlying distribution, F, is not completely specified under the null hypothesis, but rather the null hypothesis asserts that F is a member of some parametric family with one or more parameters unspecified, this is known as a *composite null hypothesis*. The statistic D, suitably modified, can be used to test a composite null

Table 11.13 Calculation of D for the Settling Velocity Data in the Case Where $F_0 = N(14, 2)$

i	$X_{(i)}$	$\frac{i}{n} - F_0(X_{(i)})$	$F_0(X_{(i)}) - \frac{(i-1)}{n}$	$\max\{\lvert\frac{i}{n} - F_0(X_{(i)})\rvert, \lvert\frac{(i-1)}{n} - F_0(X_{(i)})\rvert\}$
1	12.8	$\frac{1}{7} - .2743 = -.1314$	$.2743 - 0 = .2743$	.2743
2	12.9	$\frac{2}{7} - .2912 = -.0055$	$.2912 - \frac{1}{7} = .1483$	.1483
3	13.3	$\frac{3}{7} - .3632 = .0654$	$.3632 - \frac{2}{7} = .0775$	.0775
4	13.4	$\frac{4}{7} - .3821 = .1893$	$.3821 - \frac{3}{7} = -.0465$	.1893
5	13.7	$\frac{5}{7} - .4404 = .2739$	$.4404 - \frac{4}{7} = -.1310$	.2739
6	13.8	$\frac{6}{7} - .4602 = .3969$	$.4602 - \frac{5}{7} = -.2541$	.3969
7	14.5	$1 - .5987 = .4013$	$.5987 - \frac{6}{7} = -.2584$	.4013

hypothesis. The modified statistic is D', where

$$D' = \sup_{-\infty<x<\infty} \left|F_n(x) - \widehat{F}_0(x)\right|, \tag{11.112}$$

where $\widehat{F}_0(x)$ is an estimator of F_0 calculated using the method of maximum likelihood estimation to estimate the unspecified parameters of the hypothesized parametric family. When the underlying F is continuous, D' has a continuous distribution. The null distribution and asymptotic null distribution of D for the simple null hypothesis are no longer valid for D' in the composite case, and new results need to be derived for each parametric family. Lilliefors (1967) used simulation to obtain tables of the null distribution of D' for testing for an underlying normal distribution and more accurate tables based on simulation were provided by Dallal and Wilkinson (1986). For testing for an underlying exponential distribution, see Lilliefors (1969), Stephens (1974), and Durbin (1975); for testing for an underlying logistic distribution, see Stephens (1979); and for the extreme value and Weibull families, see Chandra, Singpurwalla, and Stephens (1981). In Comment 28, we give an illustration of the test for the normal family with the mean and standard deviation unspecified. The test is due to Lilliefors (1967, 1969). For summary articles on the Kolmogorov–Smirnov-type tests of fit, see Stephens (1983a, 1983b).

29. *Lilliefors' Test of Normality.* We wish to test if the X's come from a normal distribution (or more realistically, if their distribution can be reasonably approximated by a normal distribution). The maximum likelihood esitamtors of the mean μ and the standard deviation σ are

$$\widehat{\mu} = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{11.113}$$

and

$$\widehat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}. \tag{11.114}$$

Table 11.14 Calculation of D' for the Settling Velocity Data in the Case of Testing for an Underlying Unspecified Normal Distribution

i	$X_{(i)}$	$Z_{(i)} = \frac{(X_{(i)} - \overline{X})}{s}$	$\Phi(Z_{(i)})$	$\lvert F_7(X_{(i)}) - \Phi(Z_{(i)})\rvert$	$\lvert F_7(X_{(i)}-) - \Phi(Z_{(i)})\rvert$
1	12.8	−1.1793	.1191	.0237	.1191
2	12.9	−1.0073	.1569	.1288	.0140
3	13.3	−.3194	.3747	.0539	.0890
4	13.4	−.1474	.4414	.1300	.0128
5	13.7	.3685	.6438	.0705	.0723
6	13.8	.5405	.7056	.1516	.0087
7	14.5	1.7444	.9595	.0405	.1023

The estimator $\widehat{F}_0(x)$ for use in (11.112) is the normal distribution with mean $\overline{X}$ and standard deviation s given by (11.113) and (11.114), respectively. That is, for each x,

$$\widehat{F}_0(x) = \Phi\left(\frac{x - \overline{X}}{s}\right). \tag{11.115}$$

Letting $X_{(1)} < \cdots < X_{(n)}$ denote the ordered X's, we need to obtain the n values of $F_n(X_{(i)})$ as well as the n values of

$$\widehat{F}_0(Z_{(i)}) = \Phi(Z_{(i)}) = \Phi\left(\frac{X_{(i)} - \overline{X}}{s}\right). \tag{11.116}$$

where

$$Z_{(i)} = \frac{X_{(i)} - \overline{X}}{s}, \tag{11.117}$$

and $Z_{(1)} < \cdots < Z_{(n)}$ are the ordered Z's. Then D' can be written as

$$D' = \underset{i=1,\ldots,n}{\text{maximum}}\{|F_n(X_{(i)}) - \Phi(Z_{(i)})|, |F_n(X_{(i)}-) - \Phi(Z_{(i)})|\}, \tag{11.118}$$

where, in (11.118), $F_n(X_{(i)}-)$ denotes the height of the empirical distribution just to the left of $X_{(i)}$, that is, $F_n(X_{(i)}-)$ is $(i-1)/n$ for $i = 1, \ldots, n$. To apply Lilliefor's test, use library(nortest) to load the nortest package. Then apply the *R* command `lillie.test`.

We illustrate Lilliefors' test using the settling velocities data of Table 3.12. We are testing if the underlying distribution can be reasonably approximated by some unspecified normal distribution. Direct calculations yield

$$\overline{X} = 13.486, \quad s = .5815.$$

Table 11.14 illustrates the calculation of D'.

From Table 11.14 and (11.118), we find

$$D' = .1516.$$

Equivalently, to compute D', let `velocity<-c(12.8, 12.9, 13.3, 13.4, 13.7, 13.8, 14.5)` and apply `lillie.test(velocity)`. The output is $D' = .1516$ with a corresponding P-value of .8968. Since $n = 7$ is a

small sample size, we should not be surprised that the test based on D' does not lead to rejection of the null hypothesis at a low α value.

30. *Modifications for Discrete Data and for Censored Data.* Pettitt and Stephens (1977) define a Kolmogorov–Smirnov-type statistic for discrete or grouped data. For discrete data, the possible outcomes are divided into k cells and the null hypothesis is $H_0 : P$ (an observation falls in cell $i) = p_i$, $i = 1, \ldots, k$, where the p's are specified. For a sample of size N, their statistic is

$$PS = \max_{i=1,\ldots,k} \left| \sum_{i=1}^{j} (O_i - E_i) \right|, \tag{11.119}$$

where O_i is the observed number in the ith cell and $E_i = np_i$ is the expected number for the ith cell. The null hypothesis is rejected for significantly large values of PS. For grouped continuous data, let $x_0 < x_1 < \cdots < x_k$ define the cells in that cell i contains values of the random variable X for which $x_{i-1} < X < x_i$. If O_i and E_i are the observed and expected values in cell i, PS is defined as in (11.119).

Pettitt and Stephens provide null distribution tables of PS (the statistic is S in their notation) for sample sizes 30 or less, giving exact upper-tail probabilities in the equal-cell case (i.e., equal expected values E_i) and good approximations for other situations.

Fleming et al. (1980) provide an uncensored data modification of Kolmogorov's goodness-of-fit test, which they generalize to right-censored data. Chi-squared-type goodness-of-fit tests for censored data are given by Habib and Thomas (1986), Akritas (1988), Hjort (1990a), Hollander and Peña (1992a, 1992b), and Li and Doss (1993). (See also Section VI.3.3 of Andersen et al. (1993).)

Properties

1. *Asymptotic Distribution.* See Smirnov (1948).
2. *Closeness of Actual Coverage Probability to Nominal Coverage Probability when Using Large-Sample Approximation.* See Nair (1984).
3. *Asymptotic Efficiency.* See Nair (1984) and Hollander and Peña (1989).

Problems

29. Table 11.15 contains service-time data for a Tallahassee fast-food restaurant. The data were obtained by Schonrock (1996). The service time is defined as the time when the car pulled up to the speaker to order to the time when the car left the window with the order. The data were obtained at dinner time on a Thursday evening.

(a) Do you think the observations can be viewed as a random sample from the service-time distribution? In particular, consider the question of whether the observations are independent.

(b) Compute 90% confidence bands for the distribution of time to service.

30. Refer to Problem 29 and explain why one might, a priori, suspect that the time-to-service distribution is an IFR distribution. Apply a suitable procedure to test exponentiality versus IFR alternatives.

Table 11.15 Service Times at a Fast-Food Restaurant

Start	Finish	Time's
5:35:17	5:36:09	54
5:37:33	5:39:21	108
5:38:41	5:40:37	115
5:47:27	5:49:36	129
5:48:34	5:50:06	92
5:49:06	5:51:24	138
5:53:31	5:54:14	43
5:54:36	5:56:57	141
5:55:27	5:57:17	110
5:58:11	6:00:09	118
6:00:42	6:02:00	78
6:01:37	6:03:05	88
6:02:28	6:08:08	340
6:04:59	6:08:49	230
6:08:35	6:10:53	138
6:12:34	6:15:31	177
6:15:18	6:17:48	150
6:19:24	6:21:29	125
6:21:22	6:22:42	80
6:21:39	6:24:07	148
6:22:33	6:25:58	205
6:24:31	6:31:24	413
6:27:05	6:31:41	276
6:31:17	6:33:43	146
6:31:44	6:34:52	188
6:36:03	6:37:42	99
6:38:00	6:40:14	134
6:40:31	6:41:01	30
6:41:02	6:44:04	182
6:41:33	6:45:16	223
6:44:30	6:46:45	135
6:46:14	6:50:43	269
6:47:36	6:51:20	224
6:48:02	6:52:19	257

Source: H. Schonrock (1996).

31. With $F_n(x)$ and $F_0(x)$ plotted as ordinates against x as abscissa, the Kolmogorov statistic, D, is the value of the largest vertical distance between them. Justify that the largest vertical distance can be expressed by (11.108) and (11.109).

32. Obtain approximate 95% simultaneous confidence bands for the survival function corresponding to the leukemia data of Table 11.6.

33. Obtain approximate 95% simultaneous confidence bands for the survival function corresponding to the guinea pig data of Table 11.9.

11.6 AN ESTIMATOR OF THE DISTRIBUTION FUNCTION WHEN THE DATA ARE CENSORED (KAPLAN–MEIER)

Assumptions

B1. Let $T_1, \ldots, T_n$ be independent, each with continuous life distribution F. Let $C_1, \ldots, C_n$ be independent, each with continuous censoring distribution function G. C_i is the censoring time corresponding to T_i.

B2. We observe, for $i = 1, \ldots, n$,

$$X_i = \text{minimum}\{T_i, C_i\}$$

and δ_i, where

$$\delta_i = \begin{cases} 1, & \text{if } T_i \leq C_i, \\ 0, & \text{if } T_i > C_i. \end{cases}$$

Thus, δ_i is 1 if T_i is uncensored and we observe the true survival time, T_i, rather than the time to censorship. However, δ_i is 0 if T_i is censored and we observe C_i. In this case, we know only that the survival time T_i is greater than C_i.

B3. The T's and C's are mutually independent.

Assumption B3 in practical terms means that the censoring times provide no information about the true survival times. This would not be the case, for example, in a methadone study where patients on methadone are given weekly urine analysis tests to see if they have gone back on heroin. A patient who has started using heroin again may censor himself or herself (not show up for the weekly test) in order not to be "caught." Here X, the time to relapse, and C, the time to censorship, would not be independent.

To illustrate the notation of this model, which is known as the *randomly right-censored model*, consider the radiation of the affected node data of Table 11.16. For those data, $X_1 = T_1 = 346$ and $\delta_1 = 1$ because the 346 value is not censored as the true time to relapse is observed. However, for X_4, we have $X_4 = C_4 = 1953$ and $\delta_4 = 0$ because the true number of days to relapse is not observed. The observation is censored and we only know that the true time to relapse is greater than 1953.

In clinical trials, the T's may represent times to the occurrence of an endpoint event. For example, the endpoint event may be relapse or death. The data are typically analyzed before all the subjects have experienced the endpoint event. For example, in a fertility clinic, women may be taking hormones to enhance the chance of becoming pregnant. Suppose, when the data are analyzed, a woman has been undergoing treatment for 418 days and is not yet pregnant. If T denotes the time to pregnancy (measured from the initialization of treatment), we know that $T > 418$, but we do not at this point know the (eventual) true value of T. We call the 418 value a censored observation and denote it by 418^c. Other women in the study may have left town or stopped coming to the clinic and, again, for such women, the observations are censored.

Procedure

We let $F(x) = P(T \leq x)$ denote the distribution function and

$$\overline{F}(x) = 1 - F(x) = P(T > x) \tag{11.120}$$

denote the survival function. Furthermore, we let $t_{(1)} < \cdots < t_{(k)}$ denote the ordered distinct failure times. These are the known deaths and the censored values are not in the list of the $t_{(i)}$'s. (If F is continuous, then there will be no tied failure times, but in practice, ties occur.) We let n_i denote the number of patients *at risk* at time $t_{(i)}$. n_i is the number of patients who have not died or been censored before $t_{(i)}$. We let d_i be the

number of deaths (failures) at time $t_{(i)}$. (When there are no tied observations, all the d's are 1.) The Kaplan and Meier (1958) estimator of the survival function at time x is

$$\overline{F}_{KM}(x) = \prod_{t_{(i)} \le x} \left(1 - \frac{d_i}{n_i}\right). \tag{11.121}$$

The estimator of the distribution function at time x is, of course, $1 - \overline{F}_{KM}(x)$. We illustrate the computation of the Kaplan–Meier estimator (KME) in Example 11.6. See Comment 41 for *R* commands for the Kaplan–Meier estimator.

EXAMPLE 11.7 ***Hodgkin's Disease Data.***

The data in Table 11.16 are from a clinical trial in early Hodgkin's disease (i.e., cases in which the disease was detected at an early stage) conducted at the Stanford Medical Center by Kaplan and Rosenberg (1973). The data also appear in Chapter 14 of Brown and Hollander (2008). Hodgkin's disease is a cancer of the lymph system. The two treatments considered by Kaplan and Rosenberg (1973) were (1) radiation treatment of the affected node and (2) radiation treatment of the affected node plus all nodes in the trunk of the body (total nodal radiation). The relapse-free survival times are given in Table 11.16. If a relapse has not occurred by the date of data analysis, it is indicated by an N and a superscript c is affixed to the observation.

Figure 11.4 is a plot of the KMEs for the total nodal treatment and the affected node treatment. Table 11.17 illustrates the calculation of the KME for the affected node treatment. Note there are 16 known relapses (out of 25 patients), and thus there are $k = 16$ distinct failure times listed in Table 11.17.

Note that the KME decreases at each distinct failure time and is constant between distinct failure times (it does not decrease at censored observations). If the largest observed value is censored, the survival probability estimate $\overline{F}_{KM}$, as given by (11.121), does not decrease to zero but remains constant from that largest observed value out to ∞ (see Comment 34). To illustrate the calculations of Table 11.17 and the use of formula (11.121), consider, for example, the value of $\overline{F}_{KM}$ at $t_{(16)} = 1375$. The right-hand side of formula (11.121) is a product taken over those distinct failure times that are less or equal to 1375—that is, over the times 86, 107, 141, 292, 312, 330, 346, 364, 401, 419, 505, 570, 688, 822, 836, and 1375. The product is evaluated as follows:

$$\begin{aligned}
\overline{F}_{KM}(1375) &= \prod_{t_{(i)} \le 1375} \left(1 - \frac{d_i}{n_i}\right) \\
&= \left(1 - \frac{1}{25}\right)\left(1 - \frac{1}{24}\right)\left(1 - \frac{1}{23}\right)\left(1 - \frac{1}{22}\right)\left(1 - \frac{1}{21}\right) \\
&\quad \times \left(1 - \frac{1}{20}\right)\left(1 - \frac{1}{19}\right)\left(1 - \frac{1}{18}\right)\left(1 - \frac{1}{17}\right)\left(1 - \frac{1}{16}\right) \\
&\quad \times \left(1 - \frac{1}{15}\right)\left(1 - \frac{1}{14}\right)\left(1 - \frac{1}{13}\right)\left(1 - \frac{1}{12}\right)\left(1 - \frac{1}{11}\right)\left(1 - \frac{1}{9}\right) \\
&= .356.
\end{aligned}$$

Table 11.16 Relapse-Free Survival Times for Hodgkin's Disease Patients

Radiation of affected node		Total nodal radiation	
Relapse (Y) or not (N)	Days to relapse or to date of analysis	Relapse (Y) or not (N)	Days to relapse or to date of analysis
Y	346	N	1699[c]
Y	141	N	2177[c]
Y	296	N	1968[c]
N	1953[c]	N	1889[c]
Y	1375	Y	173
Y	822	N	2070[c]
N	2052[c]	N	1972[c]
Y	836	N	1897[c]
N	1910[c]	N	2022[c]
Y	419	N	1879[c]
Y	107	N	1726[c]
Y	570	N	1807[c]
Y	312	Y	615
N	1818[c]	Y	1408
Y	364	N	1763[c]
Y	401	N	1684[c]
N	1645[c]	N	1576[c]
Y	330	N	1572[c]
N	1540[c]	Y	498
Y	688	N	1585[c]
N	1309[c]	N	1493[c]
Y	505	Y	950
N	1378[c]	N	1242[c]
N	1446[c]	N	1190[c]
Y	86		

Source: H. S. Kaplan and S. A. Rosenberg (1973)

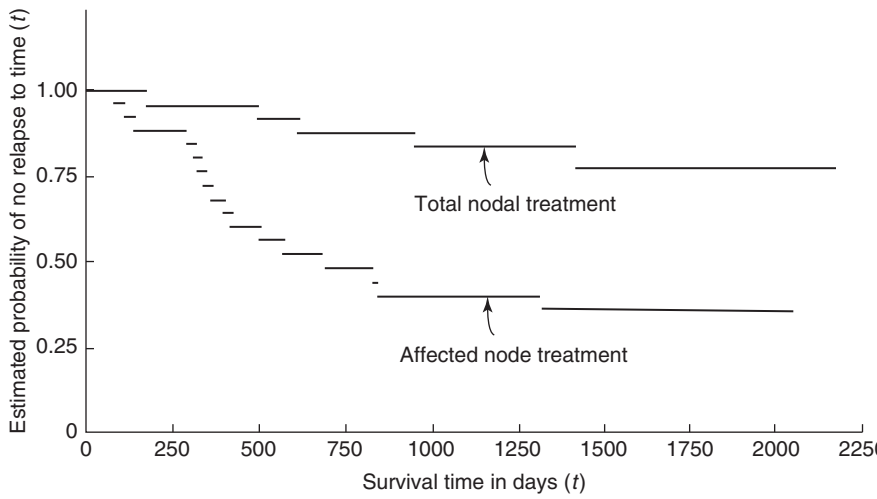

Figure 11.4 The Kaplan–Meier estimators for the total nodal and the affected-node survival distributions.

Table 11.17 Calculation of the Kaplan–Meier Estimator for the Affected Node Treatment

i	$t_{(i)}$	n_i	d_i	$\overline{F}_{KM(t_{(i)})}$
1	86	25	1	.960
2	107	24	1	.920
3	141	23	1	.880
4	296	22	1	.840
5	312	21	1	.800
6	330	20	1	.760
7	346	19	1	.720
8	364	18	1	.680
9	401	17	1	.640
10	419	16	1	.600
11	505	15	1	.560
12	570	14	1	.520
13	688	13	1	.480
14	822	12	1	.440
15	836	11	1	.400
16	1375	9	1	.356

The values in Table 11.17 can be obtained by using *R* commands (see Comment 41).

In the Kaplan and Rosenberg study, 49 patients were admitted to the study between 1967 and 1970 and randomly assigned to the affected node and the total nodal therapies. The original 49 patients were followed, and the data as of fall 1973 are given in Table 11.16. In the spring of 1970, all subsequent patients were assigned the total nodal therapy because at that point the evidence was mounting that the total nodal therapy was superior to the affected node therapy.

Note that from Table 11.16, we see that by the date of analysis in the fall of 1973, 16 of the 25 affected node patients had relapsed, whereas only 5 of the 24 total nodal patients had relapsed. Furthermore, Figure 11.4 shows that the KME for the total nodal therapy is always above that of the affected node therapy; that is, the estimated chance of relapse-free survival to any time is higher for the total nodal group than the affected node group. The statistical assessment of this difference can be formally made with a two-sample test for censored data. We describe such a test in Section 11.7.

Comments

31. *Partial Motivation for the Kaplan–Meier Estimator.* The KME (also known as the *product limit estimator*) given by (11.121) can be motivated as follows: Just before time $t_{(i)}$, there are n_i patients at risk and d_i die at time $t_{(i)}$. Thus, it is natural to estimate the probability of death at $t_{(i)}$, given that one has survived to $t_{(i)}$, as the ratio d_i/n_i, that is, the number of deaths at $t_{(i)}$ divided by the number at risk at $t_{(i)}$. Then, $(1 - (d_i/n_i))$ is the estimated conditional probability of surviving past time $t_{(i)}$, given survival up to $t_{(i)}$. The product in (11.121) corresponds to multiplying these conditional probabilities of not dying, for all known death times from zero up to the time of interest. This yields an estimate of the unconditional probability of surviving past the time of interest.

32. *Efron's Redistribute-to-the-Right Algorithm.* Suppose our data consist of the following observations: 4, 5, 5^c, 6^c, 7, 8^c, 9, 11^c, where the superscript c indicates a censored value. If the observations were all uncensored times, the empirical survival function would assign mass $\frac{1}{8}$ to each of the values. At the first censored time, 5^c, a death has not occurred but will occur somewhere to the right of 5. Efron's (1967) redistribute-to-the-right algorithm takes the mass of $\frac{1}{8}$ at 5^c and redistributes it equally among the five times 6^c, 7, 8^c, 9, 11^c to the right of 5^c, adding $\frac{1}{5}(\frac{1}{8})$ to the mass at 6^c, 7, 8^c, 9, 11^c. Now go to the next censored time 6^c and redistribute the new mass $\frac{1}{5}(\frac{1}{8}) + \frac{1}{8}$ equally among the observations to the right of 6^c. Continue this process until you reach the last observation. Efron shows this algorithm yields the KME. The algorithm is illustrated in the following display. The last value $\overline{F}_{KM}(x) = 0$ at 11^c in the southeast corner of the display is in accordance with Efron's convention. (See Comment 34.)

Observed values	Mass at start	Mass after first redistribution	Mass after second redistribution	Mass after third redistribution	$\overline{F}_{KM}(x)$
4	$\frac{1}{8}$	.125	.125	.125	.875
5	$\frac{1}{8}$	.125	.125	.125	.750
5^c	$\frac{1}{8}$	0	0	0	.750
6^c	$\frac{1}{8}$	$\frac{1}{8} + (\frac{1}{5})(\frac{1}{8}) = .150$	0	0	.750
7	$\frac{1}{8}$	.150	$.150 + (\frac{1}{4})(.150) = .1875$	.1875	.5625
8^c	$\frac{1}{8}$	.150	.1875	0	.5625
9	$\frac{1}{8}$	.150	.1875	$.150 + (\frac{1}{2})(.1875) = .28125$	.28125
11^c	$\frac{1}{8}$	.150	.1875	.28125	0

33. *Efron's Self-Consistency.* Another way to obtain the KME is via Efron's self-consistency process. If there is no censoring and we observe $T_1, \ldots, T_n$, the nonparametric estimator of $\overline{F}(x)$ is the empirical survival function

$$\overline{F}_n(x) = \frac{\sum_{i=1}^n \psi(T_i, x)}{n},$$

where $\psi(T_i, x) = 1$ if $T_i > x$, 0 otherwise. Note $n\overline{F}_n(x)$ is a sum of 0's and 1's, where 1 is scored if $T_i > x$ and 0 is scored otherwise. In the censored case, we observe

$$X_i = \min(T_i, C_i), \delta_i = \begin{cases} 1, & \text{if } X_i = T_i, \\ 0, & \text{if } X_i = C_i, \end{cases}$$

and thus for some X's we cannot tell if the corresponding T's will exceed x. If $X_i > x$, we know $T_i > x$, but if $X_i < x$ and $\delta_i = 0$, we do not know if $X_i < T_i \leq x$ or, instead, if $T_i > x$. It is thus reasonable in such a case to score (in place of 1) an estimated conditional probability $\widehat{S}(x)/\widehat{S}(X_i)$, say, an estimated chance that T_i will exceed x, given that $T_i > X_i$. Efron calls an estimator $\widehat{S}$

self-consistent if

$$\widehat{S}(x) = \frac{1}{n}\left[N(x) + \sum_{\delta_i=0, X_i \le x} \frac{\widehat{S}(x)}{\widehat{S}(X_i)}\right],$$

where $N(x)$ = number of X's $> x$. Start with some estimator (it can, e.g., be $\widehat{S}_0(x) = N(x)/n$) on the right-hand side of the preceding defining equation, calculate the left-hand side, plug the calculated value into the right-hand side, and continue this process to form a sequence of estimators

$$\widehat{S}_{j+1}(x) = \frac{1}{n}\left[N(x) + \sum_{\delta_i=0, X_i \le x} \frac{\widehat{S}_j(x)}{\widehat{S}_j(X_i)}\right].$$

Efron shows $\widehat{S}_j(x)$ converges in a finite number of steps to an estimator that will agree with $\overline{F}_{KM}(x)$ for x less than the largest observation.

34. *The Kaplan–Meier Estimated Tail Probabilities.* Let $Z_{(1)} \le \cdots \le Z_{(n)}$ denote the ordered values in the combined list of uncensored and censored values. If $Z_{(n)}$ is an uncensored value, then $\overline{F}_{KM}(x) = 0$ for all $x > Z_{(n)}$. If, however, $Z_{(n)}$ is censored, then $\overline{F}_{KM}(x)$—as defined by (11.121)—remains a nonzero constant from $Z_{(n)}$ to ∞ and thus does not have the property that $\overline{F}_{KM}(x)$ tends to be 0 as x tends to be ∞, a property that must hold for $\overline{F}(x)$, the true survival function being estimated by $\overline{F}_{KM}$. Some authors leave $\overline{F}_{KM}(x)$ undefined for $x > Z_{(n)}$ when $Z_{(n)}$ is a censored observation (as we have done in Figure 11.4, which gives the Kaplan–Meier estimates for the affected node and total nodal radiation data). Efron (1967) suggested that when $Z_{(n)}$ is censored, $\overline{F}_{KM}$ should be defined to be 0 for all $x > Z_{(n)}$. Gill (1980) suggested that when $Z_{(n)}$ is censored, one should set $\overline{F}_{KM}(x) = \overline{F}_{KM}(Z_{(n)})$ for $x > Z_{(n)}$ so that $\overline{F}_{KM}$ is a nonzero constant from $Z_{(n)}$ to ∞. Efron's convention yields underestimates of the tail survival probabilities, whereas Gill's convention yields overestimates. Brown, Hollander, and Korwar (1974) and Moeschberger and Klein (1985) suggested methods for completing the tail of the KME, which can be considered intermediate to Efron's and Gill's conventions. Brown, Hollander, and Korwar recommend fitting an exponential survival curve to complete the KME, namely, the exponential curve that agrees with $\overline{F}_{KM}(Z_{(n)})$ at $Z_{(n)}$. Determining the λ that satisfies $\exp(-\lambda Z_{(n)}) = \overline{F}_{KM}(Z_{(n)})$ yields $\lambda = -\{\ln[\overline{F}_{KM}(Z_{(n)})]\}/Z_{(n)}$, and the tail survival probabilities are defined to be

$$\exp(-\lambda x) = \exp\left(\frac{x\{\ln[\overline{F}_{KM}(Z_{(n)})]\}}{Z_{(n)}}\right) \quad \text{for } x > Z_{(n)}.$$

Moeschberger and Klein (1985) used the Weibull distribution to fit the tail survival probabilities when $Z_{(n)}$ is censored.

In Figure 11.4, we left the two Kaplan–Meier estimates undefined after the respective largest observations (which are censored in both samples). Figure 11.4 as is indicates the superiority of the total nodal treatment, and in this case, there in not a need to estimate survival probabilities in the tails. In situations where such tail estimates are desired, however, fitting the tail probabilities by the Brown, Hollander, and Korwar method or the Moeschberger and Klein method yields more realistic estimates than those obtained by the Efron or Gill conventions.

35. *Confidence Intervals for the Survival Probability at Time x.* Let

$$V(x) = \text{var}\big(\overline{F}_{KM}(x)\big), \tag{11.122}$$

the variance of the KME of surviving past time x. $V(x)$ can be estimated by

$$\widehat{V}_{(x)} = \big\{\overline{F}_{KM}(x)\big\}^2 \sum_{t_{(i)} \le x} \frac{d_i}{n_i(n_i - d_i)}. \tag{11.123}$$

The standard error given by (11.123) is known as *Greenwood's formula* (Greenwood, 1926). An asymptotic $100(1-\alpha)\%$ confidence interval for $\overline{F}_{KM}(x)$ is $(\overline{F}_L(x), \overline{F}_U(x))$, where

$$\overline{F}_L(x) = \overline{F}_{KM}(x) - z_{\alpha/2}\Big(\widehat{V}(x)\Big)^{1/2}, \quad \overline{F}_U(x) = \overline{F}_{KM}(x) + z_{\alpha/2}\Big(\widehat{V}(x)\Big)^{1/2}. \tag{11.124}$$

When no censoring and no tied observations occur, the KME reduces to the empirical survival function

$$\overline{F}_n(x) = \frac{\text{number of observations in the sample} > x}{n}, \tag{11.125}$$

where n is the sample size. Correspondingly, the asymptotic confidence interval given by (11.124) reduces, in the case of no censoring and no tied observations, to the following $100(1-\alpha)\%$ confidence interval for $\overline{F}(x)$:

$$\left(\overline{F}_n(x) - z_{\alpha/2}\left\{\frac{\overline{F}_n(x)(1-\overline{F}_n(x))}{n}\right\}^{1/2}, \right.$$
$$\left. \overline{F}_n(x) + z_{\alpha/2}\left\{\frac{\overline{F}_n(x)(1-\overline{F}_n(x))}{n}\right\}^{1/2}\right). \tag{11.126}$$

Estimators, confidence intervals, and confidence bands for $\overline{F}$ can also be obtained by exploiting the relationship between the cumulative hazard function of F and the survival function. The cumulative hazard function Λ_F corresponding to the distribution F is defined to be

$$\Lambda_F(x) = -\ln(1 - F(x)) = \int_0^x r(t)dt,$$

where r is the failure rate defined by (11.1).

$\overline{F}$ can be expressed in terms of Λ by

$$\overline{F}(x) = e^{-\Lambda(x)}.$$

If one has an estimator $\widehat{\Lambda}$ of Λ, this yields an estimator $\widehat{\overline{F}}$ of $\overline{F}$ via $\widehat{\overline{F}}(x) = e^{-\widehat{\Lambda}(x)}$. The Nelson–Aalen estimator of Λ is

$$\widehat{\Lambda}(x) = \begin{cases} 0, & x < t_{(1)}, \\ \sum_{t_{(i)} \le x} \frac{d_i}{n_i}, & x \ge t_{(1)}. \end{cases}$$

Nelson defined the estimator in an applied setting (see Nelson (1969, 1972)) and Aalen (1978) considered it in a more general theoretical setting.

Kalbfleisch and Prentice (1980) (see also Bie, Borgan, and Liestøl (1987) and Borgan and Liestøl (1990)) used a log transformation of the cumulative hazard function (or, equivalently, the transformation $g(x) = \ln(-\ln(x))$ applied to the survival function) to obtain an asymptotic confidence interval for $\overline{F}(x)$ that is a competitor of the interval given by (11.124). Thomas and Grunkemeier (1975) used an arcsine square root transformation ($g(x) = \arcsin(\sqrt{x})$) to obtain a competitor of the interval given by (11.124). The Kalbfleisch and Prentice asymptotic $100(1-\alpha)\%$ confidence interval for $\overline{F}(x)$ is

$$(\{\overline{F}_{KM}(x)\}^a, \{\overline{F}_{KM}(x)\}^{1/a}),$$

where

$$a = \exp\left[\frac{z_{\alpha/2}\sum_{t_{(i)}\le x}\frac{d_i}{n_i(n_i-d_i)}}{\ln(\overline{F}_{KM}(x))}\right].$$

Borgan and Liestøl (1990) found that the Thomas and Grunkemeier (1975) and Kalbfleisch and Prentice (1980) confidence intervals do better, in small samples, than the confidence interval given by (11.124).

36. *Exact Moments of the KME.* Chen, Hollander, and Langberg (1982) give, for $\alpha > 0$, an exact expression for $E([\overline{F}_{KM}(x)]^\alpha)$, the αth moment of the KME. They consider the case where the censoring distribution G (the distribution of each C_i) and the survival distribution F (the distribution of each T_i) satisfy a proportional hazards model (see (11.154)). Chen, Hollander, and Langberg (1982) used the Efron convention (see Comment 34) to define $\overline{F}_{KM}$ in the tail, whereas Wellner (1985) obtained closely related exact results using the Gill convention to define $\overline{F}_{KM}$ in the tail. These exact results enable one to study the exact biases of the KME and to compare the exact variances of the KME with its asymptotic variances. Wellner notes that the biases and variances of the KME based on the Gill convention are almost everywhere smaller than those for the KME based on the Efron convention. Thus Wellner advocates use of the Gill definition rather than the Efron definition.

37. *Simultaneous Confidence Bands for the Survival Function.* The interval given by (11.124) has approximate coverage probability $1-\alpha$ at a fixed value x. Suppose, instead, we desire an approximate $1-\alpha$ simultaneous confidence band for $\overline{F}$. Such a band would have the property

$$P_F(\overline{F}_\ell(x) \le \overline{F}(x) \le \overline{F}_u(x), \text{ for all } x) \doteq 1-\alpha, \tag{11.127}$$

where $\overline{F}_\ell(x)$ is the lower contour of the band and $\overline{F}_u(x)$ is the upper contour of the band. Display (11.127) is to be interpreted as meaning that simultaneously, for all x, the chance is approximately $1-\alpha$ that the random contours F_ℓ, F_u contain $\overline{F}$.

Confidence bands based on the KME have been presented by many authors including Gillespie and Fisher (1979), Gill (1980), Hall and Wellner (1980), Fleming et al. (1980), Nair (1981, 1984), Csörgő and Horváth (1986), Hollander and Peña (1989), and Hollander, McKeague and Yang (1997). (The bands in

Fleming et al. (1980) are not explicitly defined in the paper, but they are implicitly given in the appendix of the paper and can be developed as outlined in Exercise 6.5 of Fleming and Harrington (1991, pp. 397–398).) Gulati and Padgett (1996) use the Hollander and Peña (1989) approach but replace the discontinuous KME by a kernel-based continuous estimator to obtain continuous confidence bands. See also Section IV.3.3 of Andersen et al. (1993) for confidence intervals and simultaneous confidence bands for the survival function. For confidence intervals, also see Thomas and Grunkemeir (1975) and Murphy (1995).

Some of the proposed bands are contained in the family of bands, indexed by a value $c > 0$, presented by Hollander and Peña (1989). The Hollander–Peña family is of the form

$$\{[\overline{F}_\ell(x), \overline{F}_u(x)], 0 \le x \le T\}, \tag{11.128}$$

where

$$\overline{F}_\ell(x) = \overline{F}_{KM}(x)\{1 - r_n(x; c, \lambda_n)\}, \quad \overline{F}_u(x) = \overline{F}_{KM}(x)\{1 + r_n(x; c, \lambda_n)\}, \tag{11.129}$$

where

$$r_n(x; c, \lambda_n) = \frac{\lambda_n}{[(nc)^{1/2}\{1 - L_n^*(x, c)\}]}, \tag{11.130}$$

$$L_n^*(x, c) = \frac{cd_n^*(x)}{\{1 - cd_n^*(x)\}}, \tag{11.131}$$

$$d_n^*(x) = \sum_{\{t_{(i)} \le x\}} \frac{d_i}{n_i(n_i - d_i)}, \tag{11.132}$$

where, as before, d_i is the number of deaths at $t_{(i)}$ and n_i is the number of risk at $t_{(i)}$. The value of λ_n is obtained from `b.sf` (which is adapted from Table 11.1 of Hall and Wellner (1980)). Use `b.sf` with $\beta = 1 - \alpha$ and

$$a = \frac{cd_n^*(T)}{1 + cd_n^*(T)}.$$

`b.sf` gives values of λ such that $G_a(\lambda) = \beta$ for $a = .10, .25, .40, .50, .60, .75, .90, 1.00$ and $\beta = .99, .95, .90, .75, .50, .25$, where

$$G_a(\lambda) = 1 - 2\overline{\Phi}\left[\lambda\{a(1-a)\}^{-1/2}\right] + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2\lambda^2}[\overline{\Phi}(r(2k-d)\} - \overline{\Phi}(r(2k+d)\}], \tag{11.133}$$

where Φ is the standard normal distribution function, $r = \lambda\{(1-a)/a\}^{1/2}$ and $d = (1-a)^{-1}$. $G_a(\lambda) = P\{\sup_{0 \le t \le a} |\mathcal{W}_t^0| \le \lambda\}$, where $\mathcal{W}^0$ is a Brownian bridge process on $[0, 1]$ (cf. Billingsley (1968, pp. 64–65)).

The most popular confidence band in use at the time of this writing is the Hall and Wellner (1980) band. It corresponds to $c = 1$ in the Hollander–Peña family. The choice of c is discussed in Section 11.4 of Hollander and

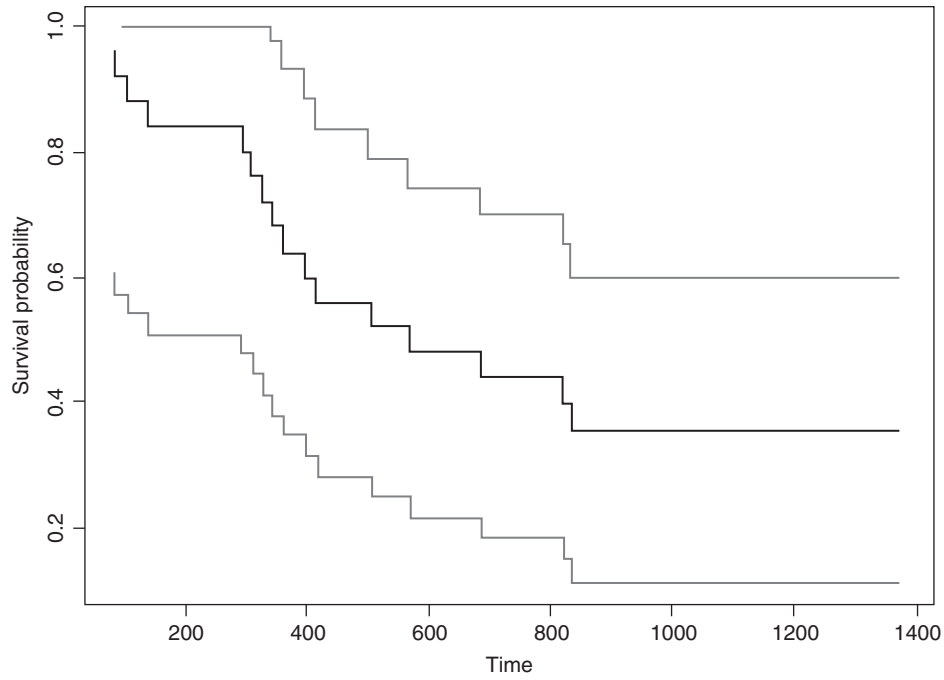

Figure 11.5 An approximate 95% confidence band for the affected-node survival distribution ($c = .5$).

Peña (1989). Although the asymptotic coverage probability of all bands in the family is $1 - \alpha$, c controls the width of the bands on various intervals. Bands corresponding to larger values of c are narrower on the left (i.e., for smaller x values). If one desires a band that is narrow at the specific point $x = x_0$, choose $c = 1/d_n^*(x_0)$.

For the affected-node data of Table 11.15, Figure 11.5, Figure 11.6, and Figure 11.7 are, respectively, plots of asymptotic 95% confidence bands for the choices $c = .5, 1$, and 2. The figures also display the KME.

38. *Monotonized Bands.* There will be some data sets for which the bands given by (11.129) can be narrowed and yet still retain the same asymptotic coverage probability. The process is called *monotonization*. Let $0 \le x \le y \le T$. From (11.129), it can be seen that the ratio of the values of the upper contour of the bands at x and y is

$$\begin{aligned}\frac{\overline{F}_u(x)}{\overline{F}_u(y)} &= \frac{\overline{F}_{KM}(x)\{1 + r_n(x; c, \lambda_n)\}}{\overline{F}_{KM}(y)\{1 + r_n(y; c, \lambda_n)\}} \\ &= \frac{\overline{F}_{KM}(x)}{\overline{F}_{KM}(y)}\left[1 - \left\{\frac{r_n(y; c, \lambda_n) - r_n(x; c, \lambda_n)}{1 + r_n(y; c, \lambda_n)}\right\}\right].\end{aligned}$$

Because $\overline{F}_{KM}(x) \ge \overline{F}_{KM}(y)$ and $r_n(y; c, \lambda_n) \ge r_n(x; c, \lambda_n)$, it may happen that, for some x and y (with $x \le y$), the ratio $\overline{F}_u(x)/\overline{F}_u(y)$ will be less than 1. In such a situation, the upper contour of the band is not monotone decreasing. This is unacceptable because $\overline{F}$ itself is monotone decreasing. For a confidence band whose upper contour is not monotone decreasing, a narrower band can be formed

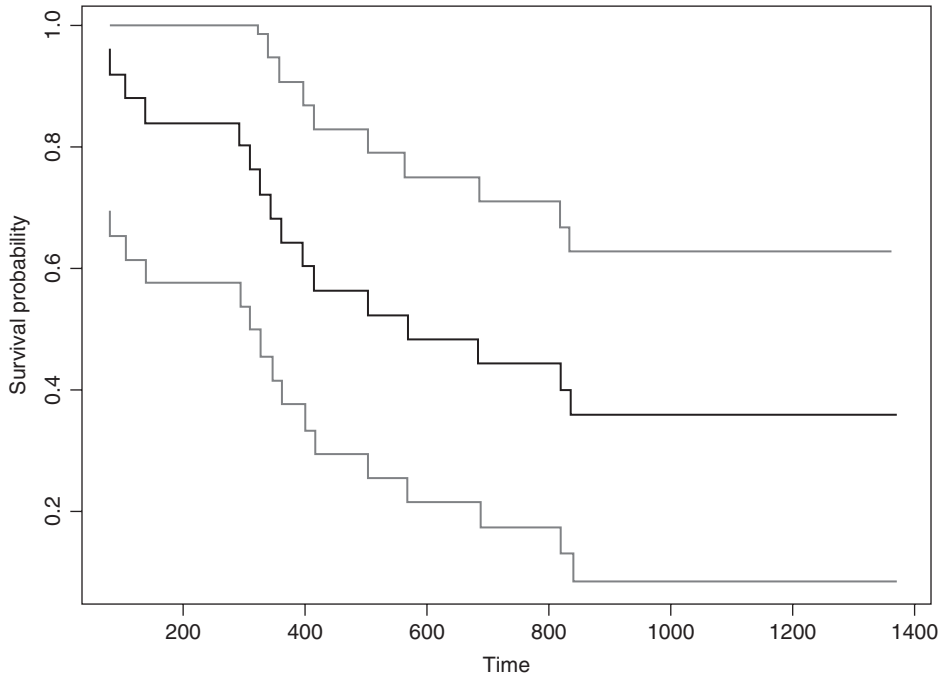

Figure 11.6 An approximate 95% confidence band for the affected-node survival distribution ($c = 1$, the Hall–Wellner band).

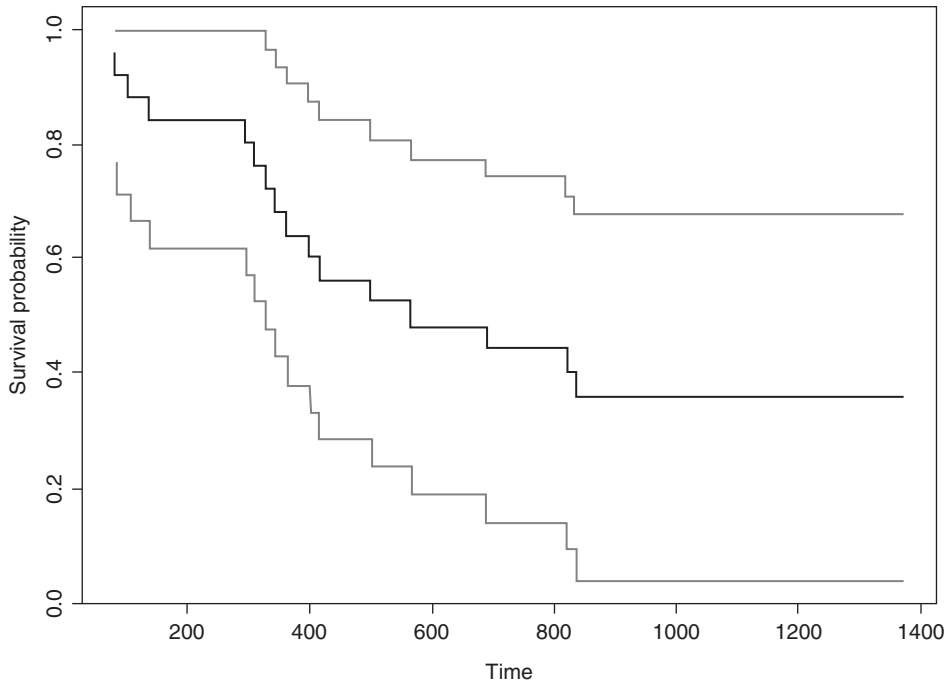

Figure 11.7 An approximate 95% confidence band for the affected-node survival distribution ($c = 2$).

that will have the monotone property and will also have the same confidence level as the original band. This narrower band is formed by *monotonizing* the upper contour of the original band. In general, suppose that $[m_n(x), M_n(x)]$ is an asymptotic $100(1-\alpha)\%$ confidence band for $\overline{F}$ on the interval $[0, T]$. Then the monotonized band $[m_n(x), \min_{a \le x} M_n(a)]$ is also an asymptotic $100(1-\alpha)\%$ confidence band for $\overline{F}$. This is true because, due to the monotonicity of $\overline{F}$, the events

$$\left\{m_n(x) \le \overline{F}(x) \le M_n(x),\ 0 \le x \le T\right\} \tag{11.134}$$

and

$$\left\{m_n(x) \le \overline{F}(x) \le \min_{a \le x} M_n(a),\ 0 \le x \le T\right\} \tag{11.135}$$

are identical and thus have the same probability of occurring. Furthermore, because $0 \le \overline{F}(x) \le 1$ for all x-values, we may use the confidence band given by

$$\left\{\max[0, m_n(x)], \min\left[1, \min_{a \le x} M_n(a)\right],\ 0 \le x \le T\right\}. \tag{11.136}$$

39. *NBU Test for Censored Data.* Chen, Hollander, and Langberg (1983a) extended the NBU test of Section 11.2 to the censored data situation. Their test statistic J_c is obtained by estimating the parameter $\Delta_{\mathrm{NBU}}(F)$ (see (11.46)) by $\Delta_{\mathrm{NBU}}(F_{KM})$, where F_{KM} is the KME. Letting $Z_{(1)} \le \cdots \le Z_{(n)}$ denote the ordered values (in the combined list of censored and uncensored values), the statistic can be written as

$$J_c = \sum_{i=1}^{n} \overline{F}_{KM}(2Z_{(i)}) d_i^2 + 2 \sum_{i<j}^{n} \overline{F}_{KM}(Z_{(i)} + Z_{(j)}) d_i d_j, \tag{11.137}$$

where

$$d_i = \overline{F}_{KM}(Z_{(i-1)}) - \overline{F}_{KM}(Z_{(i)}) \tag{11.138}$$

and $Z_{(0)}$ is defined to be 0.

The null asymptotic mean of J_c is $\frac{1}{4}$, independent of the scale parameter λ of the exponential distribution corresponding to the null hypothesis and independent of the censoring distribution governing the times to censorship. The null asymptotic variance of $n^{1/2} J_c$ does, however, depend on λ and the censoring distribution and thus must be estimated from the data. The NBU (NWU) test for censored data is based on

$$J_c^* = \frac{\sqrt{n}(J_c - \frac{1}{4})}{\widehat{\sigma}_c}, \tag{11.139}$$

where

$$\begin{aligned}\widehat{\sigma}_c^2 = (128)^{-1} + \sum_{i=1}^{n-1} n(n-i+1)^{-1}(n-i)^{-1}\big\{&(128)^{-1} \\ &- (32)^{-1} Z_{(i)}(\widehat{\lambda}) + (16)^{-1} Z_{(i)}^2(\widehat{\lambda^2})\big\} \times \exp\left\{-4Z_{(i)}\widehat{\lambda}\right\} \\ &- n\left\{(128)^{-1} - (32)^{-1} Z_{(n)}\widehat{\lambda} + (16)^{-1} Z_{(n)}^2 \widehat{\lambda}^2\right\} \exp\left\{-4Z_{(n)}\widehat{\lambda}\right\}\end{aligned}$$

and

$$\widehat{\lambda} = \frac{\text{number of uncensored observations}}{\sum_{i=1}^{n} Z_i}. \tag{11.140}$$

Under $H_0, \widehat{\lambda}$ consistently estimates λ and $\widehat{\sigma}_c^2$ consistently estimates the asymptotic variance of $\sqrt{n}J_c$. Under H_0, J_c^* is asymptotically $N(0,1)$ (see Chen, Hollander, and Langberg (1983a)).

To test H_0 against NBU alternatives at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } J_c^* \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{11.141}$$

To test H_0 against NWU alternatives, at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } J_c^* \geq z_\alpha; \quad \text{otherwise do not reject.} \tag{11.142}$$

To test H_0 against NBU and NWU alternatives, at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } |J_c^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{11.143}$$

40. *Estimation and Confidence Bands for the Quantile Function.* The quantile function is formally defined as $F^{-1}(p) = \sup\{t : F(t) \leq p\}$, $0 < p < 1$. Thus, for example, $F^{-1}(.5)$ is the median. Bootstrap confidence bands for the quantile function were given by Doss and Gill (1992). Li et al. (1996) (hereafter, Li et al. (1996)) used a likelihood ratio approach to obtain confidence bands for the quantile function. They define the likelihood function

$$L(F) = \prod_{i=1}^{n} [F(Z_i) - F(Z_i-)]^{\delta_i} [1 - F(Z_i)]^{1-\delta_i},$$

where Z_i is the observed value corresponding to patient i (Z_i, may be a censored value or a known death) and δ_i is 1 if Z_i corresponds to a known death and δ_i is 0 if Z_i corresponds to a censored observation. Let Θ be the space of all distribution functions on $[0, \infty)$. Here, F is viewed as a parameter taking values in Θ. For any $t \geq 0$ and $0 < p < 1$, Li et al. (1996) define

$$R(p,t) = \frac{\sup\{L(F) : F(t) = p \text{ and } F \in \Theta\}}{\sup\{L(F) : F \in \Theta\}}$$

and, for $0 \leq r \leq 1$,

$$C(p,r) = \{t : R(p,t) \geq r\}.$$

A large value of $R(p,t)$ can be considered evidence in favor of the hypothesis $H^* : F(t) = p$. Therefore, $C(p,r)$ can be interpreted, for each fixed p, as the set of times t for which H^* is not rejected by a test based on $R(p,t)$. Li et al. (1996) show that $C(p,r)$ is always an interval and their asymptotic confidence band is obtained by pasting together intervals of the form $C(p,r)$ with r chosen appropriately. Furthermore, they obtain an asymptotic $1-\alpha$ confidence interval for the p-quantile of F. This interval is of the form $C(p, r_\alpha^*)$, where

$r_\alpha^* = \exp\{-\mathcal{X}^2_{1,\alpha}/2\}$ and $\mathcal{X}^2_{1,\alpha}$ is the upper α-quantile of a chi-square distribution with 1 degree of freedom. Figure 11.8 contains a plot of the Kaplan–Meier quantile function, F^{-1}_{KM}, and an asymptotic 95% confidence band for the true unknown quantile function for the affected node data of Table 11.16.

41. *R commands for the KME.* For the affected node treatment data of Table 11.16, let

    days<-c(346, 141, 296, 1953, 1375, 822, 2052, 836, 1910, 419, 107, 570, 312, 1818, 364, 401, 1645, 330, 1540, 688, 1309, 505, 1378, 1446, 86).
    relapse<-c(1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1).
    Set

    ```
    my.formula = Surv(days, relapse)~1
    my.fit = survfit(my.formula)
    plot(my.fit)
    ```

    The output is the KME and 95% confidence limits. Use `km.ci` to change to a different confidence level. For example,

    ```
    library(km.ci)
    km.ci(my.fit, conf.level=.9)
    plot(km.ci(my.fit, conf.level=.9))
    ```

Properties

1. *Consistency.* See Peterson (1977), Gill (1980), Shorack and Wellner (1986, Section 7.3), Wang (1987), Ying (1989), Fleming and Harrington (1991, Section 3.4), and Andersen et al. (1993, Section IV.3.2).
2. *Asymptotic Distributional Properties.* See Kaplan and Meier (1958), Efron (1967), Breslow and Crowley (1974), Meier (1975), Gill (1980, 1983), Fleming and Harrington (1991, Chapter 6), and Andersen et al. (1993, Section IV.3.2).
3. *Asymptotic Optimality.* See Wellner (1982), van der Vaart (1988, 1991), and Andersen et al. (1993, Chapter VIII).
4. *Efficiency.* See Wellner (1982), Miller (1983), Hollander, Proschan, and Sconing (1985), Gill (1989), and Andersen et al. (1993, Chapter VIII).
5. *Nonparametric Maximum Likelihood Estimator.* See Johansen (1978) for a proof that the Kalpan–Meier estimator is the nonparametric MLE of F in the sense of Kiefer and Wolfowitz (1956).

Problems

34. The data in Table 11.18 are from Hollander, McKeague, and Yang (1997) and concern 432 manuscripts submitted for publication to the Theory and Methods Section of the *Journal of the American Statistical Association* in the period January 1, 1994–December 13, 1994. Of interest is the distribution of the time (in days) to first review. When the data were studied on December 13, 1994, 158 papers were still awaiting the first review. Thus, there are 158 censored times and 274 uncensored times. In Table 11.18, the variable X_i = minimum (T_i, C_i), where T_i is the time to first review and C_i is the time to censorship, and the indicator variable δ_i is 1 if the ith observation is uncensored and 0 if it is censored. Compute the Kaplan–Meier estimate of the survival function.

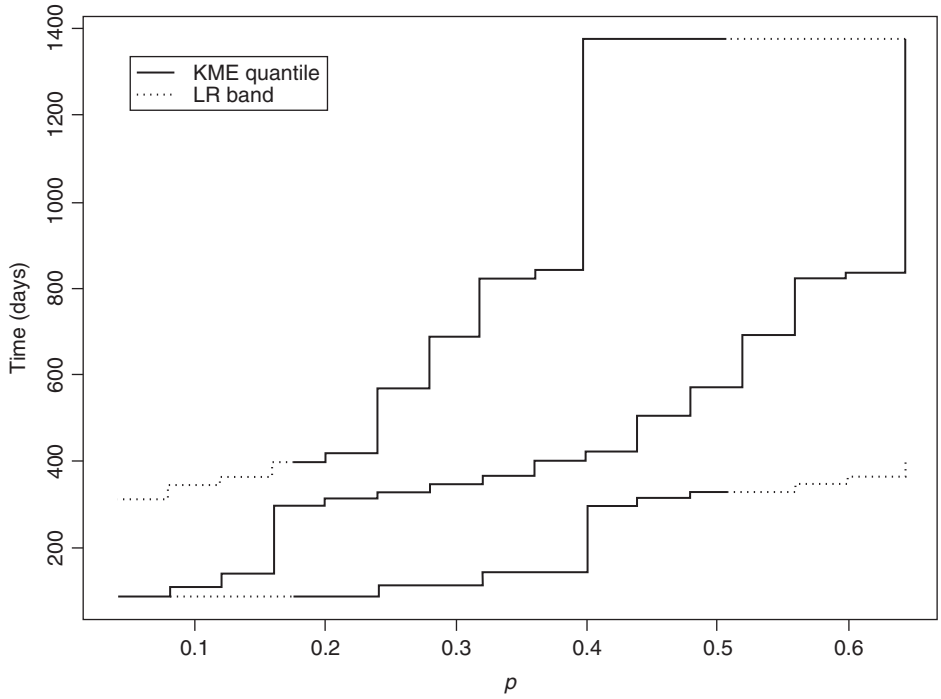

Figure 11.8 An approximate 95% confidence band for the affected-node quantile function.

35. For the review times data of Table 11.18, compute asymptotic 95% confidence bands for the survival function.

36. For the review times data of Table 11.18, compute an asymptotic 95% confidence interval for the probability that the time to first review will exceed 150 days.

37. The data in Table 11.19 were provided by Koziol and Green (1978) and can be found in Hollander and Proschan (1979). The data correspond to 211 patients with stage IV prostate cancer who were treated with estrogen in a Veterans Administration Cooperative Urological Research Group (1967) study. The observations over the years 1967 through March 1977. At the March 1977 closing date, there were 90 patients who had died of prostate cancer, 105 who had died of other diseases, and 16 who were still alive. Those observations corresponding to deaths due to other causes and those corresponding to the 16 survivors are considered censored observations (withdrawals). Compute the KME of the survival distribution for deaths from cancer of the prostate.

38. For the prostate cancer data of Table 11.19, compute asymptotic 90% confidence bands for the survival function.

39. For the prostate cancer data of Table 11.19, compute an asymptotic 90% confidence interval for the probability of surviving more than 100 months.

40. Verify that, in the case of no censored observations, the confidence interval given by (11.124) reduces to that given by (11.126).

41. Verify that, due to monotonicity of $\overline{F}$ (i.e., $\overline{F}(x) \geq \overline{F}(y)$ whenever $x < y$), the events given by (11.134) and (11.135) are identical.

42. Apply the NBU test for censored data to the prostate cancer of Table 11.19. What is your conclusion?

43. Apply the NBU test for censored data to the affected node data of Table 11.16. What is your conclusion?

Table 11.18 The Times to First Review of 1994 JASA Theory and Methods Papers

X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i
214	1	201	1	28	1	252	0	118	1	187	0	28	1	28	1	76	1	56	0	28	0
184	1	274	1	287	0	96	1	33	1	152	1	21	1	118	0	18	1	21	1	27	0
150	1	265	1	195	1	175	1	69	1	46	1	1	1	40	1	88	0	55	0	27	0
70	1	120	1	86	1	54	1	133	1	103	1	0	1	6	1	85	0	55	0	25	0
16	1	141	1	137	1	167	1	126	1	37	1	144	0	91	1	85	0	54	0	25	0
141	1	48	1	74	1	150	1	84	1	170	1	144	0	34	1	85	0	18	1	22	0
210	1	204	1	71	1	219	1	197	1	64	1	140	0	21	1	20	1	54	0	22	0
132	1	312	0	140	1	86	1	85	1	182	0	14	1	1	1	83	0	53	0	21	0
30	1	220	1	22	1	1	1	15	1	180	0	0	1	111	0	82	0	50	0	21	0
204	1	188	1	120	1	111	1	206	1	176	0	27	1	111	0	81	0	50	0	15	0
84	1	84	1	176	1	128	1	125	1	175	0	23	1	1	1	81	0	1	1	15	0
36	1	84	1	181	1	178	1	57	1	64	1	126	1	48	1	11	1	15	1	15	0
38	1	215	1	155	1	40	1	181	1	42	1	139	0	110	0	77	0	50	0	1	1
69	1	33	1	74	1	131	1	215	0	175	0	55	1	47	1	77	0	47	0	1	1
33	1	55	1	29	1	20	1	3	1	149	1	137	0	68	1	70	1	47	0	14	0
49	1	140	1	100	1	220	1	13	1	158	1	114	1	74	1	74	0	46	0	12	0
203	1	147	1	195	1	84	1	175	1	169	0	56	1	98	1	71	0	16	1	12	0
203	1	41	1	127	1	32	1	37	1	169	0	124	1	105	0	23	1	43	0	12	0
218	1	94	1	34	1	95	1	182	1	22	0	121	1	104	0	28	1	43	0	8	0
267	1	292	1	177	1	188	1	210	0	168	0	1	1	104	0	70	0	18	1	8	0
99	1	131	1	150	1	115	1	92	1	157	1	27	1	103	0	44	1	43	0	8	0
21	1	221	1	265	0	238	0	208	0	89	1	130	0	90	1	69	0	42	0	8	0
78	1	39	1	174	1	1	1	30	1	165	1	130	0	98	0	68	0	42	0	7	0
150	1	3	1	104	1	187	1	28	1	14	1	130	0	98	0	67	0	40	0	7	0
237	1	16	1	203	1	125	1	168	1	161	0	127	0	98	0	64	1	0	1	7	0
91	1	129	1	109	1	110	1	202	0	161	0	100	1	97	0	30	1	12	1	7	0
21	1	210	1	217	1	32	1	114	1	159	0	126	0	96	1	41	1	39	0	6	0
224	1	240	1	238	1	32	1	105	1	91	1	126	0	97	0	62	0	35	0	5	0
126	1	141	1	210	1	228	1	196	0	146	1	28	1	18	1	61	0	35	0	5	0
167	1	231	1	22	1	80	1	195	0	159	0	125	0	96	0	20	1	35	0	4	0
105	1	119	1	148	1	64	1	114	1	134	1	125	0	92	0	57	0	30	1	1	0
146	1	291	0	142	1	231	0	75	1	13	1	125	0	91	0	57	0	35	0	1	0
50	1	199	1	126	1	64	1	194	0	159	0	95	1	91	0	57	0	34	0		
28	1	67	1	220	1	228	0	143	1	18	1	95	1	91	0	57	0	34	0		
288	1	263	1	145	1	18	1	106	1	155	0	123	0	31	1	57	0	34	0		
37	1	155	1	21	1	55	1	128	1	154	0	123	0	27	1	57	0	33	0		
18	1	189	1	256	0	154	1	200	0	124	1	123	0	83	1	42	1	33	0		
113	1	0	1	253	0	139	1	129	1	73	1	109	1	33	1	57	0	29	1		
22	1	209	1	22	1	91	1	138	1	51	1	119	0	11	1	57	0	28	0		
234	1	223	1	80	1	196	1	152	1	21	1	6	1	88	0	57	1	27	0		

Source: M. Hollander, I. W. McKeague, and J. Yang (1997).

11.7 A TWO-SAMPLE TEST FOR CENSORED DATA (MANTEL)

For the Hodgkin's survival time data of Table 11.15, the two samples, corresponding to the affected node and total nodal radiation treatments, contain censored observations. Figure 11.4 indicates that the underlying survival distributions corresponding to the two

Table 11.19 Survival Times and Withdrawal Times in Months for 211 State IV Patients (with Number of Ties Given in Parentheses)

Survival times: 0(3), 2, 3, 4, 6, 7(2), 8, 9(2), 11(3), 12(3), 15(2), 16(3), 17(2), 18, 19(2), 20, 21, 22(2), 23, 24, 25(2), 26(3), 27(2), 28(2), 29(2), 30, 31, 32(3), 33(2), 34, 35, 36, 37(2), 38, 40, 41(2), 42(2), 43, 45(3), 46, 47(2), 48(2), 51, 53(2), 54(2), 57, 60, 61, 62(2), 67, 69, 87, 97(2), 100, 145, 158.

Withdrawal times: 0(6), 1(5), 2(4), 3(3), 4, 6(5), 7(5), 8, 9(2), 10, 11, 12(3), 13(3), 14(2), 15(2), 16, 17(2), 18(2), 19(3), 21, 23, 25, 27, 28, 31, 32, 34, 35, 37, 38(4), 39(2), 44(3), 46, 47, 48, 49, 50, 53(2), 55, 56, 59, 61, 62, 65, 66(2), 72(2), 74, 78, 79, 81, 89, 93, 99, 102, 104(2), 106, 109, 119(2), 125, 127, 129, 131, 133(2), 135, 136(2), 138, 141, 142, 143, 144, 148, 160, 164(3).

Source: J. A. Koziol and S. B. Green (1978).

treatments are not equal and the survival times in the samples are better under total nodal radiation than under radiation of the affected node. This, however, is only a visual assessment, and censoring complicates the picture. A two-sample test for censored data is needed to make an objective assessment. One commonly used test is Mantel's (1966) test, also known as the *logrank test*.

Assumptions

C1. For sample 1, let $T_1, \ldots, T_m$ be independent, each with continuous life distribution function F_1. Let $C_1, \ldots, C_m$ be independent, each with continuous censoring distribution function G_1. C_i is the censoring time corresponding to T_i. For sample 2, let $U_1, \ldots, U_n$ be independent, each with continuous life distribution function F_2. Let $D_1, \ldots, D_n$, be independent, each with continuous censoring distribution function G_2. D_j is the censoring time associated with U_j.

C2. For sample 1, we observe, for $i = 1, \ldots, m$,

$$X_i = \text{ minimum } \{T_i, C_i\}$$

and δ_i, where

$$\delta_i = \begin{cases} 1, & \text{if } T_i \leq C_i, \\ 0, & \text{if } T_i > C_i. \end{cases}$$

For sample 2, we observe, for $j = 1, \ldots, n$,

$$Y_j = \text{ minimum } \{U_j, D_j\}$$

and

$$\varepsilon_j = \begin{cases} 1, & \text{if } U_j \leq D_j, \\ 0, & \text{if } U_j > D_j. \end{cases}$$

Thus, δ_i is 1 if T_i is uncensored and we observe the true survival time T_i rather than the time to censorship C_i. However, δ_i is 0 if T_i is censored and we observe C_i. In this case, we only know that the true survival time T_i is greater than C_i. Similarly, ε_j is 1 if U_j is uncensored and we observe the true survival time U_j rather than the time to censorship D_j. However, ε_j is 0 if U_j is censored and we observe D_j. In this case, we only know that the true survival time U_j is greater than D_j.

C3. The T's, C's, U's, and D's are mutually independent.

To illustrate the notation of this model, which is known as the *randomly right-censored model*, consider the data of Table 11.16. For those data, where $m = 25$, $n = 24$, we have $X_1 = T_1 = 346$ and $\delta_1 = 1$ because the 346 value is not censored as the true time to relapse is observed. However, for X_4, we have $X_4 = C_4 = 1953$ and $\delta_4 = 0$ because the true number of days to relapse is not observed. The observation is censored and we know only that the true time to relapse is greater than 1953. Similarly, $Y_1 = D_1 = 1699, \varepsilon_1 = 0$, and $Y_5 = U_5 = 173, \varepsilon_5 = 1$.

Procedure

Combine the two samples and let k denote the number of distinct failure times. Denote these distinct failure times by $w_{(1)} < w_{(2)} < \cdots < w_{(k)}$. Let n_{ij}, $j = 1, 2$, denote the number of patients from sample j at risk at $w_{(i)}$. That is, n_{ij} is the number of patients who have not experienced the endpoint event (death, relapse, etc.) or been censored before time $w_{(i)}$. Let d_{ij}, $j = 1, 2$, be the number of failures from sample j at time $w_{(i)}$. Correspondingly, let n_i be the combined number of patients from both samples who are at risk at $w_{(i)}$ and let d_i be the combined (from both samples) number of failures at $w_{(i)}$.

Mantel's (1966) test statistic M_c for two-sample censored data is

$$M_c = \frac{\sum_{i=1}^{k}(d_{i1} - E_{i1})}{\sqrt{\sum_{i=1}^{k} V_{i1}}}, \tag{11.144}$$

where

$$E_{i1} = \frac{d_i n_{i1}}{n_i} \tag{11.145}$$

and

$$V_{i1} = \frac{d_i(n_i - d_i)n_{i1}n_{i2}}{n_i^2(n_i - 1)} \tag{11.146}$$

are, respectively, the conditional mean and variance of d_{i1}. (See Comment 42.)

Let F_1 denote the unknown life distribution of group 1 and F_2 the unknown life distribution of group 2. The null hypothesis is

$$H_0 : F_1 = F_2.$$

The test is performed by treating M_c as having an approximate $N(0, 1)$ distribution under H_0. See Comment 46 for an *R* program for computing M_c.

a. *One-Sided Test of H_0 against Alternatives for Which Treatment 2 Is Better.* To test

$$H_0 : F_1 = F_2$$

versus

H_1 : Survival times for treatment 2 tend to be longer than those for treatment 1,

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } M_c \geq z_\alpha; \quad \text{otherwise do not reject.} \tag{11.147}$$

b. *One-Sided Test of H_0 against Alternatives for Which Treatment 1 is Better.* To test

$$H_0 : F_1 = F_2$$

versus

H_2 : Survival times for treatment 1 tend to be longer than those for treatment 2,

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } M_c \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{11.148}$$

c. *Two-Sided Test against Alternatives for which the Two Treatments Differ.* To test

$$H_0 : F_1 = F_2$$

versus

H_3 : Survival times for treatment 2 have a
different distribution than that for treatment 1,

at the approximate α level of significance,

$$\text{Reject } H_0 \text{ if } |M_c| \geq z_{\frac{\alpha}{2}}; \quad \text{otherwise do not reject.} \tag{11.149}$$

EXAMPLE 11.8 ***Example 11.7 Continued.***

We return to the Hodgkin's disease data of Table 11.16. Drs. Kaplan and Rosenberg believed the total nodal radiation treatment to be superior to the radiation of the affected node treatment. A hypothesis test provides an assessment in terms of P-values. We will apply Mantel's test to the data. In the combined sample there are $k = 21$ distinct failure times $w_{(1)} < \cdots < w_{(21)}$. They are

(1)	(1)	(1)	(2)	(1)	(1)	(1)	(1)	(1)	(1)	(1)
86	107	141	173	296	312	330	346	364	401	419
(2)	(1)	(1)	(2)	(1)	(1)	(1)	(2)	(1)	(2)	
498	505	570	615	688	822	836	950	1375	1408	

In the preceding display, above each w in parentheses, we indicate if the failure is from treatment 1 (radiation of the affected node) or treatment 2 (total nodal radiation). Table 11.20 illustrates the computation of M_c. Summing columns 2, 7, and 8, respectively, of Table 11.20 yields

$$\sum_{i=1}^{21} d_{i1} = 16, \quad \sum_{i=1}^{21} E_{i1} = 8.7220, \quad \sum_{i=1}^{21} V_{i1} = 5.0146.$$

Table 11.20 Computation of M_c for the Hodgkin's Disease Data

w_i	d_{i1}	d_i	n_{i1}	n_{i2}	n_i	E_{i1}	V_{i1}
86	1	1	25	24	49	.5102	.2499
107	1	1	24	24	48	.5000	.2500
141	1	1	23	24	47	.4894	.2499
173	0	1	22	24	46	.4783	.2495
296	1	1	22	23	45	.4889	.2499
312	1	1	21	23	44	.4773	.2495
330	1	1	20	23	43	.4651	.2488
346	1	1	19	23	42	.4524	.2477
364	1	1	18	23	41	.4390	.2463
401	1	1	17	23	40	.4250	.2444
419	1	1	16	23	39	.4103	.2420
498	0	1	15	23	38	.3947	.2389
505	1	1	15	22	37	.4054	.2411
570	1	1	14	22	36	.3889	.2377
615	0	1	13	22	35	.3714	.2335
688	1	1	13	21	34	.3824	.2362
822	1	1	12	21	33	.3636	.2314
836	1	1	11	21	32	.3438	.2256
950	0	1	10	21	31	.3226	.2185
1,375	1	1	9	18	27	.3333	.2222
1,408	0	1	7	18	25	.2800	.2016

Then, from (11.144), we obtain

$$M_c = \frac{16 - 8.7220}{\sqrt{5.0146}} = 3.25.$$

Using 1-`pnorm`, (3.25) gives an approximate one-sided P-value of .0006. Thus, there is strong evidence that total nodal radiation is more effective than the radiation of affected nodes in preventing or delaying the recurrence of early stage Hodgkin's disease. For an *R* program for Mantel's test, see Comment 46.

Comments

42. *Motivation for Mantel's Test*. The development is similar to that used in Section 10.4, where success probabilities are compared in k 2×2 tables. Here, k 2×2 tables are formed, one at each known failure time $w_{(i)}$, as in Table 11.21. (Note, however, that while k was fixed in Section 10.4, here k is the random number of observed failures in the combined sample.)

    Conditioning on the marginal totals in each of the k 2×2 tables, the mean and variance of d_{ij} are, respectively

$$E_{ij} = \frac{d_i n_{ij}}{n_i} \tag{11.150}$$

and

$$V_{ij} = \frac{d_i(n_i - d_i)n_{i1}n_{i2}}{n_i^2(n_i - 1)}. \tag{11.151}$$

Table 11.21 2×2 Table of Failure and Numbers at Risk at Failure Time $w_{(i)}$

	Failures	Not Failures	Totals
Sample 1	d_{i1}	$n_{i1} - d_{i1}$	n_{i1}
Sample 2	d_{i2}	$n_{i2} - d_{i2}$	n_{i2}
Totals	d_i	$n_i - d_i$	n_i

The k "observed minus expected" differences $d_{11} - E_{11}, \ldots, d_{1k} - E_{1k}$ are not independent (in contrast to the k, "observed minus expected" differences $\mathcal{O}_{111} - E_{111}, \ldots, \mathcal{O}_{11k} - E_{11k}$ of Section 10.4, which are independent because the k 2×2 tables of that section are assumed to be independent). Due to the lack of independence, the central limit theorem, which is used to establish asymptotic normality of the MH statistic of Section 10.4, cannot be applied in this case, and other approaches are necessary to establish asymptotic normality of M_c. Proofs of asymptotic normality of M_c, as a consequence of more general results that establish asymptotic normality of classes of two-sample statistics that include M_c, can be found in Gill (1980), Fleming and Harrington (1991), and Andersen et al. (1993).

Mantel's (1966) nonrigorous development of his test for the censored case is via analogy to the Mantel–Haenszel (1959) test for k 2×2 tables presented in Section 10.4 (also see Mantel (1963)). For further discussion, see Miller (1998), Kalbfleisch and Prentice (2002), Fleming and Harrington (1991), Andersen et al. (1993), and Klein and Moeschberger (2003). Mantel's test is closely related to tests proposed by Peto and Peto (1972) and Cox (1972). (Cox (1972) is a seminal paper on nonparametric regression methods for censored data. For nonparametric regression methods for censored data, also see Kalbfleisch and Prentice (2002), Miller (1998), Cox and Oakes (1984), Fleming and Harrington (1991), Crowder et al. (1991), Andersen et al. (1993), and Klein and Moeschberger (2003).) Mantel's test, also known as the *logrank test* (a name first used by Peto and Peto (1972)), can be viewed as a generalization of a test due to Savage (1956) for uncensored data (see Kalbfleisch and Prentice (2002)). Fleming et al. (1980) provide a two-sample test for censored data that generalizes the Kolmogorov–Smirnov test of Section 5.4. They found that their test tends to do better than Mantel's test and Gehan's test when the failure rates (recall (11.1)) of the two underlying distributions cross.

43. *Choice of the Variance for the Logrank Test.* When there are no ties among the k uncensored observations in the combined sample, $d_i = 1$, $i = 1, \ldots, k$, and M_c can be written as

$$M_c = \frac{M}{\sigma_M},$$

where

$$M = \sum_{i=1}^{k} \left(d_{i1} - \frac{n_{i1}}{n_i} \right)$$

and

$$\sigma_M^2 = \sum_{i=1}^{k} \frac{n_{i1} n_{i2}}{n_i^2}.$$

σ_M^2 is the Mantel variance (also known as the Mantel–Haenszel variance). Under H_0, σ_M^2 is an unbiased estimator of the variance of M, independently of differences between the censoring distributions of the two groups. Brown (1984) shows, however, that with equal sample sizes, when M is large in absolute value, σ_M^2 tends to underestimate the true variance. Large values of $|M|$ tend to be accompanied by small values of σ_M^2. This results in exaggeratedly large values of $|M|/\sigma_M$ and P-values that are too small. This is also discussed by Morton (1978). Brown suggests that when the sample sizes are approximately equal, one should instead base tests on M/σ_P, where σ_P^2 is the permutation variance (see Peto and Peto (1972) and Brown (1984)), which assumes equal censoring distributions. The permutation variance is

$$\sigma_P^2 = \frac{mn}{(m+n)(m+n-1)}\left(k - \sum_{i=1}^{k} n_i^{-1}\right).$$

Brown shows that under equal sample sizes, the permutation variance tends to overstate the true variance when censoring is unequal.

Although Brown's results highlight the weakness of Mantel's procedure in certain situations, the assumption of equal censoring distributions required by σ_P^2 is too restrictive and thus, we have used the Mantel variance in our presentation of Mantel's test. See Mantel (1985) for a brief rebuttal to Brown (1984).

44. *Tarone–Ware Tests*. The Mantel test described in this section assigns equal weights to the 2×2 tables formed at each known failure time $w_{(i)}$. Tarone and Ware (1977) define a class of two-sample tests for censored data by allowing different weights for the tables. Their general test statistic, for weights $b_1, \ldots, b_k$, is

$$\text{TW} = \frac{\sum_{i=1}^{k} b_i (d_{i1} - E_{i1})}{\sqrt{\sum_{i=1}^{k} b_i^2 V_{i1}}}, \tag{11.152}$$

where E_{i1}, V_{i1} are given by (11.145) and (11.146), respectively. Under H_0, TW is asymptotically $N(0,1)$. Significantly large values of TW indicate survival times for treatment 2 tend to be larger than those for treatment 1. Significantly small values of TW indicate survival times for treatment 1 tend to be larger than those for treatment 2. The choice $b_i = 1$ yields Mantel's statistic. The choice $b_i = n_i$ yields a test due to Gehan (1965) (although Gehan derived his test from a different approach). Tarone and Ware advocate the choice $b_i = \sqrt{n_i}$ based on efficiency considerations.

45. *Other Two-Sample Tests for Censored Data.* There are many two-sample tests for censored data. Gehan (1965) proposed a generalization of Wilcoxon's two-sample test for uncensored data. Efron (1967) proposed a different generalization of Wilcoxon's test. Tarone and Ware (1977) defined a class of tests that include Mantel's test and Gehan's test. Prentice (1978) proposed a family of linear rank tests. Other very general families have been proposed and studied by Gill (1980) and Harrington and Fleming (1982). Leurgans (1983, 1984) gives efficiency and small-sample Monte Carlo power results for many of these tests. Fleming and

Harrington (1991, Chapter 7) provide a comprehensive treatment of Gill's 1980 $\mathcal{K}$ class.

46. *R program for Two-Sample Censored Data Tests*. We use the survival package's function `survdiff` that performs two-sample censored data tests and obtains corresponding approximate P-values for statistics in the Harrington and Fleming (1982) $\mathcal{G}^p$ class of tests. By setting $\rho = 0$, one gets Mantel's test. To illustrate the use of `survdiff`, we consider the relapse-free times of Table 11.16. We create three R objects corresponding to (i) the X's of Table 11.16, (ii) the indicators (1 for observed relapse, 0 for censored) of whether the X is a true relapse time or a censored time, and (iii) indicators (1 for sample 1, 2 for sample 2) that designate if X is from sample 1 or sample 2. Let

    rad<-c(346, 141, 296, 1953, 1375, 822, 2052, 836, 1910, 419, 107, 570, 312, 1818, 364, 401, 1645, 330, 1540, 688, 1309, 505, 1378, 1446, 86, 1699, 2177, 1968, 1889, 173, 2070, 1972, 1897, 2022, 1879, 1726, 1807, 615, 1408, 1763, 1684, 1576, 1572, 498, 1585, 1493, 950, 1242, 1190),

    cen<-c(1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0),

    smp<-c(rep(1, 25), (0, 24)).

Then, creating a survival formula,

    Surv(data, event = optional censors)~x,

we set

    formula = Surv(rad, event = cen)~smp

and with `rho = 0` corresponding to Mantel's test, the R function `survdiff(formula, rho = 0)` yields chisquare = 10.6 on 1 degree of freedom, $P = .00115$. This agrees (allowing for rounding) with what we obtained in Example 11.7.

Recall that in Example 11.7, we found $M_c = 3.25$ with an approximate one-sided P-value of .0006 based on the normal approximation. Note $(3.25)^2 = 10.56$ and $2(.0006) = .0012$.

47. *The Cox Proportional Hazards Model.* In addition to Assumptions B1–B3, we assume that each subject has a vector of non-time-dependent covariates with the ith subject having covariate vector $\mathbf{z}_i = (z_{1i}, \ldots, z_{p_i})'$. The covariate vector $\mathbf{z}$ is often called the *risk vector*. For example, in a study of coronary heart disease (CHD),, the covariates could include total cholesterol, high density lipoprotein, age, systolic blood pressure, diabetic status, and smoking status. One of the goals is to assess the influence of each covariate on the time to death from CHD.

The failure rate function at time x is modeled by Cox (1972) as

$$r(x|\mathbf{z}) = r_0(x) \exp\left(\sum_{i=1}^{p} \beta_i z_i\right),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients. The parameter β_i is a measure of the importance of the ith covariate to the failure rate. The function $r_0(x)$ is known as the *baseline failure rate* and is unspecified. The term proportional hazards is used because for two subjects with covariate $\mathbf{z}$ and $\mathbf{z}^*$ say,

$$r(x|\mathbf{z})/r(x|\mathbf{z}^*) = \exp\left\{\sum_{i=1}^{p} \beta_i (z_i - z_i^*)\right\}, \tag{11.153}$$

which is constant.

Assuming there are no ties, let $x_{(1)} < \ldots < x_{(n)}$ denote the ordered observed times. Let $R_{(i)}$ be the set of patients at risk just before time $x_{(i)}$. Let $\mathbf{z}_{(i)}$ be the covariate associated with $x_{(i)}$. For an uncensored time $x_{(i)}$, the conditional probability of death of the patient with covariate $\mathbf{z}_{(i)}$ at time $x_{(i)}$, given survival up to time $x_{(i)}$, is $\exp(\boldsymbol{\beta}'\mathbf{z}_{(i)})/\sum_{j \in R_{(i)}} \exp(\boldsymbol{\beta}'\mathbf{z}_j)$. Then, Cox's partial likelihood is the product, over the uncensored times, of these conditional probabilities, namely,

$$\mathcal{L}_c(\boldsymbol{\beta}) = \prod_{\substack{\text{uncensored} \\ \text{times}}} \frac{\exp(\boldsymbol{\beta}'\mathbf{z}_{(i)})}{\sum_{j \in R_{(i)}} \exp(\boldsymbol{\beta}'\mathbf{z}_{(j)})}. \tag{11.154}$$

Maximum likelihood estimates for the β's can be obtained numerically, for example, by a Newton–Raphson algorithm or other numerical techniques.

Hypothesis tests concerning the β's include Wald's test, the likelihood ratio test based on $\ln \mathcal{L}_c$, and the scores test. For details, including the cases where ties are present and the covariates may be time-dependent, see Klein and Moeschberger (2003). The *R* package survival contains a number of procedures related to life testing and estimation for the Cox model including `cox.zph`, which tests the proportional hazards assumption, and `coxph`, which fits a proportional hazards model. Ng'andu (1997) compares various tests for assessing the validity of the proportional hazards assumptions.

Properties

1. *Consistency*. See Gill (1980, Section 4.1) and Fleming and Harrington (1991, Section 7.3).
2. *Asymptotic Normality*. See Prentice (1978), Gill (1980), and Fleming and Harrington (1991, Section 7.2).
3. *Efficiency*. See Peto and Peto (1972), Prentice (1978), Gill (1980, Chapter 5), Harrington and Fleming (1982), Leurgans (1983, 1984), and Fleming and Harrington (1991, Section 7.4).
4. *Power and Sample-Size Calculations*. See Latta (1981), Schoenfeld (1981), Leurgans (1983, 1984), Hsieh (1987, 1992), Sposto and Krailo (1987), Lakatos and Lan (1992), and Strawderman (1997).

Table 11.22 Severe Viral Hepatitis Study

Patient	Treatment (D = drug, P = placebo)	Length of observation, weeks	Status (A = alive, D = dead)
1	D	6	D
2	P	16	A
3	P	2	D
4	D	1	D
5	D	1	D
6	P	4	A
7	D	16	A
8	D	3	A
9	P	16	A
10	D	8	D
11	P	16	A
12	P	2	D
13	D	4	D
14	P	16	A
15	P	5	A
16	D	1	D
17	D	10	A
18	P	2	A
19	P	2	A
20	D	10	D
21	P	16	A
22	D	1	A
23	P	16	A
24	D	15	A

Source: P. B. Gregory (1974).

Problems

44. The data in Table 11.22, obtained by Gregory (1974) of the Stanford University, originally appeared in Brown and Hollander (1977). The data are from a clinical trial conducted to study the efficiency of a new drug thought to be helpful for treating patients with a particular type of serious liver disease. Is there evidence that the new drug does significantly better (or significantly worse) than the placebo in terms of survival times?

45. Give an intuitive explanation why the 2×2 tables formed at the uncensored times $w_{(i)}, \ldots, w_{(k)}$ are not independent.

46. Apply Gehan's test (see Comment 43) to the hepatitis data of Table 11.22. Compare your results with those of Problem 44.

47. Apply the Tarone–Ware test with weights $b_i = \sqrt{n_i}$ (see Comment 43) to the hepatitis data of Table 11.22. Compare your results with those of Problems 44 and 46.

48. The data in Table 11.23 are from Hollander (1996) and concern 444 manuscripts submitted for publication to "Theory and Methods" Section of the *Journal of the American Statistical Associations* in the period January 1, 1995–December 15, 1995. Of interest is the distribution of the time (in days) to first review. When the data were studied on December 15, 1995, 173 papers were still awaiting the first review. Thus, there are 173 censored times and 271 uncensored times. In Table 11.23, the variable $X_i = \text{minimum}(T_i, C_i)$, where T_i is the time to first review and C_i is the time to censorship, and the indicator variable δ_i is 1 if the ith observation is uncensored and 0 if it is censored. Use the data in Table 11.18 and Table 11.23 to test if there is a significant difference between the 1994 times to first review and the 1995 times to first review.

Table 11.23 The Times to First Review of the 1995 JASA Theory and Methods Papers

X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i	X_i	δ_i
141	1	37	1	16	1	93	1	77	1	134	1	64	1	29	1	88	1	58	0	22	1
140	1	42	1	245	1	240	0	101	1	171	1	56	1	116	0	87	0	56	0	30	0
161	1	126	1	162	1	95	1	175	1	119	1	106	1	115	0	85	0	56	0	30	0
23	1	88	1	42	1	221	1	162	1	155	1	87	1	115	0	15	1	50	0	19	1
14	1	27	1	31	1	130	1	206	0	28	1	103	1	0	1	81	0	50	0	29	0
134	1	64	1	98	1	21	1	132	1	112	1	137	0	114	0	81	0	50	0	29	0
83	1	260	1	127	1	66	1	205	0	169	0	137	0	113	0	81	0	39	1	28	0
145	1	114	1	37	1	90	1	90	1	169	0	38	1	113	0	1	1	33	1	25	0
97	1	114	1	171	1	159	1	19	1	125	1	17	1	113	0	80	1	50	0	25	1
119	1	252	1	118	1	232	0	32	1	162	0	95	1	14	1	79	0	49	0	25	0
255	1	49	1	0	1	70	1	150	1	162	0	133	0	109	0	78	0	22	1	25	0
181	1	154	1	188	1	74	1	190	1	118	1	120	1	109	0	78	0	49	0	25	0
140	1	54	1	257	1	100	1	56	1	96	1	130	0	89	1	71	1	23	1	25	0
209	1	159	1	197	1	188	1	196	0	47	1	53	1	2	1	78	0	45	0	25	0
194	1	38	1	21	1	211	1	82	1	73	1	19	1	45	1	77	0	45	0	25	0
297	1	246	1	263	0	177	1	196	0	150	1	59	1	23	1	77	0	38	1	23	0
46	1	91	1	143	1	221	0	58	1	40	1	93	1	107	0	2	1	44	0	23	0
47	1	197	1	262	0	204	1	191	1	161	0	129	0	107	0	73	0	44	0	16	1
44	1	32	1	197	1	122	1	42	1	14	1	1	1	106	0	73	0	44	0	18	1
85	1	32	1	145	1	14	1	14	1	161	0	119	1	93	1	73	0	43	0	18	0
155	1	104	1	35	1	99	1	21	1	157	0	128	0	105	0	70	0	43	0	18	0
153	1	32	1	45	1	28	1	54	1	157	1	90	1	85	1	70	0	43	1	18	0
48	1	101	1	201	1	218	0	104	1	148	1	127	0	101	0	70	0	31	1	17	0
48	1	291	0	204	1	217	0	151	1	153	1	123	0	101	0	70	0	42	0	17	0
187	1	102	1	126	1	3	1	189	0	157	0	123	1	101	0	70	0	39	0	17	0
99	1	288	0	255	0	28	1	39	1	157	0	123	0	100	0	67	0	39	0	16	0
21	1	225	1	255	0	129	1	78	1	157	0	46	1	100	0	67	0	39	0	16	0
190	1	147	1	99	1	170	1	109	1	41	1	66	1	100	0	64	0	39	0	15	0
319	0	160	1	233	1	93	1	186	0	151	0	43	1	100	0	64	0	14	1	15	0
108	1	284	0	197	1	135	1	158	1	46	1	123	0	0	1	64	0	27	1	11	0
44	1	253	1	42	1	169	1	44	1	151	0	121	0	20	1	47	1	38	0	11	0
127	1	174	1	170	1	28	1	179	1	71	1	34	1	47	1	63	0	38	0	11	0
155	1	180	1	71	1	79	1	64	1	151	0	121	0	99	0	60	0	37	0	9	0
186	1	22	1	76	1	81	1	28	1	150	0	110	1	98	0	59	1	37	0	9	0
215	1	237	1	26	1	211	1	17	1	150	0	120	0	21	1	59	0	33	1	9	0
39	1	218	1	12	1	33	1	177	0	148	0	80	1	95	0	34	1	27	1	8	0
118	1	127	1	187	1	117	1	29	1	12	1	119	0	94	0	59	0	37	0	4	0
22	1	148	1	130	1	41	1	47	1	18	1	31	1	95	0	59	0	18	1	3	0
32	1	100	1	28	1	24	1	23	1	70	1	73	1	93	0	59	0	32	0	3	0
74	1	63	1	1	1	0	1	167	1	47	1	43	1	43	1	59	0	32	0	2	0
														28	1	58	0	31	0	2	0

Source: M. Hollander (1996).

Table 11.24 Asymptotic Relative Efficiencies of A, B, V', $\mathcal{E}$

	A	B	V'	$\mathcal{E}$	c^2_{MAX}
F_1 (linear failure rate)	0.44	0.31	1.00	0.91	0.820
F_2 (Makeham)	0.70	0.70	0.70	1.00	0.083
F_3 (Pareto)	0.44	0.31	1.00	0.91	0.820
F_4 (Weibull)	0.51	0.87	0.49	1.00	1.441
F_5 (gamma)	0.39	1.00	0.28	0.90	0.498

49. For the data of Table 11.22, compute the Nelson–Aalen estimator of the cumulative hazard function for the drug and the Nelson–Aalen estimator of the cumulative hazard function for the placebo. What differences are indicated?

11.8 EFFICIENCIES

The entries in Table 11.24 are given by Klefsjö (1983) and are based on efficiency calculations reported in Bickel and Doksum (1969), Hollander and Proschan (1975), and Klefsjö (1983). The statistics considered in Table 11.24 are the total-time-on-test statistic $\mathcal{E}$ of Section 11.1 given by (11.9) (also see Comment 14), the IFR statistic A (see Comment 5), the IFRA statistic B (see Comment 5), and the DMRL statistic V' (11.58). Table 11.24 gives, for the distributions F_1 (linear failure rate), F_2 (Makeham), F_3 (Pareto), F_4 (Weibull), and F_5 (gamma), the asymptotic relative efficiencies of $A, B, V', \mathcal{E}$ relative to the statistic (among $A, B, V', \mathcal{E}$) having the largest efficacy for that particular F. The c^2_{MAX} column of Table 11.24 gives, for a given F, the largest squared efficacy for the four included statistics.

For the NBU statistic T given by (11.34), Hollander and Proschan (1972) found the asymptotic relative efficiency of T with respect to $\mathcal{E}$ for Weibull and linear failure rate distributions. The values are $e_{F_4}(T, \mathcal{E}) = .937$ and $e_{F_1}(T, \mathcal{E}) = .45$. Other efficiency values for T are given in Koul (1978b) and Deshpande (1983). Other efficiency values for $\mathcal{E}$ are given by Bickel and Doksum (1969) and Borges, Proschan, and Rodrigues (1984).

To our knowledge, asymptotic relative efficiency results have not been obtained for the IDMRL and DIMRL tests of Section 11.4. Hawkins, Kochar, and Loader (1992), however, have performed a limited Monte Carlo power comparison of their IDMRL tests $T^{(1)}$ and $T^{(2)}$ (see Comment 25) for the case where the turning point τ is unknown versus the Guess–Hollander–Proschan IDMRL test for the case where τ is known (given by (11.83) and denoted by HKL as GHP_1) and the Guess–Hollander–Proschan IDMRL test when the proportion $\rho = F(\tau)$ is known. See Comment 24. The latter test is denoted by HKL as GHP_2. The distribution used by HKL in their Monte Carlo study is $F_{\alpha,\beta,\gamma}(x) = 1 - \overline{F}_{\alpha,\beta,\gamma}(x), \alpha > 0, \beta > 0, \gamma > 0$, where $\overline{F}_{\alpha,\beta,\gamma}(x)$ is given in Section 3 of HKL (1992). The distribution has mrl function.

$$m_{\alpha,\beta,\gamma}(x) = \beta + \gamma e^{-\alpha x}(1 - e^{-\alpha x}), \quad x \geq 0. \tag{11.155}$$

As γ tends to be 0, $m_{\alpha,\beta,\gamma}(x)$ tends to be the constant mrl function of the exponential distribution (11.3) with $\lambda = 1/\beta$. Each member of the $F_{\alpha,\beta,\gamma}$ family is an IDMRL distribution with turning point $\tau = \alpha^{-1}\log(2)$. HKL found their $T^{(1)}$ test and GHP_1 to be slightly conservative. They found GHP_1 generally dominates their $T^{(1)}$ and $T^{(2)}$ tests

except when $\rho \leq .5$, where $T^{(2)}$ seems to dominate. When $\rho \leq .75$, $T^{(2)}$ dominates $T^{(1)}$ and compares well with GHP_1 and GHP_2. For ρ in the neighborhood of .90, HKL found $T^{(1)}$ dominates $T^{(2)}$, but both $T^{(1)}$ and $T^{(2)}$ are considerably dominated by GHP_2. As HKL point out, it is not surprising that the GHP tests outperform the HKL in their study, because in their comparisons, the GHP tests are allowed to use information about τ, and such information is not required by the HKL tests.

Nair (1984) defined a family of confidence bands for the survival function. Nair's family is indexed by parameters a, b with $0 < a < b < 1$. Each band in his family is an equal-precision band in the sense that its width is proportional to its standard deviation. Nair compared his equal-precision (EP) bands to the Hall–Wellner (HW) band defined in Comment 34. His efficiency criterion is the ratio of the limiting squared widths of the bands. In the absence of censoring, the HW band reduces to the Kolmogorov band of Section 11.5. Thus, Nair's comparisons are relevant to Section 11.5 as well as Section 11.6. Nair concluded that the HW and EP bands are competitive, with the HW band being narrower in the middle and the EP bands narrower in the tails. He also found the relative performance of the EP bands to the HW band gets better as censoring increases. Hollander and Peña (HP) (1989) used the same efficiency criterion to compute the efficiency $e(x; c)$ of the HP band (see Comment 37) for a general c with respect to the HW band (which is the HP band corresponding to $\boldsymbol{c} = 1$). They showed, under certain conditions, $e(x; c)$ tends to c as x tends to 0 and $e(x; c)$ tends to c^{-1} tends to infinity.

The KME of Section 11.6 is the best estimator of the survival function $\overline{F}$ in the fully nonparametric model when the underlying life distribution F and the censoring distribution G (that governs the censoring patterns) are unspecified (see Andersen et al. (1993, Chapter VIII)). Efficiency robustness properties of the KME can be studied by seeing how well the KME does in models for which it is not optimal. Miller (1983) studied the KME's efficiency loss when compared to the MLE in parametric models. Not surprisingly, the KME performs poorly relative to the MLE in a fully parametric setting. For example, when F and G are exponential, Miller (1983) showed that the asymptotic efficiency of the KME $\overline{F}_{KM}(x)$ (see (11.121)) with respect to the MLE tends to zero as x tends to zero and also tends to zero as x tends to infinity.

Hollander, Proschan, and Sconing (1985) considered the efficiency properties of the KME in the proportional hazards model, where the censoring distribution G and the life distribution F satisfy

$$1 - G(x) = \{1 - F(x)\}^{\beta}, \quad x > 0, \ \beta > 0. \tag{11.156}$$

In this model, the cumulative hazard functions $\Lambda_G = -\log(1 - G)$ and $\Lambda_F = -\log(1 - F)$ are proportional. Also, the expected proportion of uncensored observations is $1/(1 + \beta)$. The model is also known as the *Koziol–Green model* (see Koziol and Green (1976)). Table 11.25 contains asymptotic efficiency values of the KME with respect to the MLE $\widetilde{\overline{F}}$ in model (11.154) with β unknown. The estimator $\widetilde{\overline{F}}$ was independently proposed by Abdushukurov (1987), Cheng and Lin (1987), and Hollander, Proschan, and Sconing (1985) (also see Csörgő (1988) and Csörgő and Faraway (1998)). The latter paper bares a paradox that shows the proportional hazards model is "too good to be true." For estimating F, sometimes it is better to have a censored sample than an uncensored sample! The estimator $\widetilde{\overline{F}}$ is

$$\widetilde{\overline{F}}(x) = \{\widehat{H}(x)\}^{d}, \tag{11.157}$$

Table 11.25 Values of a(x)

x	β:	.1	.2	$\frac{1}{3}$	.5	1.0	2.0
.1		.9186	.8522	.7812	.7130	.5903	.5048
.5		.9466	.9063	.8672	.8345	.7910	.7642
1.0		.9640	.9368	.9091	.8821	.8130	.6477
1.5		.9679	.9403	.9065	.8655	.7358	.4850
2.0		.9639	.9291	.8828	.8239	.6492	.3930

where, letting $Z_{(1)} \le Z_{(2)} \le \cdots \le Z_{(n)}$ denote the ordered values in the combined list of censored and uncensored values, $\widehat{H}(x) =$ (number of $Z's > x)/n$ and $d =$ (number of uncensored values)$/n$. Table 11.25, part of a larger table in HPS (1985), gives values of the asymptotic efficiency

$$a(x) = e(\overline{F}_{\mathrm{KM}}, \widetilde{\overline{F}})$$

of the KME with respect to $\widetilde{\overline{F}}$ when F is exponential with parameter 1 and G is exponential with parameter β. For this choice of F, G, it can be shown that $a(x)$ initially increases and then decreases. The value of x for which this change occurs is given by the solution to the equation $(\beta + 1)x = 2[1 - \exp\{-x(\beta + 1)\}]$. Table 11.25 suggests that $a(x)$ decreases as β increases. β increasing is equivalent to censoring increasing stochastically. Thus, Table 11.25 suggests that $a(x)$ decreases as censoring increases stochastically.

HPS (1985) also studied, in the fully nonparametric model where the life distribution F and the censoring distribution G are arbitrary, the asymptotic efficiency $b(x)$,

$$b(x) = e(\overline{F}_n, \overline{F}_{\mathrm{KM}}),$$

of the empirical survival function (11.125) with respect to the KME; $b(x)$ has the following interpretation. Roughly speaking, the KME requires $nb(x)$ observations in the censored model to do as well as the empirical survival function does with n observations from the noncensored model. HPS (1985) showed (1) as x tends to 0, $b(x)$ tends to 1; (2) as x tends to infinity, $b(x)$ tends to infinity; (3) $b(x)$ is increasing in x; and (4) $b(x)$ increases as censoring increases stochastically. The results show that when x is small, the KME's efficiency loss is small, but for large values of x, the KME should be used with caution, particularly in cases of heavy censoring.

Harrington and Fleming (HF) (1982) obtained, using results of Gill (1980), asymptotic relative efficiencies of their class $\mathcal{G}^\rho$ of tests. The $\mathcal{G}^\rho$ class is a special case of Gill's (1980) $\mathcal{K}$ class. They take the censoring distributions G_1, G_2 to be equal and consider the family of survival functions $\mathcal{H}_\rho(x)$ given by

$$\mathcal{H}_0(x) = e^{-e^x} \qquad (\rho = 0), \tag{11.158}$$

$$\mathcal{H}_\rho(x) = (1 + \rho e^x)^{-1/\rho} \quad (\rho > 0). \tag{11.159}$$

They consider the time-transformed location alternatives of survival functions

$$S^\rho(x) = \mathcal{H}_\rho(g(x) + \theta),$$

where g is an arbitrary monotonically increasing time transformation. The two samples can be viewed as being modeled with the survival functions $\overline{F}_i(x) = \mathcal{H}_\rho(g(x) + \theta_i)$, $i = 1, 2$, with $\Delta = \theta_2 - \theta_1$ and H_0 corresponding to $\Delta = 0$. HF (1982) obtained the asymptotic efficiency of the $\mathcal{G}^\rho$ statistic versus the $\mathcal{G}^{\rho^*}$ statistic for survival alternatives $\overline{F}_i(x) = \mathcal{H}_{\rho^*}(g(x) + \theta_i^N)$. The statistic $\mathcal{G}^{\rho^*}$ is fully efficient against $\mathcal{H}_{\rho^*}$. Expression (11.160) gives the asymptotic efficiencies $e(\mathcal{G}^\rho, \mathcal{G}^{\rho^*})$ in the special case of proportional hazards, where $\overline{G}_1(x) = \overline{G}_2(x) = \{\overline{F}(x)\}^\beta$:

$$e(\mathcal{G}^\rho, \mathcal{G}^{\rho^*}) = \frac{(2\rho + \beta + 1)(2\rho^* + \beta + 1)}{(\rho^* + \rho + \beta + 1)^2}. \tag{11.160}$$

For $\rho = 0, \mathcal{G}^\rho$ is the logrank test, and it is optimal against shifts of the extreme value distribution given by (11.156). See HF (1982), Leurgans (1983, 1984), and Fleming and Harrington (1991, Section 7.4) for further efficiency results concerning Mantel's logrank test and its competitors.

Chapter 12

Density Estimation

INTRODUCTION

A common assumption in the previous chapters has been that the sampled data comes from a continuous distribution. This chapter examines methods of estimating the distribution of the population from which the data is sampled in such cases. If an estimate of the density of a continuous population is available, one can find estimates of population statistics such as the mode, range, and quantiles and estimate probabilities associated with the population, as well as make subjective determinations of whether data appears to have a symmetric distribution or not, or whether two distributions appear to be of the same general form.

The methods of estimating densities are typically computationally intensive and require the use of software for even small sets of data. Accordingly, this chapter will rely on the use of software for examples. Section 12.1 provides an introduction to the density functions and gives a popular, commonly used estimate, the histogram. In Section 12.2, the idea of kernel estimation is introduced and several kernels are examined, and Section 12.3 discusses bandwidth selection methods.

Data. There are n observations $X_1, X_2, \ldots, X_n$.

Assumptions

A1. The observations $X_1, X_2, \ldots, X_n$ are a random sample from a continuous population. That is, the X's are mutually independent and identically distributed.

12.1 DENSITY FUNCTIONS AND HISTOGRAMS

Properties of Densities

The probabilities associated with a real-valued continuous random variable X are evaluated through the use of a probability density function (pdf) f through the relation

$$P(a < X < b) = \int_a^b f(x)dx \tag{12.1}$$

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

for any pair of real numbers $a \leq b$. One may substitute inequalities that are not strict in the above equation without affecting the evaluation of the probability. The density function f must satisfy two additional properties:

D1. $f(x) \geq 0$ for all x and

D2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

The pdf f is related to the cumulative distribution function (cdf) F by

$$F(b) = P(X \leq b) = \int_{-\infty}^{b} f(x)dx \tag{12.2}$$

which implies

$$P(a < X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a). \tag{12.3}$$

In a parametric setting, the density function relies on the parameters that specify the distribution within its family. For example, data following a normal distribution with mean μ and variance σ^2 has its probabilities completely determined through its pdf:

$$f(x) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty, \tag{12.4}$$

where μ is any real number and $\sigma > 0$. Estimating the density at (12.4) only requires estimation of the parameters μ and σ^2. Popular methods to do this include maximum likelihood estimation and the method of moments estimator. Of course, the estimate obtained is valid only if the data does, in fact, follow the normal distribution.

It is more likely that someone needing to estimate a density does not know ahead of time that the density belongs to a certain class or family of parametric distribution functions. In this case, nonparametric estimation is required. We now present one such estimator.

Using (12.3), we note that for a continuous random variable X with cdf F and $h > 0$

$$P(x - h/2 < X \leq x + h/2) = F(x + h/2) - F(x - h/2) = \int_{x-h/2}^{x+h/2} f(y)dy. \tag{12.5}$$

If f is smooth and h is small,

$$\int_{x-h/2}^{x+h/2} f(y)dy \approx h \cdot f(x).$$

A reasonable estimate for $f(x)$ is then

$$\hat{f}(x) = \frac{F(x + h/2) - F(x - h/2)}{h}. \tag{12.6}$$

From (12.2), it is clear that if f is not known, then neither is F. So we have just traded the problem of estimating f with that of estimating F. Fortunately, we already have an estimate for the cdf. The empirical cdf of a sample was used in Chapter 5 in association with the Kolmogorov–Smirnov test. The empirical cdf is

$$F_n(t) = \frac{\#\text{ of } X_i \leq t}{n} \tag{12.7}$$

for any real t. This leads to a step-function estimate for F.

EXAMPLE 12.1 ***Spatial Ability Scores.***

In a study examining the relation between student mathematical performance and their preference for solving problems, Haciomeroglu and Chicken (2011) gathered data on a student's spatial ability using four tests of visualization. For each student, their data were combined into a single score giving an overall measure of spatial ability. These scores for 68 female high school students enrolled in advanced placement (AP) calculus classes in Florida are given in Table 12.1. High scores are associated with students with strong spatialization skills. The empirical cdf for these 68 observations is shown in Figure 12.1. This figure is a plot of the output from the `R` function `ecdf`. The circles on the plot represent an end point of an interval that is closed, that is, the point is in the interval. The interval in this case is the interval on which the estimate of F is constant. These intervals are closed on the right and open on the left. Thus, for a value of x that occurs in a circle, the estimate $F_n(x)$ is always shown by the circle at x, not the solid line at x.

The Histogram

One estimate of the density function f at a point x is found by combining (12.6) and (12.7)

$$\hat{f}(x) = \frac{F_n(x - h/2) - F_n(x + h/2)}{nh} = \frac{\#\text{ of } X_i \text{ in } (x - h/2, x + h/2]}{nh}. \tag{12.8}$$

The histogram estimate of a density function simplifies the estimate at (12.8) by removing the requirement that the center of the intervals used for estimating $\hat{f}(x)$ is x. Instead, a number of centering points based on the range of the data is used. The value for h is chosen so that if the $c_j, j = 1, 2, \ldots, m$ are a collection of fixed centering points, then the intervals $I_j = (c_j - h/2, c_j + h/2]$ are nonoverlapping and the set of intervals I_j covers the range of the random sample. We will refer to these intervals I_j as bins (they are sometimes labeled classes), and we call h the *bin width*.

The histogram counts the number of X's in the bin I_j for any x in I_j, rather than using the bins defined in (12.8):

$$\hat{f}(x) = \frac{\#\text{ of } X_i \text{ in } I_j}{nh} = \frac{n_j}{nh}, \quad x \text{ in } I_j. \tag{12.9}$$

Note that this histogram estimate of f integrates to 1, as required of a density (D2).

Table 12.1 Discrepancy Scores for 68 Female AP Calculus Students

0.129	0.242	0.262	0.284	0.300	0.317	0.324	0.330	0.339	0.353
0.359	0.369	0.377	0.382	0.424	0.425	0.429	0.451	0.453	0.471
0.477	0.479	0.480	0.480	0.483	0.487	0.489	0.501	0.501	0.502
0.503	0.507	0.511	0.520	0.522	0.530	0.532	0.535	0.536	0.540
0.547	0.548	0.551	0.551	0.554	0.556	0.557	0.558	0.581	0.590
0.596	0.604	0.616	0.623	0.627	0.628	0.654	0.663	0.680	0.691
0.744	0.751	0.790	0.806	0.813	0.818	0.830	0.860		

Source: E. Haciomeroglu and E. Chicken (2011).

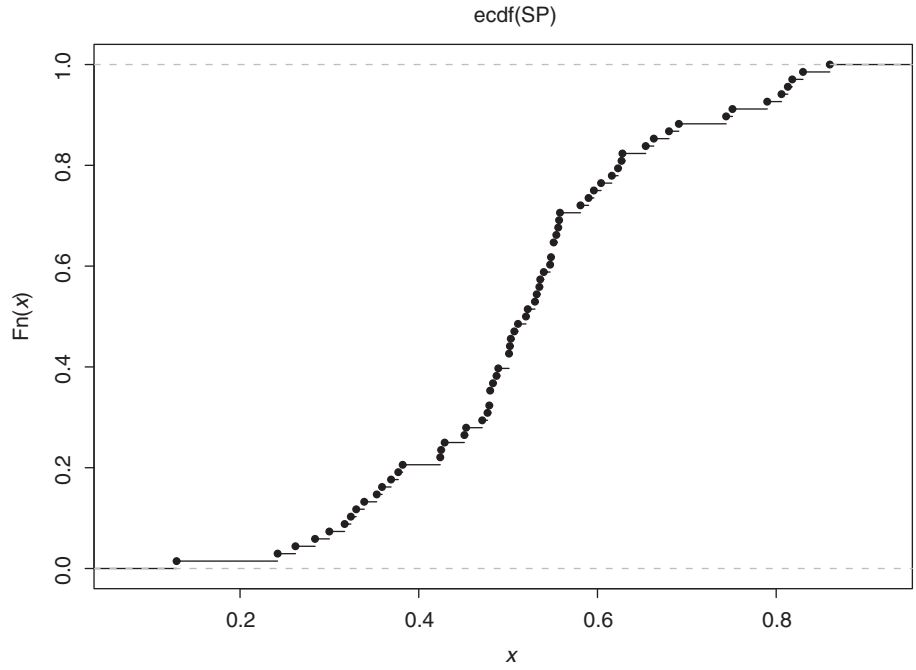

Figure 12.1 The empirical cdf for the data in Example 12.1.

Generally, a vertical line is added to a graph of the histogram at each bin end point. An alternative to (12.9) is to use only the numerator in the estimate. In this case, the height of the estimate is just the integer count in each bin. This will have the same shape as the histogram but is not a density estimator because it no longer integrates to 1.

Procedure

Suppose `x` is a random sample of size n from a continuous population, $\mathtt{x} = \{X_1, X_2, \ldots, X_n\}$. To estimate a histogram of `x` using `R`, the command `hist(x)` is used. The argument `freq` is used either to set the vertical axis to denote the number of X_i in a particular bin (`freq=TRUE`) or to set the bin height proportional to the number of X_i in a bin (`freq=FALSE`). The latter option gives a true density estimate, because only it will integrate to 1. The argument `breaks="FD"` will use the rule of Freedman and Diaconis (1981) for determining the number of bins m and the bin width h. This method selects the bin width to be

$$h = 2 \cdot \text{IQR} \cdot n^{-1/3},$$

where IQR is the interquartile range of the data and n is the number of samples. The command `hist` then divides the range of the data by h to determine m. Additionally, it may modify the number of bins and the bin width slightly (using the `pretty` function) to ensure visually pleasing bin end points.

Unlike the other methods to be discussed later in this chapter, the histogram is simple enough to be implemented manually. After selection of the number of bins m and bin width h, as well as the placement of the end points or midpoints of the bins, one sets the

value of the estimate over a bin interval to be the number of observed values X_i that fall into that bin, divided by nh. For large data sets, this manual procedure becomes unwieldy.

EXAMPLE 12.2 ***Histogram for Spatial Ability Data.***

For the data in Table 12.1, the IQR is 0.1525 and $n = 68$. This gives a bin width of $h = 0.075$ using the rule of Freedman–Diaconis. The maximum value of the spatial ability data is 0.860 and the minimum is 0.129, so the number of bins is $m = (0.860 - 0.129)/0.075 = 9.747$. This should be rounded up to next integer to ensure coverage of the range. Thus, the histogram will use $m = 10$ bins of size 0.075 to estimate the density of the data. If the histogram was to be manually drawn with these values, then these m bins are

$$I_j = \begin{cases} [0.129, 0.129 + 0.075], & j = 1, \\ (0.129 + 0.075 \cdot (j-1), 0.129 + 0.075 \cdot j], & j = 2, \ldots, 10. \end{cases}$$

The first bin, I_1, is closed on the left in order to ensure the minimum value of the data is included in the histogram. `R` takes the number of bins and the bin width m provided by the Freedman–Diaconis rule as a suggestion rather than a strict requirement. It modifies the number of bins and the bin widths with the `pretty` function to create a visually pleasing graph. The function `pretty` forces values that are one, two, or five times a power of 10. For this data, `R` has changed m to 8 and h to 0.1. Additionally, it starts the bins at 0.1, rather than the minimum data value of 0.129. Using these values, the height of the histogram for the discrepancy data is given in Table 12.2. This data is plotted in Figure 12.2a. The histogram was produced by the `R` command `hist` using arguments `freq=F` and `breaks="FD"`. Examining the plot of the histogram, possible interpretations are that the distribution of the population from which the data was sampled may be bimodal and skewed to the left.

Modifying the number of bins may have a significant effect on the histogram estimate. Choosing too large a bin width results in an oversmoothed estimate of the density, while too small a bin width gives an estimate that is overly sensitive to the sample rather than to the underlying distribution. This is equivalent to a problem in variance-bias tradeoff. As h increases, the estimate displays increasing bias and decreasing variability. As h decreases, the opposite occurs. (See Comment 1.)

EXAMPLE 12.3 ***Effect of Changing Bin Width.***

Figure 12.2 shows the histogram for the spatial ability data in Table 12.1 with different choices of bin width h. The number of bins will change as h does in order to ensure

Table 12.2 Histogram Data for Example 12.1

j	I_j	n_j	n_j/nh	j	I_j	n_j	n_j/nh
1	[0.1, 0.2]	1	1/6.8	5	(0.5, 0.6]	24	24/6.8
2	(0.2, 0.3]	4	4/6.8	6	(0.6, 0.7]	9	9/6.8
3	(0.3, 0.4]	9	9/6.8	7	(0.7, 0.8]	3	3/6.8
4	(0.4, 0.5]	13	13/6.8	8	(0.8, 0.9]	5	5/6.8

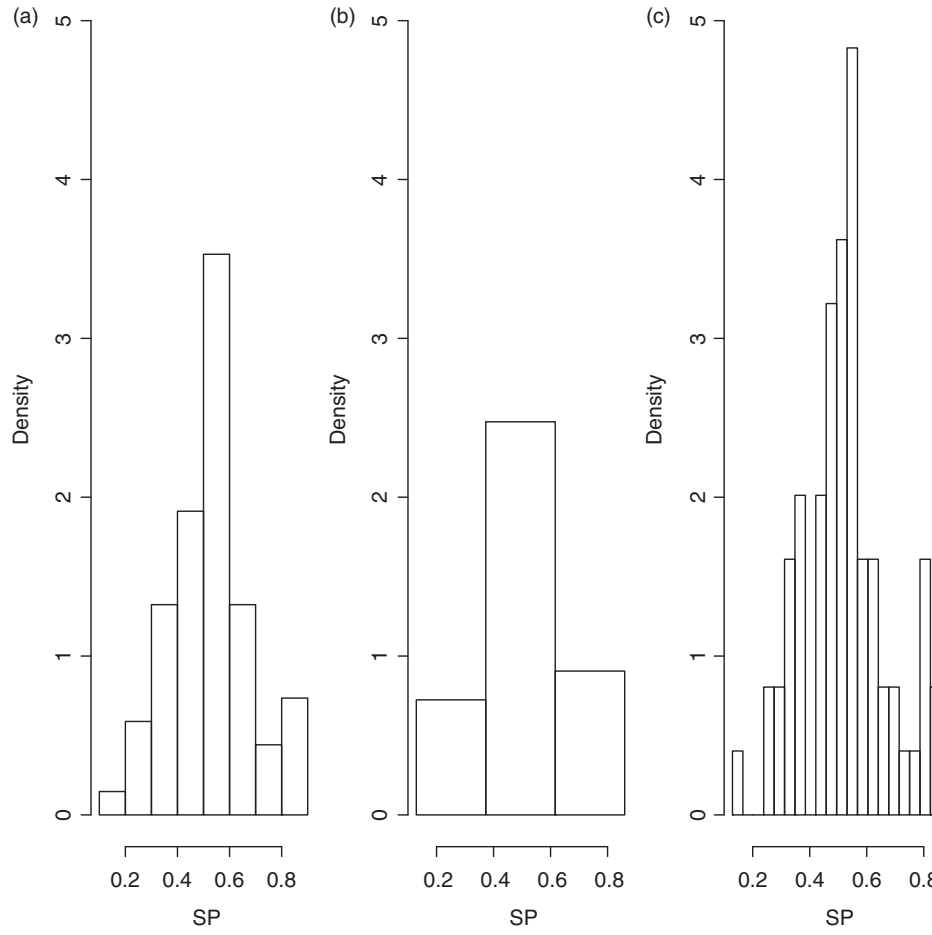

Figure 12.2 Histograms for the data in Example 12.1.

the range of the data is covered by the histogram. The histogram in Figure 12.2b has $m = 3$ bins of width 0.244. Compared to Figure 12.2a, this histogram is smoother and the distribution of the population appears much more symmetric. The histogram in Figure 12.2c has $m = 20$ bins of width 0.037. The distribution now appears to be bimodal.

A drawback to having too many bins is the introduction of regions of zero probability. The histogram in Figure 12.2c has one such area. This corresponds to a gap in the sample data. It is unlikely that the true density of the spatial ability data has regions of zero probability. Additionally, it is undersmoothed. This estimate is too variable for an assumed continuous distribution.

Comments

1. *Bias, Variance, and Integrated Mean Squared Error of the Histogram.* If the underlying density f from which the population is drawn is continuous with two continuous, bounded derivatives, Scott (1979) shows that the bias of the histogram at a point x is

$$hf'(x)/2 - f'(x)(x - t_x) + O(h^2),$$

where h is the bin width, t_x is the left end point of the bin containing x, and the term $O(h^2)$ is on the order of h^2. The variance of the histogram at a point x is given by

$$f(x)/(nh) + O(n^{-1}),$$

where $O(n^{-1})$ is on the order of n^{-1}. It is evident that as h increases, the bias increases and the variance decreases. The reverse is true as h decreases. In order to have both a small variance and bias, it is desirable to have h small and nh large. This requires that $nh \to \infty$, $n \to \infty$, and $h \to 0$. In light of this, the final terms in the above expressions for the bias and variance are ignored and the remainders are referred to as the *asymptotic bias and variance*.

To find a useful bin width for the entire histogram, the integrated mean squared error (IMSE) is used. For the histogram, the asymptotic IMSE is

$$(nh)^{-1} + h^2/12 \cdot \int (f')^2. \tag{12.10}$$

This expression is analyzed to find both optimal bandwidths and kernels. See Comment 2.

2. *Rules for Number of Bins and Bin Width*. There are a number of rules for selecting the number bins m, or equivalently, the bin width h. The `hist` function in `R` implements the Freedman–Diaconis rule discussed above (`breaks="FD"`) as well as a few others. The argument `breaks="Scott"` uses

$$h = \frac{3.5\hat{\sigma}}{n^{1/3}},$$

where $\hat{\sigma}$ is the sample standard deviation. This bin width is found by minimizing the asymptotic mean integrated squared error (MISE) (12.10) under the assumption that the population is normal. The Freedman–Diaconis (1981) rule is derived similarly to Scott's rule, but uses $2 \cdot IQR$ in place of $3.5\hat{\sigma}$.

Another method is given by Sturges (1926) and implemented with `breaks="Sturges"`. Here, the number of bins is determined by

$$m = \log_2 n + 1,$$

and the value is rounded up to the next integer. This number is found by assuming underlying normality. For nonnormal data, Sturges' rule may oversmooth the histogram density estimator by specifying too few bins. `R` uses Sturges' method as the default.

3. *End Point Determination*. The intervals used for the bins are closed on the right, that is, of the form $(a, b]$. This is the default for `R` and seems natural given (12.7). However, `R` allows for the option of having the intervals closed on the left. This is accomplished with the argument `right=TRUE` or `right=FALSE` in the `hist` function. Whichever method is selected, the bins at the extreme left and right are closed on the left and right, respectively.

4. *Unequal Bin Widths*. We have assumed that the bins are all the same size. This is not necessary. The bin widths may be $h_1, h_2, \ldots, h_m$, an arbitrary set of positive values, provided the resulting bins are nonoverlapping and cover the complete range of the data. All that is necessary to determine the histogram in this case

is to substitute h_j for h in (12.9). In R, this can be implemented by specifying a vector of unequal bin widths through the argument `breaks`.

5. *Average Shifted Histograms*. A variant of the histogram described in Scott (1992) is the average shifted histogram (ASH). In this method, m different histograms are calculated, each with fixed bin width h. However, the leftmost bin end point is varied for each histogram: if the first histogram starts at a point x_0, then the remaining $m-1$ histograms begin at $x_0 + h(i/m)$, where i runs from 1 to $m-1$. Weighted averages of these m histograms are taken as the ASH density estimate. In R, ASH is implemented with commands available in the package `ash` developed by Gebhardt (2009).

Properties

1. *Consistency*. For conditions under which consistency is obtained, See Scott (1979, 1992).

Problems

1. Using the data from Table 12.1, calculate the bin width and number of bins using Scott's rule and Sturges' rule. Compare histograms created for the spatial ability data with these values and the values obtained from the Freedman–Diaconis rule. Are different characteristics of the underlying distribution evident as the bin width changes?

2. For the data in Table 12.1, can the bin width be changed to give a density estimate that appears to be symmetric? Multimodal? Skewed left or right?

3. The data in Table 12.3 contains 82 male spatial ability studies from Haciomeroglu and Chicken (2011). Find the bin widths and number of bins using the three different rules. Compare the histograms created using these three methods. Are different characteristics evident?

4. Do the male and female spatial ability data from Table 12.1 and Table 12.3 appear to come from the same underlying distribution when using one of the three bin width rules? Can a bin width be chosen that maximizes (minimizes) the similarity of the two histograms?

5. Combine the spatial ability data from Tables 12.1 and 12.3 into a single sample of size 150. Does the male or female data appear to come from the same distribution as the combined distribution based on the shape of the histogram? Can the bin width be changed to change this conclusion?

6. Use the data in Table 5.8 to create a histogram for each group of subjects. Do the histograms appear to represent the same underlying distribution?

Table 12.3 Discrepancy Scores for 82 Male AP Calculus Students

-0.147	0.075	0.244	0.275	0.326	0.329	0.335	0.353	0.369	0.411
0.416	0.419	0.427	0.427	0.432	0.442	0.443	0.447	0.451	0.473
0.475	0.476	0.489	0.490	0.501	0.505	0.505	0.513	0.517	0.519
0.530	0.531	0.537	0.550	0.553	0.559	0.568	0.573	0.576	0.577
0.578	0.582	0.582	0.603	0.619	0.624	0.625	0.628	0.629	0.632
0.636	0.641	0.646	0.655	0.662	0.662	0.663	0.667	0.668	0.676
0.677	0.677	0.683	0.684	0.693	0.696	0.699	0.699	0.707	0.711
0.727	0.738	0.749	0.749	0.749	0.772	0.794	0.797	0.810	0.822
0.859	0.859								

Source: E. Haciomeroglu and E. Chicken (2011).

12.2 KERNEL DENSITY ESTIMATION

While a histogram provides a simple estimate of a density function, it does have some drawbacks. First, if the data comes from a continuous distribution, the histogram cannot provide a continuous density estimate. There will be jumps in the data corresponding to the end points of the bins unless two adjoining bins have the same number of observations in those bins. Even then, there are jumps at the beginning and end of the histogram. Second, the estimator is constant over intervals. Only the number of observations in a bin have an effect on the density estimate, not the placement of the observed data within each bin. Thus, for all points x in a bin, the value of the estimate is the same. To overcome these issues, we use kernel functions to estimate the densities. We introduce this idea through the use of centered histograms.

Centered Histogram

In a histogram, each sample point in an interval I_j adds $1/(nh)$ to the height of the estimate for all points x in I_j. This results in the blocky, discontinuous shape of histograms. To get around this, we drop the fixed intervals I_j in favor of the x-centered intervals in (12.8).

The bin width will remain h, a constant. The estimate of the modified histogram at a point x will be proportional to the number of observed data points X_i that fall in the interval $(x - h/2, x + h/2]$. For each observed data point in this interval, $1/(nh)$ is added to the estimate. This is the same amount added to the usual histogram estimate with fixed bins. The difference is that now the bins are centered at each point x where a density estimate is desired. The height of the modified histogram does not depend on the fixed bins used in the usual histogram but rather on the number of data points near x. In this modified histogram, points are considered to be near x by being no more than $h/2$ away from x, while in the usual histogram, a point is near x if it is in the same, predefined bin.

EXAMPLE 12.4 ***Histogram with Centered Bins.***

As an illustration, take $h = 0.5$ and use the spatial ability data from Table 12.1. The estimate at (12.8) becomes

$$\hat{f}(x) = \frac{\#\text{ of } X_i \text{ in } (x - 0.25, x + 0.25]}{34} \tag{12.11}$$

For any $x < 0.129 - h/2 = -1.121$, the number of points X_i in $(x - h/2, x + h/2]$ is zero. This is also true for any $x \geq 0.860 + h/2 = 1.110$. The estimate of the density is therefore $\hat{f}(x) = 0$ for these values of x.

For $x \in [0.129, 0.860)$, the x-centered bins will contain at least one data point X_i. For example, if $x = 0$, the number of data points in $(x - h/2, x + h/2] = (-0.25, 0.25]$ is 2 and $\hat{f}(0) = 12/nh = 2/34$. Similarly, for $x = 0.2$, there are 17 points in the x-centered interval and $\hat{f}(0.2) = 17/34$.

This modified histogram estimator integrates to 1 just as before and is always greater or equal to 0. Similar to the histogram, a value for $\hat{f}$ is available for every x. However, this value depends on the point x more so than in the histogram. With the histogram, all that counted was which bin x was in. Now, the actual value of x is important.

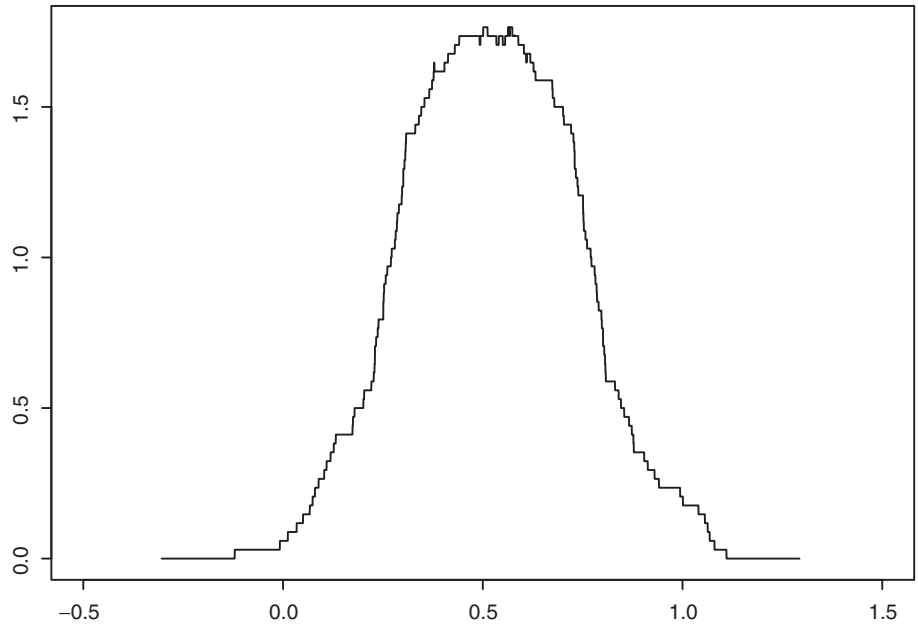

Figure 12.3 Centered histogram estimate of the spatial ability data.

Evaluating the centered histogram is more time intensive than a histogram, and it is not recommended that it be found by hand. A centered histogram may be created in R with the function `density`. Figure 12.3 shows the centered histogram using the spatial ability data. By default, R will evaluate $\hat{f}$ at each of the n data points. Users may specify other ranges of x for which $\hat{f}$ is evaluated.

The reliance of the histogram on the bin end points has thus been addressed, but Figure 12.3 shows that (12.8) is still a discontinuous estimate and constant over intervals (albeit small intervals). This is overcome by using a kernel function to smooth out the estimator.

Kernels

A kernel function K is a function such that

$$K(x) \geq 0, \quad -\infty < x < \infty,$$
$$K(-x) = K(x),$$

and

$$\int_{-\infty}^{\infty} K(x)dx = 1.$$

Thus, K is a nonnegative function symmetric about 0, which integrates to 1. Numerous kernel functions have been proposed for use with density estimation. A simple one is the uniform, or box, kernel. This kernel is defined as

$$K(x) = \begin{cases} 1, & -1/2 \leq x < 1/2, \\ 0, & \text{otherwise.} \end{cases} \tag{12.12}$$

and clearly meets the three restrictions given above. Another common kernel function is the normal kernel (12.4) with μ set to 0 to ensure symmetry and σ a constant. If K is a kernel function, then so is the scaled version

$$\frac{1}{h} \cdot K\left(\frac{x}{h}\right) \tag{12.13}$$

for $h > 0$. This scaled kernel can be centered at any data point X_i, and symmetry is not around 0 as above, but around X_i

$$\frac{1}{h} \cdot K\left(\frac{x - X_i}{h}\right).$$

Given the data, the kernel density estimate is

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{12.14}$$

When using kernel density estimators, it is customary to refer to h as the bandwidth, rather than the bin width.

If the kernel function is taken to be the box kernel (12.12), then (12.14) and (12.8) are identical. In this case, the kernel K takes on the value 1 whenever an observation X_i is near x. The notion of nearness is determined by the bandwidth h and the form of K. Taking the sum, the estimate at x is the number of data points X_i near x multiplied by the factor $1/(nh)$. As with the centered histogram, the height of the box kernel density estimator increments by $1/(nh)$ for every X_i in the interval centered at x.

EXAMPLE 12.5 ***Spatial Ability Kernel Density Estimate.***

Using the box kernel (12.12) with the spatial ability data and $h = 0.5$, the density estimate at a point x is given by

$$\hat{f}(x) = \frac{1}{34} \sum_{i=1}^{68} K\left(\frac{x - X_i}{1/2}\right).$$

$K((x - X_i)/(1/2))$ will equal 1 whenever X_i is in $(x - h/2, x + h/2] = (x - 0.25, x + 0.25]$ and 0 otherwise. Each time it is 1, an increment of $1/34$ is added to the estimate of f at x. This is, of course, the same as the centered histogram estimate. The plot in Figure 12.3 was created by plotting the output from the function call

```
density(SP, kernel="r", bw=1/(4 * sqrt(3)), n=2^14)
```

where `SP` is the data, `kernel="r"` specifies using the box ("rectangular") kernel, `bw` specifies the bandwidth and `n` tells `R` at how many values of x the density will be evaluated. The unusual value for the bandwidth is a result of the manner in which the `density` function scales the kernel function. The function `density` scales the kernel K to K^* such that

$$\texttt{bw} = \sqrt{\int x^2 K^*(x) \mathrm{d}x}$$

for a user-specified value of the argument bw. Because K is a density, this is equivalent to scaling the kernel so that a random variable with density K^* has standard deviation equal to bw. The value of $\mathtt{bw} = 1/(4\sqrt{3})$ will result in the scaled kernel (12.13) with $h = 0.5$. For a general h, one would set $\mathtt{bw} = h/(2\sqrt{3})$.

As with the histograms, changing the bandwidth h will affect the shape of the estimated density. In Figure 12.3, one might believe that the bandwidth h is too large, resulting in the loss of the bimodal and skew features evident in the histogram estimate. Figure 12.4 shows estimates with the box kernel, but $h = 0.1$ ($\mathtt{bw} = 0.1/(2\sqrt{3})$, solid line) and $h = 0.2$ ($\mathtt{bw} = 0.2/(2\sqrt{3})$, dashed line). The dotted line is the estimate from Figure 12.3 ($h = 1/2$). Increasing the bandwidth provides a smoother estimate over a wider range of values.

R provides several additional options for the choice of the kernel function. Three of these are displayed in Figure 12.5. Figure 12.5a is the normal, or Gaussian, kernel. This uses the form given at (12.4) with μ set to 0 and $\sigma = 1$. The normal kernel is smoother than the box kernel, and its support is infinite, rather than finite. However, because it is a symmetric density, it does meet the above requirements for a kernel function. An advantage to using a smooth kernel like the normal is that it will provide a density estimate that is smooth. The box kernel contains jumps that were carried over to the estimate it produced. The normal kernel has no such problems.

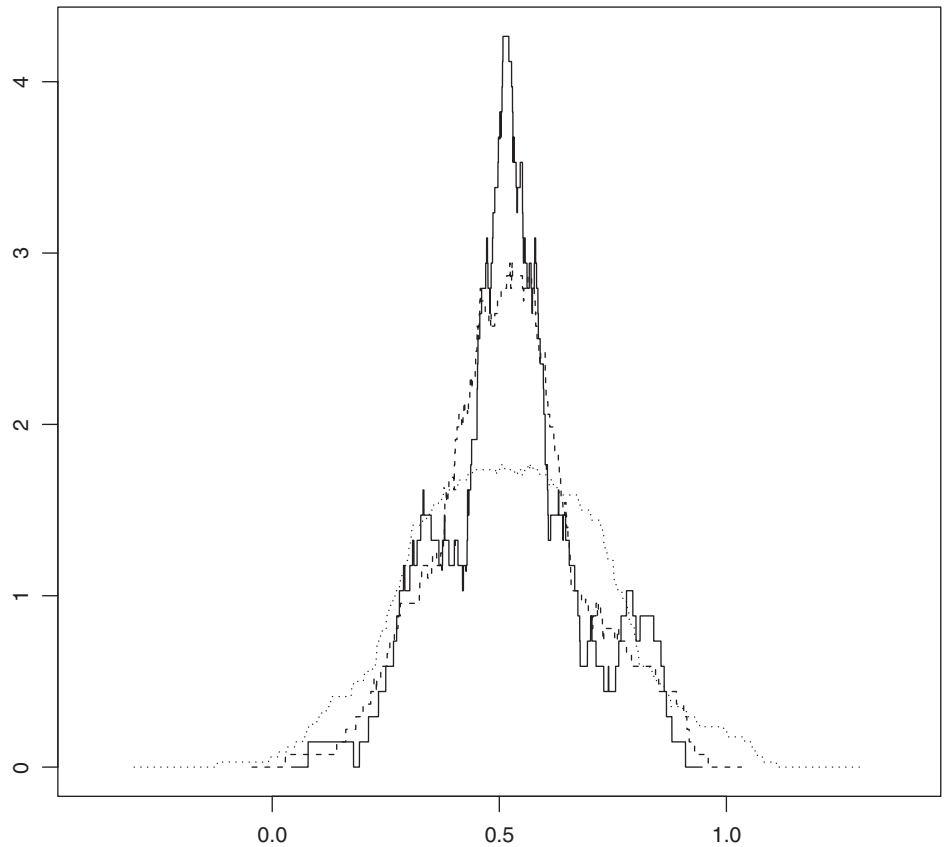

Figure 12.4 Box kernel estimate with differing bandwidths.

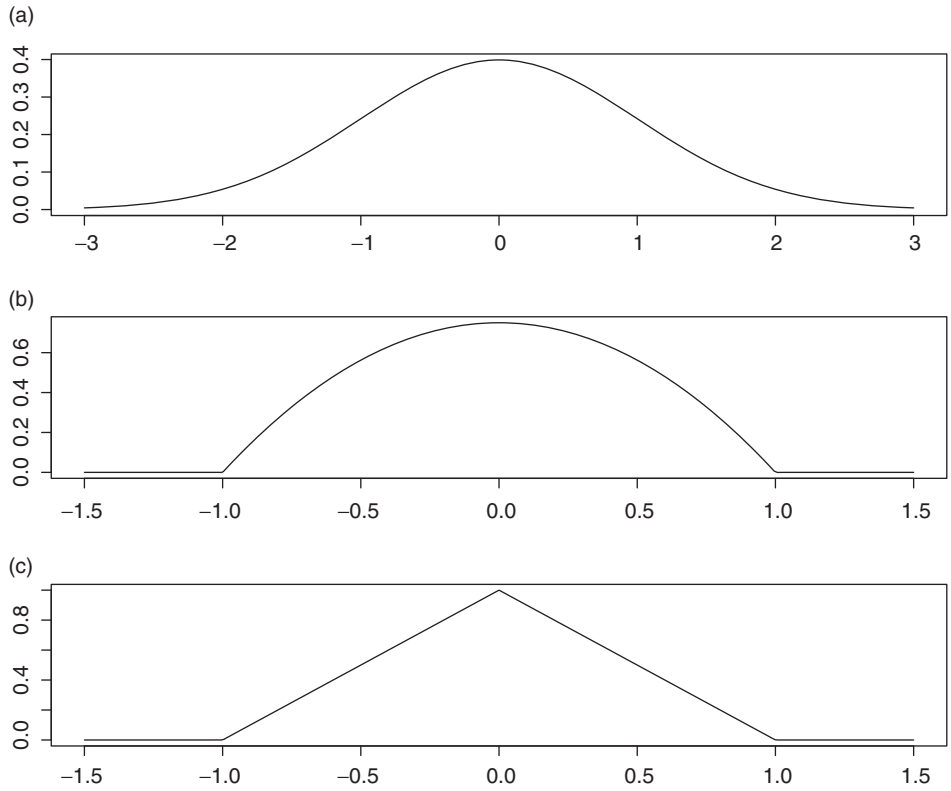

Figure 12.5 Kernel functions.

Figure 12.5b is the Epanechnikov kernel. This kernel is of the form

$$K(x) = \frac{3(1-x^2)}{4}, \quad -1 \le x < 1. \tag{12.15}$$

Figure 12.5c is the triangle kernel:

$$K(x) = 1 - |x|, \quad -1 \le x < 1.$$

If the kernel function is not the box kernel, the kernel density estimator is still interpreted in the same way. Like before, $K((x - X_i)/h)$ provides a measure of the number of points near x. Instead of being a 1 if near x and a 0 if not near x, which is the same as counting the number of X_i near x, the kernel now assigns a measure that reflects the degree of nearness to x for the points X_i. The kernel K integrates to 1 and is nonnegative, so its value approaches 0 when its argument is far from 0. If the kernel has a single mode at 0, then low values of $K((x - X_i)/h)$ correspond to X_i that are far from x, while high values are close to x. For the box kernel, which is not unimodal, all data points are either equally close or equally far from x. There is no degree of closeness or farness.

One may also think of the kernel as providing a weighted value for the number of X_i near x. In particular, data values near x give more weight than those further away.

Because the normal kernel has infinite support, the bandwidth h does not provide a fixed window like the box kernel. Every point X_i gives some weight to the estimate at x. Recalling that 99.7% of the mass of a normal density is within three standard deviations

of the mean, most of the observed data points that provide weight in the kernel estimate at a point x will be close.

Unlike the box kernel, the relation between the argument `bw` and the bandwidth h is simple. Because the standard deviation is σ, one just equates h with σ.

EXAMPLE 12.6 ***Kernel Choice.***

The density estimate using the normal, Epanechnikov, and triangle kernels on the spatial ability data is shown in Figure 12.6. Note that using these kernels provide a much smoother estimate than that obtained with the box kernel in Figure 12.3. In general, for a given data set, the smoother the kernel, the smoother the estimate. The three kernel density estimates in Figure 12.6 each use `bw=0.25`.

Comments

6. *Equivalency of Histogram and Kernel Estimator*. We have seen that the box kernel and the centered histogram are equivalent. This equivalency carries over to the histogram as well. The centered histogram is

$$\hat{f}(x) = \frac{\#\text{ of } X_i \text{ in } (x - h/2, x + h/2]}{nh}.$$

This is the same as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

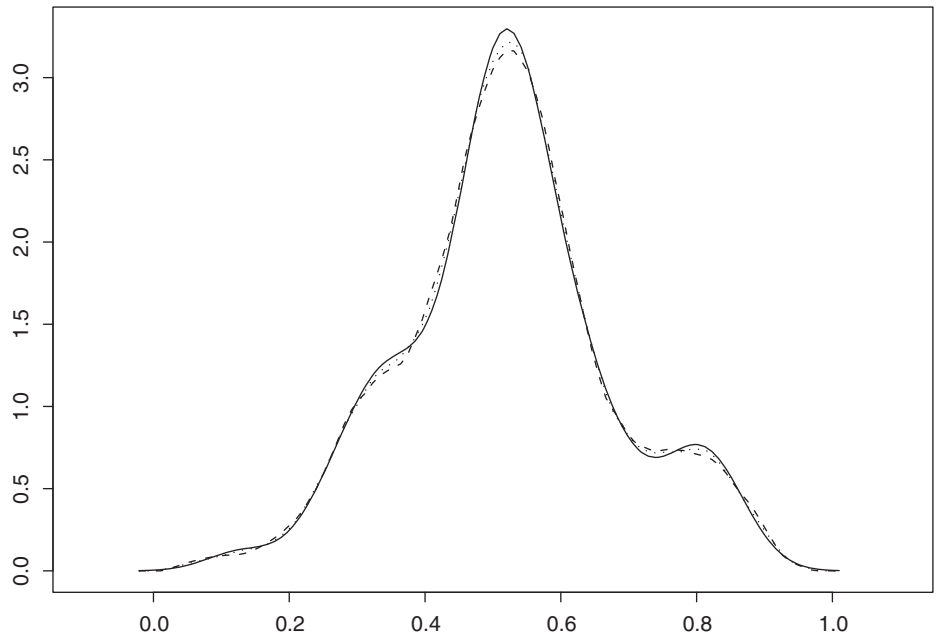

Figure 12.6 Density estimated of spatial ability data using differing kernel functions.

provided the box kernel (12.12) is used. This counts the number of points X_i in the interval centered at x. Let c_j be the centers of the histogram bins I_j. For any x in $\cup_j I_j$, set c_x to be c_k that is closest to x, that is, $c^* = \arg\min |x - c_k|$. If x is equidistant from two c_k, take c_x to be the larger. Then the histogram estimate $\hat{f}_H$ may be written as

$$\hat{f}_H(x) = \hat{f}(c_x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{c_x - X_i}{h}\right) = \frac{\#\text{ of } X_i \text{ in } (c_x - h/2, c_x + h/2]}{nh}$$
$$= \frac{\#\text{ of } X_i \text{ in } I_j}{nh}, \quad x \text{ in } I_j.$$

7. *Asymptotic Mean Integrated Squared Error (MISE) and the Epanechnikov kernel.* The MISE between a function f and its estimate $\hat{f}$ is given by (Scott and Terrell, 1987).

$$MISE(f, \hat{f}) = E\left(\int (f - \hat{f})^2\right).$$

Asymptotically, for a kernel function K this is

$$\frac{\int K^2}{nh} + \frac{\sigma_K^4 h^4 \int (f'')^2}{4}, \tag{12.16}$$

where $\sigma_K^2 = \int x^2 K(x) dx$.

Epanechnikov (1969) minimized (12.16) with respect to the choice of kernel function K, resulting in (12.15). The reference density f is again assumed to be normal.

8. *Kernel Choice*. Rosenblatt (1956) and Parzen (1962) pioneered development of nonparametric kernel density estimation. In these papers they discussed properties of density estimates based on using weight functions (kernels), including the box and triangle kernels described in this section. While differing choices for kernel functions result in visual changes to the estimate (i.e., smooth kernels give smooth estimates), Epanechnikov (1969) shows that the kernel choice does not have a significant impact on the statistical properties of kernel density estimates. While the choice of kernel is not of great importance, the choice of bandwidth is a crucial part of kernel density estimation. Bandwidth selection will be explored in the subsequent section.

9. *Binned kernel estimates*. Binning the observed data increases the speed at which a kernel density estimate can be obtained. A regularly spaced grid covering the range of the data is created and a binned data value at each grid point is determined by weighting the observed data. The binned value of the data at g_j is a weighted combination of the observed data X_i. Observed points X_i near a particular grid point g_j provide more information and weight than those further away. Using a normal kernel, Silverman (1982) provided very fast algorithm for calculating the density estimate. Hall and Wand (1996) show the bias of the binned density estimate is more sensitive to binning than the variance and that this sensitivity is reduced for smoother kernels. They also provide some guidelines for the size of the grid.

Properties

1. *Consistency*. For conditions under which consistency is obtained, see Parzen (1962).

Problems

7. Find the kernel density estimate of the female spatial ability data from Table 12.1. What differences are evident as the kernel function K is changed? And as the bandwidth changes? For a fixed bandwidth, how do the characteristics of the estimate change as the kernel function is varied?

8. Find the kernel density estimate of the male spatial ability data from Table 12.3. Does the underlying distribution of this data appear to come from the same distribution as the female data? Can changing the kernel function change this conclusion? Explain.

9. Find the kernel estimate of the combined male and female spatial ability date. Do the distributions for the male or female data appear to come from the common distribution of all the students? Can choice of bandwidth or kernel change this? Explain.

10. Using the box kernel (12.12), show that the expressions at (12.8) and (12.14) are equivalent.

11. Show that the kernel density estimate K at (12.14) satisfies the properties **D1** and **D2**.

12. Show that for h small enough, the box kernel density estimate becomes $1/(nh)$ on intervals of length h centered at each data point X_i, and 0 everywhere else. Show this estimate has integral 1. *Hint*: Let $h < \min_{i,j} |X_i, X_j|$.

12.3 BANDWIDTH SELECTION

In the implementation of the kernel density estimator in Section 12.2, the bandwidth was a user-specified numeric value. However, choosing a bandwidth in such a way is not ideal. Indeed, the examples make it clear that one can choose a bandwidth that will provide as smooth (or unsmooth) and estimate as desired. Instead, some automated methods for bandwidth are desired. Such bandwidth selection methods should rely on either properties derived from the data, as was done with bin width determination with a histogram, or be based on some theoretical concerns.

Fixed Bandwidth

A common bandwidth derived by minimizing the MISE (12.16) is

$$h = 1.06n^{-1/5} \cdot \min\{\hat{\sigma}, IQR/1.34\}, \tag{12.17}$$

where $\hat{\sigma}$ is the estimate of the standard deviation using the data. See Comment 10. If the data is normal, then the IQR is close to 1.34σ. Rather than specifying a numeric value to the argument `bw` in `R`, a method is provided. In this case, it is implemented via the argument `bw="nrd"`. A variant of this bandwidth described in Silverman (1986) is obtained by using the argument `bw="nrd0"` in the call to `density`. This has the effect of changing the constant 1.06 to 0.90. This is the default bandwidth selection rule in the `R`.

The above bandwidths are found under the assumption that the underlying density is normal. When the data is not normal, they still provide reasonable bandwidth choices.

However, other bandwidth selection methods are available that are data driven rather than dependent on the assumption of normality.

Two such data-driven bandwidths are derived using cross-validation methods in Scott and Terrell (1987). These methods select a value of h that minimizes the MISE. Cross-validation is part of the process in that for each candidate value of h, the method compares the estimated density using all the data to an estimate that leaves out the point X_i when estimating the density at that point. The bandwidths differ in that one provides a function in h that is unbiased estimate of the MISE, while the other produces a biased estimate of the MISE but with smaller variability. These two cross-validation methods are implemented in the function `density` through the argument `bw="ucv"` and `bw="bcv"`.

Another data-driven method is based on the expression for the MISE in (12.16). The bandwidth h in (12.17) is calculated by using a normal density for f. In contrast, Sheather and Jones (1991) replace $\int (f'')^2$ in (12.17) with an estimate using the data instead of assuming a particular form for the density's second derivative. This method is implemented using `bw="SJ"`. Estimating $\int (f'')^2$, perhaps with another kernel estimator, evaluating (12.16) with this estimate and minimizing is referred to as a *plug-in* estimator.

Variable Bandwidth

The previous estimators used a constant bandwidth h for all points x. A modification to this is to use a bandwidth that varies from point to point. If the number of observed data points near a particular observed value of X_i is large, one would expect this dense placement of data to result in an improved density estimate at X_i. In this case, a small bandwidth would be reasonable. If, on the other hand, there are very few observed data points near X_i, a larger bandwidth may be needed in order to incorporate a sufficient number of observations to accurately estimate the density at X_i. The bandwidth becomes locality dependent. Small bandwidths are associated with areas of dense data, while large bandwidths are desirable in areas of sparse data. Thus, h is replaced with h_i:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x - X_i}{h_i}\right). \tag{12.18}$$

Following Silverman (1986), the bandwidths h_i are proportional to a fixed bandwidth h used to construct a pilot estimate $\hat{f}_p$ of the density f:

$$h_i = \lambda_i h. \tag{12.19}$$

The bandwidth factors λ_i are given by

$$\lambda_i = \left(\frac{\hat{f}_p(X_i)}{\left(\prod_{j=1}^{n} \hat{f}_p(X_j)\right)^{1/n}} \right)^{-\alpha}.$$

The parameter α provides a level of sensitivity of the bandwidth to the pilot density estimate.

R implements the adaptive kernel with the function call `akj` in package `quantreg` developed by Koenker (2011). It uses the normal kernel by default. The initial h is

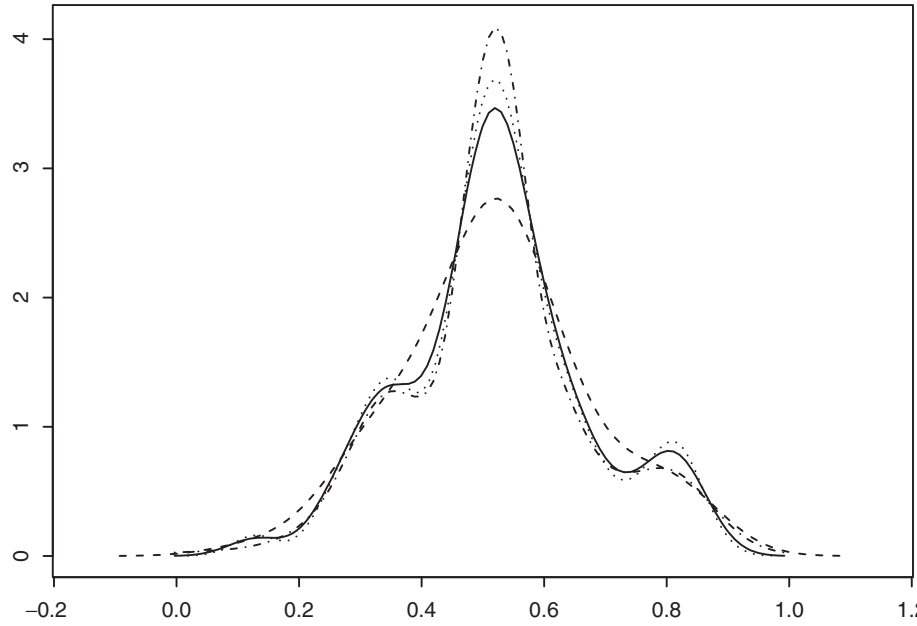

Figure 12.7 Kernel density estimates using automatic bandwidth selection methods and normal kernel.

selected according to (12.17) with Silverman's 0.9 modifier in place of 1.06. Portnoy and Koenker (1989) recommend setting the sensitivity parameter α to be $1/2$ (the default value in R).

EXAMPLE 12.7 ***Bandwidth Choice.***

Figure 12.7 shows the results of applying different automated bandwidth selection methods to the spatial ability data. The solid line uses `bw="nrd0"`, the dashed line is `bw="bcv"`, the dotted line is `bw="SJ"`, and the mixed line uses the variable bandwidth method. The bandwidths for the first three methods are shown in Table 12.4. The value of h in (12.19) is in the final row of the table.

Comments

10. *Optimal Bandwidth*. The Epanechnikov kernel is found by minimizing (12.16) with respect to K. If (12.16) is instead minimized with respect to h, the optimal

Table 12.4 Automated Bandwidths for the Spatial Ability Data

Method	Bandwidth
`"nrd0"`	0.044
`"bcv"`	0.074
`"SJ"`	0.037
variable	0.042

bandwidth is the value of h, which minimizes the MISE:

$$h = \left(\frac{\int K^2}{n\sigma_K^4 \int (f'')^2} \right)^{1/5}. \tag{12.20}$$

Using the normal kernel K with $\sigma_K = 1$ and assuming the underlying data come from a normal population with mean 0 and variance σ^2, the optimal bandwidth is

$$h = (4/3)^{1/5} \sigma n^{-1/5} \approx 1.06 \cdot \sigma n^{-1/5}.$$

11. *Balloon Estimators*. In (12.18), there is a local bandwidth for each observed data point X_i. One could also have the bandwidth varying by location x, the point at which the density is to be estimated. Such a density estimator is called a *balloon estimate*. However, this type of estimate leads to poor results in terms of asymptotic behavior in the univariate case. See Terrell and Scott (1992).
12. *Bootstrap Method*. Faraway and Jhun (1990) proposed a resampling via bootstrap to determine the value of a fixed bandwidth. In simulations, this methods outperforms fixed bandwidth cross-validation methods in terms of squared errors but is computationally expensive. Hall and Kang (2001) further examine such bandwidth methods and compare the computational expense with the improvements in the estimates.

Problems

13. For the female spatial ability data in Table 12.1, find the kernel density estimate using several of the bandwidth selection rules. Compare these bandwidths to those calculated for the male spatial ability data from Table 12.3.

14. Using the female spatial ability data, do the automatically selected bandwidths remain stable as the kernel function changes? Is this true for the male data?

15. Use the data in Table 5.8 to create a kernel density estimate for each group of subjects. Do the estimated densities appear to represent the same underlying distribution?

16. For the box kernel K, find the standard deviation of a random variable with density $h^{-1}K(x/h)$ for a fixed value $h > 0$.

17. Show that the asymptotic MISE is minimized when

$$h = \left(\frac{\int K^2}{n\sigma_K^4 \int (f'')^2} \right)^{1/5}.$$

18. If f is the density for a normal random variable with mean 0 and variance σ^2, show

$$\int (f'')^2 = 3/\left(8\sqrt{\pi}\sigma^5\right).$$

19. Using the normal kernel K with $\sigma_K = 1$ and assuming the underlying data come from a normal population with mean 0 and variance σ^2, show the optimal bandwidth is

$$h = (4/3)^{1/5} \sigma n^{-1/5}.$$

12.4 OTHER METHODS

Other methods of density estimation besides kernel methods are available, though less commonly used. Orthogonal series estimates of densities were proposed by Cencov (1962). These estimators represent the density function f in terms of a series expansion involving a set of specified basis functions and a set of scalar coefficients. Because f is a density, the scalars are estimated from the observed data by equating the coefficients with the expectation of the basis functions evaluated at the data. Common choices for the orthogonal bases are the Fourier series and, more recently, wavelet bases. See Kronmal and Tarter (1968) and Donoho et al. (1996). In both cases, the number of scalar coefficients is theoretically infinite, but a practical estimate will use only a finite number of them. Orthogonal series estimates then become a problem is selecting the appropriate subset of coefficients that best represent the distribution of the underlying observed data.

Another class of estimators are nearest neighbor estimates introduced in Loftsgaarden and Quesenberry (1965). In these estimators, a neighborhood around the point at which the density is to be estimated is created that holds a specified number of data values. In regions where the data is dense, such a neighborhood window will be small, while in sparse areas, it will by necessity be larger. This is similar to the variable bandwidth kernel estimates where the size of the window (neighborhood) varies with x. But where the variable bandwidth methods will have differing numbers of observation data in each window, nearest neighbor methods keep this number fixed. A drawback to this type of estimator is that it may not integrate to 1, violating **D2**. Additionally, these estimates have poor asymptotic behavior in terms of mean squared errors. See Terrell and Scott (1992) for a discussion of these limitations.

Chapter 13

Wavelets

INTRODUCTION

In Chapters 9 and 11, methods were presented that estimated specific types of function: linear functions in Chapter 9 and survival curves in Chapter 11. Each of these methods made use of assumptions that specified the nature of the shape of the function to be estimated. This shape restriction is not discussed in this chapter. Here, the unknown function is not assumed to have any particular parametric shape or representation but rather the function belongs to a class of functions possessing more general characteristics, such as a certain level of smoothness. Using the observed data, one may estimate such a function by representing the function in another domain. One common way to approach this is to use an orthogonal series representation of the function. This shifts the estimation problem from directly trying to estimate the unknown function f, to estimating a set of scalar coefficients that represent f in the orthogonal series domain. An efficient method for estimating such functions involves the use of wavelets. Wavelets are strong tool in such methods because they concentrate most of the information about the function in a much reduced set of data and have the ability to estimate both global and local features in the underlying function.

In Section 13.1, wavelet methods are introduced. Section 13.2 discusses one of the main tools in wavelet analysis, thresholding. Thresholding provides a significant level of data reduction for the problem.

Data. There are n pairs of observations $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$.

Assumptions

A1. The observations are related through the expression

$$y_i = f(x_i) + \varepsilon_i, i = 1, 2, \ldots, n.$$

A2. The ε_i are independent and identically distributed.

A3. The function f is square integrable, that is, $\int f^2 < \infty$. It is defined on a closed interval $[a, b]$. For simplicity, it assumed that $[a, b] = [0, 1]$.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

13.1 WAVELET REPRESENTATION OF A FUNCTION

Basis Functions

A set of functions $\Psi = \{\psi_1, \psi_2, \ldots\}$ is called a basis for a class of functions $\mathcal{F}$ if any function $f \in \mathcal{F}$ can be represented as a linear combination of the basis functions ψ_i. This is written as

$$f(x) = \sum_{i=1}^{\infty} \theta_i \psi_i(x), \tag{13.1}$$

where the θ_i are scalar constants, usually referred to as coefficients. The equality in (13.1) is understood in the sense of some particular measure; in this case, we take equality to mean

$$\int \left[f(x) - \sum_{i=1}^{\infty} \theta_i \psi_i(x) \right]^2 dx = 0. \tag{13.2}$$

The constants θ_i are the inner product of the function f and the basis functions ψ_i

$$\theta_i = \langle f, \psi_i \rangle = \int f(x)\psi_i(x)dx.$$

The basis functions are orthogonal if $\langle \psi_i, \psi_j \rangle = 0$ for $i \neq j$. They are orthonormal if they are orthogonal and $\langle \psi_i, \psi_i \rangle = 1$.

There are many sets of basis functions available to estimate functions in a variety of classes $\mathcal{F}$. In this chapter, we consider orthonormal wavelet bases. A simple wavelet function ψ first appeared in Haar (1910), but more flexible and powerful wavelets were developed by Daubechies (1992) and many others (see Vidakovic (1999)). The Haar wavelet and a "D2" Daubechies wavelet are depicted in Figure 13.1. If ψ is a wavelet function, then the collection of functions

$$\Psi = \{\psi_{jk} : j, k \text{ integers}\}, \tag{13.3}$$

where

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k), \tag{13.4}$$

forms a basis for square-integrable functions. Ψ is the collection of translations and dilations of ψ. Figure 13.1 shows some translations and dilations of the wavelet functions. The function ψ may be constructed to ensure that the set Ψ is orthonormal. Although the function ψ may be defined on the entire real line, the property that $\int \psi^2 = 1$ implies that the value of ψ is near 0 except over a small range. This, combined with (13.4), means that as j increases, ψ_{jk} becomes increasingly localized. Often, it is desirable that the function ψ have finite support, that is, ψ is nonzero on an interval of finite length. Constructive methods exist to ensure this property in addition to orthonormality.

Multiresolution Analysis

Careful construction of the wavelet function ψ leads to a multiresolution analysis (MRA). As described in Mallat (1989a), the MRA provides an interpretation of the wavelet representation of f in terms of location and scale. Rewriting (13.1) with the dual indexing

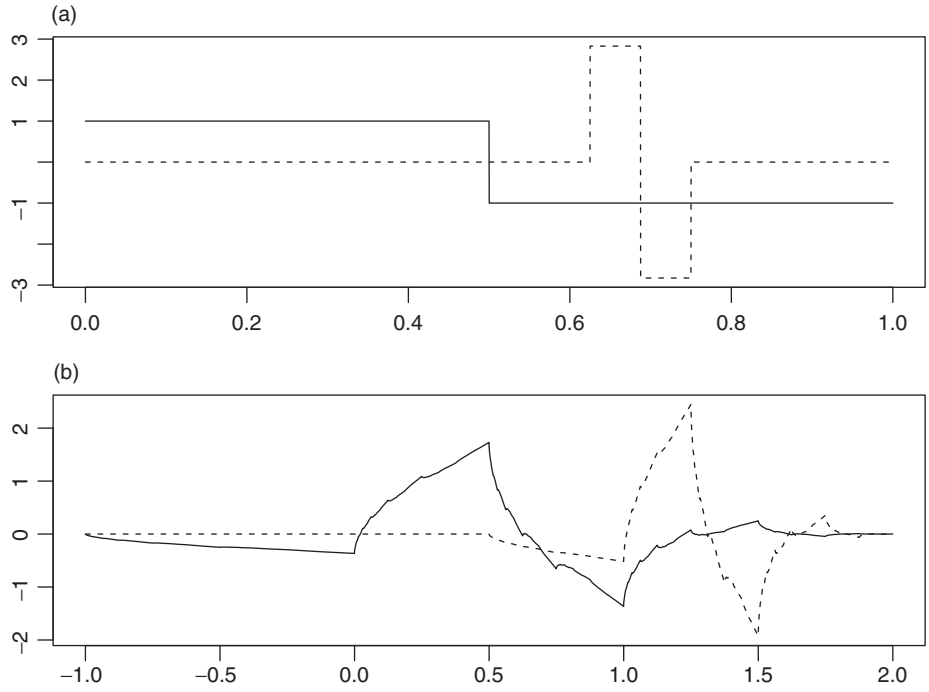

Figure 13.1 (a) Haar wavelet $\psi = \psi_{0,0}$ (solid line) and $\psi_{3,5}$ (dashed line). (b) D2 wavelet. $\psi = \psi_{0,0}$ (solid line) and $\psi_{1,1}$ (dashed line).

of wavelets provided by the translations and dilations of ψ gives

$$f(x) = \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} \theta_{jk} \psi_{jk}(x), \tag{13.5}$$

where $\mathbb{Z}$ is the set of integers. Under the conditions specified for an MRA, this may be interpreted as a series of approximations of f, where each approximation is at a different scale level j. In this context, scale may be thought of as synonymous with frequency. With wavelets, the term resolution is generally used when referring to scale or frequency. For a fixed j, the index k represents the behavior of f at resolution scale j and a particular location. One can then interpret the wavelet representation at (13.5) as giving information about the function f at differing resolution (scale, frequency) levels j and locations k, hence the term MRA. If using finitely supported wavelets, this interpretation is strengthened because the locality indexed by k becomes more pronounced.

Consider a cumulative approximation of f using all the values of j less than some integer J,

$$f_J(x) = \sum_{j < J} \sum_{k \in \mathbb{Z}} \theta_{jk} \psi_{jk}(x). \tag{13.6}$$

As J increases, f_J is able to model smaller scale (higher frequency) behavior of f. These correspond to changes in f that occur over a small interval of the x-axis. As J decreases, f_J models larger scale (lower frequency) behavior of f. The complete representation of f is the limit of the f_J. The limit of these approximations f_J get closer, in the sense of (13.2), to the function f as J increases. See Comments 2 and 3.

It is common to truncate the MRA. One may write (13.5) as

$$f(x) = \sum_{k\in\mathbb{Z}} \xi_{j_0 k}\phi_{j_0 k}(x) + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{jk}(x). \tag{13.7}$$

The first term on the right side of (13.7) is the cumulative approximation f_{j_0} using all resolution levels $j < j_0$. It makes use of a function ϕ related to the wavelet ψ (see Comment 3). The second term on the right side of (13.7) is a set of series, one for each resolution level $j \geq j_0$. Each of these series, when added to f_{j_0}, allows for modeling higher scale-frequency behavior of f and brings the approximation closer to f. When a function f is written as at (13.7), f_{j_0} is the approximation at the "smooth" or "coarse" resolution level and each of the remaining resolution level series is a "detail" level. The MRA then gives a smooth, cumulative approximation of f and several detail levels to increase the modeling accuracy.

For $J \geq j_0$, the cumulative approximation f_J at (13.6) becomes

$$f_J(x) = \sum_{k\in\mathbb{Z}} \xi_{j_0 k}\phi_{j_0 k}(x) + \sum_{j_0\leq j<J}\sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{jk}(x), \tag{13.8}$$

See Comment 4. The functions ϕ and ψ are sometimes referred to as the scaling function and the wavelet function, respectively. They are also called the father wavelet and mother wavelet.

EXAMPLE 13.1 ***MRA Using the Haar Wavelet.***

Using the wavelet representation of a function at (13.7) the MRA for the function

$$f(x) = x, \quad x \in [0, 1)$$

can be found analytically when ψ is the Haar wavelet. The wavelet functions ψ and ϕ are given by

$$\psi(x) = \begin{cases} 1, & x \in [0, 1/2), \\ -1, & x \in [1/2, 1), \end{cases}$$

and

$$\phi(x) = 1, \ x \in [0, 1).$$

Setting J to 0 in (13.6), or, equivalently, $J = j_0 = 0$ in (13.8) results in $f_0(x) = 1/2$. This approximation of f is shown as a solid line in Figure 13.2a. The function f is displayed as a dotted line. Figure 13.2b shows f_1, the approximation of f up to resolution level $j = 0$. Figure 13.3a shows $f_1 - f_0$, the change in the estimate of f obtained by adding in detail resolution level 1.

Figure 13.2c and d shows the cumulative approximations for resolution levels f_2 and f_3. The nature of the Haar wavelet is easily seen in these approximations. As j increases, it is clear that the stair-like approximations will get arbitrarily close to the function f. Details on evaluating these approximations are given in Comment 9 and the problems.

In Figure 13.3, note how increasing the resolution level j allows for increased modeling of high frequency behavior. The larger values of j enable modeling of changes occurring over smaller intervals of the x-axis. Each panel in Figure 13.3 displays the

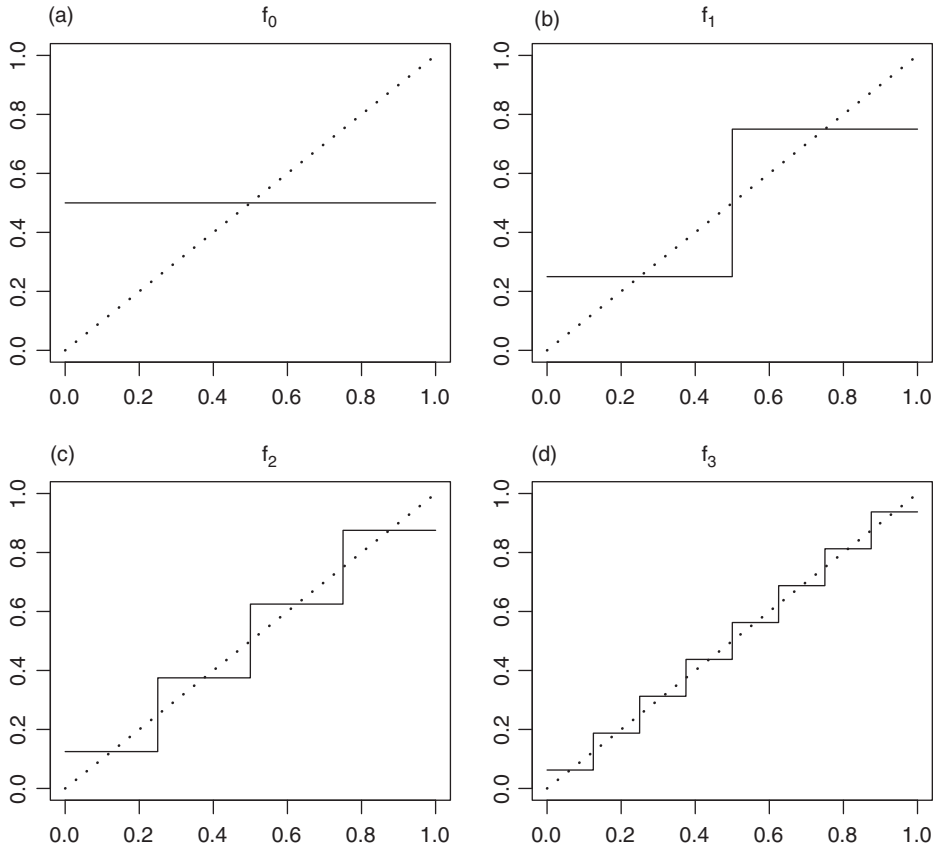

Figure 13.2 Cumulative approximations up to resolution levels $J = -1, 0, 1, 2$ from Example 13.1 using the Haar wavelet. The underlying function is $f(x) = x$, shown with a dotted line.

approximation of f at a single detail resolution level, while in Figure 13.2 each panel displays the cumulative approximation of f up to a particular level.

Figure 13.4a displays the cumulative approximations for the same function f but using the D2 wavelet shown in Figure 13.1. Note that the approximations display some issues at the boundary value. This is due to using periodic wavelets. See Comment 5. The original wavelet is of length 3. When periodized, certain translation indices k will cause the wavelet functions to "wraparound" the end points. The Haar wavelet does not display this property because ψ and all its translations and dilations fit within $[0, 1]$, that is, the Haar wavelet is already periodic with respect to $[0, 1]$. To avoid this, one may specify using reflection at the boundaries, rather than periodicity. However, this will increase the number of indices k that must be considered at each resolution level j. The cumulative approximations when using reflection at the boundaries is shown in Figure 13.4b.

Discrete Wavelet Transform

In Example 13.1, the simple form of the Haar wavelet allows exact determination of the wavelet coefficients θ_{jk}. Other wavelets do not have analytic forms that can be integrated. In these cases, a numeric algorithm must be used to estimate the coefficients.

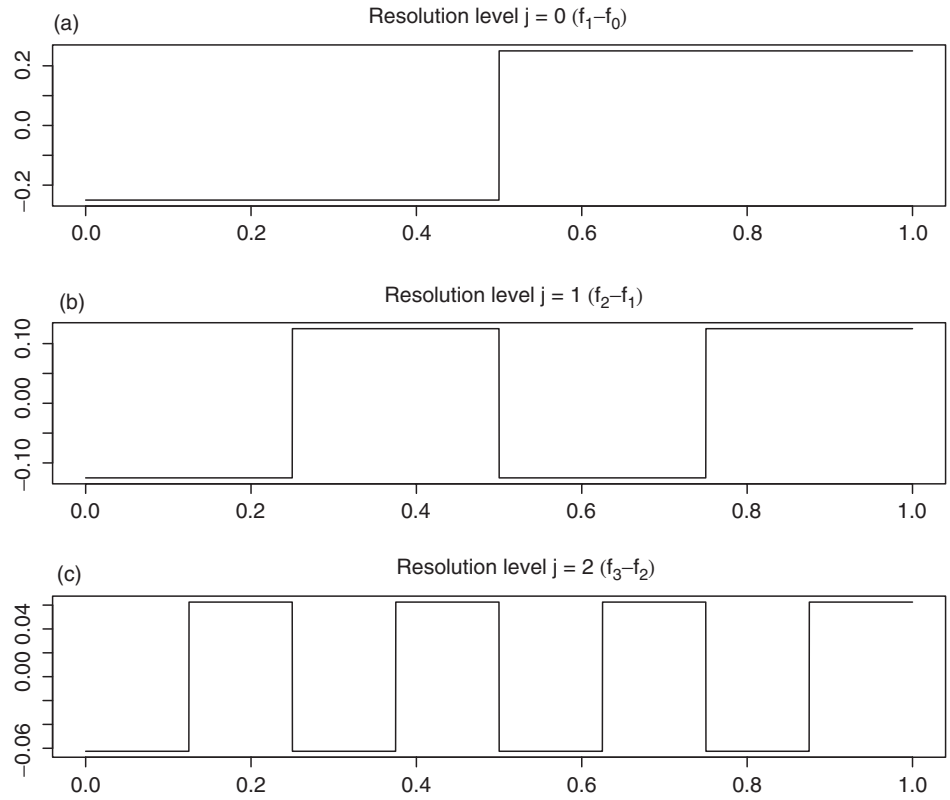

Figure 13.3 Resolution levels $j = 0, 1, 2$ from Example 13.1 using the Haar wavelet. The underlying function is $f(x) = x$, shown with a dotted line. Note the differing vertical scales.

One such method is the cascade algorithm of Mallat (1989b). If the observed data is the vector $y = (y_1, y_2, \ldots, y_n)$ sampled from the function f, the cascade algorithm recursively estimates the wavelet coefficients from the data. The sample size needs to be a power of 2 for this algorithm, $n = 2^J$ for some positive integer J. Given a set of wavelet coefficients, the cascade algorithm works in reverse, as well. It can convert the coefficients back into the sample data. Creating wavelet coefficients from data is called *decomposition*, while the reverse is known as *reconstruction*. See Comment 10.

Use of the data to estimate the wavelet coefficients restricts the upper level of summation in (13.7) to $J - 1$, where $J = \log_2(n)$. Thus, the number of resolution levels in the wavelet series is truncated both above and below in practice, resulting in $J - j_0 + 1$ series, each representing a resolution level:

$$f(x) = \sum_{k \in \mathbb{Z}} \xi_{j_0 k} \phi_{j_0 k}(x) + \sum_{j=j_0}^{J-1} \sum_{k \in \mathbb{Z}} \theta_{jk} \psi_{jk}(x). \tag{13.9}$$

In addition, the cascade algorithm cannot determine wavelet coefficients below $j = 0$, so a lower bound for j_0 is 0.

Implementation of the cascade algorithm is efficiently carried out with a transformation matrix whose entries are determined by the choice of wavelet basis and the sample size n. This transformation is referred to as the discrete wavelet transform (DWT). Three

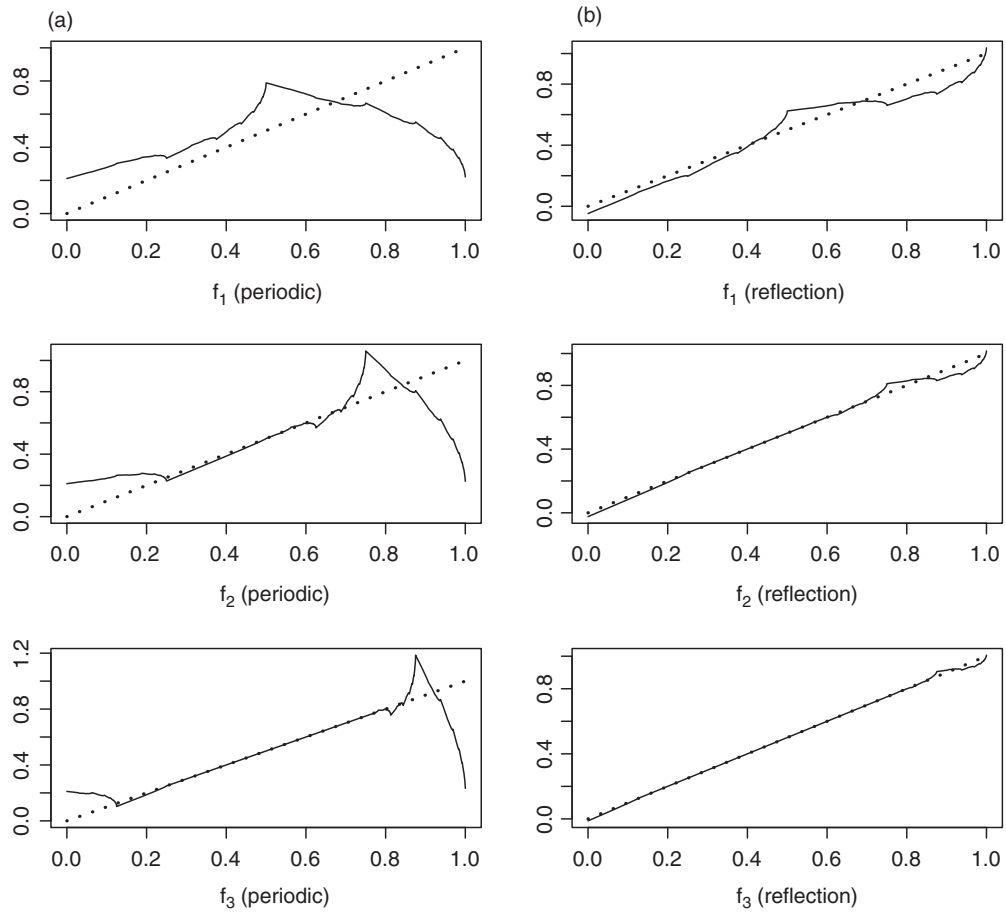

Figure 13.4 Approximations up to resolution levels $j = 0, 1, 2$ from Example 13.1 using the D2 wavelet. The left panels use periodic boundary handling, the right panel use reflection. The underlying function is $f(x) = x$, shown with a dotted line.

commands in R that make use of the DWT are `dwt`, `idwt`, and `mra`. All these functions are in the package `waveslim` (Whitcher (2010)).

EXAMPLE 13.2 ***Implementing the MRA.***

The command `mra` uses the DWT to provide the approximations of the sampled data y at various resolution levels. This command requires arguments for the sampled data `y`, `method="dwt"`, the wavelet basis `wf`, and the number of resolution levels to determine, `J`. If the sample size is $n = 2^J$, then possible values for `J` are from 1 to J.

Suppose the function f from Example 13.1 is sampled at $n = 2^{12}$ equally spaced points on the interval $[0, 1)$, say,

$$y_i = x_i = (i - 1)/n, i = 1, 2, \ldots, n.$$

To decompose the sampled data y into the maximum possible resolution levels with the Haar basis, the following is used:

```
mra(y, method="dwt", wf="haar", J=12)
```

Periodic wavelets are the default setting. Otherwise, the argument `boundary="reflection"` would be added to the above command. This command provides a list of 13 vectors. The first vector is the change necessary to go from the approximation f_{12} to f_{13}: $f_{13} - f_{12}$. This is the approximation at the highest detail resolution level. The next vector in the list is $f_{12} - f_{11}$. The next to last vector is $f_1 - f_0$. The final, thirteenth vector is the smooth approximation f_0. Summing the thirteenth vector and the twelfth vector results in f_1, summing the thirteenth, twelfth, and eleventh vectors results in f_2 and so on.

Setting the argument `J` to be a smaller integer than 12 in this example will decompose the sampled data into fewer levels. This is equivalent to setting j_0 in (13.7) to a value greater than 0. For example, if `J` is set to 3, then four vectors will be created. The first three of these vectors are the detail approximations $f_{12} - f_{11}$ through $f_{10} - f_9$ and the final vector is the smooth approximation $f_{j_0} = f_9$.

The solid line in Figure 13.2b is a result of the command

```
mra(y, method="dwt", wf="haar", J=12)[[13]]
```

The argument `J=12` will decompose the observed data `y` into 13 approximations of f at consecutive resolution levels. Specifying `[[13]]` gives the lowest level approximation, the smooth approximation f_0. The other 12 approximations correspond to the detail resolution levels. The plots in Figure 13.3 are of the vectors

```
mra(y, method="dwt", wf="haar", J=12)[[12]]
mra(y, method="dwt", wf="haar", J=12)[[11]]
mra(y, method="dwt", wf="haar", J=12)[[10]]
```

In the above three commands, the data is again decomposed into the maximum number of resolution levels by setting `J=12`. The three lowest detail resolution levels are provided by specifying `[[12]]` ($f_1 - f_0$), `[[11]]` ($f_2 - f_1$) and `[[10]]` ($f_3 - f_2$). The smooth approximations f_1,f_2 and f_3 in Figure 13.2 are plots of

```
mra(y, method="dwt", wf="haar", J=11)[[12]]
mra(y, method="dwt", wf="haar", J=10)[[11]]
mra(y, method="dwt", wf="haar", J=9)[[10]]
```

In the above three `R` commands, note that the data is being decomposed into fewer than the maximum number of levels. The final vector in the produced list is the smooth approximation. For example, the first command produces 12 vectors: 11 detail approximations $f_{12} - f_{11}$ through $f_2 - f_1$ and the smooth approximation f_1. The plots in Figure 13.4a are

```
mra(y, method="dwt", wf="d4", J=11)[[12]]
mra(y, method="dwt", wf="d4", J=10)[[11]]
mra(y, method="dwt", wf="d4", J=9)[[10]]
```

while those in Figure 13.4b add the argument `boundary="reflection"` to the previous three commands. `R` labels the Daubechies wavelets with twice the number of vanishing moments. See Comment 7. Thus, `d4` refers to the wavelet D2.

Instead of using the double bracketing to index the vectors produced by `mra`, the output is also indexed by names. If a vector is decomposed into, say, 6 detail resolution levels and a single smooth approximation, the vectors provided by `mra` are named `D1`, `D2`, ..., `D6` and `S6` with `D1` being the highest detail resolution level. So, for example, the following two commands are equivalent:

```
mra(y, method="dwt", wf="haar", J=6)[[1]]
mra(y, method="dwt", wf="haar", J=6)$D1
```

The `mra` command provides approximations of f at differing resolution levels. In contrast, the `dwt` command determines the wavelet coefficients at each resolution level. This command has arguments for the sampled data `y`, the wavelet basis `wf` and the number of resolution levels to determine, `n.levels`. As with `mra`, an optional `boundary` argument can be added. Use of `dwt` results in a list of `n.levels` + 1 vectors. The first vector in the list is the highest resolution level of detail coefficients, the lower detail level coefficients follow, and the final vector consists of the wavelet coefficients corresponding to the smooth approximation. There are $n/2 = 2^{J-1}$ coefficients in the first vector, $n/4 = 2^{J-2}$ in the second, and so on. The final vector of smooth coefficients has the same length as the lowest detail level of coefficients, $2^{J-\texttt{n.levels}}$. These decreasing vector sizes are a result of the increasing support of the wavelet basis functions as the resolution level j decreases. The total number of coefficients in all the vectors provided by `dwt` is the same as the length of the vector of sampled data, n.

If `n.levels` is set to its largest possible value, `n.levels` $= \log_2(n)$, the maximum number of decomposition levels is generated. The sampled data `y` is assumed to be equally spaced and dyadic in number. The list of vectors generated from `dwt` is indexed in the same way as the `mra` output. One may either use the double bracket indexing or the names of the vectors. Unlike `mra`, though, `dwt` uses lower case letters in the names, that is, `d1` instead of `D1`.

Once a sampled function has been decomposed via `dwt`, the resulting `R` list of coefficients may be used to reconstruct the original vector of sampled data y. The command for this is `idwt`. If `y` is the sampled data as above,

```
y.dwt <- dwt(y, wf="haar", n.levels=12)
```

generates the list of wavelet coefficients and assigns it to the `R` object `y.dwt`. The original data `y` is reconstructed with

```
idwt(y.dwt)
```

The `idwt` command only needs the `R` object containing the list of coefficient vectors. The other necessary information for reconstruction, number of levels and wavelet basis, is contained in the object `y.dwt`.

Comments

1. *Wavelet Function*. This chapter describes wavelets in terms of generating a basis for square-integrable functions. A more basic definition of a wavelet function ψ can be given in terms of its Fourier transform:

$$\hat{\psi}(\omega) = \frac{1}{\sqrt{2\pi}} \int e^{-i\omega t} \psi(t) dt.$$

A function ψ is a wavelet if it satisfies the admissibility condition (Daubechies (1992)).

$$\int \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty.$$

If working in the Fourier domain is not desired, then ψ is a wavelet, provided

$$\int \psi(t)dt = 0$$

and

$$\int (1 + |t|^{\alpha})|\psi(t)|dt < \infty$$

for some positive α.

2. *Interpreting Resolution Levels via the MRA*. The MRA relates the different resolution levels to approximations of f through (13.6),

$$f_j(x) - f_{j-1}(x) = \sum_{k\in\mathbb{Z}} \theta_{j-1,k}\psi_{j-1,k}(x). \tag{13.10}$$

Thus, the series at a single index j is the difference between two approximations at adjacent resolution levels.

In (13.6), increasing J allows approximation of functions possessing higher scale behavior. This follows from above by observing that $f_{J+1} - f_J = \sum_{k\in\mathbb{Z}} \theta_{Jk}\psi_{Jk}$ adds to the approximation f_J using basis functions with half the support of those used in f_J. Equivalently, decreasing J will restrict the approximations to lower scales.

One can extend (13.10) to a range of resolution levels rather than just a single level. For any integers $j_1 < j_2$,

$$f_{j_2}(x) - f_{j_1}(x) = \sum_{j_1 \le j < j_2} \sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{j,k}(x)$$

represents the approximation of f over several scales.

3. *MRA*. The MRA is a sequence of nested subspaces

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots$$

such that the intersection of all these subspaces is $\{0\}$, the closure of their union is the space of square-integrable functions, and

$$f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}.$$

It is additionally assumed that there is a function $\phi \in V_0$ such that every function $f_0 \in V_0$ can be written as

$$f_0(x) = \sum_k \xi_{0k}\phi_{0k}(x),$$

where ξ_{jk} and ϕ_{jk} are defined as at (13.13). Using this definition, due to Mallat (1989a), it can be seen that as j increases, the functions residing in the spaces V_j are allowed to display higher frequency behavior. Functions in lower spaces V_j are smoother than those in higher spaces. The functions in V_{j+1} that are not in V_j are the detail functions needed to go from a lower space to a higher space.

Writing W_j as this set, then V_{j+1} is the sum of functions in W_j and functions in V_j: $V_{j+1} = V_j \oplus W_j$. Continuing, this results in

$$V_J = V_{j_0} \oplus W_{j_0} \oplus W_{j_0+1} \oplus \cdots \oplus W_{J-1}.$$

Wavelet bases represent functions in V_j as series involving ξ_{jk} and ϕ_{jk}, while functions in W_j may be written as a series involving θ_{jk} and ψ_{jk}. With the orthonormal wavelet basis used in this chapter, this implies that functions in W_j are orthogonal to functions in W_k for any $k \neq j$ and V_k for any $k \leq j$. The decomposition provided by the cascade algorithm at (13.16) is represented by the ladder of function spaces above.

4. *Truncating the MRA from Below*. One may write (13.5) as

$$f(x) = \sum_{j<j_0} \sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{jk}(x) + \sum_{j\geq j_0} \sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{jk}(x) \tag{13.11}$$

for some fixed integer j_0. The properties of the MRA allow the first set of series on the right side of (13.11) to be written as a single series

$$f_{j_0}(x) = \sum_{j<j_0} \sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{jk}(x) = \sum_{k\in\mathbb{Z}} \xi_{j_0k}\phi_{j_0k}(x), \tag{13.12}$$

where

$$\phi_{j_0k}(x) = 2^{j_0/2}\phi(2^{j_0}x - k) \text{ and } \xi_{j_0k} = \langle f, \phi_{j_0k}\rangle \tag{13.13}$$

for some function ϕ dependent on ψ. See Comment 3. This approximation f_{j_0} is considered "smooth" or "coarse" with respect to an approximation that would include higher resolution levels $j \geq j_0$. The expression at (13.5) is therefore

$$f(x) = \sum_{k\in\mathbb{Z}} \xi_{j_0k}\phi_{j_0k}(x) + \sum_{j\geq j_0} \sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{jk}(x),$$

where the first series on the right represents the smooth approximation f_{j_0} of f and the second set of series allows for modeling higher scale-frequency behavior of f. The set of basis functions becomes

$$\Psi = \{\psi_{jk} : j \geq j_0, k \in \mathbb{Z}\} \cup \{\phi_{j_0k} : k \in \mathbb{Z}\}.$$

The single series involving ξ and ϕ is a smooth approximation of f up to a particular resolution level j_0. The series involving θ and ψ represent "details," which when added to f_{j_0} create a less smooth, more detailed approximation of f at higher resolution levels.

The smooth approximation f_J may also be written in terms of the wavelet functions ψ and ϕ. From (13.11) and (13.12)

$$\begin{aligned} f_J(x) &= \sum_{j<j_0} \sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{jk}(x) + \sum_{j_0\leq j<J} \sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{jk}(x) \\ &= \sum_{k\in\mathbb{Z}} \xi_{j_0k}\phi_{j_0k}(x) + \sum_{j_0\leq j<J} \sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{jk}(x) \end{aligned}$$

for $J \geq j_0$. If $J = j_0$, just the first series using ϕ is used.

5. *Periodic Wavelets*. If a function is defined on the interval $[0, 1]$, it may be preferable to estimate it with basis functions restricted to the same interval. One way to implement this is by periodizing the wavelet functions. If ϕ and ψ are father and mother wavelets (see Comment 4), then their periodized versions are given by Vidakovic (1999).

$$\phi_{jk}^{p}(x) = \sum_{l=-\infty}^{\infty} \phi_{jk}(x - l)$$

and

$$\psi_{jk}^{p}(x) = \sum_{l=-\infty}^{\infty} \psi_{jk}(x - l).$$

If the wavelets are finitely supported, then for each j there are only a finite number of indices k that result in nonzero coefficients θ_{jk} for a function f supported on $[0, 1]$. If the wavelets are periodized to the interval $[0, 1]$, then the number of coefficients at each nonnegative level resolution level j is 2^j. This is a result of the dilation in (13.4).

6. *Using Sampled Data as Coefficients*. The cascade algorithm begins the decomposition process by equating the sample vector of data $y = (y_1, y_2, \ldots, y_n)$ with the coefficients ξ_{Jk}. See Comment 3. If the function f is sampled at the equally spaced points $k/n, k = 1, 2, \ldots, n$ on the interval $[0, 1]$, then ξ_{Jk} are taken to be $f(k/n)$. However, $\xi_{Jk} \approx \sqrt{n} f(k/n)$ for smooth f (Daubechies (1992)). So, the coefficients estimated with these initial values by the cascade algorithm are too large by the factor $\sqrt{n}$. The `dwt` command does not use the factor $\sqrt{n}$. This is not an issue because the reconstruction command `idwt` takes this into account. `dwt` also reverses the sign of the coefficients. Again, this is accounted for in both the decomposition and the reconstruction and does not pose any problems in analyzing the functions.

7. *Daubechies Wavelets*. Daubechies (1992) generated families of compactly supported orthonormal wavelet bases that are commonly used in many applications. Two such categories of wavelets, determined by the choice of roots of a polynomial used in the wavelet construction, are extremal phase wavelets and least asymmetric wavelets. The latter are often referred to as *symmlets*. Each of these categories is further characterized by the number of vanishing moments. A wavelet function ψ has N vanishing moments if

$$\int x^l \psi(x) dx = 0$$

for $l = 0, 1, \ldots, N - 1$. A wavelet basis in which ψ has N vanishing moments can model a polynomial up to degree $N - 1$ with only the smooth approximation based on ϕ and the associated ξ. See Comment 4.

R implements both types of wavelets in `dwt` and `mra`. This is done via the argument `wf`. The extremal phase wavelets are indicated by `wf="dN"`, where `N=4,6,8,16`, is twice the number of vanishing moments. The support length for these wavelets is $2N - 1$. The least asymmetric wavelets are indexed in R by `wf="laN"`, where `N=8,16,20` is again twice the number of vanishing moments.

8. *Nondyadic Length Sampled Functions*. If the length n of a sampled function is not dyadic, the DWT may not be directly implemented as described above. However, the observed vector can be lengthened or shortened to give a length that is a dyadic integer. Shortening the vector is not desirable because this ignores data that may be important. To lengthen the data, there are two common methods. The simplest is to add data values to the end of the observed vector to bring the total length to the closest dyadic number. The DWT is applied to this extended data vector. In R, this is implemented with `dwt.nondyadic`. This will pad the observed vector in the right with a string of zeros.

   A second method to bring the vector size up to a dyadic length is through the use of reflection. This is not the same as using reflection to handle boundary issues. Instead, an observed vector is reflected about its end points. For example, if the length n of the observed data y is k less than the first dyadic integer after n, then the DWT is performed on

$$y^* = (y_1, y_2, \ldots, y_{n-1}, y_n, y_{n-1}, y_{n-2}, \ldots, y_{n-k}).$$

   A modification of this is obtained using double reflection. First, the observed vector is reflected about the end point as above. This is a horizontal reflection. Then, the added data is reflected vertically about this same end point. Using this method, no discontinuities are introduced into the extended vector at the original vector's end point.

9. *Determining the Approximations with the Haar Basis*. To find the approximations f_j for Example 13.1, Problem 1 gives

$$\psi_{jk}(x) = \begin{cases} 2^{j/2}, & x \in [k2^{-j}, (k+1/2)2^{-j}), \\ -2^{j/2}, & x \in [(k+1/2)2^{-j}, (k+1)2^{-j}). \end{cases} \tag{13.14}$$

   To evaluate the inner product of f and ψ, the supports of these functions must be carefully considered. The support of f is $[0, 1]$, the support of ψ_{jk} is seen from above to be $[k2^{-j}, (k+1)2^{-j})$. For $j < 0$, the only translation k that has a nonzero length overlap with $[0, 1]$ is $k = 0$. For $k = 0$ and $j < 0$, the value of ψ_{jk} for any x in the interval $[0, 1)$ is $2^{j/2}$. The inner product for $j < 0$ and $k = 0$ is then

$$\theta_{j0} = \langle f, \psi_{j0} \rangle = \int f(x)\psi_{j0}(x)dx = \int_0^1 x2^{j/2}dx = 2^{j/2-1}.$$

   Only considering negative values of j in (13.5) results in

$$f_0(x) = \sum_{j=-\infty}^{-1} \sum_{k \in \mathbb{Z}} \theta_{jk}\psi_{jk}(x) = \sum_{j=-\infty}^{-1} \theta_{j0}\psi_{j0}(x) = \sum_{j=-\infty}^{-1} 2^{j/2-1}2^{j/2}$$
$$= \frac{1}{2} \sum_{j=-\infty}^{-1} 2^j = \frac{1}{2}.$$

   The smooth approximation using all negative resolution levels is just a constant, $f_0 = 1/2$.

Although f_0 was evaluated above as a series involving ψ, it could also have been found using the father wavelet ϕ. See Problem 4.

The support of the Haar wavelet results in 2^j nonzero wavelet coefficients θ_{jk} for each nonnegative resolution level j. The only translations k that have nonzero length overlap with the support of the function f are $k = 0, 1, \ldots, 2^j - 1$. From Problem 5, the inner product when $j \geq 0$ is

$$\theta_{jk} = \langle f, \psi_{jk} \rangle = -2^{(-3j/2-2)}.$$

The function f is written with (13.5) as

$$\begin{aligned} f(x) &= \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} \theta_{jk} \psi_{jk}(x) \\ &= f_0(x) + \sum_{j \geq 0} \sum_{k=0}^{2^j - 1} \theta_{jk} \psi_{jk}(x) \\ &= 1/2 + \sum_{j \geq 0} \sum_{k=0}^{2^j - 1} -2^{(-3j/2-2)} \psi_{jk}(x). \end{aligned} \tag{13.15}$$

When $j = 0$, the only value of k is 0. From Problem 6, the series at this detail resolution level is

$$\theta_{0,0}\psi_{0,0}(x) = \begin{cases} -1/4, & x \in [0, 1/2), \\ 1/4, & x \in [1/2, 1). \end{cases}$$

When this is added to the smooth approximation f_0, it lowers the approximation of f by $1/4$ on the interval $[0, 1/2)$ and raises the approximation of f by $1/4$ on $[1/2, 1)$. Figure 13.2b shows f_1, the smooth approximation of f through resolution level $j = 0$. Figure 13.3a shows $f_1 - f_0$, the change in the estimate of f obtained by adding in detail resolution level 1.

When $j = 1$, two values of k are relevant, $k = 1$ and $k = 2$, and $\theta_{1k} = -2^{(-3 \cdot 1/2-2)} = -2^{-7/2}$. From Problem 7, the series at this detail resolution level is

$$\sum_{k=0}^{1} \theta_{1,k} \psi_{1,k}(x) = \begin{cases} -1/8, & x \in [0, 1/4) \text{ or } x \in [1/2, 3/4), \\ 1/8, & x \in [1/4, 1/2) \text{ or } x \in [3/4, 1). \end{cases}$$

This detail resolution level series is shown in Figure 13.3b, while Figure 13.3c shows the detail resolution level series at $j = 2$.

10. *Estimating Coefficients with the Cascade Algorithm.* Assume the data is an equally spaced vector $y = (y_1, y_2, \ldots, y_n)$ sampled from the function f. The cascade algorithm of Mallat (1989b) recursively estimates the θ_{jk} and ξ_{jk}. Using periodic wavelets and assuming that n is a dyadic number, that is, $n = 2^J$ for some integer J, the sampled data y is treated as though it is the set of n coefficients ξ_{Jk} needed for the smooth wavelet approximation f_J. The cascade algorithm then uses these coefficients to estimate the coefficients at the next lower resolution level. It will create the $n/2$ $\theta_{J-1,k}$ coefficients and $n/2$ $\xi_{J-1,k}$ coefficients for a particular choice of wavelet basis functions ψ.

After one step of the algorithm, f is written as a detail series at resolution level $J-1$ and the smooth approximation f_{J-1}. Using (13.6) and (13.7) with $j_0 = J-1$, this is

$$f_J(x) = \sum_{k\in\mathbb{Z}} \xi_{J-1,k}\phi_{J-1,k}(x) + \sum_{j=J-1}\sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{jk}(x).$$

Note the use of f_J rather than f. This is due to using the sample y, rather than f itself. The data y is assumed to be at resolution level J, so it is not possible to find a higher resolution level approximation than this.

The next step of the algorithm operates in the same manner, but on the coefficients $\xi_{J-1,k}$ instead of ξ_{Jk}. At each step, the number of coefficients θ or ξ estimated at a resolution level j is half the number of the previously estimated coefficients at resolution level $j+1$. After J steps, only one of each coefficient may be estimated: $\theta_{0,0}$ and $\xi_{0,0}$. The algorithm may not proceed any further. This process of recursively finding lower resolution level coefficients is called decomposition.

It is not necessary to decompose the sampled function data y to the lowest possible level. A decomposition may stop at any fixed level j_0, where $0 \le j_0 \le J-1$. In this case, the cascade algorithm provides coefficients for the following decomposition:

$$f_J(x) = \sum_{k\in\mathbb{Z}} \xi_{j_0k}\phi_{j_0k}(x) + \sum_{j=j_0}^{J-1}\sum_{k\in\mathbb{Z}} \theta_{jk}\psi_{jk}(x). \tag{13.16}$$

The cascade algorithm works in reverse, as well. Given sets of coefficients ξ_{j_0k} and $\theta_{jk}, j_0 \le j < J$, the algorithm will combine ξ_{j_0k} and θ_{j_0k} into $\xi_{j_0+1,k}$, then combine $\xi_{j_0+1,k}$ and $\theta_{j_0+1,k}$ into $\xi_{j_0+2,k}$, and so on. It will stop at ξ_{Jk}, which is the original data y. Using the algorithm in this fashion is called *reconstruction*.

Implementation of the cascade algorithm is efficiently carried out with a transformation matrix whose entries are determined by the choice of wavelet basis and the sample size n.

Properties

1. *Approximation Accuracy of the MRA*. See Mallat (1989a, 1989b, 2009).
2. *Characteristics of Wavelet Bases*. See Daubechies (1992).

Problems

1. The Haar wavelet $\psi = \psi_{0,0}$ is given by

$$\psi(x) = \begin{cases} 1, & x \in [0, 1/2), \\ -1, & x \in [1/2, 1). \end{cases}$$

Show that

$$\psi_{jk}(x) = \begin{cases} 2^{j/2}, & x \in [k2^{-j}, (k+1/2)2^{-j}), \\ -2^{j/2}, & x \in [(k+1/2)2^{-j}, (k+1)2^{-j}). \end{cases}$$

2. Using the Haar wavelet, verify that $\langle \psi_{ij}, \psi_{kl} \rangle = 1$ when $i = k$ and $j = l$, and 0 otherwise.

3. If the basis at (13.3) is truncated below by $j = 0$, then the associated function ϕ when ψ is the Haar wavelet is given by

$$\phi(x) = \phi_{0,0}(x) = 1, \quad x \in [0, 1).$$

Show that

(a) $\langle \phi_{0j}, \psi_{kl} \rangle = 0$ for any $k \geq 0$ and j and l .
(b) $\langle \phi_{0j}, \phi_{0k} \rangle = 1$ when $j = k$ and 0 otherwise.

4. In Example 13.1, the mother wavelet ψ was used to find the smooth approximation f_0. Show $f_0 = 1/2$ using the father wavelet ϕ for the Haar basis and $f(x) = x, x \in [0, 1]$.

5. Let $f(x) = x, x \in [0, 1]$, and ψ be the Haar wavelet.

(a) Show that for $j \geq 0$ and $k = 0, 1, \ldots, 2^j - 1$, $\theta_{jk} = -2^{(-3j/2-2)}$.
(b) Use (13.14) and (13.15) to show the series at detail resolution level $j = 0$ is

$$\theta_{0,0}\psi_{0,0}(x) = \begin{cases} -1/4, & x \in [0, 1/2), \\ 1/4, & x \in [1/2, 1). \end{cases}$$

(c) Show the series at detail resolution level $j = 1$ is

$$\sum_{k=0}^{1} \theta_{1,k}\psi_{1,k}(x) = \begin{cases} -1/8, & x \in [0, 1/4) \text{ or } x \in [1/2, 3/4), \\ 1/8, & x \in [1/4, 1/2) \text{ or } x \in [3/4, 1). \end{cases}$$

6. Let y be the first 512 components in the `sunspots` data from package `datasets`.

(a) Use the `mra` command to plot f_4, f_5 and f_6 using the Haar wavelet. Describe the different characteristics of each of these three smooth approximations. For example, use `J=5` and either `[[6]]` or `$S5` to find f_4.
(b) Repeat the above problem using the wavelet basis indexed by `wf="la8"`. This wavelet is from the family of "least asymmetric" wavelets (Vidakovic (1999)). Describe the differences between the smooth approximations using the different wavelet bases.

13.2 WAVELET THRESHOLDING

Section 13.1 explained how a function f may be represented with a wavelet basis. Using the DWT, a sample of length n from f may be decomposed into n wavelet coefficients making up a single smooth approximation and up to $J = \log_2(n)$ detail resolution levels. The inverse DWT can be applied to these resulting wavelet coefficients to reconstruct the original sample data.

In this section, the sparsity property of wavelets is described and thresholding, the statistical method based on sparsity, is given. Sparsity refers to the ability of wavelets to represent a function by concentrating or compressing the information about f into a few large magnitude coefficients and many small magnitude coefficients. In general, as f gets smoother, the amount of compression attained becomes greater. Compression is applied to the wavelet coefficients of a sampled function f prior to its reconstruction.

A concept related to compression is thresholding. Sparsity tells us that only a few coefficients are needed to represent a function reasonably well. Thresholding tells us how to find those few coefficients. Thresholding is an essential step in nearly all statistical function estimation procedures involving wavelets. Thresholding will reduce the dimension of the sample vector based on some theoretical concerns, rather than just compressing the data to a specified amount. The amount of reduction is often significant,

sometimes just a few coefficients out of the original n are retained in the thresholding of the wavelet series. The most common use of wavelet thresholding is to obtain a reconstruction of an unknown function f when the Assumptions A1–A3 hold and only the noisy data y is observed.

Sparsity

The MRA represents a function f as a smooth approximation and several series of increasingly detailed resolution levels. The basis functions at the lower resolution levels have wider support than those at higher levels, so fewer of these functions and their associated coefficients are needed at lower levels when compared to higher levels. This is very clear when using the DWT with a sampled vector of data from f. There are $n/2$ coefficients at the highest level, $n/4$ at the next, and so on.

If f is smooth, the lower resolution levels will provide a reasonable approximation to f without the need to include higher resolution levels. These higher levels are modeling high frequency behavior, therefore a smooth function will not have need of them. The majority of the wavelet coefficients reside in these higher levels, so being able to ignore them means a small subset of the coefficients is sufficient to model f. This is the sparseness property.

EXAMPLE 13.3 ***Sparsity of the Wavelet Representation.***

Andrews and Herzberg (1985) provide data on mean monthly sunspot observations collected at the Swiss Federal Observatory in Zurich and the Tokyo Astronomical Observatory from 1749 to 1983. The data displays excessive variability over time, obscuring any underlying trend in the cycle of sunspot appearances. The top panel in Figure 13.5 shows monthly sunspot data from January 1749 through July 1919. This data is named `sunspots` and is found in the package `datasets` in R. The data `sunspots` has length 2820, but only the first 2048 are used here because that is a dyadic number. The DWT is applied to this data resulting in 2048 coefficients. This decomposition was accomplished using the command

```
dwt(sunspots[1:2048])
```

The wavelet basis and the number of decomposition levels were not specified, so the default values of `wf="la8"` and `n.levels=4` are used. This results in 2048 wavelet coefficients. These coefficients are sorted in magnitude and the smallest 50% (1024) are set to 0. The inverse DWT is applied to this compressed set of coefficients, resulting in the reconstruction shown in Figure 13.5b. Figure 13.5c is the result of setting the smallest 95% of the coefficients to 0 prior to reconstruction. As the figure shows, compressing the data by half provides a reconstruction nearly indistinguishable from the original data. Keeping only 5% (102) of the coefficients results in a reconstruction with the basic shape of the original data, but with the very localized variability mostly removed. This final reconstruction make the underlying trend apparent without the obscuring short-term variation evident in the original data set.

Thresholding

A drawback to compression as used in Example 13.3 is the need to specify the amount of reduction. Choosing a certain percentage reduction becomes a subjective decision,

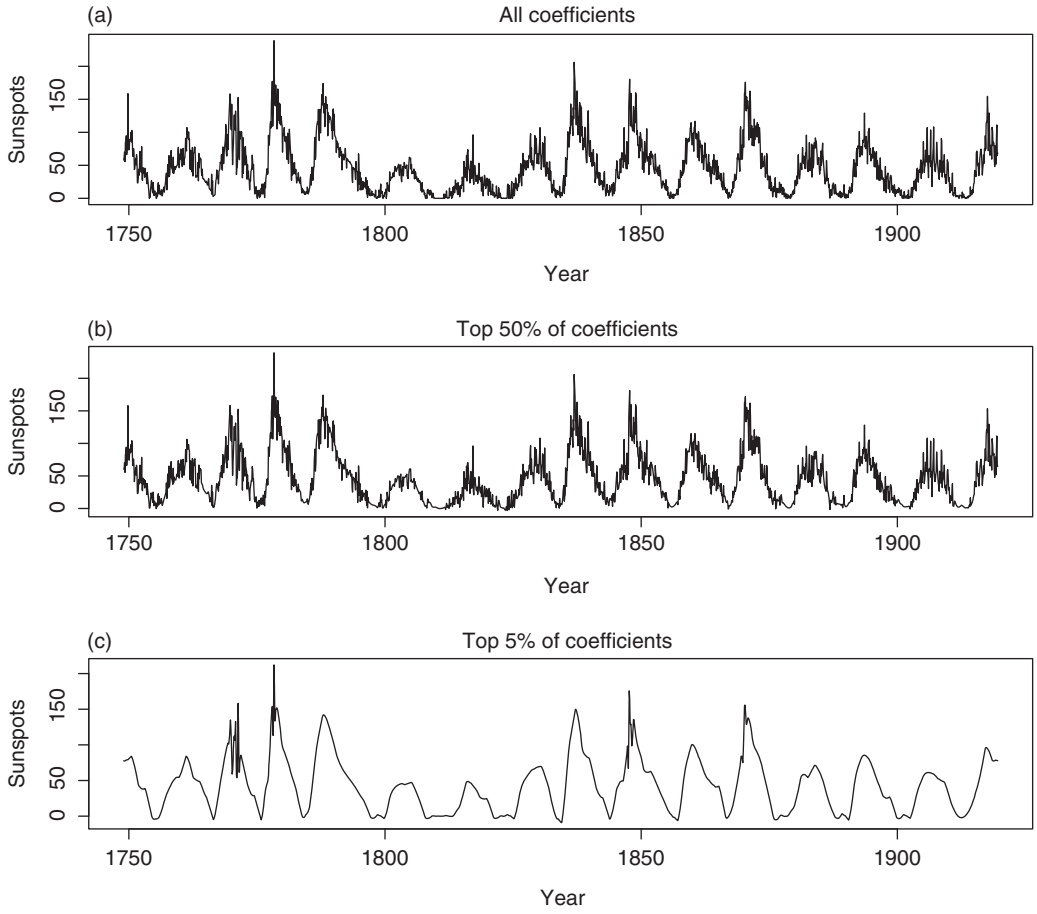

Figure 13.5 Monthly sunspot data from Example 13.3. (a) displays 2048 observations from January 1749 to July 1919. (b) the reconstruction using the 1024 largest (in magnitude) wavelet coefficients. (c) the reconstruction using the 102 largest (in magnitude) wavelet coefficients.

perhaps not adequately reflecting the nature of the underlying function f well. For example, one may choose to keep a very small percentage of the wavelet coefficients resulting in a smooth estimate of f, when in fact f may not be smooth at all.

In contrast, thresholding methods specify a value based on theoretical or data-driven considerations. No subjectivity is required to specify which coefficients to set to 0 and which to keep. Many methods of thresholding are based on assuming that the errors in Assumption A2 are normally distributed. Some of these methods are given in Donoho and Johnstone (1994). They make use of two rules for thresholding wavelet coefficients:

$$\eta_H(\theta, \lambda) = \theta \cdot I\,(|\theta| > \lambda) \tag{13.17}$$

$$\eta_S(\theta, \lambda) = \text{sgn}(\theta)\,(|\theta| - \lambda)_+\,. \tag{13.18}$$

The first of these rules is called *hard thresholding*, while the second is *soft*. In each rule, θ is a coefficient estimated with the DWT and λ is a specified threshold value. The

hard threshold sets a coefficient to 0 if it has small magnitude and leaves the coefficient unmodified otherwise. The soft threshold sets small coefficients to 0 and shrinks the large ones by λ toward 0. The compression method discussed in Example 13.3 uses hard thresholding with the threshold determined by sorting the coefficients in absolute magnitude.

The DWT operation may be represented as a matrix operator. If W represents this matrix and y is the observed data, then

$$\tilde{\theta} = Wy = Wf + W\varepsilon$$

is the vector of wavelet coefficients based on the observed noisy data as in Assumption A1. The DWT is a linear transformation, so this can also be written as

$$\tilde{\theta} = \theta + \tilde{\varepsilon},$$

where $\theta = Wf$ represents the wavelet coefficients of the unobserved sampled function f (without errors) and $\tilde{\varepsilon} = W\varepsilon$ represents the coefficients of the errors. The DWT matrix W is orthogonal, so the $\tilde{\varepsilon}$ are normally distributed, provided the original errors ε were normal. Unless the size of noise contaminating the underlying function f is excessive, the $\tilde{\varepsilon}$ are generally smaller in magnitude than θ. In this case, the sparsity property of wavelet coefficients implies that coefficients representing error may safely be ignored. Donoho and Johnstone make use of this in their "VisuShrink" estimator. VisuShrink is the result of applying the soft threshold rule (13.18) to $\tilde{\theta}$ using the threshold

$$\lambda_v = \sqrt{2\sigma^2 \log(n)},$$

where σ^2 is the variance of the errors ε. The value of σ^2 is not generally known and must be estimated. See Comment 11. If $\hat{\theta} = \eta_S(\tilde{\theta}, \lambda_v)$, the thresholded coefficients, then the VisuShrink estimate of f is

$$\hat{f}_v = W^{-1}\hat{\theta}, \tag{13.19}$$

where W^{-1} represents the inverse DWT.

In general, analyzing an observed sample of data y via wavelets refers to the process of decomposing the data via the DWT, applying some method of thresholding, and then reconstructing the thresholded coefficients using the inverse DWT. Letting $\hat{f}$ be the reconstructed estimate of the unknown function f, this is written as

$$\hat{f} = W^{-1}\eta\left(Wy, \lambda\right). \tag{13.20}$$

Note that Wy are the observed coefficients $\tilde{\theta}$, the result of applying the DWT W. This is followed by applying the threshold rule η with some threshold λ, resulting in $\hat{\theta}$. Finally, the inverse DWT is applied to the $\hat{\theta}$.

The soft threshold rule η_S in (13.19) can be replaced with the hard threshold rule η_H without affecting the asymptotic performance of the VisuShrink estimate in terms of mean squared error (MSE):

$$E\left\{\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - \hat{f}_v(x_i)\right)^2\right\}.$$

The expectation is needed to account for the randomness from the error term ε. See Comment 12.

When applying thresholding to wavelet coefficients, not all coefficients are considered. Generally, only the wavelet coefficients in the detail resolution levels above the smooth approximation are subjected to thresholding. The justification for this is that the smooth approximation is a coarse estimate of f and none of this information should be discarded. Recalling the discussion on the sparsity of the wavelet representation of a function, the majority of the coefficients are in high detail resolution levels. Relatively few coefficients are in the smooth approximation level. So, only a few wavelet coefficients are exempt from thresholding.

EXAMPLE 13.4 ***Thresholding the Sunspot Data.***

The first 2048 observations from the `sunspot` data will be thresholded using VisuShrink. Note that VisuShrink provides an optimal reconstruction of f only in the case of normally distributed errors. Ignoring this for the moment, the analysis begins by obtaining the DWT of the observed data. The next step is to threshold the coefficients. This is done in `R` via the `universal.thresh` command. This command requires a list of wavelet coefficients from applying the DWT to the observed data y, the number of levels to threshold, and an indication of whether to use the hard or soft threshold. The value of σ is estimated within `universal.thresh` using (13.21) with $m = 0$. See Comment 11. The default value for the number of levels to threshold is `max.level=4`. This coincides with the default number of levels for decomposition in the command `DWT`: `n.levels = 4`. The default rule for thresholding is `hard = TRUE`. The final step, after thresholding, is to reconstruct the thresholded coefficients using `idwt`. The following commands will implement these three steps:

```
y <- sunspots[1:2048]
y.dwt <- dwt(sunspots[1:2048])
y.thresh <- universal.thresh(y.dwt)
y.idwt <- idwt(y.thresh)
```

The `dwt` and `universal.thresh` commands used default values, so four detail resolution levels and a smooth approximation are provided by `dwt` and only the four detail levels are thresholded. The rule used is the hard rule at (13.17).

Figure 13.6 shows the plots of reconstructions of the `sunspot` data using both hard and soft thresholding rules. Figure 13.6a displays the observed data, Figure 13.6b the VisuShrink estimate using the soft threshold rule, and Figure 13.6c the VisuShrink estimate using the hard threshold. The default setting in `DWT` decomposed the observed data into four detail resolution levels of lengths $n/2 = 1024, n/4 = 512, n/8 = 256$, and $n/16 = 128$. The smooth approximation level is therefore of length $n/16 = 128$. The default setting in `universal.thresh` only thresholds the wavelet coefficients in the four detail resolution levels. Of these 1920 detail coefficients, each rule thresholded 1840 of them. Only 80 detail coefficients are not set to 0. With hard thresholding, these 80 coefficients are unmodified. With soft thresholding, these 80 coefficients are shrunk toward 0 by $\lambda = 31.78$. For comparison with the compressed reconstructions in Figure 13.5, 89.84% of the coefficients are set to 0 using VisuShrink.

Although both hard and soft rules set the same subset of coefficients to 0, the reconstructions shown in Figure 13.6 are very different. Any coefficient larger than $\lambda = 31.78$ is reduced toward 0 by λ in the soft thresholded estimate. This results in

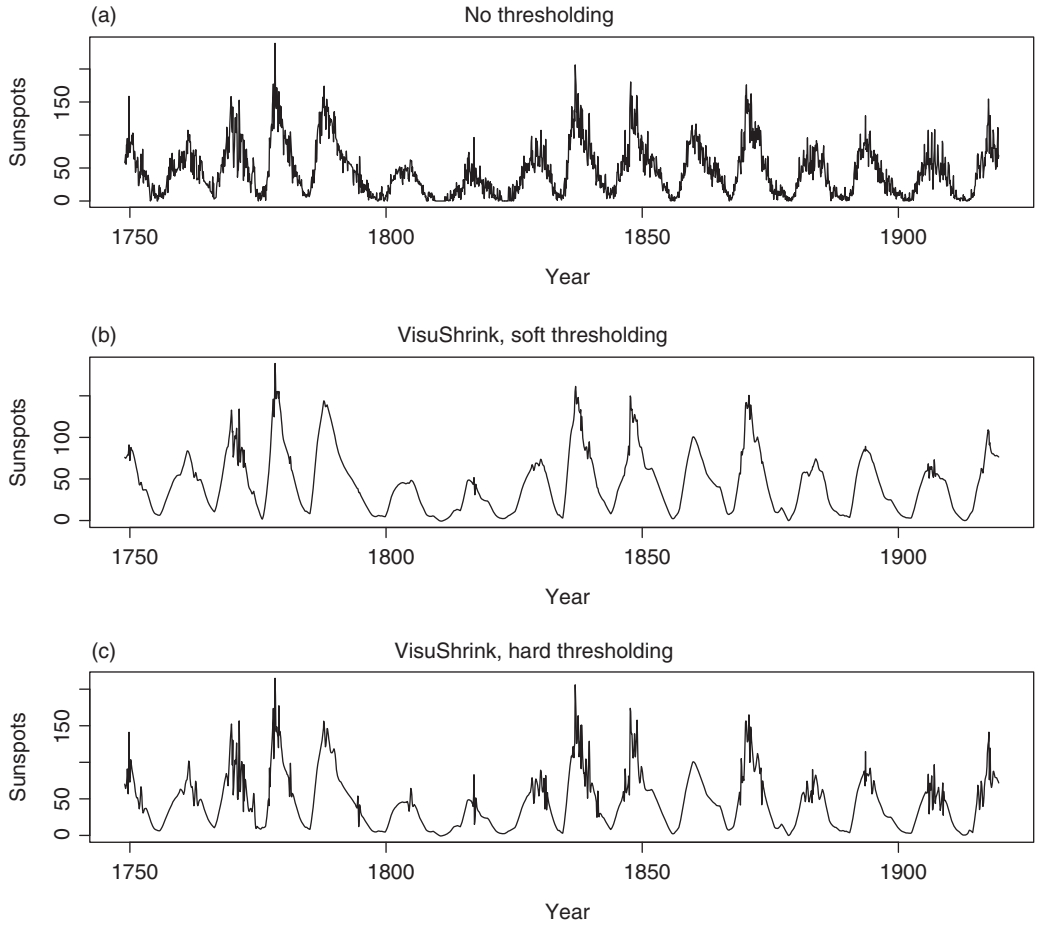

Figure 13.6 VisuShrink estimates of the monthly sunspot data from Example 13.4. (a) the original 2048 observations from January 1749 to July 1919. (b) the reconstruction using soft thresholding. (c) the reconstruction using hard thresholding.

the smoother estimate that is evident in the figure. In fact, one of the desirable properties of VisuShrink with soft thresholding is its ability to produce visually pleasing results (hence the "visu" in VisuShrink). As mentioned above, both hard and soft thresholding with VisuShrink result in the same MSE performance asymptotically.

The threshold used in VisuShrink is a universal, or global, threshold. The same threshold is applied to every coefficient. Another method of thresholding uses a different threshold at each resolution level of the wavelet decomposition of f. One such example is the SureShrink method of Donoho and Johnstone (1995). SureShrink chooses a threshold at each resolution level by minimizing the risk at that level as given by Stein (1981). See Comment 17. Unlike VisuShrink, SureShrink is only intended to be used with the soft threshold rule. SureShrink outperforms VisuShrink in terms of lower MSE (see Comment 12), but it does not always provide the visually pleasing reconstructions which some find desirable.

SureShrink is actually a hybrid threshold method. While sparsity of the wavelet representation of a function is considered to be a beneficial property, there is a concern

when implementing SureShrink that the wavelet coefficients at certain resolution levels can be too sparse. There may not enough information in a resolution level to determine what the threshold should be. Only a few wavelet coefficients in a resolution level may be representing function, while the rest of the coefficients are transformations of the errors ε. In this case, SureShrink will revert to using the universal threshold of VisuShrink at the resolution level in question.

EXAMPLE 13.5 ***SureShrink Thresholding.***

R implements the SureShrink threshold with command `hybrid.thresh`. The arguments are the DWT of the observed data and the number of levels to threshold. Additionally, `hybrid.thresh` makes use of a sampling scheme. At each resolution level, one half of the wavelet coefficients are sampled and used to estimate the threshold for the other half.

Figure 13.7 shows the results of applying SureShrink to an observed data vector sampled from a function. This data is in R as the object `blocks`. This object is part of the package `waveslim` and is made available for use with the command `data(blocks)`. Figure 13.7a shows the a sample of size $n = 512$ from the `blocks` function f with errors added to it. Figure 13.7b is obtained using VisuShrink with soft thresholding,

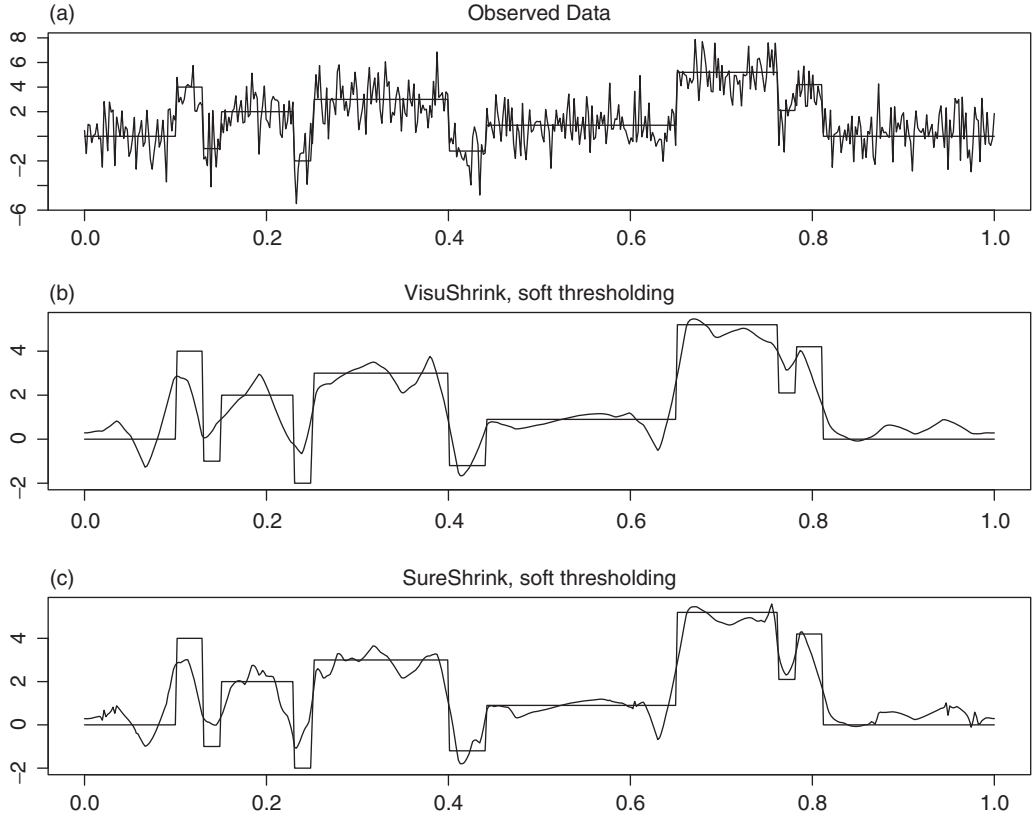

Figure 13.7 Reconstructions using VisuShrink and SureShrink from Example 13.5. (a) displays the original 512 noisy observations and the true, underlying function. (b) is the VisuShrink reconstruction using soft thresholding. (c) is the SureShrink reconstruction.

Figure 13.7c is the reconstruction using SureShrink. In all three panels of Figure 13.7, the piecewise linear function is the true underlying function f.

In this example, the errors added to the function `blocks` are independent, identically distributed normal errors with mean 0 and standard deviation $\sigma = 1.5$. The wavelet decomposition went down four levels, providing a smooth approximation and four detail resolution levels. VisuShrink set 477 of the 480 detail coefficients to 0. SureShrink set 421 of the 480 detail coefficients to 0. The MSE of the reconstructions is 0.78 and 0.64 for VisuShrink and SureShrink, respectively. While SureShrink had the lower reconstruction error, Figure 13.7c clearly exhibit the ability of VisuShrink to provide a more visually pleasing reconstruction than SureShrink. VisuShrink is probably oversmoothing the data. Note that all but three of the detail coefficients were set to 0, while 59 of these coefficients were not removed in SureShrink.

The coefficients at the second highest level of detail coefficients were deemed too sparse. See Comment 17. The hybrid nature of SureShrink requires that the universal threshold of VisuShrink must be used for this level, while the risk minimization thresholds were used in the remaining three detail resolution levels. This is done automatically with `hybrid.thresh`.

Other Thresholding Methods

The thresholding methods discussed above make decisions about whether or not to shrink a coefficient by considering one coefficient at a time. A threshold is determined by VisuShrink, SureShrink, or the top percentage method and each coefficient is compared against this threshold. Several methods (Pensky and Vidakovic (1999); Hall et al. (1998, 1999); Cai and Silverman (2001)) have been proposed to threshold groups of coefficients simultaneously. In these methods, subset of the coefficients are grouped together and a function of these coefficients is evaluated. A decision to threshold or not is made for all coefficients in the subset, and this rule is applied to all of them at once. One motivation for such methods is that neighboring coefficients may contain information about each other that will improve the thresholding decision. See Comment 14.

Thresholding the wavelet coefficients by keeping a certain percentage of the largest coefficients did not require distributional assumptions on the coefficients, but both VisuShrink and SureShrink were designed for normally distributed errors. Thresholding without strong distributional assumptions on the errors may be implemented using a cross-validation threshold rule. See Nason (1996), for example, and Comment 15.

These alternative methods of thresholding are discussed in detail in Nason (2008).

Comments

11. *Estimating* σ. The threshold λ_v requires σ^2. In practice, this is not known and must be estimated. For a vector y sampled from an observed noisy function $f + \varepsilon$, the highest level of detail coefficients are used to estimate σ^2. This is a sound strategy if one believes the estimated coefficients in this detail level represent the observed error, rather than function. This is not an unreasonable assumption. If the sample is of size $n = 2^J$, this highest detail resolution level only serves to bridge the gap between two successive smooth approximations at levels $J - 1$ and J. For large n, or for smooth f, this difference between the two approximations may

indeed be just noise. However, it is possible that some functional component of f may be present in the highest detail resolution level of coefficients. In this case, a robust estimator of σ^2 is desired. Donoho and Johnstone use the median absolute deviation estimator. Because they assume the errors to be normally distributed, their estimate of σ is

$$\hat{\sigma} = \left\{\text{median}\left(|\tilde{\theta}_{J-1,k} - m|\right)\right\} \cdot \left(\Phi^{-1}(3/4)\right)^{-1}, \tag{13.21}$$

where $m = \text{median}(\tilde{\theta}_{J-1,l})$ and Φ is the normal cumulative distribution function. The constant involving Φ provides an unbiased estimate of σ in the event of normality.

12. *Rates of Convergence*. The measure of error for an estimate $\hat{f}$ of a function f when using wavelet methods is generally the MSE,

$$E\left\{\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - \hat{f}_v(x_i)\right)^2\right\}.$$

If the x_i are equally spaced over $[0, 1]$ with distance between points equal to n^{-1}, the MSE above is approximately the same as

$$E\left\{\int_0^1 \left(f(x) - \hat{f}_v(x)\right)^2 dx\right\}.$$

If $f \in \mathcal{F}$ has smoothness parameter α (roughly the number of derivatives possessed by f), then the VisuShrink estimate $\hat{f}_v$ has error rate

$$\sup_{f\in\mathcal{F}} E\left\{\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - \hat{f}_v(x_i)\right)^2\right\} \le C\left(\frac{\log n}{n}\right)^{\frac{2\alpha}{2\alpha+1}},$$

where C is some unknown finite constant which does not depend on n. The SureShrink estimator and block threshold estimators remove the log term from the above rate. For details, see Donoho and Johnstone (1994, 1995), Hall et al. (1999), and Cai (1999).

13. *The Translation Invariant Estimator*. One of the results of using compactly supported wavelets is that the end points of the support of the basis functions may adversely affect estimation at those points. In particular, if there are jumps in the function that do not occur where the wavelet basis functions begin or end, these jumps will be poorly modeled. A fix to this was proposed in Coifman and Donoho (1995) and Nason and Silverman (1995). Their method, the translation invariant estimate, applies wavelet decomposition, thresholding, and reconstruction to all possible shifts of the original observed data $y = f + \varepsilon$. For a signal of length n, there are n possible shifts of the data. Each shift cycles the data one position to the right. Data moving past the end point is cycled back to the starting point. After the shifted data is thresholded and reconstructed, it is "unshifted" back to the original positions. Then, these n estimates of the underlying function are averaged to provide a final estimate. The hope is that shifting the data removes the dependency on the estimation at a point from the location of the wavelet basis

function end points. If S_l shifts the data by l positions and S_l^{-1} shifts the data back to its original position, then the translation invariant estimator is given by

$$f_{TI} = \frac{1}{n}\sum_{l=1}^{n} S_l^{-1} W^{-1} \eta\left(WS_l y, \lambda\right).$$

Cuevas and Chicken (2012) extend this idea by considering only select shifts at each point. Both these methods provide increased ability to model jumps in the underlying functions f. The translation invariant estimate is implemented in `R` with the commands `modwt` and `imodwt`. It is also an option in the `mra` command: `method="modwt"`.

14. *Thresholding Multiple Wavelet Coefficients Simultaneously*. The threshold methods presented in this chapter are term-by-term methods. They consider whether or not to shrink one coefficient at a time. Each decision is made independently of the others. Other thresholding methods have been developed to threshold several coefficients at once. Pensky and Vidakovic (1999) proposed using a global threshold which simultaneously considers every coefficient in a resolution level j. If the sum of the squared coefficients in a resolution level are larger than some specified threshold, then all coefficients are kept, otherwise, all are set to 0. Hall, Kerkyacharian, and Picard (1998, 1999) and Cai (1999) proposed and examined block thresholding methods that consider groups of coefficients within a resolution level. Again, sums of squared coefficients in a block are compared to a threshold and all coefficients in a block are shrunk by the same amount. Note that a block size of 1 is term-by-term thresholding, while a block that encompasses the entire resolution level is the global threshold of Pensky. Variations on these ideas are the neighbor methods of Cai and Silverman (2001). These methods are similar to term-by-term and block thresholding but use additional coefficients outside the range of coefficients to be thresholded to increase the precision of the decision on whether or not to shrink the coefficients.
15. *Cross-Validation Thresholding*. VisuShrink and SureShrink determine the values of their thresholds under the assumption that the errors in Assumption A2 are normally distributed. Nason (1996) removes this assumption by using a cross-validation threshold. An observed sample vector y of length n is split into even and odd components of length $n/2$. For each candidate value of the threshold λ, the wavelet estimate (decomposition, thresholding, reconstruction) of the even (odd) data is compared to the observed odd (even) data using the MSE. The λ that minimizes this error is the chosen threshold. This threshold is chosen with respect to sampled functions of length $n/2$, so a multiplier is applied to make it suitable for the original observed data of length n. This type of thresholding is implemented in the package `wavethresh` (Nason, 2010).
16. *Unequally Spaced Sample Points*. It is not necessary that the vector y be sampled at equally spaced points. Cai and Brown (1998, 1999) and Kovac and Silverman (2000) show that term-by-term thresholding methods may be applied to sample points that are fixed and irregularly spaced or follow a uniform placement on the support interval of the underlying f without affecting the estimation properties of wavelet analysis. Chicken (2003, 2005) shows similar results for block thresholding methods. In some cases, the wavelet transform may be applied directly to

nonequispaced data, but in other situations, a transformation must be applied to the observed data y before and after the wavelet analysis.

17. *SureShrink Threshold*. The word "sure" in SureShrink refers to "Stein's unbiased risk estimate." For a fixed resolution level j and threshold λ, this risk is

$$\text{SURE}(\theta_j., \lambda) = n_j - 2 \cdot \#\{k : |\theta_{jk}| \leq \lambda\sigma_j\} + \sum_{k=1}^{n_j} \left(\min(|\theta_{jk}|, \lambda)\right)^2,$$

where $\theta_j.$ is the set of coefficients at resolution level j and n_j is the number of such coefficients. The threshold at resolution level j is then

$$\lambda_j = \underset{0 \leq \lambda \leq \lambda_v}{\arg\min}\ \text{SURE}(\theta_j., \lambda),$$

where λ_v is the VisuShrink threshold using n_j. SureShrink will use λ_v if a level is too sparse:

$$\frac{1}{n_j}\sum_k(\theta_{jk}^2 - 1) \leq n_j^{-1/2}(\log_2 n_j)^{3/2}.$$

For details, see Donoho and Johnstone (1995).

R allows the use of SureShrink without the sparsity condition through the command `sure.thresh`. This command also allows hard thresholding, which SureShrink is not designed for. The sampling scheme used to determine the threshold for half the data based on the other half is not implemented in `sure.thresh`.

Properties

1. *Convergence Rates*. For rates of convergence of wavelet-based estimates in a variety of settings, see Donoho and Johnstone (1994, 1995), Hall et al. (1999), Cai (1999), Vidakovic (1999) and Chicken (2003, 2005).

Problems

7. Use the `dwt` command to obtain the wavelet coefficients of the first 512 components of the `sunspots` data. Use `n.levels=9`.

(a) Use the `unlist` command on this object to create a single vector of coefficients. Make two histograms: one of the coefficient vector and one of the untransformed data. How do these histogram shapes illustrate the sparsity of the wavelet representation of the data?

(b) Create a histogram using only the highest level of detail coefficients (use `[[1]]` or `$d1` to access these coefficients from the output of `dwt`). Do these coefficients appear to be symmetric about 0? Normal?

(c) Threshold the wavelet coefficients for this data using both SureShrink and VisuShrink. Describe the differences in the reconstructions.

8. Create vectors of noisy `blocks` data with errors having mean 0 and variance 1 using the following R commands:

```
y1 <- blocks + rnorm(512)
y2 <- blocks + rexp(512) - rep(1, 512)
y3 <- blocks + runif(512, -sqrt(3), sqrt(3))
```

(a) Apply VisuShrink separately to each of the vectors `y1`, `y2`, and `y3`. Find and compare the MSEs for each reconstruction with respect to $f =$ `blocks`. Comment on the visual quality of the reconstructions.
(b) Repeat the above tasks using SureShrink in place of VisuShrink.
(c) How do the MSEs compare for the VisuShrink and SureShrink estimates from the above two problems? Comment on any differences in the visual quality of the estimates.
(d) Construct a histogram using only the highest level of detail wavelet coefficients from the DWT of $y1$. Do the same for $y3$. Do either sets of these coefficients appear to be normal?

9. Using the initial 2^{10} components of the `sunspots` data, find VisuShrink estimates of the data using `n.levels=6` in `dwt` and `max.level=4, 5, 6`, and `7` in `universal.thresh` (setting `max.level=7` will threshold the wavelet coefficients used in the smooth approximation of f). Compare the VisuShrink estimates using both MSE and visual fit.

13.3 OTHER USES OF WAVELETS IN STATISTICS

The methods discussed in this chapter use wavelets to estimate a function f when it is observed under Assumptions A1 to A3. However, this is by no means the only use of wavelets. In fact, wavelet methods have been applied successfully to many statistical problems other than those presented in Sections 13.1 and 13.2. As mentioned in Chapter 12, wavelets may be used in nonparametric density estimation. See Vidakovic (1999) and references therein for examples of density estimators based on wavelets. Wavelets are also useful in understanding the properties of time series and random processes. See Percival and Walden (2000) and Craigmile and Percival (2005) for results and references on the use of wavelets in these cases.

Many engineering applications consist of observing sequences of large data sets. For example, each observation $W_i, i = 1, 2, \ldots, m$ may be a set of bivariate data $\{(x_j, y_j)\}_{j=1}^{n}$ where x and y are related functionally as specified in Assumptions A1 to A3: for each i, W_i is an observation consisting of n bivariate pairs of data satisfying

$$y_{ij} = f(x_{ij}) + \varepsilon_{ij}, \quad j = 1, 2, \ldots, n.$$

The wavelet property of sparsity becomes useful to reduce the size of the data in these problems. Jin and Shi (2001) and Chicken et al. (2009) provide examples where wavelets are used as a dimension reduction tool to monitor sequences of such observations under the assumption of normality on the errors, while McGinnity et al. (2013) developed a method to monitor such sequences under general conditions for the observed errors. See Chicken (2011) for additional references and more examples of the utility of wavelets in such situations.

In addition to the three examples described above, wavelets are applicable to many other statistical problems. Vidakovic (1999) and Nason (2008) provide additional details, examples and references for the use of wavelets in statistical methods beyond those presented in this chapter. In particular, Nason provides extensive examples implemented with the R software package.

Chapter 14

Smoothing

INTRODUCTION

The wavelet methods from Chapter 13 are useful at estimating a function from a sample of bivariate data (x, y) because these methods do not rely on specific assumptions about the functional relation underlying the data. In contrast, the functional estimation problems considered in Chapters 9 and 11 were designed to estimate very specific types of functions. In this chapter, we continue to pursue methods to estimate a general function from a collection of bivariate observations.

Wavelet methods projected the data into resolution levels at various scales (frequencies) through the use of a set of special basis functions. The properties of wavelets lead to representing the data by a greatly reduced set of objects through the nonlinear process of thresholding. The methods presented here are "smoothers." Similar to wavelet methods, these smoothers may make use of external functions to model the functional relation between y and x. These functions, however, neither form a basis for a space of functions nor do they provide a dimension reduction property or analyze an observed function in terms of scale and location. Additionally, the external functions used in smoothers are typically very simple: lines or low order polynomial functions. Most wavelet functions are very complex.

We refer to these methods as smoothers, but another common term used is nonparametric regression. The term *nonparametric* in this case refers not to distributional assumptions on observed errors but to the lack of a specific, parametric form assumed for the function being estimated. For example, in Chapter 9, we assumed that the function to be estimated was a linear function. Here, no such assumptions are made. In addition, we also make no strong distributional assumptions on the errors.

In this chapter, three types of smoothing methods are discussed. All are linear smoothers: the estimates obtained are linear combinations of the observed data. The first section introduces a local averaging estimator. Next, we discuss a local linear estimate based on regressing the observed data in localized windows. In the final section, kernel methods, similar to those used for density estimation found in Chapter 12, are given.

Data. There are n pairs of observations $(x_1, y_1), (x_2, y_2) \ldots, (x_n, y_n)$. Without loss of generality, assume the x_i are ordered, $x_1 \leq x_2 \leq \cdots \leq x_n$.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

Assumptions

A1. The observations are related through the expression

$$y_i = f(x_i) + \varepsilon_i, i = 1, 2, \ldots, n.$$

A2. The errors ε_i are independent and identically distributed from a continuous population, centered at 0.

14.1 LOCAL AVERAGING (FRIEDMAN)

In local averaging, the estimate of f at the point x_i is taken to be the average of observed values y_j corresponding to values x_j in some vicinity of x_i. In Friedman (1984), this neighborhood of x_i is chosen to be the smallest symmetric window about x_i containing s observations. Since the average is a linear combination of the points in the neighborhood, the fit is a linear smoother. The term *span* is used to denote the number of points in the window. The window is sized to include s points, including and centered at x_i. The window size will change for different values of x_i, but always includes the same number of points. This is a nearest neighbor method, as opposed to a method where the window is always the same size.

The value of the span s is critical. Recall from Chapter 12 that selecting an appropriate bandwidth drives the usefulness of the density estimation procedure. The choice of the span is similar in this case. Too large a span will oversmooth the data resulting in a large bias in the estimate. Too small a span provides an estimate that is undersmoothed. In this second case, we have large variability. The correct span will find a balance between the bias and variance.

Friedman proposed using a cross-validation method to choose the span. This method uses a leave-one-out approach to find the best span. One point at a time is removed from the set of n bivariate observations (x, y). The local average is determined for the point (x_i, y_i) using the $n - 1$ remaining points, and the error between this point and its estimate is noted. This process is done for a variety of candidate values for the span s, and the selected value is that which minimizes the average of these errors squared. See Comment 2.

Rather than using a single span for the entire estimate of f, Friedman makes use of variable spans. A variable span is analogous to the variable bandwidths used in density estimation in Chapter 12. In a region where the function f is irregular or displays much variability, it might be beneficial to have a large span. Smooth regions might require fewer points in the estimate. The variable span depends on x_i and is denoted as $s(x_i)$. See Comment 2.

If it is known (or suspected) that the underlying function f is relatively smooth curve, an additional smoothing parameter may be used with the variable span. The "bass" parameter resides in the interval $[0, 10]$. Small values of the bass have little or no additional smoothing effect on the estimated functions. Large values of the bass will drive the span close to a window containing $n/2$ of the data points. See Comment 3.

EXAMPLE 14.1 ***Nitrogen Oxide Concentrations.***

Cleveland (1979) examines data gathered by Brinkman (1981) on the nitrogen oxide concentrations found in engine exhaust for ethanol engines with various equivalence

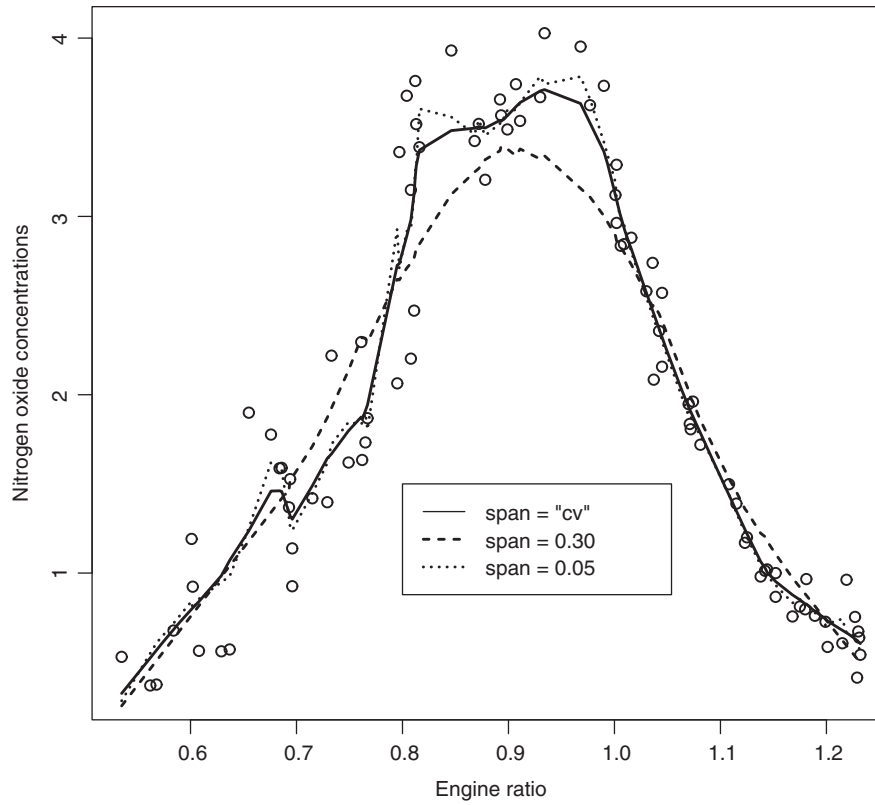

Figure 14.1 Smoothing of the `ethanol` data from Brinkman (1981) using Friedman's local averaging smoother. The lines are the smoothed data with differing spans, the points are the original observed pairs (x, y). The smoothing parameter `bass` is set to 0 (the default value in `supsmu`).

ratios. These 88 pairs of data are in the R data set `ethanol`. They may be smoothed by Friedman's local averaging method with the command `supsmu`. The command requires arguments for the x and y vectors, a span and a bass. The cross-validated variable span is implemented using the argument `span="cv"`, the default value. For a constant span of size pn where $p \in [0, 1]$, use `span=p`. The default bass value is 0, but this may be changed with the argument `bass`. Figure 14.1 displays the 88 observations from the ethanol data set as points and the smoothed data (solid line) using these default values for `bass` and `span`, as well as two additional constant spans. For the cross-validated span (solid line), the overall shape of the data is maintained, with the short-term irregularities smoothed over. For the specified constant span of 0.30 (dashed line), the estimator is too smooth (biased). In particular, it misses the data peak in the center. Setting the constant span lower to 0.05 (dotted line) will pick up the peak, but there is excess variability in the smoothed estimate. Using the cross-validation variable span provides a more reasonable estimate in that it is neither oversmoothing nor too irregular. It is striking a reasonable balance between bias and variance.

The smoothed estimates in Figure 14.1 are generated with the following command:

```
supsmu(x=ethanol$E, y=ethanol$NOx, bass=0, span="cv")
```

where `E` is the engine equivalence ratio and `NOx` the nitrogen oxide concentration. The argument `span` may be changed to 0.30 and 0.05 for comparisons. If the results of the

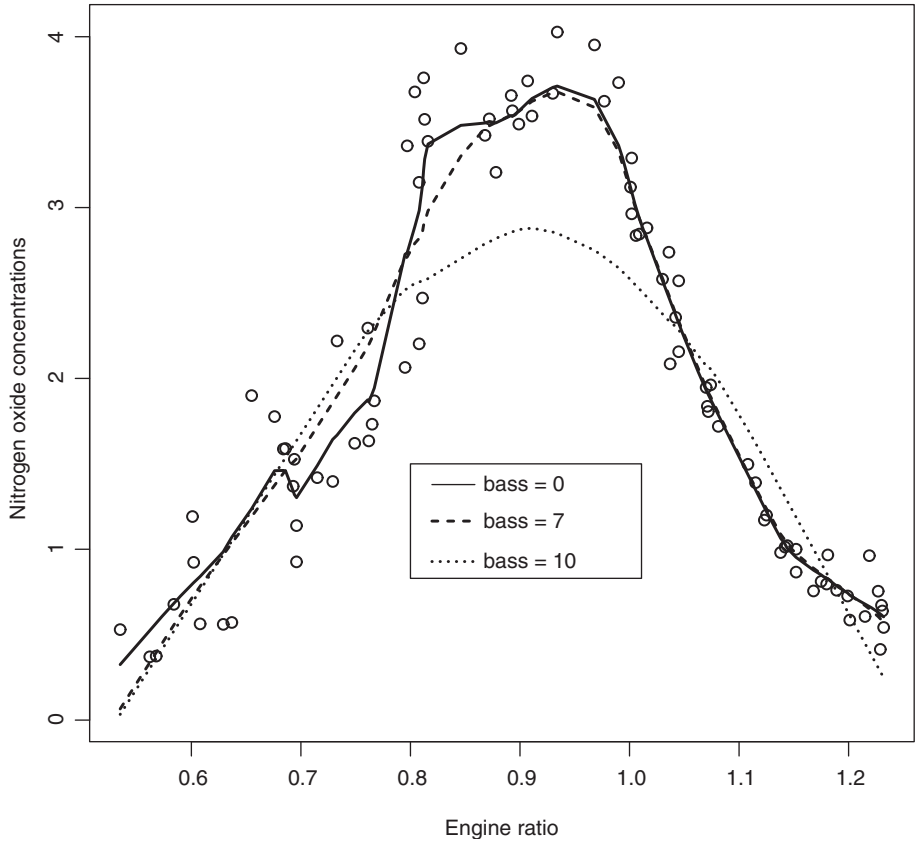

Figure 14.2 Smoothing of the `ethanol` data from Brinkman (1981) using Friedman's local averaging smoother and varying the `bass` argument in `supsmu`.

above call are saved in an R object, say `ethanol.supsmu`, then `ethanol.supsmu$y` is the vector of the fitted smoothed values $\hat{y}$.

Increasing the bass will impose additional smoothing on the data. This is illustrated in Figure 14.2. The span is chosen again using the cross-validation method described earlier. The three values for `bass` are 0 (default), 7, and 10. The solid line in Figure 14.2 is the same as that in Figure 14.1. The dashed line shows the smoothed data using `bass=7`. Note that the resulting smoothed data with this value is close to the default case with a notable exception near the equivalence ratio of 0.7. Setting `bass=10`, the maximum value, results in the smoothed data shown with the dotted line. With this `bass` value, the span is quite far from the cross-validated span. See Comment 3. It is clear that this estimate (`bass=10`) is oversmoothing the data. The center portion of the estimate is biased downward significantly. The variance of the estimate is reduced but at the expense of excessive bias.

Comments

1. *Choice of Window.* Friedman's local averaging smoother uses the smallest symmetric window centered at each x_i and containing s points as the neighborhood for smoothing. An alternative is to use the smallest (not necessarily symmetric)

window that contains the specified number of observations, s. This alternative method has the advantage that it generalizes to smoothing higher dimensional functions. However, the symmetric window approach is simpler from a computational standpoint. In particular, for the x_i being equally spaced over the support of the function f, adjacent neighborhoods are identical except for one entering and one leaving point. See Friedman (1984).

2. *Cross-Validation.* For an index i, let $\hat{y}_{(i)}$ be the local averaging estimate of the function at the point x_i determined by using all the observed data points except for (x_i, y_i) and a particular specified value of the span s. Then,

$$e_i(s) = y_i - \hat{y}_{(i)}, i = 1, 2 \ldots, n$$

is the error based on this leave-one-out approach for a single point (x_i, y_i). The choice for a span is the value of s that minimizes the average of these errors:

$$\hat{s} = \arg\min_{s \in S} \frac{1}{n} \sum_{i=1}^{n} e_i^2(s)$$

for some set S of candidate values of s. The span selected in such a way is a global span. It is used for every point x_i.

If using a variable span (the default method in `supsmu`), a span $s(x_i)$ is found for each x_i. The set S is taken to be $S = \{.05n, .2n, .5n\}$. These three spans are referred to as the tweeter, midrange, and woofer smoothers. The woofer is global in nature and provides a high level of smoothing; the tweeter is very localized and give the least amount of smoothing. The midrange is in between the two. Rather than minimizing the sum of squared errors as in the constant span case, the errors associated with a particular candidate span are first smoothed with the midrange smoother. The error at x_i is taken to be the smoothed value at x_i. The value of the variable span can change abruptly for adjacent points x_i. To account for this, the function $s(x)$ is subjected to further smoothing using the midrange smoother and an interpolation step is used to determine the final fitted value $\hat{y}_i$. See Friedman (1984) for details on the mechanics of variable span selection. For more details on cross-validation, see, for example, Allen (1974), Stone (1974), or Golub et al. (1979).

3. *Additional Smoothing via the Bass Parameter.* The woofer (see Comment 2) in Friedman (1984) provides a significant amount of smoothing due to its large size: $n/2$. The bass parameter is used to force the span closer to the woofer through user input. For a point x_i, the bass modifies the span through the relation

$$s(x_i) = s_{cv}(x_i) + (s_w - s_{cv}(x_i)) \cdot \left(\frac{e_{cv,i}}{e_w} \right)^{10-\alpha},$$

where $\alpha \in [0, 10]$ is the user-specified value of the bass, s_{cv} is the optimal span for x_i, and s_w is the fixed woofer span. The fraction is the ratio of two fitted errors: $e_{cv,i}$ represents the error between the fitted point $\hat{y}_i$ and the observed point y_i using the optimal span, e_w is this error when fitting with the woofer. Note that setting $\alpha = 10$ results in $s(x_i) = s_w$, while $\alpha = 0$ has $s(x_i)$ very close to $s_{cv}(x_i)$ (the fraction should be less than equal to 1). The bass is applied to the data before the optimal spans are smoothed. See Comment 2.

4. *Dependence of Errors.* If Assumption A2 is invalid, in particular with positive or negative correlation among the errors, then using cross-validation to choose the optimal span will result in poor estimates. Positive correlation will result in underestimating the underlying function f, while negatively correlated errors will lead to overestimation. See Friedman (1984).
5. *Sample Size n.* The use of woofer, midrange, and tweeter values as candidate spans in the variable span selection method may be too limited in certain situations. For example, in data sets where the irregular features of the underlying function f to be smoothed require a span less than the tweeter, $s = .05n$, oversmoothing will occur. In these cases, manual values of smoothing spans may be preferable.

Problems

1. The data set `cars` from Ezekiel (1930) contains stopping distances for various speeds. Smooth the data using Friedman's smoother by choosing your own value of the `span`. Use `dist` as the dependent (response) variable and `speed` as the independent (predictor) variable. Using trial and error, what seems to be a reasonable span? Comment on the graphical comparison between the estimate using your choice of span with the estimate using the span chosen by cross-validation.

2. Consider the data set `sunspots` from Andrews and Herzberg (1985) as a response variable. For the predictor data x, use

```
x <- c(1:length(sunspots))
```

Apply Friedman's smoother using trial and error to find a span that seems to work well with the data. Then find an estimate using the span determined by cross-validation. Describe the results (taking into account Comment 5).

3. Reduce the size of the `sunspots` data from Problem 2 by taking only the first m observations:

```
y <- sunspots[1:m]
x <- c(1:m)
```

At what value of m does the cross-validation method begin to provide a reasonable estimate? Why is this occurring?

4. Using the first $m = 128$ observations from the `sunspots` data in Problem 2, obtain the wavelet estimate using the VisuShrink method from Chapter 13. Comment on the graphical comparison between the VisuShrink estimate and the estimate you would have gotten in Problem 3 by setting $m = 128$. The thresholded wavelet estimate requires a dyadic number n of points x_i, say $n = 2^J$ for some integer J. Additionally, it assumes that these points are equally spaced apart. These two conditions are met by the truncated `sunspots` data. However, VisuShrink is designed for normal errors ε, while Friedman's estimator makes no such assumption.

5. Pagan and Ullah (1999) reported data from the 1971 Canadian census on wages for male high school graduates. This data in the `R` object `cps71`. Use the logarithm of wages `logwage` as the response and `age` as the predictor. Smooth this data using Friedman's method with the span found by trial and error. Comment on the graphical comparison between this estimate to that obtained when using the span selected by cross-validation.

6. Using the Canadian census wage data from Problem 5, smooth the data with Friedman's method using the cross-validation span and `bass=0`. Vary the bass parameter to obtain a reasonable fit. Comment on the graphical comparison of the estimates with varying bass values. Which value of the bass seems most reasonable?

7. Using the data from Table 9.3 on body weight and total surface area of squirrel monkeys, find and plot the estimated line using Theil's method. Use surface area as the response and weight as the predictor. Then find the estimate using Friedman's method with the span chosen by cross-validation. Comment on the graphical comparison between these two estimates.

14.2 LOCAL REGRESSION (CLEVELAND)

Instead of estimating the function f at a point x_i by using local averaging, Cleveland (1979) proposed a method to estimate this value by performing a local linear regression on the observations (x, y) near x_i. The resulting regression model's fitted value at x_i is then used as the smoothed estimate at this point. The regression performed is a weighted regression, where the weights are related to the distance of the points used in the regression to the point x_i.

Like Friedman's local average smoother, this is a nearest neighbor method. The weight function is scaled so that only a specified number of points are used in a local regression. See Comment 6. The size of the window of points used will vary from one x_i to another, but each window will contain the same proportion αn of the total number of points, where α is some number in $(0, 1]$.

Once the initial fit at each point x_i is found using the above weighted regression, additional regressions are performed to reduce the impact of outliers on the smoothed estimate. This is done to make the estimator robust against such points. This second weighting scheme uses weights based not on the distance between the points in a neighborhood of x_i but on the residual errors between the fitted points $\hat{y}_j$ found in the initial local weighted regression and the corresponding observed values y_j. This residual weighted regression is performed iteratively, each time determining new residual-based weights on the previous local residual weighted regression. See Comment 7.

The span may also be chosen automatically. Cleveland suggests implementing cross-validation by using a modification of the PRESS (predictive residual sums of squares) method of Allen (1974). This method was further modified by Golub et al. (1979), where it is called *generalized cross-validation*. See Comment 9.

EXAMPLE 14.2 ***Nitrogen Oxide Concentrations.***

Continuing with Example 14.1, we apply Cleveland's method to the ethanol data from Brinkman (1981). Cleveland's local regression smoother is implemented in `R` using the command `loess`. This command requires arguments for a model, the `span` α and the `degree` of the polynomial to use in the local regression. The ethanol data is not sorted by equivalence ratios. The `loess` command requires the data to be sorted with respect to the x variable. This is easily accomplished with the commands

```
ethanol$NOx <- ethanol$NOx[order(ethanol$E)]
ethanol$E <- sort(ethanol$E)
```

For the local linear regression described here, `degree=1` is needed. See Comment 8. The model is specified using the notation `y ~ x`, where `x` and `y` are vectors containing the observations. The `span` argument serves the same purpose as in `supsmu`. It gives the proportion of the n observations that will be included in the local regression window. The weight functions are the tricube and biweight functions, see Comments 6 and 7. The number of iterations for the residual-based weighted regressions is by default equal to 4. This may be changed with the argument `loess.control(iterations=m)`, where `m` is the desired number of iterations.

The smooth is obtained with the call

```
loess(ethanol$NOx ~ ethanol$E, degree=1, span=0.75)
```

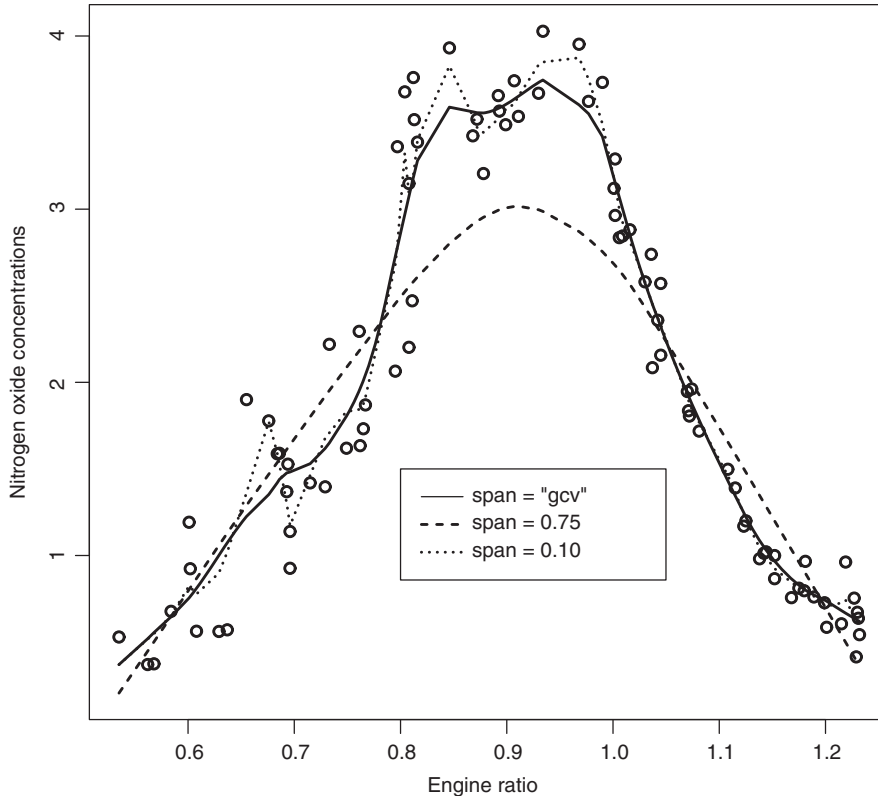

Figure 14.3 Smoothing of the `ethanol` data from Brinkman (1981) using Cleveland's local regression smoother. The estimates are obtained from `loess` for spans 0.75 and 0.10 and from `loess.as` for the cross-validation span.

or, equivalently

```
loess(NOx ~ E, data=ethanol, degree=1, span=0.75)
```

If the output to this call is saved to an `R` object, say `ethanol.loess`, then `ethanol.loess$fitted` is the vector of the fitted values $\hat{y}$. Figure 14.3 shows the different smooth estimates for span values $\alpha = 0.75$ (the default value) and $\alpha = 0.10$. The estimate using the span 0.75 (dashed line) appears to be oversmoothing the data (increased bias). This is most apparent in the center of the data where this span underestimated the peak in the data. The estimate using the span 0.10 (dashed line) displays too much irregularity (increased variance). To select a span with generalized cross-validation, the command `loess.as` from the `R` package fANCOVA (Wang (2010)) is called with arguments for the observed data `x` and `y`, and specifying `criterion="gcv"`:

```
loess.as(x=ethanol$NOx, y=ethanol$E, criterion="gcv")
```

Unlike `loess`, the default `degree` is 1 for this command. If the output to this call is saved to an `R` object, say `ethanol.loess.as`, then `ethanol.loess.as$fitted` is the fitted data and `summary(ethanol.loess.as)` will provide the selected span. For the ethanol data, generalized cross-validation chooses the span to be $\alpha = 0.19$. The solid line in Figure 14.3 is the smooth obtained with this span. Considering all

three spans, 0.10, 0.19 (the generalized cross-validation span), and 0.75 (the default span), it is clear that as the span increases the estimates become smoother. The estimate using the cross-validation span seems preferable to the other two. It displays neither oversmoothing nor excess variability. Additionally, it possesses the added benefit of not being chosen subjectively.

It is possible to set a span that will result in error messages within `loess` and `loess.as`. These are generally a result of specifying too small a span. If so, the problem is easily remedied by increasing the span incrementally until the errors no longer occur. This problem will not occur when using an automatically selected method.

Comments

6. *Weights for the Weighted Local Regression.* The function for determining the weights in the initial weighted regression model are defined by a function W and the proportion of points α to include in a local regression. In Cleveland (1979), the function W is a symmetric function, nonincreasing for nonnegative x, with the following properties:

$$W(x) > 0, |x| < 1,$$
$$W(x) = 0, |x| \geq 1.$$

This W is then modified for each index $i = 1, 2, \ldots, n$ by centering W at x_i and scaling W such that the first value at which it is zero is at the αn nearest neighbor of x_i. One such W suggested by Cleveland and implemented in the `R` command `loess` is the tricube function:

$$W(x) = \left(1 - |x|^3\right)^3, |x| < 1,$$

and $W(x) = 0$ elsewhere. Letting $w_1^i, w_2^i, \ldots, w_n^i$ be the weights determined by the centered and scaled W for a particular point x_i, the weighted local regression is found by minimizing the weighted least squares

$$\sum_{j=1}^{n} w_j^i \left(y_j - \beta_0^i - \beta_1^i x_j\right)^2, \tag{14.1}$$

where β_0^i and β_1^i are the intercept and slope of the linear relation between x and y in the neighborhood of x_i.

7. *Residual-Based Weights.* Once the initial weighted local regression is performed (see Comment 6), one may find the residual errors between each observed value y_i and the local weighted fitted value

$$\hat{y}_i = \hat{\beta}_0^i + \hat{\beta}_1^i x_i$$

where the $\hat{\beta}^i$ are the estimates of the parameters β^i using local weighted least squares regression. Residual-based weights are found by defining the bisquare weight function B,

$$(Bx) = \left(1 - x^2\right)^2, |x| < 1,$$

and $B(x) = 0$, elsewhere. Denoting the residuals by

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n,$$

the residual-based weights are

$$\delta_i = B\left(e_i\left(6 \cdot \underset{1\le j\le n}{\text{median}}|e_j|\right)^{-1}\right).$$

The weights used are the product of the original distance-based weights and the residual-based weights,

$$\sum_{j=1}^{n} \delta_j w_j^i \left(y_k - \beta_0^{i\prime} - \beta_1^{i\prime} x_j\right)^2$$

where $\beta_0^{i\prime}$ and $\beta_1^{i\prime}$ are the parameters for this new local regression fit in the neighborhood of the point x_i. Using residual-based weights to reduce the effect of outliers in the fitted models is considered a robust method. Cleveland refers to his method as robust locally weighted regression.

This is an iterative procedure. The residual weights may be recomputed using the new fit from the previous residual-based weight regression. Cleveland's use of this iteration is a modification of iterated weighted least squares. See, for example, Beaton and Tukey (1974) and Andrews (1974).

8. *Local Polynomial Regression.* The method in this section focuses on simple linear regression, that is, the degree is 1. However, the method may be implemented with a local polynomial regression in place of linear regression. R allows for degrees of $d = 0, 1$, or 2. The local polynomial regression model minimizes

$$\sum_{j=1}^{n} w_j^i \left(y_k - \beta_0^i - \beta_1^i x_j - \cdots - \beta_d^i x_j^d\right)^2.$$

For a discussion of the choice of polynomial degree, see Cleveland (1979).

9. *Automatic Span Selection.* R provides two methods of automatic span selection through the command `loess.as`. The generalized cross-validation method of Golub et al. (1979) is a modification of the ordinary cross-validation discussed in Comment 2 of Section 14.1. Ordinary cross-validation minimizes the sum of squared residuals over the candidate spans. In the notation from the above referenced comment,

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2$$

is the error based on the leave-one-out approach for a single point (x_i, y_i) and a specific value of the span. For generalized cross-validation, this is rewritten as

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_{ii}}\right)^2,$$

where the h_{ii} are the diagonal entries of the matrix H that determines the linear regression estimates. The h_{ii} are then replaced with their average value

$$\frac{1}{n}\sum_{i=1}^{n} h_{ii} = \text{trace}(H)/n,$$

and generalized cross-validation chooses the span that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i}{1-\text{trace}(H)/n}\right)^2.$$

A second method for span selection is based on using a modification of Akaike (1974). In a parametric regression setting, one compares several models and selects the one with the minimum value of Akaike information criterion (AIC). Hurvich and Tsai (1989) and Hurvich et al. (1998) modified this to apply to more general models and showed their corrected AIC is equivalent to

$$\ln\left(\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\right)+1+\frac{2(\text{trace}(H)+1)}{n-\text{trace}(H)-2}.$$

The above expression is evaluated at candidate spans, each choice of span resulting in different fitted values $\hat{y}_i$ and H. The span that minimizes the expression is chosen. This method is implemented in `loess.as` with the argument `criterion="aicc"`.

10. *Local Multivariate Regression.* Local weighted regression may be extended to the multivariate case. The fitted or smoothed values $\hat{y}_i$ are now determined by a multiple polynomial regression on a vector of p independent variable $x_1, x_2, \ldots, x_p$. The regressions are local, as before, and the nearest neighborhood method requires a p-dimensional measure of distance. See Cleveland and Devlin (1988) for details.
11. *Regression Estimation.* The regression in Cleveland's smoother is performed with weighted least squares using the argument `family="gaussian"`, the default value in `loess` and `loess.as`. Although no assumptions of normality were made on the errors ε in Assumption A2, this does not present a difficulty. While inference on usual regression (weighted or not) depends heavily on normality, the estimation of the parameters β does not. The estimates found by least squares regression are a form of robust estimation known as *M-estimation*. For details on regression as a robust M-estimator, see Huber and Ronchetti (2009).
12. *Sample Size n*. The automatic span selection method based on generalized cross-validation searches for values of α in the interval [0.05, 0.95]. However, cases may occur where the variability of the underlying function f requires a span below 0.05. In such situations, manual selection of the smoothing span α may be necessary. See Comment 5.

Properties

1. For the sampling distributions associated with varying assumptions on the errors ε, see Cleveland (1979).

Problems

8. Smooth the data set `cars` from Problem 1 using Cleveland's smoother. Choose a reasonable value for the span using trial and error. Why does this span appear to be a good choice?

Comment on the graphical comparison of the estimate using the trial and error span with the estimates found using the span chosen by the two cross-validation validation methods (`aicc` and `gcv`).

9. Comment on the graphical comparison between the estimates obtained with the automatically selected spans in Problem 8 and the estimate obtained from Friedman's smoother using the cross-validation span in Problem 1.

10. Through trial and error, find a span that seems to work well with the `sunspots` data from Problem 2 using Cleveland's method. Then find the estimates using the spans determined by the two cross-validation methods. Describe the results (taking into account Comment 12).

11. Reduce the size of the `sunspots` data set by taking only the first m observations as was done in Problem 3. At what value of m do the cross-validation methods begin to provide reasonable estimates? Why is this occurring?

12. Using the first $m = 128$ observations from the `sunspots` data in Problem 2, obtain the wavelet estimate using the VisuShrink method from Chapter 13. Comment on the graphical comparison between the VisuShrink estimate and the estimate you would have gotten in Problem 11 by setting $m = 128$. The thresholded wavelet estimate requires a dyadic number n of points x_i, say $n = 2^J$ for some integer J. Additionally, it assumes that these points are equally spaced apart. These two conditions are met by the truncated `sunspots` data. However, VisuShrink is designed for normal errors ε, while Clevelands's estimator makes no such assumption.

13. Apply Cleveland's method to Canadian wage data from Problem 5. Obtain an estimate using a span found by trial and error. Then obtain estimates using the two cross-validation span selectors. Comment on the graphical comparison between these estimates.

14. Comment on the graphical comparison of the estimates obtained with the two automatically selected spans in Problem 13 to the estimate obtained in Problems 5 and 6 using Friedman's method with cross-validation span and an appropriate value for the bass.

15. Change the degree of the local polynomial in Cleveland's method to 0 (locally constant) using the argument `degree` (see Comment 8). Using the `cars` data from Problem 1, how does changing the degree affect the estimates compared to the locally linear estimates (`degree=1`) from Problem 8? Change the degree to 2 (locally quadratic) and compare graphically the estimate obtained to those found using degrees 0 and 1.

16. Change the degree of the local polynomial in Cleveland's method as in Problem 15, but use the data Canadian census data `cps71` from Problem 5. Comment on the graphical comparison between these estimates.

17. Using the data from Table 9.3 on body weight and total surface area of squirrel monkeys, find and plot the estimated line using Theil's method. Use surface area as the response and weight as the predictor. Then find the estimate using Cleveland's method with the spans chosen by cross-validation. Comment on the graphical comparison between these estimates.

14.3 KERNEL SMOOTHING

In Chapter 12, we introduced the kernel function as a tool for density estimation. We return to the idea of kernels, but now in the context of estimating a general function f rather than a density function. Nadaraya (1964, 1965) and Watson (1964) independently introduced the kernel regression estimate

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\dfrac{x - x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\dfrac{x - x_j}{h}\right)} \tag{14.2}$$

for observations gathered under this chapter's assumptions. The kernel function K obeys the same restrictions as those set forth in Chapter 12. The parameter h is the bandwidth. The estimate (14.2) can be rewritten as

$$\hat{f}(x) = \sum_{i=1}^{n} y_i \left[\frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)} \right] = \sum_{i=1}^{n} y_i w_i$$

illustrating that this estimate is a linear estimate of the observed data y_i. The weights w_i are dependent on the choice of kernel function K, the bandwidth h, and the distance between the observed data x_i and the point x. An estimate of the fitted value $\hat{y}_i$ is given by replacing x with x_i in (14.2),

$$\hat{y}_i = \hat{f}(x_i).$$

Unlike the methods in the previous two sections, this is not a nearest neighbor method for a given kernel K and bandwidth h. The number of observations used in the estimate at any point x is not fixed but the window size is. See Comment 13.

Other kernel regression methods have been proposed, see Comment 16. These methods are similar to the Nadaraya–Watson estimator in that the kernel is used to weight the observed data y_i. Other kernel methods have been developed where the kernel is used as weights for local linear least squares. See Comment 15.

As in the case of density estimation, the importance of the bandwidth h in obtaining a good estimate of f is critical. Instead of relying on choosing h based on subjective means, an automated method for bandwidth selection is desired. One such method is the AIC method of Hurvich et al. (1998) used earlier with Cleveland's local weighted regression smoother. See Comment 9. A second method is the simple leave-one-out least squares cross-validation similar to that in Comment 2. In this case, the bandwidth h is the changing value over which the least squares are minimized, rather than the span s.

EXAMPLE 14.3 ***Nitrogen Oxide Concentrations.***

The R command `npreg` from the Hayfield and Racine (2008) package `np` will implement the Nadaraya–Watson kernel regression estimator. It requires arguments for the observed data x and y and a bandwidth. As with `loess`, the data must be sorted with respect to the x variable, which in this case is the engine ratio variable. For the ethanol data from the previous examples, the estimate is obtained and stored in the R object `ethanol.npreg` by

```
ethanol.npreg <- npreg(bws=.09, txdat=ethanol$E,
  tydat=ethanol$NOx)
```

The arguments `txdata` and `tydata` identify the observed data x and y. The argument `bws` specifies the bandwidth. By default, the kernel used is the normal density centered at 0 with standard deviation σ equal to the bandwidth h. The results of running the above command with three different bandwidths ($h = 0.01, 0.03, 0.05$) are shown in Figure 14.4. The estimated function is obtained with the command `fitted(ethanol.npreg)`. The figure clearly displays how increasing the bandwidth results in smoother estimates. However, the two larger bandwidths seem to be oversmoothing the data, while with $h = 0.01$, the estimate appears to be overly

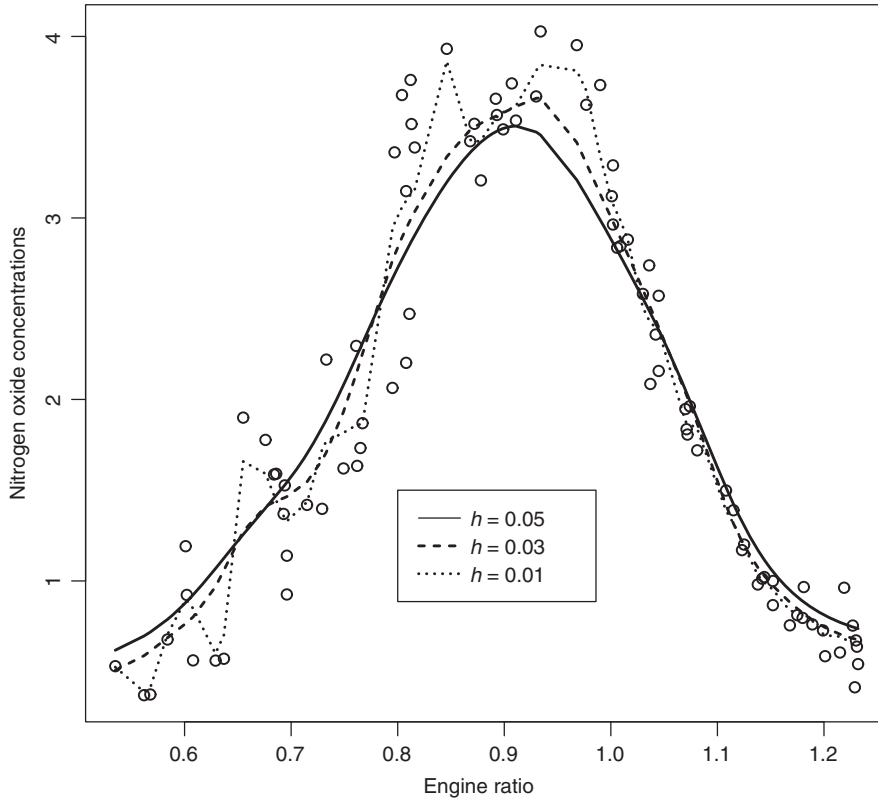

Figure 14.4 Smoothing of the `ethanol` data from Brinkman (1981) using the Nadaraya–Watson estimator with varying bandwidths h and the normal kernel.

influenced by the variability of the observed data. Which bandwidth is preferable is once again a choice in weighing bias versus variance.

The kernel K may be changed, as well. The argument setting the kernel in `npreg` is `ckertype`, which refers to the continuous kernel type. Choices available are `ckertype="gaussian"`, `"epanechnikov"` or `"uniform"`. In Chapter 12, this last kernel is called the `"rectangular"` when performing kernel density estimation with `density`. Additionally, one may specify the order of the kernel. See Comment 14. Figure 14.5 shows the results of using the three different kernels with a fixed bandwidth of $h = 0.03$. Each estimate is approximately the same shape, following the same general pattern. Differences are apparent, though. In particular, the uniform kernel provides a less smooth estimate than the other two kernels. As in the case of density estimation, the choice of bandwidth is more important than the choice of kernel.

The two methods of automatic bandwidth selection are implemented with the argument `bwmethod`. Setting `bwmethod="cv.aic"` or `"cv.ls"` will choose the AIC or simple least squares cross-validation method, respectively. The bandwidth chosen by this method for the Nadaraya–Watson regression estimate of the ethanol data is $h = 0.019$. The fit with this bandwidth is shown in Figure 14.6 as the solid line. For comparison, finding the estimate using `cv.ls` results in a bandwidth of $h = 0.016$ and is shown as the dashed line. The estimators using the cross-validation bandwidths do not appear to be oversmoothing the data or including unnecessary variability, both of which are problems with the arbitrarily selected bandwidths in Figure 14.4.

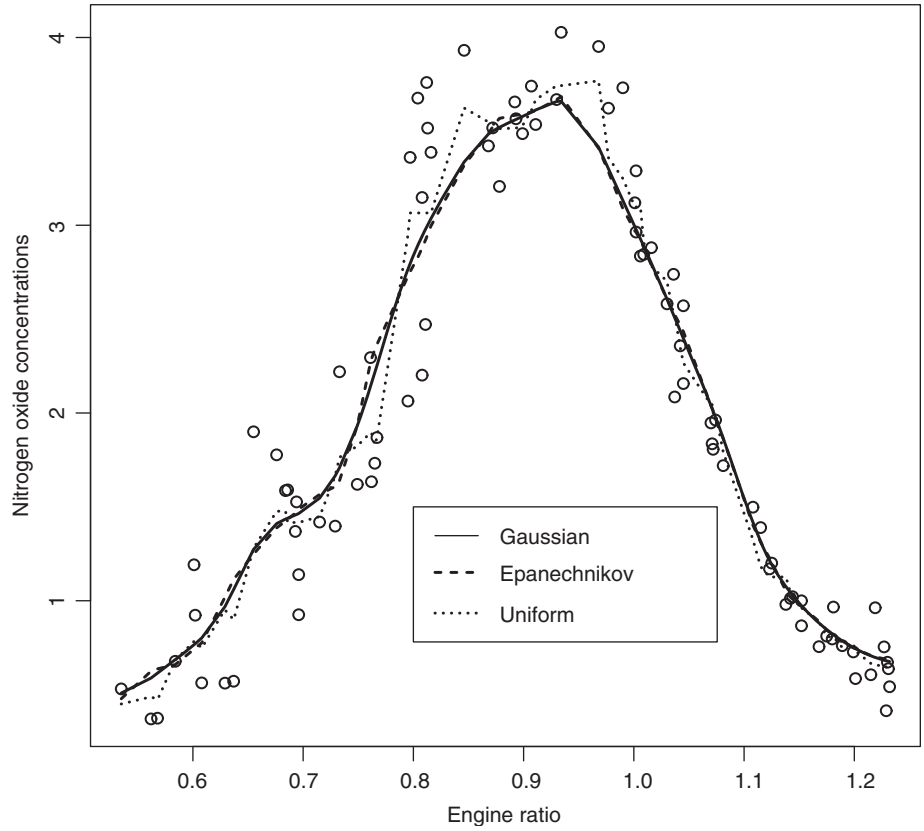

Figure 14.5 Smoothing of the `ethanol` data from Brinkman (1981) using the Nadaraya–Watson estimator with varying kernel functions K and fixed bandwidth $h = 0.03$.

Figure 14.7 displays a comparison of estimates of the relation between engine ratio and nitrous oxide concentrations for each of the methods discussed in this chapter. All methods are shown using cross-validation span or bandwidth selection. For Cleveland's estimate, generalized cross-validation is used. For the Nadaraya–Watson estimate, least squares cross-validation is employed with the normal kernel and fixed bandwidth. The estimates are quite similar. The main difference appears in the region around an engine ratio of about 0.7. Cleveland's robust estimator smooths over a feature that each of the other two methods pick up.

Comments

13. *Nearest Neighbor Estimates.* The kernel regression estimator can become a nearest neighbor method by letting the bandwidth h vary to include a specified number of points in the window. There are two methods by which this may be accomplished. An adaptive bandwidth varies h_i, a bandwidth for each point x_i in the set of observed data. A generalized nearest neighbor approach varies h_x, a bandwidth for each point x at which one wishes to estimate the function f. In (14.2), for example, replace h with either h_i or h_x.

    The `R` command `npreg` implements this type of bandwidth through the argument `bwtype`. The default value is `bwtype="fixed"`, a single

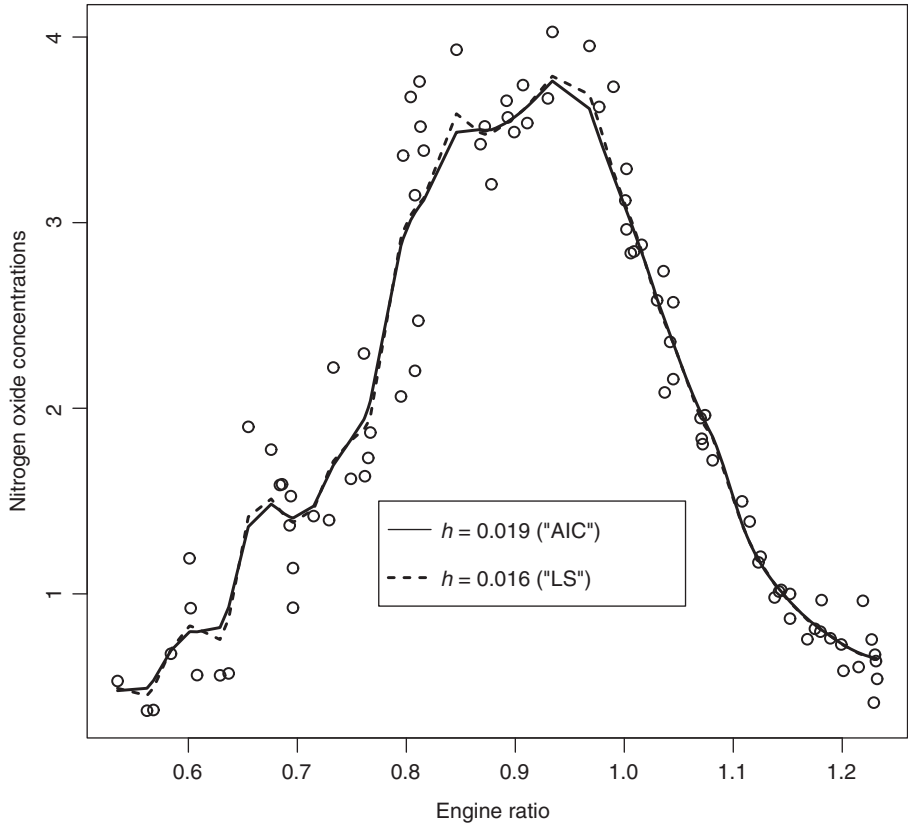

Figure 14.6 Smoothing of the `ethanol` data from Brinkman (1981) using the Nadaraya–Watson estimator with Gaussian kernel functions K and bandwidth chosen as in Hurvich et al. (1998).

bandwidth h for the entire estimate. Using `bwtype="adaptive_nn"` or `"generalized_nn"` will provide the number of nearest neighboring points to include in the corresponding kernel window.

14. *Kernel Order.* The order of a kernel K is defined in terms of its moments m_K^i:

$$m_K^i = \int_{-\infty}^{\infty} x^i K(x) dx.$$

For a symmetric kernel, $m_K^i = 0$ for all odd integers i. The order of K is i, where i is the smallest integer such that $m_K^i > 0$. The normal, uniform, and Epanechnikov kernels are all of order 2. If $m_K^i > 2$, the kernel is said to be a higher order kernel. Higher order kernels are formed from lower order kernels through multiplication by an appropriate polynomial. These higher order kernels are useful in kernel regression in that for large n the error between the true underlying function f and the kernel regression estimate is smaller than that for low order kernels. Additionally, there is some reduction in the bias of the estimate when using higher order kernels. A drawback to using higher order kernels is the introduction of negative weights. For a more complete discussion of higher order kernel regression, see Marron (1994). The command `npreg` allows for modifying the order of the kernel function. This is done with the

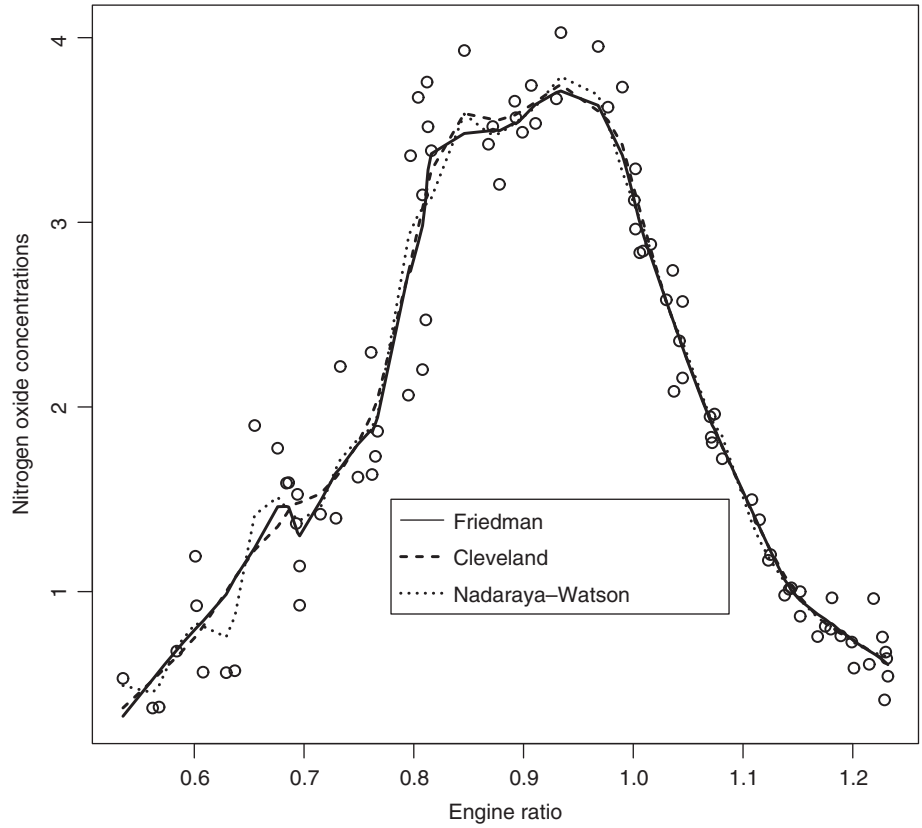

Figure 14.7 Comparison of the three estimates discussed in Sections 14.1–14.3.

argument `ckerorder` (continuous kernel order). Allowable values are $2, 4, 6$, or 8, with 2 as the default value.

15. *Local Linear Kernel Regression.* The kernel regression estimator may be extended to a local linear kernel estimator. The idea is to find values for β_0 and β_1 that minimize the equation

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1(x - x_i))^2 K\left(\frac{x - x_i}{h}\right). \tag{14.3}$$

This is a weighted local linear regression problem, similar to that specified in Cleveland (1979) and Cleveland and Devlin (1988) (compare to (14.1) in Comment 6). If β_1 is set to 0, the solution to (14.3) is the Nadaraya–Watson estimator (14.2). Thus, the Nadaraya–Watson kernel regression estimator is a locally weighted *constant* regression, rather than locally weighted *linear* regression. In the context of Cleveland (1979), the locally weighted constant regression is akin to setting the degree of the polynomial to 0. Of course, Cleveland's estimator differs substantially in that it iterates many weighted regressions in order to be robust against overly influential observations. For details and properties of

local linear kernel regression estimators, see Fan (1992, 1993), Li and Racine (2004), Racine and Li (2004), and Ruppert and Wand (1994).

The `R` command `npreg` allows for local linear kernel regression. One can specify either the Nadaraya–Watson estimator using the argument `regtype="lc"` (local constant) or `regtype="ll"` (local linear). The independent (predictor) variables x_i are assumed to be continuous in Assumption A2. However, other types of predictor variables are considered and implemented in the `R` package `np`. For details on these noncontinuous x_i and the cross-validation bandwidth selection methods associated with these `R` commands, see Li and Racine (2004) and Racine and Li (2004).

16. *Other Kernel Estimators.* Priestley and Chao (1972) proposed a kernel estimator that incorporates the distance between adjacent observed points x_i in addition to weighting with a kernel function K,

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{y_i(x_i - x_{i-1})}{h} K\left(\frac{x - x_i}{h}\right),$$

where it is assumed that $x_{i-1} \le x_i$. For convergence results, asymptotic properties and a choice of optimal kernel, see Priestley and Chao (1972) and Benedetti (1977).

Gasser and Müller (1979) proposed a method that integrated the kernel in a small neighborhood of x_i,

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{y_i}{h} \int_{s_{i-1}}^{s_i} K\left(\frac{x - t}{h}\right) dt,$$

where $x_i \le s_i \le x_{i+1}$. A simple choice for the s_i are the midpoints,

$$s_i = \frac{x_i + x_{i+1}}{2}.$$

Properties and optimal choices for kernels are found in Gasser and Müller (1979) and Müller (1984). The Gasser–Müller kernel estimator has larger variance than the Nadaraya–Watson estimator but smaller bias. See Fan (1992).

17. *Derivation of the Nadaraya–Watson Estimator.* Although the Nadaraya–Watson estimator can be thought of as a special case of a locally weighted linear kernel estimator (see Comment 15), it was derived in a different manner. If (X_i, Y_i) are independent pairs of random variables, then the solution to the nonparametric regression problem

$$Y_i = m(X_i) + \varepsilon_i$$

is

$$\hat{m}(x) = E(Y|X = x) = \int y f_1(y|x)dy = \frac{\int y f(x,y)dy}{f_2(x)}, \tag{14.4}$$

where $f(x,y)$ is the joint density of X and Y, $f_1(y|x)$ is the conditional density of Y given $X = x$ and $f_2(x)$ is the marginal density of X. The denominator $f_2(x)$

may be estimated from the observed data with a kernel density estimator (12.14) as

$$\hat{f}_2(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

The joint density f_1 may be estimated with

$$\hat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) K\left(\frac{y - Y_i}{h}\right),$$

see Scott (1992). The Nadaraya–Watson estimator is found by taking the ratio. See Problem 18. The local average smoother and local linear estimators may also be considered to be solutions of (14.4). See Friedman (1984) and Fan (1992), for example. Because the Nadaraya–Watson estimator is the ratio of two correlated density estimators, its properties are more difficult to determine, at least when compared to the univariate density estimates in Chapter 12.

Properties

1. For the bias, variance, and asymptotic properties of the Nadaraya–Watson estimator, see Nadaraya (1964, 1965), Watson (1964), Bierens (1987), and Fan (1992).

Problems

18. Using the kernel properties specified in Chapter 12, show that (14.4) in Comment 17 is equivalent to the Nadaraya–Watson estimator, that is, show

$$\frac{\sum_{i=1}^{n} y_i K\left(\frac{x - x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)} = \frac{\int y \hat{f}(x, y) dy}{\hat{f}_2(x)}.$$

19. Using the `cars` data from Problem 1, find the Nadaraya–Watson estimate with the normal kernel and a variety of bandwidths. By trial and error, determine a bandwidth that seems most reasonable. Then find an estimate using a bandwidth selected by cross-validation. Comment on the graphical comparison of these estimates.

20. Change the kernel K in Problem 19 to the uniform kernel, but continue using a cross-validation bandwidth selector. Then change K to the Epanechnikov kernel. Comment on the graphical comparison of these three estimates using different kernels.

21. Find the Nadaraya–Watson estimate using the normal kernel on the `sunspots` data from Problem 2. Use bandwidths selected by trial and error and then using the cross-validation methods. Comment on the graphical comparison of these estimates. Comment on the cross-validation estimates from this problem with the cross-validation estimates found in Problems 2 and 10.

22. Find the local linear kernel regression estimate using the `cars` data from Problem 1 by changing the regression type with the argument `regtype="ll"`. Use cross-validation bandwidths. Comment on the graphical comparison of this estimate to the estimates obtained in Problem 19 with cross-validation bandwidths.

23. Comment on the graphical comparison of the estimates from Problem 22 to the local linear estimates found using Cleveland's method in Problem 8.

24. Using the first $m = 2048$ observations from the `sunspots` data in Problem 2, obtain the wavelet estimate using the VisuShrink method from Chapter 13. Compare the VisuShrink estimate to the estimate you would have gotten in Problem 21 by using only the first 2048 observations:

```
y <- sunspots[1:2048]
x <- c(1:2048)
```

The basic wavelet threshold estimate requires a dyadic number of points x_i ($n = 2^J$ for some integer J) and that these points be equally spaced apart. These conditions are met by the truncated `sunspots` data. However, VisuShrink is designed for normal errors ε, while the Nadaraya–Watson estimator makes no such assumption.

25. Using the Canadian census data from Problem 5, find the Nadaraya–Watson estimate of the relation between the dependent variable `logwage` and the independent variable `age`. Use nearest neighbor bandwidth selection methods (see Comment 13) and fixed bandwidth selection methods. Comment on the graphical comparison of these estimates. Also compare these estimates to those obtained in Problems 5 and 13.

26. Using the data from Table 9.3 on body weight and total surface area of squirrel monkeys, find and plot the estimated line using Theil's method. Use surface area as the response and weight as the predictor. Then find the Nadaraya–Watson estimate of this relation with the span chosen by cross-validation. Comment on the graphical comparison between these estimates.

14.4 OTHER METHODS OF SMOOTHING

This chapter discussed three methods of smoothing: Friedman's local averaging smoother, Cleveland's locally weighted linear regression estimator, and the Nadaraya–Watson kernel estimator. However, there are many other types of smoothing estimators currently in use. Tukey (1977) proposed smoothing the data using running medians. The fitted value $\hat{y}_i$ is estimated using the median of the values within a small neighborhood. The use of medians was implemented in order to be robust against outliers, much as Cleveland's estimator attempts to be robust through iterative weighted regressions. Splines, introduced by Schoenberg (1946), are polynomial functions which have found wide use in the problem of smoothing. For details on the many methods of statistical spline smoothing, see Wegman and Wright (1983), Wahba (1990), Eubank (1999), and Wang (2011), for example. Orthogonal series methods such as wavelets and Fourier series are also used for smoothing. See Chapter 13 for a discussion of wavelets.

The methods presented in this chapter have been univariate: the response y depends only on a single x. More generally, one may consider y to depend on $p > 1$ independent variables $x_1, x_2, \ldots, x_p$. Cleveland and Devlin (1988) and Ruppert and Wand (1994) proposed an extension of locally weighted regression. See Ruppert and Wand (1994) and Härdle and Müller (2000) for a discussion of multivariate extension of the univariate kernel regression methods described in this chapter. Multivariate spline methods exist, as well. See, for example, Cox (1984) and Friedman (1991). And, finally, Hastie and Tibshirani (1986) introduced a multivariate method using multiple smoothers, one for each of the p predictors. The method is iterative, using the backfitting algorithm of Friedman and Stuetzle (1981) to estimate the p individual smoothers.

Chapter 15

Ranked Set Sampling

INTRODUCTION

This chapter is devoted to the concept of ranked set sampling, a technique for data collection that generally leads to more efficient statistical procedures than competitors based on simple random samples. The rationale behind ranked set sampling and its historical development is provided in Section 15.1. In Section 15.2, we describe how to collect a ranked set sample and discuss (see Comment 1) some of the structural differences between ranked set samples and simple random samples. We illustrate the application of ranked set sampling to the estimation of a population mean in Section 15.3 and present the ranked set sample analog of the two-sample Mann–Whitney–Wilcoxon test procedure (see Section 4.1) in Section 15.4. Other important issues for ranked set sampling are discussed in Section 15.5, and a number of recent developments in statistical inference with similarities to ranked set sampling are discussed briefly in Section 15.6.

15.1 RATIONALE AND HISTORICAL DEVELOPMENT

When we collect a simple random sample (SRS) from a population, what makes associated statistical inference procedures appropriate is not the fact that each individual measurement in the sample is likely to be representative of the population characteristic of interest. Rather, it is through the concept of the sampling distributions of the relevant statistics that we should, "on the average," obtain a set of sample observations that are truly representative of the full population. In practice, however, we obtain only a single random sample and the on-the-average concept does not help much if the particular population items selected for our sample are, in fact, not really very representative of the full population.

There are a number of ways to address the concerns associated with the possibility of obtaining an unrepresentative SRS. Most of them involve using additional information about the population to *a priori* partition it into more homogeneous subgroups that are designed to more fully cover the entire population. SRSs are then collected independently from these subgroups to form a more structured overall set of sample data that, by design, will more likely be representative of the entire population. Such approaches include stratified, cluster, and proportional sampling schemes.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

Ranked set sampling (RSS) is a sampling approach that also uses additional information to provide more structure to the collected sample items. In RSS, however, this additional information is not used to partition the full population prior to the collection of appropriate SRSs. Rather, in RSS, potential SRSs are selected directly from the full population and then auxiliary population information is used to impose an "artificially poststratified" structure that enables us to collect measurements from units that are more likely to represent the full spectrum of values in the population.

The concept of RSS was first proposed by McIntyre (1952) (reprinted in 2005) for situations where taking the actual measurements for sample observations is difficult (e.g., costly, destructive, time consuming), but mechanisms for either informally or formally ranking a set of sample units with regard to the aspect of interest are relatively easy and reliable. Takahasi and Wakimoto (1968) and Dell and Clutter (1972) were the first to provide some of the basic properties for statistical procedures based on RSS data. For additional information about the historical development of RSS methodology, see the review paper by Patil (1995).

15.2 COLLECTING A RANKED SET SAMPLE

The goal of RSS is to collect observations from a population that are more likely to span the full range of values in the population (and, therefore, be more representative of it) than the same number of observations obtained via simple random sampling. What is a ranked set sample (RSS) and how do we collect it? For ease of discussion, we assume throughout this chapter that all sampling is from an infinite population or with replacement from a finite population. (For information about RSS without replacement from a finite population, see, for example, Patil, Sinha, and Taillie (1995, 1999).)

To obtain an RSS of k observations from a population, we proceed as follows. First, an initial SRS of k units is selected from the population and rank-ordered on the attribute of interest. A variety of mechanisms can be used to obtain this ranking, including visual comparisons, expert opinion, or through the use of auxiliary variables, but it cannot involve actual measurements of the attribute of interest on the sample units. The unit that is judged to be the smallest in this ranking is included as the first item in the RSS and the attribute of interest is formally measured for the unit. This initial measurement is called the *first judgment order statistic* and is denoted by $X_{[1]}$, where square brackets are used instead of the usual parentheses (1) for the smallest order statistic because $X_{[1]}$ may or may not actually have the smallest attribute measurement among the k units in the SRS, even though our ranking judged it to be the smallest. The remaining $k - 1$ units (other than $X_{[1]}$) in our initial SRS are not considered further in making inferences about the population—their role was solely to assist in the selection of the smallest ranked unit for measurement.

Following the selection of $X_{[1]}$, a second SRS (independent of the first SRS) of size k is selected from the population and ranked in the same manner as the first SRS. From this second SRS we select the item ranked as the second smallest of the k units (i.e., the second judgment order statistic) and add its attribute measurement, $X_{[2]}$, to the RSS. From a third SRS (independent of both the previous SRSs) of size k, we select the unit ranked to be the third smallest (i.e., the third judgment order statistic) and include its attribute measurement, $X_{[3]}$, in the RSS. This process continues until we have selected

the unit ranked to be the largest of the k units in the kth independent SRS and included its attribute measurement, $X_{[k]}$, in our RSS.

This entire process results in the k measured observations $X_{[1]}, \ldots, X_{[k]}$ and is called a *cycle*. The number of units, k, in each SRS is called the *set size*. Thus to complete a single ranked set cycle, we need to use a total of k^2 units from the population to separately rank k independent SRSs of size k each. The measured observations $X_{[1]}, \ldots, X_{[k]}$ constitute a *balanced ranked set sample of size* k, where the descriptor "balanced" refers to the fact that we have collected one judgment order statistic for each of the ranks $1, 2, \ldots, k$.

To obtain a balanced RSS with a desired total number of measured observations (i.e., sample size) $n = km$, we repeat the entire process for m independent cycles, yielding the following balanced RSS of size n:

Cycle 1	$X_{[1]1}$	$X_{[2]1}$	$X_{[3]1}$	$\ldots$	$X_{[k]1}$
Cycle 2	$X_{[1]2}$	$X_{[2]2}$	$X_{[3]2}$	$\ldots$	$X_{[k]2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Cycle m	$X_{[1]m}$	$X_{[2]m}$	$X_{[3]m}$	$\ldots$	$X_{[k]m}$

EXAMPLE 15.1 ***Obtaining a Ranked Set Sample—Gasoline Reid Vapor Pressure.***

Unburned hydrocarbons emitted from automobile tailpipes and via evaporation from manifolds are among the primary contributors to ground-level ozone and smog levels in large cities. One way to reduce the effect of this factor on air pollution is through the use of reformulated gasoline, designed to reduce its volatility, as measured by the Reid Vapor Pressure (RVP) value. To assure that gasoline stations in metropolitan areas are selling gasoline that complies with clean air regulations, regular samples of reformulated gasoline from the pumps at these stations are collected and RVP values are measured.

The RVP value for a sample can be measured either by a crude field technique right after collection at the gasoline pump or via a more sophisticated analysis after the sample has been shipped to a government laboratory. While the actual laboratory analysis of RVP is not overly expensive, it is costly to ship these gasoline samples to the laboratory, since they must be packed to prevent gaseous hydrocarbons from escaping en route and special transport measures are required for flammable liquids such as gasoline. It would be beneficial to use these cruder, less expensive, field RVP measurements as reliable surrogates for the more expensive laboratory RVP measurements in order to reduce the required number of formal laboratory tests without significant loss of accuracy, resulting in considerable cost savings.

Nussbaum and Sinha (1997) suggested the use of RSS as an aid in achieving this goal. Thirty-six of the field RVP measurements (collected from the pumps) considered by Nussbaum and Sinha are given in the following table.

Sample number	Field RVP value	Sample number	Field RVP value
1	7.60	19	7.85
2	9.25	20	7.86
3	7.73	21	7.92
4	7.88	22	7.95
5	8.89	23	7.85
6	8.88	24	7.95
7	9.14	25	7.98
8	9.15	26	7.80
9	8.25	27	7.80
10	8.98	28	8.01
11	8.63	29	7.96
12	8.62	30	7.86
13	7.90	31	8.89
14	8.01	32	7.89
15	8.28	33	7.73
16	8.25	34	9.21
17	8.17	35	8.01
18	10.72	36	8.32

Source: B. D. Nussbaum and B. K. Sinha (1997).

Nussbaum and Sinha recommended using these field RVP values (highly correlated with the more precise laboratory measurements) to provide the ranking mechanism for selection of a much smaller subgroup of gasoline samples to submit for follow-up laboratory analysis. They considered a set size of $k = 3$ with $m = 4$ cycles, which leads to an RSS of only $n = 12$ gasoline samples to send for full laboratory RVP measurement.

To select this RSS, using a set size of $k = 3$, of 12 gasoline samples to be sent to the laboratory for more precise RVP measurements, the first thing we must do is to randomly divide the 36 gasoline samples into 12 sets of three each. For this purpose, we use the R command `sample(1:36, 36, replace = F)` to obtain the following random ordering of the sample numbers 1–36 clustered into 12 sets of size $k = 3$ each based on their order of appearance:

(10, 13, 23) (11, 12, 17) (16, 2, 21) (15, 18, 36) (34, 27, 5) (24, 31, 14)
(22, 35, 19) (30, 9, 4) (28, 32, 8) (33, 26, 29) (7, 1, 6) (3, 20, 25)

Next we must decide which four sets will be used to obtain the smallest judgment ordered units, which four will be used to obtain the median judgment ordered units, and which four will be used to obtain the largest judgment ordered units. There is complete flexibility here, but these decisions must be made without knowledge of the actual field RVP values in the 12 sets. For the sake of illustration here, we choose to select the minimum judgment ordered unit from the first four sets, the median judgment ordered unit from the second four sets, and the largest judgment ordered unit from the final four sets.

The 12 sets of three RVP values each that result from our sampling process are given in the following table:

8.98	7.90	7.85	8.63	8.62	8.17	8.25	9.25	7.92
8.28	10.72	8.32	9.21	7.80	8.89	7.95	8.89	8.01
7.95	8.01	7.85	7.86	8.25	7.88	8.01	7.89	9.15
7.73	7.80	7.96	9.14	7.60	8.88	7.73	7.86	7.98

Using our chosen criteria for selecting the judgment ordered units, we see that the units selected by our RSS scheme for shipment to the laboratory for precise RVP measurements are those gasoline samples corresponding to the bold field RVP values in the following table:

8.98	7.90	**7.85**	8.63	8.62	**8.17**	8.25	9.25	**7.92**
8.28	10.72	8.32	9.21	7.80	**8.89**	7.95	8.89	**8.01**
7.95	8.01	7.85	7.86	8.25	**7.88**	8.01	7.89	**9.15**
7.73	7.80	**7.96**	**9.14**	7.60	8.88	7.73	7.86	**7.98**

Thus we will send gasoline samples 23, 17, 21, 15, 5, 14, 22, 4, 8, 29, 7, and 25 to the laboratory for more precise RVP determinations, and the resulting laboratory RVP values will constitute our balanced RSS of size $n = 12$ based on a set size of $k = 3$ with $m = 4$ cycles, using the field RVP value as our auxiliary ranking variable.

Comments

1. *Comparison of Ranked Set Samples and Simple Random Samples.* A balanced RSS of size n differs from an SRS of the same size in a number of important ways. An SRS is designed so that the n observations in the sample are mutually independent and identically distributed. Probabilistically speaking, it means that each of the individual sample items represents a typical value chosen from the underlying population. That is not the case for a balanced RSS of size n. While the individual observations in a balanced RSS remain mutually independent, they are clearly not identically distributed, so that individual observations in a balanced RSS do not represent typical values from the underlying population. In fact, the individual judgment order statistics represent very distinctly different portions of the underlying population. This is a very important feature of an RSS, as the items in the sample are designed in such a way as to provide greater assurance that the entire range of population values are represented.

   This is best illustrated by considering an example. Suppose that X has a standard normal distribution and let X_1, X_2, X_3, X_4, X_5 be a random sample of size 5 from this distribution. Let $X_{(1)} \le X_{(2)} \le X_{(3)} \le X_{(4)} \le X_{(5)}$ be the associated order statistics. In Figure 15.1, we plot the underlying $N(0, 1)$ density as well as the **marginal distributions** for the five individual order statistics $X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}$, and $X_{(5)}$.

   If we use perfect rankings to collect an RSS of size 5 from the standard normal distribution, then these five RSS observations behave like **mutually independent** order statistics from the standard normal and their densities are represented by the five individual marginal density curves in Figure 15.1. While these

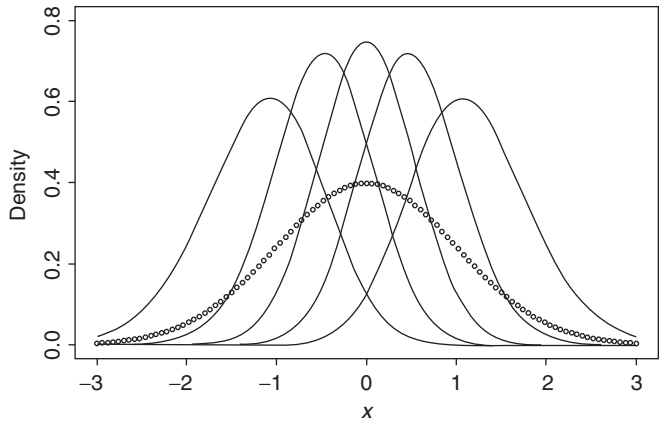

Figure 15.1 Standard normal density (dotted curve) and the individual marginal densities of the five order statistics $X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}$, and $X_{(5)}$ (solid curves, in order of peaks, from the minimum, $X_{(1)}$, on the left to the maximum, $X_{(5)}$, on the right) for a random sample of size 5 from the standard normal distribution.

five densities certainly overlap, they assign the bulk of their individual marginal probabilities to five subregions of the standard normal domain. As a result, the five RSS observations are much more likely to represent the full range of values for the standard normal distribution than would a SRS of size five; that is, the probability that the five SRS observations fail to represent the full range of the standard normal distribution is greater than the corresponding probability for the five RSS observations. As we shall see in the next section, this feature enables RSS to be more effective than SRS in the estimation of a population mean.

Problems

1. Gemayel et al. (2011) demonstrated that RSS can be an effective way to reduce auditing costs when assessing the true value of an account is time consuming (and, consequently, expensive) by allowing for smaller sample sizes than would be necessary for a SRS approach with the same precision. They point out that this is commonly the case for accounts such as inventory; accounts receivable; property, plant, and equipment; and accounts payable. In these settings, the auditor will draw a sample from a set of accounts through their subsidiary ledgers, and then proceed with on-site inspections, recalculations, confirmations, and other auditing procedures when necessary.

Tackett (2012) provided simulated sales invoice records data for an electrical/plumbing distribution center, constructed in such a way that 15% of the recorded sales invoices are fraudulent, with stated book values larger than the true audited values for the materials sold in those transactions. Table 15.1 in the NSM Third Edition R Package (sample subset included here) provides the stated book values and the true audited values for a population of 12,557 such sales invoice records, with 15% (1884) of them being fraudulently recorded. (In practice, of course, all that would be available for the auditor would be the stated book values, and the true audited values would be obtained only for those accounts that were selected for inclusion in the RSS. The auditor would not need to find the audited values for the entire population. However, we have included the true audited values in Table 15.1 for all 12,557 sales invoices so that the data set can be used to illustrate the application of RSS in such a setting.)

Select an SRS of 20 invoices from the population of 15% overstated book value accounting data given in Table 15.1. Be explicit about how you choose the invoices to include in your SRS.

Table 15.1 15% Overstated Book Value (BV) and Audited Value (AV)

Invoice number	15% Overstated BV, $	AV, $
1	889.80	889.80
2	111.94	111.94
3	259.08	259.08
4	5326.21	2322.23
5	51.55	51.55
6	310.54	310.54
7	64.70	64.70
8	1207.81	1207.81
9	193.68	72.27
10	1581.39	1581.39
11	2013.36	2013.36
12	2151.67	2151.67
13	508.60	508.60
14	196.80	196.80
15	62.30	62.30
16	5495.16	5495.16
17	582.44	582.44
18	71.27	71.27
19	1531.85	1531.85
20	1631.94	1631.94

Source: Personal Communication from J. A. Tackett (2012).
The complete population of 12,557 sales invoice records is available in the NSM Third Edition R Package.

2. Consider the population of 15% overstated book value accounting data given in Table 15.1. (See Problem 1 for more discussion about these data.) Using the stated book values to perform your judgment rankings, select a balanced RSS of 20 invoices for auditing using a set size of $k = 5$. How many cycles m did you need to obtain the RSS? How many invoice book values did you use in total for the judgment rankings leading to the selection of the invoices to audit?

3. Consider the SRS and RSS of size 20 each obtained in Problems 1 and 2, respectively, and list the audited (true) invoice value for each of these units. Compare and contrast the SRS and RSS as far as being representative of the entire population of audited (true) values.

4. Consider the population of 15% overstated book value accounting data given in Table 15.1. (See Problem 1 for more discussion about these data.) Using book values to perform your judgment rankings, select a balanced RSS of 96 invoices for auditing using a set size of $k = 3$. How many cycles m did you need to obtain the RSS? How many invoice book values did you use in total for the judgment rankings that led to the selection of the invoices to audit?

5. Consider the population of 15% overstated book value accounting data given in Table 15.1. (See Problem 1 for more discussion about these data.) Using book values to perform your judgment rankings, select a balanced RSS of 96 invoices for auditing using a set size of $k = 4$. How many cycles m did you need to obtain the RSS? How many invoice book values did you use in total for the judgment rankings leading to the selection of the invoices to audit?

6. Consider the population of 15% overstated book value accounting data given in Table 15.1. (See Problem 1 for more discussion about these data.) Using book values to perform your judgment rankings, select a balanced RSS of 96 invoices for auditing using a set size of

$k = 6$. How many cycles m did you need to obtain the RSS? How many invoice book values did you use in total for the judgment rankings leading to the selection of the invoices to audit?

7. We would like to collect a balanced RSS of 24 observations from a population. List all of the (set size, cycle size) combinations that could be used to obtain this balanced RSS of size 24. For each of these combinations, calculate the total number of sample units that need to be included in the judgment rankings. Discuss the pros and cons of these options.

8. You have been asked to justify the use of a balanced RSS as opposed to an SRS of the same size to obtain information about a population. Discuss the pros and cons of these two approaches to data collection.

9. The Third National Health and Nutrition Examination Survey (NHANES III, 1988–1994) was conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention. This survey was designed to obtain nationally representative information on the health and nutritional status of the population of the United States. Specifically, it contains various body measurements and information on other health-related variables for the respondents. We consider here a portion of this data set for those 13,267 NHANES III participants who were at least 21 years old and not pregnant at the time they completed their surveys.

Body mass index (BMI), which is commonly used to classify an adult's weight status, will be the variable of interest in this problem. It is calculated as the ratio of weight (kg) to height squared (m^2). Table 15.2 in the NSM Third Edition R package (sample subset included here) provides the following NHANES III data for each of these 13,267 subjects: Gender, Age, BMI, Arm Circumference (ArmCir), Buttocks Circumference (ButtocksCir), and Thigh Circumference (ThighCir).

Select an SRS of 25 subjects from this population of NHANES III participants represented in Table 15.2. Be explicit about how you choose the individuals to include in your SRS.

10. Consider the population of NHANES III data given in Table 15.2. (See Problem 9 for more discussion about the NHANES III study.) Using arm circumference as the auxiliary variable to perform your judgment rankings, select a balanced RSS of 15 subjects using a set size of $k = 3$. How many cycles m did you need to obtain the RSS? How many subjects did you use in total for the judgment rankings leading to the selection of the subjects to include in your RSS?

11. Consider the population of NHANES III data given in Table 15.2. (See Problem 9 for more discussion about the NHANES III study.) Using buttocks circumference as the auxiliary variable to perform your judgment rankings, select a balanced RSS of 24 subjects using a set size of $k = 8$. How many cycles m did you need to obtain the RSS? How many subjects did you

Table 15.2 NHANES III Data[a]

Subject	Gender	Age	BMI	ArmCir	ButtocksCir	ThighCir
1	1	21	25.5	34.9	97.7	53.5
2	2	32	23.4	32.8	98.9	46.9
3	2	48	27.6	33.3	106.3	51.1
4	1	35	29.4	36.1	105.9	57.0
5	1	48	25.0	31.2	93.6	46.5

Source: The Third National Health and Nutrition Examination Survey (NHANES III), 1988–1994, conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention.

[a]The code for gender is 1 for Male and 2 for Female, BMI is given in kilograms per square meter, and Arm Circumference (ArmCir), Buttocks Circumference (ButtocksCir), and Thigh Circumference (ThighCir) are all measured in centimeters.

Data for the complete population of 13,267 NHANES III respondents is available in the NSM Third Edition R package.

use in total for the judgment rankings leading to the selection of the subjects to include in your RSS?

12. Consider the population of NHANES III data given in Table 15.2. (See Problem 9 for more discussion about the NHANES III study.) Using buttocks circumference as the auxiliary variable to perform your judgment rankings, select a balanced RSS of 25 subjects using a set size of $k = 5$. How many cycles m did you need to obtain the RSS? How many subjects did you use in total for the judgment rankings leading to the selection of the subjects to include in your RSS?

13. Consider the SRS and RSS of size 25 each obtained in Problems 9 and 12, respectively, and list the BMI value for each of these individuals. Compare and contrast the SRS and RSS as far as being representative of the full population of BMI values.

14. Consider the population of NHANES III data given in Table 15.2. (See Problem 9 for more discussion about the NHANES III study.) Using arm circumference as the auxiliary variable to perform your judgment rankings, select a balanced RSS of 96 using a set size of $k = 6$. How many cycles m did you need to obtain the RSS? How many subjects did you use in total for the judgment rankings leading to the selection of the subjects to include in your RSS?

15. Consider the population of NHANES III data given in Table 15.2. (See Problem 9 for more discussion about the NHANES III study.) Using arm circumference as the auxiliary variable to perform your judgment rankings, select a balanced RSS of 96 using a set size of $k = 8$. How many cycles m did you need to obtain the RSS? How many subjects did you use in total for the judgment rankings leading to the selection of the subjects to include in your RSS?

16. Consider the population of NHANES III data given in Table 15.2. (See Problem 9 for more discussion about the NHANES III study.) Using buttocks circumference as the auxiliary variable to perform your judgment rankings, select a balanced RSS of 96 using a set size of $k = 6$. How many cycles m did you need to obtain the RSS? How many subjects did you use in total for the judgment rankings leading to the selection of the subjects to include in your RSS?

17. Consider the population of NHANES III data given in Table 15.2. (See Problem 9 for more discussion about the NHANES III study.) Using buttocks circumference as the auxiliary variable to perform your judgment rankings, select a balanced RSS of 96 using a set size of $k = 8$. How many cycles m did you need to obtain the RSS? How many subjects did you use in total for the judgment rankings leading to the selection of the subjects to include in your RSS?

18. We would like to collect a balanced RSS of 48 observations from a population. List all of the (set size, cycle size) combinations that could be used to obtain this balanced RSS of size 48. For each of these combinations, calculate the total number of sample units that need to be included to obtain the required judgment rankings. Discuss the pros and cons of these options.

19. Consider the RVP data presented in Table 15.3. Using the RSS approach discussed in Example 15.1 with set size $k = 2$, how many of the 36 samples obtained from the pump would we have needed to send to the laboratory for more formal RVP measurements?

20. Consider the first four (Arm Circumference, BMI) pairs in the NHANES III data given in Table 15.2, namely,

$$(34.9, 25.5), (32.8, 23.4), (33.3, 27.6), (36.1, 29.4).$$

(a) How many different SRSs of size $n = 2$ can be selected from this subset of four BMI values? List each of these possible SRSs.

(b) Using a set size of $k = 2$ and arm circumference as the auxiliary variable for the ranking process, how many different RSSs of size $n = 2$ can be selected from this subset of four BMI values? List each of these possible RSSs.

Table 15.3 Reid Vapor Pressure (RVP)

Pump sample number	Field RVP value	Laboratory RVP value
1	7.60	7.59
2	9.25	9.33
3	7.73	7.76
4	7.88	7.98
5	8.89	9.02
6	8.88	8.92
7	9.14	9.28
8	9.15	9.28
9	8.25	8.60
10	8.98	8.56
11	8.63	8.64
12	8.62	8.70
13	7.90	7.83
14	8.01	8.15
15	8.28	8.03
16	8.25	8.29
17	8.17	8.21
18	10.72	10.67
19	7.85	7.86
20	7.86	7.86
21	7.92	7.83
22	7.95	7.83
23	7.85	7.79
24	7.95	7.85
25	7.98	7.83
26	7.80	7.80
27	7.80	7.77
28	8.01	7.73
29	7.96	7.73
30	7.86	8.28
31	8.89	9.01
32	7.89	7.98
33	7.73	7.78
34	9.21	9.30
35	8.01	8.11
36	8.32	7.99

Source: B. D. Nussbaum and B. K. Sinha (1997).

15.3 RANKED SET SAMPLING ESTIMATION OF A POPULATION MEAN

Data. We obtain two sets of n observations each from a population. One set of n observations, $X_1, \ldots, X_n$, is collected as a SRS and the second set of n observations is collected as a balanced RSS, corresponding to set size k and m cycles, with $n = km$. The RSS observations from cycle 1 are denoted by $(X_{[1]1}, X_{[2]1}, \ldots, X_{[k]1})$, the RSS observations from cycle 2 are denoted by $(X_{[1]2}, X_{[2]2}, \ldots, X_{[k]2}), \ldots$, and the RSS observations from the final cycle m are denoted by $(X_{[1]m}, X_{[2]m}, \ldots, X_{[k]m})$.

Assumptions

A1. The underlying population is continuous with distribution function F, density function f, and it has finite mean, μ, and finite variance, σ^2.

A2. All $2n$ observations (n SRS observations and n RSS observations) are mutually independent.

Procedure

The natural RSS estimator, $\hat{\mu}_{\text{RSS}}$, for the population mean μ based on the balanced RSS $(X_{[1]1}, \ldots, X_{[k]1}; X_{[1]2}, \ldots, X_{[k]2}; \ldots; X_{[1]m}, \ldots, X_{[k]m})$ is simply the average of the sample observations, namely,

$$\hat{\mu}_{\text{RSS}} = \overline{X}_{\text{RSS}} = \sum_{j=1}^{m} \sum_{i=1}^{k} \frac{X_{[i]j}}{km}. \tag{15.1}$$

Properties of $\hat{\mu}_{\text{RSS}}$

Result. The balanced RSS estimator $\hat{\mu}_{\text{RSS}}$ (15.1) is an unbiased estimator for the population mean μ regardless of whether the judgment rankings are perfect or imperfect.

Dell and Clutter (1972) established this result in the general setting for set size k and m cycles without any restriction on the accuracy of the judgment rankings. We demonstrate the argument under the more restrictive additional Assumption A3 that the judgment rankings are perfect.

A3. The judgment ranking process used to obtain the RSS is perfect, so that the RSS observations are, in fact, true order statistics from the underlying population.

For simplicity in the argument, we consider only the case of a single cycle ($m = 1$), so that the total sample size n is equal to the set size k. Under the Assumption A3 of perfect rankings, we can represent the RSS observations for this setting by $X^*_{(1)}, \ldots, X^*_{(k)}$, where these k variables are mutually independent and $X^*_{(i)}, \mathrm{i} = 1, \ldots, k$, is distributed like the ith order statistic for a random sample of size k from a continuous distribution with distribution function $F(x)$ and density $f(x)$.

It follows immediately from the properties of a simple average that

$$E[\hat{\mu}_{\text{RSS}}] = E[\overline{X}_{\text{RSS}}] = \frac{1}{k} \sum_{i=1}^{k} E[X^*_{(i)}]. \tag{15.2}$$

Moreover, since $X^*_{(i)}$ is distributed like the ith order statistic for a random sample of size k from a continuous distribution with distribution function $F(x)$ and density $f(x)$ under perfect rankings, we have

$$E[X^*_{(i)}] = \int_{-\infty}^{\infty} x \frac{k!}{(i-1)!(k-i)!} [F(x)]^{i-1} [1 - F(x)]^{k-i} f(x) dx, \tag{15.3}$$

for $i = 1, \ldots, k$. Combining (15.2) and (15.3), we obtain

$$E[\overline{X}_{\text{RSS}}] = \frac{1}{k}\sum_{i=1}^{k}\left\{\int_{-\infty}^{\infty} kx\binom{k-1}{i-1}[F(x)]^{i-1}[1-F(x)]^{k-i}f(x)dx\right\}$$
$$= \int_{-\infty}^{\infty} xf(x)\left\{\sum_{i=1}^{k}\binom{k-1}{i-1}[F(x)]^{i-1}[1-F(x)]^{k-i}\right\}dx. \tag{15.4}$$

Letting $q = i - 1$ in the summation in (15.4) we see that

$$\sum_{i=1}^{k}\binom{k-1}{i-1}[F(x)]^{i-1}[1-F(x)]^{k-i} = \sum_{q=0}^{k-1}\binom{k-1}{q}[F(x)]^{q}[1-F(x)]^{(k-1)-q} = 1,$$

since the latter expression is just the sum over the entire sample space of the probabilities for a binomial random variable with parameters $k - 1$ and $p = F(x)$.

Using this fact in (15.3) we obtain

$$E[\hat{\mu}_{\text{RSS}}] = E[\overline{X}_{\text{RSS}}] = \int_{-\infty}^{\infty} xf(x)dx = \mu, \tag{15.5}$$

establishing the fact that $\hat{\mu}_{\text{RSS}}$ is an unbiased estimator for μ.

To obtain the variance of the RSS estimator $\hat{\mu}_{\text{RSS}}$, we note that the mutual independence of the $X_{(i)}^{*}{}'s, i = 1, \ldots, k$, enables us to write

$$\text{Var}(\overline{X}_{\text{RSS}}) = \frac{1}{k^2}\sum_{i=1}^{k}\text{Var}(X_{(i)}^{*}). \tag{15.6}$$

Letting $\mu_{(i)}^{*} = E[X_{(i)}^{*}]$, for $i = 1, \ldots, k$, we note that

$$E[(X_{(i)}^{*} - \mu)^2] = E[(X_{(i)}^{*} - \mu_{(i)}^{*} + \mu_{(i)}^{*} - \mu)^2]$$
$$= E[(X_{(i)}^{*} - \mu_{(i)}^{*})^2] + (\mu_{(i)}^{*} - \mu)^2$$
$$= \text{Var}(X_{(i)}^{*}) + (\mu_{(i)}^{*} - \mu)^2, \tag{15.7}$$

since the cross-product terms are zero. Combining (15.6) and (15.7) yields the expression

$$\text{Var}(\overline{X}_{\text{RSS}}) = \frac{1}{k^2}\sum_{i=1}^{k}E[(X_{(i)}^{*} - \mu)^2] - \frac{1}{k^2}\sum_{i=1}^{k}(\mu_{(i)}^{*} - \mu)^2. \tag{15.8}$$

Now, proceeding as we did with $E[\overline{X}_{\text{RSS}}]$, we see that

$$\sum_{i=1}^{k}E[(X_{(i)}^{*} - \mu)^2] = \sum_{i=1}^{k}\int_{-\infty}^{\infty} k(x-\mu)^2\binom{k-1}{i-1}[F(x)]^{i-1}[1-F(x)]^{k-i}f(x)dx$$
$$= k\int_{-\infty}^{\infty}(x-\mu)^2 f(x)\left\{\sum_{i=1}^{k}\binom{k-1}{i-1}[F(x)]^{i-1}[1-F(x)]^{k-i}\right\}dx.$$

Once again using the binomial distribution, the interior sum is equal to 1 and we obtain

$$\sum_{i=1}^{k} E[(X_{(i)}^{*} - \mu)^2] = k \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = k\sigma^2. \tag{15.9}$$

Combining (15.8) and (15.9), it follows that

$$\begin{aligned} \mathrm{Var}(\overline{X}_{\mathrm{RSS}}) &= \frac{1}{k^2} \left\{ k\sigma^2 - \sum_{i=1}^{k} (\mu_{(i)}^{*} - \mu)^2 \right\} \\ &= \frac{\sigma^2}{k} - \frac{1}{k^2} \sum_{i=1}^{k} (\mu_{(i)}^{*} - \mu)^2. \end{aligned} \tag{15.10}$$

Comparison of SRS and RSS Estimators

The SRS estimator for the population mean μ is just the sample mean $\hat{\mu}_{\mathrm{SRS}} = \overline{X} = \frac{1}{k} \sum_{j=1}^{k} X_j$ and it is well known that $E[\hat{\mu}_{\mathrm{SRS}}] = \mu$ and $\mathrm{Var}(\hat{\mu}_{\mathrm{SRS}}) = \frac{\sigma^2}{k}$. Thus, both $\hat{\mu}_{\mathrm{SRS}}$ and $\hat{\mu}_{\mathrm{RSS}}$ are unbiased estimators for the population mean. Moreover, from (15.10), it follows that

$$\begin{aligned} \mathrm{Var}(\overline{X}_{\mathrm{RSS}}) &= \frac{\sigma^2}{k} - \frac{1}{k^2} \sum_{i=1}^{k} (\mu_{(i)}^{*} - \mu)^2 \\ &= \mathrm{Var}(\overline{X}) - \frac{1}{k^2} \sum_{i=1}^{k} (\mu_{(i)}^{*} - \mu)^2 \leq \mathrm{Var}(\overline{X}), \end{aligned} \tag{15.11}$$

since $\sum_{i=1}^{k} (\mu_{(i)}^{*} - \mu)^2 \geq 0$.

Hence, in the case of perfect rankings, not only is $\overline{X}_{\mathrm{RSS}}$ an unbiased estimator but its variance is also never larger than the variance of the SRS estimator $\overline{X}$ based on the same number of measured observations. In fact, this is a strict inequality unless $\mu_{(i)}^{*} = \mu$ for all $i = 1, \ldots, k$, which is the case only if the judgment rankings are purely random.

EXAMPLE 15.2 ***Quinine Content in Cinchona Plants.***

Cinchona plants are a primary source of quinine for use in the treatment of malaria. One source for these plants is the steep hills in southern India. Sengupta, Chakravarti, and Sarkar (1951) conducted a study for the Indian government with the goal of estimating dry bark and quinine content of these plants. However, the measurement of dry bark quinine yield requires uprooting the plant, stripping the bark, and drying it until a reliable weight can be obtained. This can be a time-consuming and costly operation. Fortunately, a simultaneous investigation demonstrated that dry bark quinine yield was highly correlated with the volume of bark in the growing plant, computed as a function of the total height of the plant and girths and thicknesses of the bark at a number of different heights above ground level. All of these measurements can be made with little cost and without uprooting and destruction of the plant. This is an ideal setting for the use of RSS with

the goal of estimating the average dry bark weight (in ounces), μ_W, for cinchona plants produced in this area of southern India.

The raw data from the study by Sengupta, Chakravarti, and Sarkar (1951) are no longer available. Thus we follow the lead of Stokes (1980) and Chen, Bai, and Sinha (2004) and simulate data pairs (V, W) of cinchona plant volume V and dry bark weight W (in ounces) that are roughly representative of the cinchona plants in southern India, where the mean cinchona plant bark volume is $\mu_V = 200$ with standard deviation $\sigma_V = 10$, the mean cinchona plant dry bark weight is $\mu_W = 20.08$ with standard deviation $\sigma_W = 7.79$, and the correlation between cinchona plant bark volume and cinchona plant dry bark weight is $\rho = 0.9$. Two hundred simulated (V, W) pairs from the bivariate normal distribution with parameters $\mu_V = 200$, $\sigma_V = 10$, $\mu_W = 20.08$, $\sigma_W = 7.79$, and $\rho = 0.9$ are given in Table 15.4. We will use a balanced RSS of size $n = 20$ based on a set size of $k = 4$ with $m = 5$ cycles to obtain our RSS estimate of μ_W. Using the `R` command `sample(1:200, 80, replace = F)`, we obtain the following set of 80 random numbers selected without replacement from $1, \ldots, 200$, inclusive, listed in their order of selection:

129 151 167 114 100 155 118 128 149 184 171 119 191 002 088 106 143 012
175 061 148 003 104 197 042 107 161 105 181 127 164 157 140 133 110 048
177 056 176 044 138 053 192 008 120 113 131 031 035 116 134 126 166 160
141 011 188 025 101 071 027 081 135 153 189 182 058 152 085 196 019 159
041 144 115 162 075 111 070 091

Using this order of selection, we group these 80 random numbers into 20 sets of set size $k = 4$ each as follows:

(129, 151, 167, 114) (100, 155, 118, 128) (149, 184, 171, 119) (191, 002, 088, 106)
(143, 012, 175, 061) (148, 003, 104, 197) (042, 107, 161, 105) (181, 127, 164, 157)
(140, 133, 110, 048) (177, 056, 176, 044) (138, 053, 192, 008) (120, 113, 131, 031)
(035, 116, 134, 126) (166, 160, 141, 011) (188, 025, 101, 071) (027, 081, 135, 153)
(189, 182, 058, 152) (085, 196, 019, 159) (041, 144, 115, 162) (075, 111, 070, 091)

Without knowledge of the actual volumes for the 80 sampled cinchona plants, we decide without prejudice to select the smallest judgment order statistics from the first five of these sets, the second judgment order statistics from the second five of these sets, the third judgment order statistics from the third five of these sets, and, finally, the largest judgment order statistics from the final five of these sets.

The volumes for the 80 cinchona plants in these 20 sets of four each are as follows:

(192.56, 201.41, 196.67, 208.26)	(181.25, 190.65, 200.16, 196.85)
(204.67, 200.97, 188.99, 197.17)	(211.99, 197.96, 196.15, 185.82)
(186.95, 190.89, 188.79, 192.20)	(201.45, 194.29, 199.72, 209.02)
(207.86, 198.41, 186.93, 205.14)	(201.25, 184.23, 196.60, 211.21)
(201.76, 195.09, 198.44, 205.59)	(206.04, 198.42, 216.59, 189.79)
(204.04, 175.77, 186.73, 199.17)	(194.82, 194.73, 188.54, 201.87)
(191.85, 190.20, 202.48, 201.12)	(206.88, 198.80, 214.41, 181.43)
(179.59, 195.77, 189.78, 192.33)	(178.88, 193.30, 198.33, 188.83)
(194.37, 200.40, 197.23, 184.96)	(199.05, 197.31, 187.73, 186.87)
(208.90, 220.42, 211.56, 196.22)	(206.68, 207.87, 192.17, 224.41)

Table 15.4 Cinchona Plant Volume and Dry Bark Weight (oz)[a]

Plant number	Volume (V)	Dry bark weight (W)
1	188.58	15.29
2	197.96	20.72
3	194.29	21.20
4	202.61	23.92
5	205.20	15.36
6	165.15	22.03
7	204.67	22.80
8	199.17	19.91
9	207.49	15.28
10	197.51	21.33
11	181.43	13.10
12	190.89	36.66
13	199.93	26.87
14	192.93	20.10
15	210.27	33.11
16	180.83	18.44
17	204.18	42.06
18	207.97	16.43
19	187.73	9.93
20	207.30	25.47
21	208.16	18.75
22	173.90	22.34
23	197.18	14.74
24	200.05	25.36
25	195.77	14.22
26	201.04	15.57
27	178.88	17.67
28	188.82	31.54
29	191.46	6.13
30	186.48	2.06
31	201.87	21.93
32	193.83	15.39
33	187.87	11.58
34	188.78	32.66
35	191.85	29.37
36	191.33	22.81
37	192.83	19.15
38	196.86	17.11
39	200.50	25.83
40	197.67	25.82
41	208.90	24.76
42	207.86	19.88
43	226.36	17.70
44	189.79	14.91
45	200.42	27.12
46	194.58	20.36
47	187.85	19.49
48	205.59	28.57
49	186.94	20.42
50	205.34	17.02
51	204.51	30.19
52	217.73	24.93
53	175.77	18.74
54	218.39	22.57

(*continued*)

Table 15.4 (*Continued*)

Plant number	Volume (V)	Dry bark weight (W)
55	219.53	17.90
56	198.42	17.47
57	212.63	24.24
58	197.23	19.05
59	208.94	5.18
60	199.46	11.05
61	192.20	24.62
62	191.23	25.38
63	187.78	12.85
64	173.04	16.21
65	187.08	13.89
66	183.86	9.13
67	207.97	18.72
68	185.22	11.75
69	225.00	16.87
70	192.17	13.02
71	192.33	22.99
72	195.27	12.02
73	209.31	12.16
74	190.59	20.41
75	206.68	9.27
76	203.39	23.58
77	210.76	14.27
78	187.93	24.04
79	196.74	30.67
80	213.38	18.12
81	193.30	2.36
82	189.55	12.22
83	199.32	12.73
84	202.34	18.56
85	199.05	14.59
86	199.41	11.00
87	201.04	24.07
88	196.15	31.64
89	190.97	12.50
90	223.14	25.13
91	224.41	15.88
92	199.94	21.64
93	201.11	24.92
94	195.06	19.99
95	197.75	3.41
96	206.80	22.98
97	200.78	13.77
98	209.33	12.05
99	217.98	24.31
100	181.25	24.58
101	189.78	21.24
102	209.30	27.89
103	206.79	28.70
104	199.72	12.24
105	205.14	22.96
106	185.82	31.81
107	198.41	10.37
108	192.57	19.97

Table 15.4 (*Continued*)

Plant number	Volume (V)	Dry bark weight (W)
109	212.32	22.51
110	198.44	20.20
111	207.87	19.79
112	205.76	13.80
113	194.73	19.67
114	208.26	21.52
115	211.56	29.64
116	190.20	17.62
117	203.41	11.63
118	200.16	16.21
119	197.17	11.46
120	194.82	13.19
121	202.44	21.80
122	208.63	11.61
123	201.12	32.64
124	208.10	11.80
125	196.29	22.43
126	201.12	27.41
127	184.23	5.63
128	196.85	31.98
129	192.56	36.89
130	187.54	28.94
131	188.54	16.03
132	192.61	20.87
133	195.09	9.76
134	202.48	3.82
135	198.33	24.95
136	200.26	21.63
137	189.70	25.70
138	204.04	14.05
139	210.52	18.78
140	201.76	16.42
141	214.41	16.16
142	199.07	29.25
143	186.95	23.46
144	220.42	10.06
145	201.76	17.04
146	211.69	26.11
147	192.65	20.86
148	201.45	28.41
149	204.67	19.55
150	188.45	26.46
151	201.41	16.22
152	184.96	28.78
153	188.83	2.72
154	178.42	24.36
155	190.65	10.66
156	198.36	1.12
157	211.21	14.00
158	212.38	26.71
159	186.87	25.57
160	198.80	23.50
161	186.93	8.04
162	196.22	13.91

(*continued*)

Table 15.4 (*Continued*)

Plant number	Volume (V)	Dry bark weight (W)
163	180.64	30.92
164	196.60	19.84
165	185.15	25.42
166	206.88	23.78
167	196.67	32.51
168	211.62	9.09
169	217.01	27.42
170	198.72	16.35
171	188.99	13.58
172	190.51	13.77
173	186.20	20.40
174	202.73	3.20
175	188.79	8.89
176	216.59	12.54
177	206.04	26.13
178	200.13	21.14
179	188.41	20.14
180	194.41	19.41
181	201.25	27.31
182	200.40	4.04
183	187.00	19.66
184	200.97	11.02
185	197.08	8.23
186	199.69	18.98
187	201.69	19.35
188	179.59	27.69
189	194.37	21.49
190	186.07	7.65
191	211.99	22.54
192	186.73	10.83
193	198.61	13.51
194	196.40	20.21
195	209.28	22.78
196	197.31	22.96
197	209.02	15.46
198	195.40	26.92
199	215.47	26.72
200	202.77	27.20

Source: J. M. Sengupta, I. M. Chakravarti, and D. Sarkar (1951), S. L. Stokes (1980), and Z. Chen, Z. D. Bai, and B. K. Sinha (2004).

[a]Random sample of 200 observations from a bivariate normal distribution with parameters $\mu_V = 200$, $\sigma_V = 10$, $\mu_W = 20.08$, $\sigma_W = 7.79$, and $\rho_{V,W} = 0.9$. This bivariate normal distribution provides an approximate representation of the joint distribution of $V =$ cinchona plant bark volume and $W =$ cinchona plant dry bark weight for cinchona plants grown in the hills of southern India.

Given our stated criterion for selecting the judgment ordered units, we see that the cinchona plants designated by our RSS scheme for which dry bark measurements are to be determined correspond to the bold cinchona plant volume measurements in the following display:

(**192.56**, 201.41, 196.67, 208.26)	(**181.25**, 190.65, 200.16, 196.85)
(204.67, 200.97, **188.99**, 197.17)	(211.99, 197.96, 196.15, **185.82**)
(**186.95**, 190.89, 188.79, 192.20)	(201.45,194.29, **199.72**, 209.02)
(207.86, **198.41**, 186.93, 205.14)	(201.25, 184.23, **196.60**, 211.21)
(201.76, 195.09, **198.44**, 205.59)	(206.04, **198.42**, 216.59, 189.79)
(204.04, 175.77, 186.73, **199.17**)	(**194.82**, 194.73, 188.54, 201.87)
(191.85, 190.20, 202.48, **201.12**)	(**206.88**, 198.80, 214.41, 181.43)
(179.59, 195.77, 189.78, **192.33**)	(178.88, 193.30, **198.33**, 188.83)
(194.37, **200.40**, 197.23, 184.96)	(**199.05**, 197.31, 187.73, 186.87)
(208.90, **220.42**, 211.56, 196.22)	(206.68, 207.87, 192.17, **224.41**)

The 20 cinchona plants with bold volume values in this display correspond to plant numbers

$(129, 100, 171, 106, 143)$ for the smallest judgment order statistics

$(104, 107, 164, 110, 056)$ for the second smallest judgment order statistics

$(008, 120, 126, 166, 071)$ for the third largest judgment order statistics

$(135, 182, 085, 144, 091)$ for the largest judgment order statistics.

In practice, we would then need to measure the dry bark weights for these 20 plants. In this example, of course, these 20 dry bark weights are simply obtained from the complete listing in Table 15.4, and the corresponding RSS of judgment order statistics are given by

$$
\begin{array}{lllll}
X_{[1]1} = 36.89 & X_{[1]2} = 24.58 & X_{[1]3} = 13.58 & X_{[1]4} = 31.81 & X_{[1]5} = 23.46 \\
X_{[2]1} = 12.24 & X_{[2]2} = 10.37 & X_{[2]3} = 19.84 & X_{[2]4} = 20.20 & X_{[2]5} = 17.47 \\
X_{[3]1} = 19.91 & X_{[3]2} = 13.19 & X_{[3]3} = 27.41 & X_{[3]4} = 23.78 & X_{[3]5} = 22.99 \\
X_{[4]1} = 24.95 & X_{[4]2} = 4.04 & X_{[4]3} = 14.59 & X_{[4]4} = 10.06 & X_{[4]5} = 15.88
\end{array}
$$

Hence, from (15.1) it follows that the RSS estimate of the average dry bark weight $\mu_{\text{dry bark weight}}$ of cinchona plants raised in this area of southern India is given by

$$
\begin{aligned}
\hat{\mu}_{\text{dry bark weight, RSS}} &= \frac{1}{20}\sum_{j=1}^{5}\sum_{i=1}^{4} X_{[i]j} \\
&= (36.89 + 24.58 + 13.58 + 31.81 + 23.46 + 12.24 + 10.37 \\
&\quad + 19.84 + 20.20 + 17.47 + 19.91 + 13.19 + 27.41 + 23.78 \\
&\quad + 22.99 + 24.95 + 4.04 + 14.59 + 10.06 + 15.88)/20 \\
&= 387.24/20 = 19.362 \text{ ounces},
\end{aligned}
$$

which is in good agreement with the average dry bark weight of 20.08 ounces that was used in the simulation of the data pairs (V, W) of cinchona plant volume V and dry bark weight W, as given in Table 15.4.

As a naive comparison, we note that if we had simply taken the first 20 cinchona plants selected in our RSS process, namely, plants 002 012 061 088 100 106 114 118 119 128 129 143 149 151 155 167 171 175 184, and 191, to be our SRS of size $n = 20$, the corresponding SRS estimate of the average dry bark weight $\mu_{\text{dry bark weight}}$ of cinchona plants raised in this area of southern India would have been

$$\begin{aligned}\hat{\mu}_{\text{dry bark weight, SRS}} &= (20.72 + 36.66 + 24.62 + 31.64 + 24.58 + 31.81 + 21.52 + 16.21 \\ &\quad + 11.46 + 31.98 + 36.89 + 23.46 + 19.55 + 16.22 + 10.66 + 32.51 \\ &\quad + 13.58 + 8.89 + 11.02 + 22.54)/20 = 446.52/20 = 22.326 \text{ ounces},\end{aligned}$$

which is not as accurate an estimate of the true simulation mean dry bark weight of 20.08 ounces as that provided by the RSS estimator.

EXAMPLE 15.3 ***Auditing to Detect Fraud.***

Gemayel et al. (2011) demonstrated that RSS can be an effective way to reduce auditing costs when assessing the true value of an account is time consuming (and, consequently, expensive) by allowing for smaller sample sizes than would be required for an SRS approach to have the same precision. They point out that this is commonly the case for accounts such as inventory; accounts receivable; property, plant, and equipment; and accounts payable. In these settings, the auditor will draw a sample from a set of accounts through their subsidiary ledgers, and then proceed with on-site inspections, recalculations, confirmations, and other auditing procedures when necessary.

Tackett (2012) provided simulated sales invoice records data for an electrical/plumbing distribution center, constructed in such a way that 15% of the recorded sales invoices are fraudulent, with stated book values larger than the true audited values for the materials sold in those transactions. Table 15.1 in the NSM Third Edition R package (sample subset included here) provides the stated book values and the true audited values for a population of 12,557 such sales invoice records, with 15% (1884) of them having fraudulently recorded book values. (In practice, of course, all that would be available for the auditor would be the stated book values, and the true audited values would be obtained only for those accounts that were selected for inclusion in the RSS. The auditor would not need to find the audited values for the entire population. However, we have included the true audited values in Table 15.1 for all 12,557 sales invoices so that the data set can be used to illustrate the application of RSS in such a setting.)

We will use the stated book value (readily available in the company's electronic ledgers) as the auxiliary ranking variable to obtain an RSS of 200 sales invoices for auditing purposes. For the sake of illustration, we take our set size to be $k = 10$ and employ $m = 20$ cycles to achieve our overall sample size of $n = km = 10(20) = 200$. (We could, of course, have used other (k, m) combinations such as $(4, 50)$, $(5, 40)$, or $(8, 25)$ to obtain our overall sample size of 200.)

With $k = 10$ and $m = 20$, we must first obtain $mk = 20(10) = 200$ SRSs of size 10 each (without replacement) from the book values listed in Table 15.1. Then we use the book values to judgment rank the invoices within each of these 200 sets of 10 sales invoices and select the smallest judgment-ordered sales invoice from 20 of these sets, the second smallest judgment-ordered sales invoice from a second group of 20 sets, the third judgment-ordered sales invoice from a third group of 20 sets, etc., until, finally, we select the largest judgment-ordered sales invoices from the remaining group of 20 sets. Then the true audited value will be obtained for the 200 sales invoices selected in this manner.

Fortunately, the entire process of selecting the units to include in the RSS can be accomplished by simply applying the R command RSS(*k, m, x*) to the sales invoice population, where k is the set size, m is the number of cycles, and x is the label for the auxiliary variable to be used for the judgment ranking. Thus, applying RSS(10, 20, **book.value. audited. value[,2]**) to the book value sales invoice data in Table 15.1, we obtain the following 200 SRSs of 10 sales invoices each, ordered separately by their book values within each of the random samples:

200 Random Samples of 10 Book Values Each, Ordered from Least to Greatest Within Each of the Random Samples

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
84.16	35.84	24	44.52	1.84	38.46	29.89	2.98	42.61	4.18
119.54	37.47	62.1	47.33	21.65	81.76	66.19	46.29	87.55	56.17
138.02	60.95	101.93	67.86	25.13	139.08	418	57.97	89.76	84.6
150.55	83.25	176.01	72.86	30.52	318.51	504.06	80.89	345.37	106.86
209	83.69	228.02	79.35	45.21	338.51	1117.84	90.19	642.78	149.07
217.91	173.63	234.18	89.06	77.35	384.37	1391.44	184.28	674.14	221.96
438.2	241.41	259.82	211.46	245.56	1379.63	1450.12	304.46	940.72	292.23
976.28	1055.25	349.17	354.76	445.24	1727.92	1569.27	678.9	958.53	360.5
1819.27	1124.2	425	450.68	567.42	2024.36	2377.07	1074.38	1388.55	1239.07
3448.28	1430.48	1000.84	473.28	1513.33	2838.76	3095.73	4332.08	41488.06	3196.05
11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
54.34	31.56	50.6	29.32	30.4	159.88	28.8	37.77	58.62	49.14
59.92	40.99	130.19	47.11	111.64	274.44	72.42	122.35	152.69	145.95
264.17	155.38	159.14	54.82	112.39	416.87	116.55	197.39	170.61	211.41
287.45	186.56	184.57	58.62	916.14	422.35	117.64	806	202.25	229.13
813.1	277.39	413.86	401.88	1513.73	588.45	172.41	841.83	238.74	561.01
858.41	286.16	517.93	410.48	1904.01	1019.67	352.92	1147.21	312.06	567.1
897.01	465.35	677.7	1013.34	3578.18	1439.55	889.8	1521.58	846.28	942.35
2621.47	920.43	784.84	1040.13	4316.92	1552.04	976.11	1844.69	1061.44	1285.18
8497.63	5862.7	2709.18	1183.22	9213.44	2422.35	1758.17	2157.08	2317.06	2538.26
236345.1	18066.54	5733.49	1313.5	9249.18	5827.13	3271.78	86153.45	8629.54	4076.38
21	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
4.3	21.1	71.43	34.86	51.37	74.9	54.43	32.02	30.07	26.71
34.67	66.29	83.2	106.04	62.4	90.42	98.83	49.99	149.32	38.03
49.18	68.94	201.89	148.55	82.29	103	130.21	60.97	178.42	201.53
115.07	265.11	241.72	530.59	165.78	112.39	868.85	78.15	347.3	223.02
273.02	322.61	288.69	566.1	255.11	165.46	1190.58	176.85	439.55	224.83
674	485.97	290.92	685.92	330.2	245.44	1251.93	229.58	909.64	450.79
944.18	539.44	913.72	1179.26	367.46	304.5	1680.82	247.51	1587.07	787.01
986.88	679.74	1036.41	1425.45	573.07	383.39	2352.46	772.61	1995.59	1024.45
1542.49	1004.08	1952.96	3567.94	804.36	1054.68	2566.22	8618.81	2034.76	1271.18
2092.18	17919.76	2806.22	6150.92	1450.9	1564.28	4085.44	9389.97	6228.45	1440.78

(*continued*)

200 Random Samples of 10 Book Values Each, Ordered from Least to Greatest Within Each of the Random Samples (*Continued*)

31	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
40.82	67.66	136.97	24.44	29.25	23.93	170.32	1.96	28.08	61.73
67.57	88.27	288.3	38.38	51.4	160.76	269.36	25.81	134.58	122.05
155.17	105.55	518.75	51.15	99.39	163.64	292.56	89.69	385.84	126.71
178.31	112.6	721.64	58.91	283.65	217.46	299.67	310.31	422.62	190.52
374.4	251.67	958	134.76	325.87	274.67	544.59	348.69	583.61	334.15
478.77	664.42	1228.66	187.15	454.43	334.67	555.47	487.82	779.09	475.12
776.83	731.43	1409.18	281.06	571.78	1234.11	717.02	736.31	1155.62	802.29
909.73	903.59	2590.92	503.5	797.74	1266.71	1299.02	3935.95	1763.31	3776.49
1709.93	1051.7	3826.11	589.46	1166.8	2023.97	1660.5	4252.76	2324.99	8106.33
8059.29	5341.06	4604.8	1527.72	2026.32	4686.67	1705.56	196374.2	2760.85	150890.4

41	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**
1.9	45.57	26.41	37.44	39.06	31.11	26.61	73.94	23.56	48.31
39.24	86.53	117.96	92.94	69.9	31.66	41.18	74.82	46.87	71.14
40.16	150.19	249.58	171.22	117	40.85	96.47	133.63	434.81	127.05
56.44	218.65	265.17	204.44	164.89	248.87	104.13	190.97	595.09	236.57
66.61	274.75	1244.68	391.96	530.21	293	115.79	282.11	839.17	260.39
102.98	513.8	1433.07	485.58	672.84	310.7	290.01	383.7	1713.56	702.69
129.57	644.75	2226.89	813.06	740.87	373.49	293.43	486.54	1723.48	1335.5
537.19	668.72	2392.93	1359.63	808.46	1145.2	377.1	492.85	2750.03	2310.4
1274.79	844.56	2832.84	3032.98	1007.63	1306.36	1324.46	579.35	2901.52	3154.24
2298.92	1052.26	4451.9	26049.71	1401.11	6918.05	1961.79	616	3665.03	3322.79

51	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
61.32	25.15	96.27	108.89	83.6	51.82	28.3	55.68	59.83	26.41
112.21	41.41	325.76	193.79	222.98	84.41	64.76	104.46	127.26	31.86
170.91	49.41	376.22	342.86	243.5	190.89	123.8	230.55	234.07	37.04
463.94	147.86	496.48	474.8	341.68	213.49	176.6	1084.09	263.64	38.39
918.76	156.96	578.28	513.23	356.12	316.36	285.52	1649.13	543.62	45.48
1373.17	362.96	624.13	585.71	594.19	376.21	317.56	2488.8	635.74	76.63
1850.73	551	650.85	898.72	1198.57	448.45	1019.57	2767.18	720.57	166.45
3036.96	573.33	1507.11	1156.77	2606.47	702.65	1266.49	3616.39	836.91	176.46
5608.04	875.72	2338.01	2010.66	4644.18	1255.07	2774.95	7823.93	6209.07	502.05
13922.81	23450.95	5798.55	5369.7	5461.37	9117.69	3696.51	52667.94	56747.1	1440.31

61	**62**	**63**	**64**	**65**	**66**	**67**	**68**	**69**	**70**
37.67	24.45	33.53	58.92	64.88	81.87	9.08	57.3	107.19	37.35
167.05	28.67	74.92	111.57	190.44	106.65	32.35	91.43	118.03	53.19
202.86	125.38	123.29	196.83	312.44	256.05	171.72	103.46	192.43	261.84
252.32	188.11	143.47	454.59	349.36	305.7	176.34	143.9	204.37	437.19
1611.75	286.96	187.5	522.96	631.37	528.59	224.75	469.89	840.56	453.45
1636.34	576.79	336.14	1005.31	753.57	647.38	310.02	579.32	1106.81	646.09
1882.98	612.32	369.76	1348.14	848.51	1226.5	778.49	976.96	1771.6	808.11
2981.79	871.25	1517.3	3797.21	4630.35	1244.05	1809.65	996.83	1910.64	973.32
3831.65	1373.7	2463.4	17580	5107.12	4103.11	2514.48	1373.86	2168.64	1700.51
96539.69	2443.64	26796.4	82221.27	18290.83	5995.58	4593.12	1564.39	2224.44	2458.23

71	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**	**80**
24.08	153.86	21.64	6.96	35.58	49.15	90.99	42.58	32.33	27.77
52.33	157.26	121.79	56.36	37.91	62.92	110.31	48.01	35.49	101.1
75.71	188.5	129.21	63.69	48.87	95.49	148.64	132.43	48.8	109.05
111.04	632.72	206.13	85.39	52.42	102.22	182.91	235.52	63.81	110.95
217.68	641.59	462.91	269.73	69.96	127.92	203.45	779.08	134.44	144.71
247.25	1142.27	813.14	432.06	113.92	290.06	311.01	1153.21	281.87	165.18

200 Random Samples of 10 Book Values Each, Ordered from Least to Greatest Within Each of the Random Samples (*Continued*)

549.14	1213.05	969.04	636.54	278.24	398.75	330.04	1485.27	286.18	214.93
1634.24	1506.53	1439.62	1014.16	420.05	1945.04	418.09	1818.02	375.35	362.46
2278.06	1646.55	2303.61	2339.42	1450.44	2912.21	575.89	2121.94	1338.93	784.79
14717.51	2909.86	7813.33	7112.99	1857.68	5972.28	723.6	9244.39	32604.49	1089.44
81	**82**	**83**	**84**	**85**	**86**	**87**	**88**	**89**	**90**
46.67	36.31	53.18	125.07	78.11	82.54	55.3	25.45	26.13	114.01
63.83	160.72	78.3	192.74	80.1	220.62	150	73.53	34.66	145.84
148.47	273.02	653.91	349.11	177.49	530.09	154.52	449.1	207.06	283.07
706.53	462.96	802.68	377.84	234.12	582.15	197.38	504.45	240.56	310.66
817.38	486.6	827.62	475.97	330.2	673.44	308.07	779.78	364.9	482.05
1152.54	497.58	919.35	726.1	518.94	1140.68	387.03	1351.43	378.31	692.12
1162.41	693.15	1025.76	936.44	520.92	1151.37	917.02	1467.39	457.13	702.66
1721.81	1293.28	1029.33	1109.16	1140.71	1218.27	941.01	1673.19	1034.76	956.32
2135.36	1522.15	1600.07	2190.74	1258.62	1467.68	1709.09	1675.87	1062.19	1016.35
66667.65	2583.99	6785.24	4471.82	5414.46	3122.75	72710.23	1678.04	44733.81	1592.1
91	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**	**100**
35.29	2	41.1	46.35	40.39	54.23	171.49	41.05	57.9	27.51
38.32	46.73	42.71	102.92	43.44	68.95	223.71	52.74	66.34	32.15
146.41	49.98	59.21	159.42	122.97	119.48	283.12	99.52	75.74	33.7
228.12	106.79	107.06	167.74	189.88	274.18	316.53	106.96	76.87	135.24
369.68	195.37	141.88	974.73	307.17	290	475.54	115.98	166.91	165.33
857.36	229.59	376.94	983.12	458.5	466.78	657.95	280.5	292.93	249.19
1228.35	338.93	392.4	1468.73	851.53	565.56	717.46	438.87	642.86	542.27
1474.9	937.67	554.81	5983.71	1283.04	696.27	822.36	686.52	1641.53	2455.2
1793.21	1929.87	760.32	7381.72	3422.32	5149.18	844.47	8170.31	1822.78	20685.19
4627.99	1950.58	937.41	17536.25	125013.5	6891.97	846.22	66798.69	4653.38	96287.45
101	**102**	**103**	**104**	**105**	**106**	**107**	**108**	**109**	**110**
34.22	39.73	1.31	35.94	45.61	32.82	30.62	29.1	54.76	28.59
56.35	43.56	29.58	111.55	59.88	51.21	46.18	57.05	72.57	79.21
132.99	54.9	101.13	192.11	71.72	59.73	82.56	254.44	278.68	121.46
199.36	87.42	121.76	339.69	152.5	65.34	85.58	305.08	444.27	125.59
231.45	147.66	311.12	438.82	182.96	169.72	120.63	328.89	496.13	178.81
319.6	160.71	348.81	650.81	224.43	190.93	219.22	336.45	913.48	199.84
321.95	189.41	705.51	1336.97	435.88	278.21	636.59	463.49	996.96	833.21
553.72	234.09	1114.41	1466.34	806.56	484.04	1155.18	1020.08	1196.38	1686.73
2921.93	614.62	3540.56	1656.27	1707.77	935.31	1974.95	1395.27	1359.47	4766.41
12028.74	1990.78	3871.55	15191.31	1828.05	2024.08	2815.26	1611.98	9152.17	5084.95
111	**112**	**113**	**114**	**115**	**116**	**117**	**118**	**119**	**120**
57.82	157.56	34.49	3.13	22.89	45.89	65.78	35.3	24.65	103.42
227.53	375.25	81.15	36.58	34.79	87.3	115.91	38.33	63.69	105.79
397.97	398.9	100.64	42.27	89.88	310.2	119.95	39.28	125.39	598.19
413.99	561.19	102.99	84.75	376.61	526.63	226.28	98.82	180.48	636.6
503.1	777.61	165.9	88.43	463.45	575.1	467.43	222.96	332.89	776.68
983.3	785.72	293.05	128.85	738.72	753.13	468.31	226.03	501.62	1067.48
1057.24	887.91	456.99	200.45	778.92	1138.99	606.03	292.73	544.3	1793.44
1518.02	1222.23	986.06	1295.76	857.41	1236.53	613.8	297.02	1611.87	2765.62
2194.9	1729.86	1054.68	2024.99	1683.45	2102.22	765.32	753.43	2001.23	3289.82
2254.55	5934.77	1115.26	7548.64	25565.67	2797.63	2308.88	1657.63	2355.22	27066.71

(*continued*)

200 Random Samples of 10 Book Values Each, Ordered from Least to Greatest Within Each of the Random Samples (*Continued*)

121	**122**	**123**	**124**	**125**	**126**	**127**	**128**	**129**	**130**
26.05	31.74	33.04	41.82	48.39	40.98	35.63	28.89	65.48	67.57
175.27	116.97	98.47	86.23	94.25	76.21	82.84	29.07	68.83	105.39
310.97	190.66	233.47	210.65	94.26	81.49	104.6	39.57	86.76	138.49
348.5	339.46	344.23	330.4	171.74	115.38	362.01	59.08	335.69	420.1
411.63	513.42	547.41	400.68	194.12	209.43	688.96	84.06	378.97	434.03
609	1476.14	670.56	487.03	730.07	520.42	1035.43	426.01	668.08	522.74
1209.13	1528.65	738.87	968.94	1212.04	2280.45	1244.2	749	719.89	1647.61
1367.6	1791.62	834.09	1187.13	1966.33	2311.24	1617.8	1124.5	865.18	1780.87
1390.95	2151.4	1175.55	5904.52	2139.87	3785.65	4612.68	9599.66	1483.33	1901.64
13358.53	10886.61	1713.5	78885.34	46273.1	5868.22	43303.89	17623.17	1819.4	57451.82

131	**132**	**133**	**134**	**135**	**136**	**137**	**138**	**139**	**140**
61.78	74.51	66.11	30.08	43.17	41.6	46.85	12.94	87.88	37.17
441.62	148.16	68.17	50.73	48.35	63.45	147.16	58.17	146.29	122.5
786.09	284.29	77.91	53.95	65.51	143.82	388.05	218.7	151.4	204.59
1250.16	335.07	166.52	93.57	74.68	175.59	677.55	264.62	247.04	264.7
1585.72	342.03	531.02	265.01	100.55	389.69	878.45	421.21	888.85	505.44
2023.24	544.4	941.22	276.62	135.99	909.07	997.41	433.32	1609.94	643.68
2053.38	752.68	992.73	694.1	533.87	994.84	1305.35	517.48	1632.58	965.71
4881.81	1230.22	1235.28	704.85	569.66	1613.08	1770.07	1061.29	1752.71	1703.55
5504.25	1639.7	2385.17	829.68	1378.34	1675.89	1957.88	1292.66	1797.37	1763.89
9691.12	2217.98	3557.41	2422.83	2592.68	8288.31	6099.11	8025.95	2641.33	2260.17

141	**142**	**143**	**144**	**145**	**146**	**147**	**148**	**149**	**150**
27.22	29.65	303.22	29.08	70.24	41.08	50.01	1.05	49.9	31.23
41.86	41.15	352.1	43.04	111.17	83.9	104.05	56.87	151.25	143.3
74.66	46.93	364.32	68.84	152.8	140.84	158.44	139.74	151.32	157.66
130.33	58.45	506.32	83.19	198.96	195.78	184.58	276.22	220.12	232.34
363.93	237.38	543.51	98.94	210.21	234.34	466.67	390.94	454.48	269.94
522.28	298.76	572.79	158.54	287.61	480.64	560.07	563.43	530	378.53
786.43	320.9	611.59	159.94	482.15	524.95	805.55	1367.97	646.58	381.37
823.22	636.59	930.55	421.86	958.63	825.86	1556.95	1403.52	1295.77	1076.36
903.89	881.47	1757.59	2140.42	1157.68	1764.09	1802.14	2097.1	2655.41	1797.02
8282.78	911.95	1868.31	5036.25	1924.51	9498.1	6122.85	8640.57	52958.58	2758.76

151	**152**	**153**	**154**	**155**	**156**	**157**	**158**	**159**	**160**
62.88	31.15	153.96	75.85	287.69	46.61	96.03	23.71	30.93	23.84
83.75	70.88	166.75	87.49	382.06	53.14	140.14	53.17	53.04	24.92
105.01	176.68	186.69	151.2	495.07	115.61	151.05	70.38	86	70.2
341.73	249.77	320.53	151.38	665.38	123.49	318.68	115.02	165.6	94.54
412.8	345.29	522.96	161.87	1879.71	404.05	446.73	117.94	216.57	690.35
454.22	657.77	1363.86	176.01	1927.41	419.45	571.61	286.4	422.6	733.58
571.66	786.15	1386.28	182.34	2010.44	557.99	940.76	354.15	553.34	1031.12
994.54	963.67	1564.75	328.87	2409.72	575.81	1382.08	1257.55	572.69	1817.31
2751.88	1082.63	13945.1	737.95	3553.37	1668.63	2057.73	1413.81	1528.69	2216.27
8176.53	8142.2	56682.16	17085.76	9287.11	2471.45	9663.45	1725.45	2850.15	6637.89

161	**162**	**163**	**164**	**165**	**166**	**167**	**168**	**169**	**170**
60.84	63.6	33.56	90.28	60.26	76.45	24.6	114.78	1.43	49.58
69.76	98.75	41.96	130.35	81.62	260.4	27.66	121.09	161.62	52.7
128.97	290.11	119.85	280.97	85.74	271.82	35.95	134.24	199.03	60.15

200 Random Samples of 10 Book Values Each, Ordered from Least to Greatest Within Each of the Random Samples (*Continued*)

138.97	292.87	198.39	1277.23	110.23	306.58	66.22	226.18	244.48	110.05
163	360.68	407.41	1300.71	193.69	584.62	132.4	277.16	268.72	204.48
163.16	381.69	558.59	1565.84	921.41	1115.09	159.31	876.91	330.39	295.02
319.1	578.46	1600.98	1736.01	1516.19	3628.26	383.52	1242.41	696.63	322.89
481.18	2081.33	2710.99	2175.19	1828.23	3908.45	896.24	1622.5	1591.96	510.03
786.54	2689.33	4642.34	6509.82	2281.23	5128.69	1162.02	1967.99	2581.98	663.62
905.32	3702.78	12882.12	9456.03	11751.72	36107.4	5442.17	60716.14	6835.45	3966.28
171	**172**	**173**	**174**	**175**	**176**	**177**	**178**	**179**	**180**
2.87	26	49.03	39.14	58.99	47.96	31.31	1.31	82.05	22.82
92.48	81.76	74.46	51.67	66.29	67.48	38.01	69.36	106.63	45.36
151.29	179.63	80.85	331.67	100.24	98.87	72.57	120.85	181.26	48.84
157.21	224.52	199.52	634.15	147.08	202.89	125.99	132.4	208.63	112.71
224.23	305.67	200.99	686.44	154.58	343.13	164.08	199.75	240.58	116.31
410.08	307.76	336.44	820.22	154.92	814.78	216.44	746.48	291.99	127.41
410.72	393.43	557.38	1288.63	277.68	1298.55	326.41	784.35	336.26	457.99
567.84	467.61	1218.65	1990.96	808.83	2184.41	723.73	878.99	468.87	1036.91
1018.32	939.12	6580.75	5060.14	2123.95	2942.23	2312.77	1015.27	537.05	2049.97
3568.35	2002.1	11102.92	5723.42	4001.74	5579.02	18441.51	1169.51	8058.74	3941.9
181	**182**	**183**	**184**	**185**	**186**	**187**	**188**	**189**	**190**
41.72	35.28	101.29	36.07	49.62	31.35	48.2	27.84	32.03	4.97
114.28	57.6	118.68	40.91	192.49	70.46	65.84	54.58	36.35	38.57
140.08	114.84	153.57	163.14	228.08	143.9	87.43	121.35	45.82	78.04
309.63	142.16	187.03	266.87	257.58	433.98	206.83	181.77	87.07	102.65
538.01	334.57	254.13	828.74	340.19	621	882.11	219.98	113.21	141.36
606.25	355.54	402.91	901.67	383.6	659.05	1253.72	585.05	128.56	466.61
648.31	455.71	470.44	1005.83	1116.13	860.38	1386.3	602.03	712.25	643.65
1072.35	493.41	755.81	1437.76	1717.49	1161.35	1509.31	755.56	1802.02	936.67
2029.85	673.18	795.25	1449.21	2046.42	2206.49	2300.67	1176.53	2137.93	1107.56
3049.17	6569.74	18513.61	3348.4	3497.52	5262.55	3363.05	1675.16	7937.35	1638.54
191	**192**	**193**	**194**	**195**	**196**	**197**	**198**	**199**	**200**
58.66	261.89	23.01	27.3	24.75	54.27	41.51	25.05	73.88	35.41
63.8	416.88	49.66	39.87	39.08	130.31	81.69	38.1	79.13	49.73
109.44	431.34	69.46	43.23	76.76	210.44	118.4	126.94	88.04	64.79
113.83	575.84	99.78	59.63	113.55	275.57	131.45	539.89	115.08	86.95
133.86	674.52	1785.28	120.42	162.62	479.28	139.25	607.93	426.65	102.41
223.23	916.47	2087.05	315.82	257.26	618.7	178.12	968.09	634.83	285.01
283.71	965.22	2272.46	1388.44	321.72	1039.61	577.25	995.17	729.89	541.89
836.19	1343.42	5793.3	1524.38	1405.04	2121.15	1828.2	1235.88	1569.06	1094.93
1578.92	1944.74	6165.71	2025.73	1561.28	4604.91	1907.56	4398.7	5300.33	1505.08
2110.99	18940.75	48561.93	7721.78	2107.9	59865.95	5201.64	9877.81	16734.4	5075

Using our predesigned sampling plan, we select the unit with the smallest ordered book value in each of the first 20 samples of 10 sales invoices, the unit with the second smallest ordered book value in each of the second 20 samples of 10 sales invoices, the unit with the third smallest ordered book value in each of the third 20 samples of 10 sales invoices, etc., until, finally, we select the unit with the largest ordered book value in each of the remaining 20 samples of 10 sales invoices for inclusion in our RSS of $n = 200$ units. Thus our RSS consists of the following 200 sales invoices, listed from the smallest to the largest unit number:

Sales Invoice Numbers for the 200 Invoices in the Ranked Set Sample

56	2406	5054	7732	9872
165	2415	5154	7758	9885
238	2428	5245	7793	9907
250	2458	5341	7840	10055
275	2468	5430	7932	10137
454	2709	5431	7962	10266
548	2780	5452	7990	10302
565	3050	5468	8023	10340
598	3127	5618	8099	10377
632	3164	5657	8126	10393
708	3235	5698	8186	10449
775	3249	5852	8199	10525
891	3258	6004	8246	10533
952	3319	6050	8364	10608
968	3349	6106	8384	10617
1007	3433	6117	8396	10619
1040	3568	6140	8495	10779
1042	3816	6162	8562	10784
1156	3855	6196	8601	10874
1167	3888	6331	8691	10969
1235	3908	6433	8701	11184
1319	3946	6619	8815	11231
1331	3953	6655	8826	11355
1346	3983	6684	8828	11371
1379	4028	6696	8837	11386
1382	4106	6742	8862	11475
1442	4141	6762	8887	11538
1650	4150	6872	9043	11640
1729	4217	6901	9048	11665
1801	4466	6957	9119	11672
1821	4476	6998	9130	11826
2004	4482	7102	9193	11876
2005	4493	7141	9246	11893
2231	4625	7154	9353	11950
2247	4696	7263	9375	11959
2285	4763	7274	9393	12168
2315	4863	7312	9492	12293
2319	4975	7468	9527	12371
2363	5005	7538	9581	12397
2398	5018	7645	9842	12472

The final step in the process for the practicing auditor would then be to perform the required steps to obtain the true audited value for each of the 200 sales invoice units in the selected RSS. Of course, for us it is simply a matter of going to Table 15.1 and recording the listed audited values for these units. The audited values for the 200 sales

invoices in the RSS are given in the following table, in the order of increasing unit number (as in the previous table).

Audited Values for the 200 Invoices in the Ranked Set Sample

5060.14	96.47	575.81	112.48	1018.32
330.20	2349.97	663.62	5751.84	182.91
190.89	749.00	2689.33	2280.45	226.03
461.17	58.62	106.04	31.56	24.00
1076.36	7181.90	38.03	482.05	5075.00
201.57	149.32	117.00	42.35	994.84
288.30	143.47	49.41	2942.23	369.68
3497.52	85.39	421.86	2.98	166.91
293.05	111.04	357.72	1632.58	319.60
29.89	6569.74	537.05	965.71	486.60
51.40	786.54	23819.98	290.00	269.36
4.18	42.61	1001.14	454.59	3049.17
38.46	434.81	34.68	817.38	974.73
17278.42	752.68	40.16	54.34	224.43
342.86	133.63	165.33	150.19	34.67
753.13	1798.00	122.05	40.85	98.83
501.62	1638.54	58.57	90.42	785.72
525.61	28.52	28.80	533.87	2281.23
50.60	632.72	38.38	6580.75	84.16
308.07	738.87	376.22	349.81	9877.81
37.04	468.31	930.55	517.48	1564.75
572.69	3363.05	30.40	1015.27	3348.40
190.93	2049.97	80.45	25.81	35.84
73.33	22.26	18940.75	128.85	143.90
994.54	176.34	219.22	348.81	719.89
234.07	1647.61	195.37	1.84	1040.83
1067.48	7721.78	88.27	475.54	437.19
673.44	364.90	7937.35	1162.02	779.78
2107.90	1675.16	1556.95	305.70	171.22
48.94	252.32	160.76	63.81	1295.77
992.73	650.81	38.46	2110.99	2075.73
2123.95	5201.64	307.17	123.80	249.58
1209.13	134.58	206.13	1382.08	958.63
328.87	160.71	37.77	1212.04	230.55
27.78	939.12	1244.20	1817.31	1528.65
44.52	983.30	694.10	879.15	188.11
582.45	243.50	827.62	110.95	115.98
1257.55	2053.38	149.38	49.14	29.32
52.42	738.72	66.29	1743.17	968.94
170.91	770.93	825.86	636.59	235.52

We can then use these RSS observations to estimate the percentage of fraudulent sales invoices in the entire population of 12,557 sales invoices and the total amount of fraud

(overstatement) in the population. To estimate the population percentage of fraudulent sales invoices, p_F, we can simply use the sample percentage of fraudulent sales invoices in our RSS (see Comment 6), namely,

$$\hat{p}_{F,RSS} = \frac{30}{200} = 15\%.$$

This happens to be a perfect match with the known overall percentage of 15% fraudulent sales invoices in the population (but do not expect such perfection with all RSSs!).

To estimate the total amount of fraud (overstatement) in the population, we first estimate the average amount of fraud per account, μ_F, and then expand that average to the total number (12,557) of accounts in the population. For this purpose, we record the amount of overstatement for each of the units in our RSS (recording a "0" if a sales invoice is not fraudulent), yielding the following overstatements for the 200 invoices in the RSS.

Overstatements (Fraud) for the 200 Invoices in the Ranked Set Sample

0	0	0	223.97	0
0	4159.85	0	10982.56	0
0	0	0	0	0
502.50	0	0	0	0
0	11331.71	0	0	0
274.40	0	0	84.70	0
0	0	0	0	0
0	0	0	0	0
0	0	465.50	0	0
0	0	0	0	0
0	0	36045.97	0	0
0	0	1408.58	0	0
0	0	48.52	0	0
31283.51	0	0	0	0
0	0	0	0	0
0	2844.34	0	0	0
0	0	101.31	0	0
779.74	39.05	0	0	0
0	0	0	0	0
0	0	0	563.67	0
0	0	0	0	0
0	0	0	0	0
0	0	119.39	0	0
131.04	27.73	0	0	0
0	0	0	0	0
0	0	0	0	1541.15
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Overstatements (Fraud) for the 200 Invoices in the Ranked Set Sample (*Continued*)

92.94	0	0	0	0
0	0	63.76	0	3186.82
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
34.62	0	0	0	0
0	0	0	1433.62	0
821.07	0	0	0	0
0	0	199.98	0	0
0	0	0	3385.52	0
0	1197.06	0	0	0

From (15.1), the RSS estimate of μ_F is then just the average of these 200 values, namely,

$$\hat{\mu}_{F,RSS} = \frac{\$113{,}374.58}{200} = \$566.87.$$

The RSS estimate of the total amount of fraud in the entire population is thus 12,557 $(\hat{\mu}_{F,\,RSS}) = 12{,}557(\$566.87) = \$7{,}118{,}187$.

Comments

2. *Contrasting $\hat{\mu}_{SRS}$ and $\hat{\mu}_{RSS}$.* The k components of the SRS estimator $\hat{\mu}_{SRS} = \overline{X}$ are mutually independent and identically distributed and each is itself an unbiased estimator for μ. While the k components of the RSS estimator $\hat{\mu}_{RSS} = \overline{X}_{RSS}$ are also mutually independent, they are not identically distributed and none of them (except for the middle order statistic when k is odd and the underlying distribution is symmetric about μ) are individually unbiased for μ. Yet the averaging process leaves $\hat{\mu}_{RSS}$ unbiased. Interestingly, it is the additional structure associated with the nonidentical nature of the distributions for the terms in $\hat{\mu}_{RSS}$ that leads to the improvement in precision for $\hat{\mu}_{RSS}$ relative to $\hat{\mu}_{SRS}$.
3. *Imperfect Rankings.* The RSS estimator $\hat{\mu}_{RSS}$ remains unbiased for the population mean μ and at least as precise (i.e., as small or smaller variance) than the corresponding SRS estimator $\hat{\mu}_{SRS}$ based on the same number of measured observations even if the judgment ranking process does not yield perfect rankings. The relationship between the variances for the two estimators is still given by the expression in (15.10), except that the expected value $\mu^*_{(i)}$ for the ith true order statistic $X^*_{(i)}$ under perfect rankings is replaced by the expected value of the judgment order statistic $\mu_{[i]} = E[X_{[i]}]$, for i $= 1, \ldots, k$. See Dell and Clutter (1972) for a more detailed discussion of the effect of imperfect judgment rankings in this setting.
4. *Estimation of the Population Variance.* The RSS approach has also been used to estimate a population variance. Let $X_{[1]j}, \ldots, X_{[k]j}$, for $j = 1, \ldots, m$, be an RSS

(for set size k and m cycles) from a population with finite variance σ^2. Stokes (1980) proposed the following RSS estimator, $\hat{\sigma}^2_{\text{Stokes}}$, for σ^2:

$$\hat{\sigma}^2_{\text{Stokes}} = \frac{1}{mk-1}\sum_{j=1}^{m}\sum_{i=1}^{k}(X_{[i]j} - \hat{\mu}_{\text{RSS}})^2, \tag{15.12}$$

where $\hat{\mu}_{\text{RSS}}$ (15.1) is the RSS estimator for the population mean. Stokes shows that the estimator $\hat{\sigma}^2_{\text{Stokes}}$ (15.12) is asymptotically unbiased for σ^2 and, for sufficiently large m or k, at least as efficient as the standard variance estimator, $\hat{\sigma}^2_{\text{SRS}}$, based on an SRS of the same size $n = mk$. Stokes points out, however, that the estimator $\hat{\sigma}^2_{\text{Stokes}}$ does not do as well for small or moderate samples, primarily due to the fact that it can be quite biased for even moderate sample sizes.

MacEachern et al. (2002) note that the Stokes estimator $\hat{\sigma}^2_{\text{Stokes}}$ treats each observation in the RSS the same regardless of which judgment order statistic it corresponds to, thereby ignoring some of the structural information provided by the RSS design. They took advantage of this additional structure inherent in the RSS design to propose a competitor estimator

$$\hat{\sigma}^2_{\text{MOSW}} = \frac{\sum_{r\neq s=1}^{k}\sum_{i=1}^{m}\sum_{j=1}^{m}(X_{[r]i} - X_{[s]j})^2}{2m^2k^2} + \frac{\sum_{r=1}^{k}\sum_{i=1}^{m}\sum_{j=1}^{m}(X_{[r]i} - X_{[r]j})^2}{2m(m-1)k^2} \tag{15.13}$$

that incorporates both within judgment ranking and between judgment ranking information from the RSS data. MacEachern et al. show that $\hat{\sigma}^2_{\text{MOSW}}$ (15.13) is an unbiased estimator for σ^2 and that it is more efficient over a broad variety of underlying distributions for small to moderate sample sizes than the Stokes estimator $\hat{\sigma}^2_{\text{Stokes}}$(15.12). Under mild conditions, however, the asymptotic relative efficiency of $\hat{\sigma}^2_{\text{MOSW}}$ relative to $\hat{\sigma}^2_{\text{Stokes}}$ is 1 when the judgment ranking is perfect.

5. *Estimation of the Population Distribution Function.* Utilization of information obtained from rankings is clearly an integral part of the RSS concept through the judgment ranking process used to select the specific items for measurement. However, it was not until the seminal paper by Stokes and Sager (1988) that a rank-based nonparametric approach was proposed for analysis of the RSS measurements themselves. Stokes and Sager studied the use of RSS data to estimate the distribution function $F(t)$ of a population.

   Let $X_{[1]j}, \ldots, X_{[k]j}$, for j $= 1, \ldots, m$, be the RSS (for set size k and m cycles) from a distribution with distribution function $F(t)$. The natural RSS estimator for $F(t)$ considered by Stokes and Sager (1988) is the sample distribution function for the RSS data, namely,

$$\hat{F}_{\text{RSS}}(t) = \frac{1}{mk}\sum_{i=1}^{k}\sum_{j=1}^{m} I_{(-\infty,t]}(X_{[i]j}). \tag{15.14}$$

   Stokes and Sager show that $\hat{F}_{\text{RSS}}(t)$ is an unbiased estimator of $F(t)$ and that

$$\text{Var}(\hat{F}_{\text{RSS}}(t)) \leq \text{Var}(\hat{F}_{\text{SRS}}(t)) \text{ for all t,}$$

where $\hat{F}_{\text{SRS}}(t)$ is the usual sample distribution function for a SRS of the same size $n = mk$. They also demonstrate how to use $\hat{F}_{\text{RSS}}(t)$ in conjunction with the Kolmogorov–Smirnov statistic to provide simultaneous confidence bands for the distribution function $F(t)$.

Kvam and Samaniego (1993) consider competitors to $\hat{F}_{\text{RSS}}(t)$ that allow for differential weightings of the RSS observations in the averaging process. Their approach leads to more efficient estimators than $\hat{F}_{\text{RSS}}(t)$ under a variety of specific distributional assumptions about $F(t)$. Kvam and Samaniego (1994) use a similar approach to obtain a nonparametric maximum likelihood estimator for $F(t)$ based on RSS data. The estimators proposed by Kvam and Samaniego in these two papers also automatically accommodate unbalanced RSS data, where the different order statistics are not equally represented in the collected RSS (see the "Unbalanced Ranked Set Sampling" discussion in Section 15.5 for more about unbalanced RSS options). The original Stokes and Sager estimator $\hat{F}_{\text{RSS}}(t)$ does not immediately adapt to such unbalanced RSSs.

6. *Estimation of a Population Proportion.* For populations consisting of binary data corresponding to "success" or "failure," for example, the feature of interest is the proportion, p, of "successes" in the population. If we assign the numerical values of 0 and 1 to "failure" and "success," respectively, then the proportion p is nothing more than the population average μ as discussed in this section, so that one natural estimator for p is simply the sample average, $\hat{p}_{\text{RSS}}$, corresponding to the percentage of 'successes' observed in the RSS. However, this naive estimator does not fully utilize the additional information incorporated in the RSS data via the prior ranking process; that is, unlike an SRS, not all "successes" in an RSS should be treated equally.

   Taking into account this special information associated with the different RSS observations, Terpstra (2004) developed the RSS maximum likelihood estimator, $\hat{p}_{\text{RSS, MLE}}$, for a population proportion p. He showed that $\hat{p}_{\text{RSS, MLE}}$ is slightly more efficient than $\hat{p}_{\text{RSS}}$ and uniformly more efficient than the standard sample percentage of "successes," $\hat{p}_{\text{SRS}}$, for an SRS of the same size.

   Another factor that is important to consider when applying RSS methodology to the estimation of a population proportion is the curious aspect of initially "ranking" binary data to implement the RSS structure. This is not an issue if individuals are used to subjectively judgment rank the candidates within a set with respect to their relative likelihoods of being "successes." However, if we wish to use additional quantitative information from the population to aid in these within-sets binary rankings, then appropriate mechanisms are required to enable that process. Terpstra and Liudahl (2004) suggested the use of a single concomitant to facilitate the ranking of binary data, and Chen, Stasny, and Wolfe (2005) expanded on this concept through the use of logistic regression to incorporate multiple concomitants in a formal mechanism for ranking such data. They found that the use of logistic regression substantially improves the accuracy of the preliminary ranking in the RSS process, which, in turn, can lead to considerable gains in precision for estimation of the population proportion.

7. *Ordered Categorical Data.* Chen, Stasny, and Wolfe (2008) extended the application of RSS methodology to ordered categorical variables with the goal of estimating the probabilities of all of the categories. They used ordinal logistic regression to aid in the ranking of the ordinal variable of interest and proposed an optimal allocation scheme, as well as methods for implementing it.

Properties

1. *Consistency, Asymptotic Normality, and Efficiency of the RSS Mean Estimator* $\hat{\mu}_{\text{RSS}}$. See Takahasi and Wakimoto (1968), Takahasi (1970), and Dell and Clutter (1972).

Problems

21. Using the RSS of 20 invoices obtained in Problem 2, estimate the average audited (true) invoice value for the entire population of invoices. Compare this estimate with the estimate for the same quantity based on the SRS of 20 invoices obtained in Problem 1.

22. Consider the population of 15% overstated book value accounting data given in Table 15.1. (See Example 15.3 for more discussion about these data.) Select an SRS of 96 invoices from this population and use these data to estimate the average audited (true) invoice value for the entire population of invoices. Compare this value with the corresponding estimates of the same quantity using the three RSSs of 96 invoices each obtained in Problems 4–6.

23. Estimate the percentage of fraudulent invoices in the entire population using the SRS of 96 invoices from Problem 22 and, separately, using the three RSSs of 96 invoices each obtained in Problems 4–6. How do these four estimates compare with the actual 15% of fraudulent invoices for the entire population?

24. Using the SRS of 96 invoices from Problem 22 and, separately, the three RSSs of 96 invoices each from Problems 4–6, estimate the total amount of overstatement (fraud) in the entire population. Compare these four estimates.

25. Using the RSS of 25 subjects obtained in Problem 12, estimate the average BMI for the entire NHANES III population. Compare this estimate with the estimate for the same quantity based on the SRS of 25 subjects obtained in Problem 9. How do these two estimates compare with the actual average BMI for the entire population of the NHANES subjects?

26. Consider the population of NHANES III data given in Table 15.2. (See Problem 9 for more discussion about the NHANES III study.) Select an SRS of 96 subjects from this population and use these data to estimate the average BMI for the entire NHANES III population. Compare this value with the four estimates of the same quantity using the RSSs of 96 NHANES III subjects each obtained in Problems 14–17. How do these five estimates compare with the actual average BMI for the entire population?

27. An individual is considered to be overweight if his or her BMI is at least 25 (Kuczmarski et al., 1997). Estimate the percentage of overweight subjects in the NHANES III population using the SRS of 96 subjects from Problem 26 and, separately, using the four RSSs of 96 subjects each obtained in Problems 14–17. How do these five estimates compare with the actual percentage of overweight subjects in the NHANES III population?

28. An individual is considered to be obese if his or her BMI is at least 30 (Kuczmarski et al., 1997). Estimate the percentage of obese subjects in the NHANES III population using the SRS of 96 subjects from Problem 26 and, separately, using the four RSSs of 96 subjects each obtained in Problems 14–17. How do these five estimates compare with the actual percentage of obese subjects in the NHANES III population?

29. Consider the first four (Arm Circumference (ArmCir), BMI) pairs in the NHANES III data given in Table 15.2, namely,

$$(34.9, 25.5),\ (32.8, 23.4),\ (33.3, 27.6),\ (36.1, 29.4).$$

(a) In part (a) of Problem 20 you listed the possible SRSs of size $n = 2$ that could be selected from this subset of four BMI values. Calculate the sample mean for each of these potential

SRSs. Calculate the average of these SRS means and compare it with the mean of the full subset of four BMI values.

(b) In part (b) of Problem 20 you listed the possible ranked set samples of size $n = 2$ that could be selected from this subset of four BMI values using a set size of $k = 2$ and arm circumference as the auxiliary variable for the ranking process. Calculate the sample mean for each of these potential RSSs. Calculate the average of these RSS means and compare it with the mean of the full subset of four BMI values.

(c) What important properties of the SRS mean and the RSS mean are illustrated by what you found in parts (a) and (b)?

30. Consider the population of 15% overstated book value accounting data given in Table 15.1. (See Problem 1 for more discussion about these data.) Use the sample distribution function $\hat{F}_{\text{SRS}}(t)$ for the SRS of 96 audited values from this population obtained in Problem 22 to estimate the distribution function $F(t)$ for the full population of audited invoice values. Compare $\hat{F}_{\text{SRS}}(t)$ with the Stokes and Sager (1988) RSS estimator $\hat{F}_{\text{RSS}}(t)$ (15.14) given in Comment 5 using the RSS of 96 audited invoice values obtained in Problem 4.

31. Consider the population of NHANES III data given in Table 15.2. Use the RSS of 96 NHANES III subjects obtained in Problem 17 and the MacEachern et al. (2002) estimator $\hat{\sigma}^2_{\text{MOSW}}$ (15.13) given in Comment 4 to estimate the variance of the BMI values for the NHANES III population.

32. One of the research focuses of the Research Farm at Ataturk University, Erzurum, Turkey, is on increasing meat quality and production in sheep. As part of this research, the sheep are sampled periodically in order to monitor their biological growth and provide estimates for the population means of specific traits of interest. Young sheep are very active animals and it is labor intensive to hold them secure during the measurement process. As a result, it is important to minimize the number of sheep included in these samples while still obtaining reliable information about the broader population. RSS can be used effectively to reduce the number of lambs that must be measured at each of these periodic samplings by taking advantage of readily available archival information about the birth weight of a lamb and its mother's weight at the time of mating, both of which are positively correlated with the lamb's weight throughout its growth cycle. Table 15.5 contains the birth weight, mother's weight at the time of mating, and the weight at 7 months for 224 lambs at the Research Farm of Ataturk University. Use the mother's weight at the time of mating as the auxiliary ranking variable and set size $k = 3$ to select an RSS (using SRSs without replacement throughout the ranking process) of size $n = 21$. Obtain the RSS estimate of the average 7-month weight for the entire population of 224 lambs based on the 7-month weights of these 21 lambs.

33. Consider the weight data for 224 lambs at the Research Farm at Ataturk University, Erzurum, Turkey given in Table 15.5. (See Problem 32 for more discussion about these data.) Use the lamb's birth weight as the auxiliary ranking variable and set size $k = 3$ to select an RSS (using SRSs without replacement throughout the ranking process) of size $n = 21$ from the population of 224 lambs. Obtain the RSS estimate of the average 7-month weight for the entire population of lambs based on the 7-month weights of these 21 lambs. Compare this estimate with that obtained in Problem 32 using the mother's weight at the time of mating as the auxiliary ranking variable. Which of these two would you expect to be the more accurate estimate of the population average 7-month weight? Why?

34. Consider the population of 15% overstated book value accounting data given in Table 15.1. (See Example 15.3 for more discussion about these data.)

(a) Select an SRS of 200 invoices from this population and use these sample data to estimate both the percentage of fraudulent accounts in the population and the total amount of overstatement (fraud) in the population.

(b) Use book values to perform your judgment rankings and a set size of $k = 4$ to select a balanced RSS of 200 invoices for auditing. Use these RSS data to estimate both the percentage of fraudulent accounts in the population and the total amount of overstatement (fraud) in the population.

Table 15.5 Sheep Weight (kg) from the Research Farm at Ataturk University

Mother mating weight	Lamb weight	
	at Birth	at 7 Months
50.5	3.7	25.0
44.3	3.5	26.5
47.7	3.7	23.5
44.4	3.9	23.5
44.8	4.4	27.2
51.9	4.0	26.6
51.2	4.0	23.5
56.8	4.8	31.0
58.4	5.2	34.5
51.5	3.9	27.5
48.2	4.9	31.0
50.0	3.1	22.7
54.9	4.2	27.1
52.3	4.4	27.9
58.5	4.5	29.0
52.6	5.0	30.3
55.8	4.7	28.4
50.9	3.7	23.7
48.0	3.4	21.6
55.2	6.3	35.5
52.1	4.2	26.3
53.3	5.6	31.4
50.2	5.9	34.4
52.7	4.2	27.0
50.0	4.0	25.5
55.6	4.2	31.4
49.1	4.5	25.5
44.5	4.1	25.9
53.3	4.4	30.5
54.4	3.9	25.3
50.8	4.8	28.8
52.2	4.4	28.5
56.2	4.5	29.5
52.5	5.9	37.0
54.0	4.3	31.8
59.2	6.7	40.5
50.8	4.0	28.5
51.8	4.0	27.1
55.7	4.5	30.5
43.3	3.2	21.5
57.5	4.9	30.5
51.3	4.3	27.0
54.8	6.6	37.0
56.1	6.7	37.7
51.3	3.2	24.0
43.1	3.8	25.5
46.0	4.3	27.9
54.6	5.3	33.1
54.7	4.7	27.5
58.6	4.4	28.5
47.1	3.8	24.6
49.7	2.6	21.9

Table 15.5 (*Continued*)

Mother mating weight	Lamb weight	
	at Birth	at 7 Months
54.7	4.3	32.5
52.2	3.9	25.9
53.2	2.6	21.5
53.1	4.1	25.0
46.2	3.5	26.6
50.9	3.9	25.8
45.4	3.0	22.7
51.1	3.7	25.3
56.5	4.8	34.0
52.5	4.5	30.5
50.5	4.0	27.5
47.4	4.5	27.0
47.9	3.9	28.0
43.3	3.6	22.9
52.6	4.6	31.0
55.9	4.4	28.5
59.9	4.7	33.0
57.8	4.8	31.5
56.5	4.5	25.5
50.2	3.9	26.7
56.7	3.6	27.0
49.1	3.0	22.0
49.0	3.0	21.3
48.3	3.3	26.5
48.9	2.9	21.6
46.5	3.7	21.8
43.9	3.6	20.3
47.1	2.6	22.4
42.8	2.5	21.3
45.2	4.2	23.0
45.2	4.7	28.4
46.6	4.4	28.1
50.3	5.2	33.5
53.3	5.2	31.0
48.6	4.8	28.8
54.5	6.6	36.9
47.5	3.9	23.0
51.5	4.2	26.2
63.7	5.2	36.0
52.4	4.6	28.7
50.9	5.0	29.0
47.5	5.2	31.0
53.9	5.4	30.0
52.3	4.8	28.5
52.7	4.5	28.5
55.1	4.5	26.5
55.0	5.4	31.0
61.7	5.8	35.0
55.1	5.0	33.0
54.7	4.7	34.0
50.7	3.5	25.0
47.1	4.2	28.0

(*continued*)

Table 15.5 (*Continued*)

Mother mating weight	Lamb weight	
	at Birth	at 7 Months
60.8	5.1	32.1
53.5	4.5	28.3
53.2	3.2	24.0
55.0	4.4	29.6
57.2	4.8	29.0
59.6	5.0	29.5
46.5	3.8	26.5
48.4	4.3	24.0
51.3	4.4	31.0
56.7	5.3	33.4
56.3	4.0	30.0
55.2	4.6	27.0
52.1	4.6	28.5
51.4	4.6	27.4
58.7	4.7	34.0
53.6	4.3	31.6
53.4	5.0	34.5
46.6	3.8	25.5
51.2	4.7	33.5
49.8	4.8	27.2
49.5	5.1	29.5
46.8	5.3	31.7
56.8	6.1	33.5
50.7	5.7	34.0
52.0	5.2	30.0
52.3	3.3	21.0
52.4	3.1	23.5
43.8	4.0	23.9
56.2	3.9	25.9
54.7	4.9	31.0
50.0	4.2	25.0
51.9	4.8	26.0
51.6	4.9	29.9
48.3	4.2	28.5
49.8	3.6	24.5
53.9	4.5	29.0
53.5	4.0	27.6
52.0	3.3	23.6
51.1	3.3	22.3
60.2	4.7	32.5
46.0	4.2	24.5
56.3	4.1	26.5
53.8	4.3	28.5
44.1	4.4	27.2
58.8	4.6	33.0
52.6	4.3	27.2
52.9	3.9	27.6
53.8	4.4	25.9
54.1	3.4	24.2
54.1	4.1	25.5
53.5	4.1	27.0
56.0	5.3	29.5
55.1	3.8	26.3

Table 15.5 (*Continued*)

Mother mating weight	Lamb weight	
	at Birth	at 7 Months
54.9	4.7	31.5
52.2	4.5	26.0
55.7	4.1	31.0
55.7	3.7	24.0
55.0	3.2	24.2
59.6	3.6	29.0
59.8	4.1	29.0
61.3	5.7	35.6
49.3	4.4	24.9
56.1	5.6	35.5
55.5	3.1	25.0
58.8	5.1	31.5
49.9	3.2	23.5
56.8	4.1	28.5
63.2	5.2	34.0
57.6	4.3	31.0
49.9	5.6	31.0
54.9	4.6	32.0
50.0	4.5	28.0
49.6	2.9	23.4
53.3	4.2	25.0
52.3	4.7	30.5
58.0	4.7	29.5
59.9	4.2	30.5
49.0	3.6	25.9
53.3	3.4	24.5
56.0	4.2	28.6
51.5	5.0	33.5
51.3	3.6	25.6
52.3	5.1	29.0
45.9	4.2	24.0
52.7	4.4	29.7
54.5	5.0	30.3
46.7	3.8	22.0
60.3	5.4	35.1
51.4	4.7	29.5
54.2	6.0	33.0
53.9	4.4	26.2
45.4	3.4	23.5
51.1	5.1	28.0
57.0	4.5	25.5
49.0	4.9	26.0
52.2	3.2	24.6
60.0	4.9	32.2
45.3	3.7	25.4
44.2	3.5	20.5
58.2	4.9	35.0
52.8	4.7	25.5
49.6	4.8	27.0
50.9	3.7	27.6
51.4	3.5	27.9
52.3	4.9	30.2

(*continued*)

Table 15.5 (*Continued*)

Mother mating weight	Lamb weight	
	at Birth	at 7 Months
60.1	5.6	32.0
57.5	4.9	28.5
50.6	4.8	27.0
42.2	2.7	22.0
44.3	4.4	27.5
48.1	5.1	31.0
51.2	4.4	25.5
48.0	4.8	29.5
58.5	5.1	34.0
52.6	4.9	33.0
52.2	4.0	28.5
50.9	3.8	26.2
44.5	3.8	21.4
53.1	2.8	23.9
51.6	4.7	27.8

Source: Ö. Özturk, Ö. C. Bilgin, and D. A. Wolfe (2005).

(c) Compare the results obtained in parts (a) and (b) with those discussed in Example 15.3, as well as with the true percentage of fraudulent accounts in the population and the true amount of overstatement (fraud) in the population.

35. Consider the population of NHANES III data given in Table 15.2. (See Problem 9 for more discussion about the NHANES III study.)

(a) Select 20 independent SRSs of $n = 96$ subjects each from this NHANES III population and compute the average BMI for each of these SRSs. Calculate the sample standard deviation for these 20 SRS means.

(b) Using buttocks circumference as the auxiliary ranking variable and set size $k = 6$, select 20 independent RSSs of size $n = 96$ each from this NHANES III population and compute the average BMI for each of these RSSs. Calculate the sample standard deviation for these 20 RSS means.

(c) Considering the results of parts (a) and (b), discuss the relative effectiveness of using an SRS versus an RSS (with buttocks circumference as the ranking variable) of the same size to estimate the average BMI value for the entire NHANES III population.

36. The average tree height for a heavily forested area is almost impossible to fully quantify. Statistical sampling approaches can be used to obtain a reliable estimate for this attribute, but even in those approaches the size of the required sample can be prohibitive, both in terms of time commitment and expense. Fortunately, the height of a tree is highly correlated with the more readily available diameter of the tree at breast height. This feature can be used quite effectively in conjunction with an RSS approach to reduce the number of trees for which we must obtain the actual tree height measurement in order to obtain a reliable estimate of the average tree height in the entire forest.

Platt, Evans, and Rathbun (1988) studied the composition and dynamics of an old-growth longleaf pine population on the Wade Tract in southern Thomas County, Georgia. They obtained a variety of measurements for the longleaf pine trees in this population. In Table 15.6, we present the full height in feet and the diameter at breast height in centimeters for 396 of the trees in their study.

Using the easily measured tree diameter at breast height as the auxiliary ranking variable and a set size of $k = 6$, obtain an RSS (without replacement in the sampling process) of size $n = 24$ trees from this population of 396 trees. Estimate the average tree height of the entire population of 396 trees using the measured heights for the 24 trees in the RSS.

Table 15.6 Conifer Tree Diameter (cm) and Height (ft)

Tree	Diameter	Height	Tree	Diameter	Height
1	15.9	28	199	40.2	75
2	22.0	26	200	66.8	223
3	56.9	119	201	4.1	11
4	9.6	16	202	60.6	180
5	24.6	43	203	8.0	15
6	3.3	7	204	17.2	43
7	11.4	21	205	22.0	46
8	4.7	6	206	15.9	39
9	21.3	40	207	3.1	4
10	16.8	28	208	4.5	12
11	5.1	12	209	32.0	65
12	7.5	22	210	46.9	126
13	3.1	7	211	36.4	103
14	4.9	7	212	25.4	64
15	6.1	9	213	40.0	82
16	5.5	12	214	40.4	87
17	6.5	11	215	19.8	42
18	5.6	14	216	30.5	37
19	6.9	11	217	37.7	183
20	3.8	6	218	22.1	33
21	9.7	27	219	5.5	6
22	6.9	16	220	28.4	76
23	4.1	8	221	46.4	120
24	58.5	192	222	15.8	33
25	46.0	203	223	45.9	202
26	22.2	51	224	33.5	82
27	3.7	5	225	36.7	77
28	52.9	162	226	44.0	105
29	63.2	223	227	51.6	197
30	46.5	211	228	45.0	78
31	56.3	196	229	34.0	99
32	21.9	43	230	53.1	198
33	11.0	20	231	30.8	85
34	4.7	14	232	17.2	24
35	11.0	19	233	57.0	213
36	58.8	222	234	6.3	9
37	3.5	4	235	44.2	216
38	10.1	28	236	3.0	4
39	16.9	38	237	36.4	62
40	10.8	26	238	2.7	3
41	9.0	21	239	4.4	7
42	8.0	19	240	41.4	177
43	17.8	38	241	3.4	7
44	23.9	37	242	8.4	25
45	2.3	5	243	4.8	12
46	5.8	13	244	4.2	5
47	6.0	16	245	6.3	16
48	8.8	23	246	32.6	67
49	9.9	20	247	15.3	31
50	14.6	34	248	38.6	42
51	10.8	29	249	5.2	6
52	44.2	181	250	61.8	239
53	12.9	16	251	10.9	33

(*continued*)

Table 15.6 (*Continued*)

Tree	Diameter	Height	Tree	Diameter	Height
54	28.0	77	252	3.5	6
55	39.8	76	253	2.5	4
56	20.4	37	254	10.9	26
57	47.3	111	255	8.9	24
58	35.7	66	256	21.0	67
59	44.9	87	257	44.1	107
60	8.7	25	258	7.0	16
61	24.3	46	259	9.4	27
62	15.7	35	260	8.0	17
63	30.9	54	261	23.0	59
64	69.2	131	262	11.6	35
65	24.1	72	263	33.0	90
66	4.2	8	264	7.5	17
67	3.8	8	265	17.5	46
68	41.2	94	266	8.9	33
69	39.8	68	267	47.4	53
70	18.6	33	268	22.0	49
71	38.7	68	269	6.8	18
72	12.2	17	270	7.5	18
73	6.0	16	271	22.2	32
74	8.0	14	272	19.3	25
75	13.5	19	273	14.5	22
76	20.1	32	274	3.5	5
77	57.4	202	275	10.9	26
78	8.2	22	276	14.7	33
79	32.7	41	277	12.5	34
80	9.4	23	278	18.7	35
81	8.9	25	279	20.5	38
82	9.2	18	280	11.5	26
83	6.1	14	281	43.7	92
84	7.5	19	282	10.1	36
85	52.3	152	283	42.1	70
86	15.5	25	284	41.8	92
87	23.7	51	285	21.9	70
88	67.1	208	286	56.9	113
89	12.3	16	287	40.5	83
90	14.0	16	288	15.9	76
91	4.9	9	289	18.8	58
92	5.5	8	290	26.5	89
93	7.6	17	291	42.2	133
94	3.5	5	292	39.8	196
95	6.3	18	293	48.2	197
96	19.0	39	294	25.5	40
97	2.7	5	295	19.6	40
98	8.2	24	296	59.4	176
99	7.6	20	297	9.3	25
100	9.2	27	298	19.8	33
101	5.9	9	299	34.0	42
102	6.2	12	300	4.9	6
103	13.3	22	301	8.3	14
104	13.4	30	302	3.7	8
105	33.9	82	303	32.7	53
106	33.7	93	304	2.6	7
107	8.3	26	305	44.8	140
108	48.0	99	306	10.3	21

Table 15.6 (*Continued*)

Tree	Diameter	Height	Tree	Diameter	Height
109	40.4	78	307	28.5	32
110	8.6	22	308	34.0	119
111	16.0	26	309	36.6	81
112	29.1	49	310	50.8	106
113	18.4	22	311	29.2	78
114	26.8	37	312	8.5	21
115	6.2	7	313	23.4	35
116	2.9	6	314	7.9	15
117	3.0	8	315	44.6	149
118	14.6	20	316	2.5	4
119	18.4	32	317	9.4	17
120	15.0	34	318	3.0	6
121	18.4	41	319	2.8	3
122	44.5	64	320	3.0	5
123	4.5	8	321	4.1	8
124	10.4	20	322	23.4	42
125	24.0	37	323	59.0	189
126	5.1	10	324	5.2	8
127	5.3	13	325	8.5	10
128	2.5	4	326	7.8	15
129	2.2	3	327	44.9	140
130	3.1	4	328	54.4	104
131	2.6	4	329	47.9	129
132	8.1	26	330	41.3	94
133	12.4	31	331	38.8	91
134	15.1	34	332	41.1	105
135	12.7	38	333	39.0	116
136	49.0	96	334	45.4	140
137	20.8	35	335	47.9	137
138	11.9	18	336	53.7	105
139	47.6	154	337	43.5	96
140	10.6	32	338	18.7	68
141	22.9	33	339	57.8	188
142	10.6	27	340	14.9	23
143	49.7	103	341	4.5	10
144	50.6	122	342	8.8	22
145	19.1	40	343	23.6	26
146	53.0	114	344	11.5	21
147	18.0	82	345	20.0	27
148	44.4	105	346	8.3	19
149	10.8	35	347	12.6	30
150	51.7	219	348	5.8	14
151	22.6	48	349	12.9	25
152	7.7	19	350	5.4	11
153	43.5	60	351	22.5	42
154	3.1	3	352	11.8	32
155	5.0	13	353	51.2	203
156	4.4	8	354	45.3	85
157	3.3	5	355	48.7	120
158	2.6	5	356	6.6	20
159	53.5	211	357	16.7	33
160	48.9	206	358	12.3	23
161	47.8	176	359	6.5	15

(*continued*)

Table 15.6 (*Continued*)

Tree	Diameter	Height	Tree	Diameter	Height
162	17.2	37	360	53.0	106
163	28.6	45	361	18.1	21
164	10.8	31	362	2.4	5
165	50.1	212	363	5.8	11
166	4.7	8	364	27.0	29
167	5.3	10	365	19.9	24
168	10.6	19	366	17.5	22
169	3.7	6	367	62.5	232
170	3.9	8	368	44.6	92
171	5.3	12	369	38.0	167
172	2.5	3	370	3.2	2
173	13.2	38	371	13.4	21
174	17.1	37	372	5.7	14
175	13.9	33	373	3.6	5
176	8.0	21	374	2.6	3
177	8.5	27	375	75.4	244
178	50.1	109	376	2.2	5
179	6.8	18	377	3.7	7
180	19.9	55	378	3.1	4
181	17.5	47	379	7.2	26
182	6.8	21	380	8.2	20
183	10.9	33	381	3.2	5
184	11.2	23	382	2.5	3
185	20.2	38	383	4.0	11
186	19.6	26	384	1.8	1
187	18.4	46	385	2.7	3
188	50.9	84	386	9.9	21
189	17.6	42	387	6.3	11
190	44.1	113	388	3.2	11
191	17.0	31	389	3.3	5
192	46.9	135	390	5.0	12
193	2.8	6	391	3.7	6
194	25.5	40	392	2.0	5
195	14.5	28	393	5.1	13
196	14.1	40	394	6.0	12
197	47.1	85	395	3.8	8
198	42.2	93	396	3.5	9

Source: W. J. Platt, G. W. Evans, and S. L. Rathbun (1988) and Z. Chen, Z. D. Bai, and B. K. Sinha (2004).

15.4 RANKED SET SAMPLE ANALOGS OF THE MANN–WHITNEY–WILCOXON TWO-SAMPLE PROCEDURES (BOHN–WOLFE)

Data. We obtain $N = m + n$ RSS observations, m from population 1 and n from population 2.

Assumptions

A1. The m balanced RSS observations from population 1 are collected using c cycles of set size k each, where $m = kc$, and the RSS observations from cycle

1 are denoted by $(X_{[1]1}, X_{[2]1}, \ldots, X_{[k]1})$, the RSS observations from cycle 2 are denoted by $(X_{[1]2}, X_{[2]2}, \ldots, X_{[k]2}), \ldots$, and the RSS observations from the final cycle c are denoted by $(X_{[1]c}, X_{[2]c}, \ldots, X_{[k]c})$.

A2. The n balanced RSS observations from population 2 are collected using d cycles of set size q each, where $n = qd$, and the RSS observations from cycle 1 are denoted by $(Y_{[1]1}, Y_{[2]1}, \ldots, Y_{[q]1})$, the RSS observations from cycle 2 are denoted by $(Y_{[1]2}, Y_{[2]2}, \ldots, Y_{[q]2}), \ldots$, and the RSS observations from the final cycle d are denoted by $(Y_{[1]d}, Y_{[2]d}, \ldots, Y_{[q]d})$.

A3. All $N = m + n$ RSS observations are mutually independent. Thus, in addition to the independence within the RSSs from populations 1 and 2, we also have independence across the two populations.

A4. Populations 1 and 2 are both continuous populations.

A5. The ranking processes used to obtain the two RSSs are perfect, so that the RSS observations are true order statistics from their respective populations.

Hypothesis

Let F be the distribution function corresponding to population 1 and let G be the distribution function corresponding to population 2.

The null hypothesis is

$$H_0 : F(t) = G(t), \quad \text{for every } t. \tag{15.15}$$

The null hypothesis asserts that the X variable and the Y variable have the same probability distribution, but the common distribution is not specified.

The alternative hypothesis in this two-sample location setting typically specifies that Y tends to be larger (or smaller) than X. One model that is useful to describe such alternatives is the translation model—also called the *location-shift model*—that was discussed in Chapter 4. This location-shift model is

$$G(t) = F(t - \Delta), \quad \text{for every } t. \tag{15.16}$$

Model (15.16) says population 2 is the same as population 1 except it is shifted by the amount Δ. Another way of writing this is

$$Y \stackrel{d}{=} X + \Delta,$$

where the symbol $\stackrel{d}{=}$ means "has the same distribution as." The parameter Δ is called the *location shift*. It is also known as the *treatment effect*. If X is a randomly selected value from population 1, the control population, and Y is a randomly selected value from population 2, the treatment population, then Δ is the expected effect due to the treatment. If Δ is positive, it is the expected increase due to the treatment, and if Δ is negative, it is the expected decrease due to the treatment. If the mean $E(X)$ of population 1 exists, then letting $E(Y)$ denote the mean of population 2, we have

$$\Delta = E(Y) - E(X),$$

the difference in population means. In terms of the location-shift model, the null hypothesis H_0 reduces to

$$H_0 : \Delta = 0,$$

the hypothesis that asserts the population means are equal or, equivalently, that the treatment has no effect.

We note that although we find it convenient to use the "treatment" and "control" terminology, many situations will arise in which we want to compare two populations, neither of which can be described as a control population. The procedures of this section are applicable even when there are no natural control or treatment designations for the two populations.

Procedure

To compute the Bohn–Wolfe (1992) statistic BW, we follow the lead of the two-sample Mann–Whitney U statistic discussed in Comment 4.7 by computing the $mn = kcqd$ count statistics $\phi(X_{[s]t}, Y_{[u]v})$, for $s = 1, \ldots, k$; $t = 1, \ldots, c$; $u = 1, \ldots, q$; $v = 1, \ldots, d$, where

$$\phi(X_{[s]t}, Y_{[u]v}) = \begin{cases} 1, & \text{if } X_{[s]t} < Y_{[u]v} \\ 0, & \text{otherwise.} \end{cases} \tag{15.17}$$

The Bohn–Wolfe statistic is then

$$\begin{aligned} \text{BW} &= \sum_{s=1}^{k} \sum_{t=1}^{c} \sum_{u=1}^{q} \sum_{v=1}^{d} \phi(X_{[s]t}, Y_{[u]v}) \\ &= (\#\ of\ X_{[s]t}{}'s < Y_{[u]v}{}'s\ in\ the\ RSS\ data). \end{aligned} \tag{15.18}$$

a. *One-Sided Upper-Tail Test.* To test

$$H_0 : \Delta = 0$$

versus

$$H_1 : \Delta > 0,$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if BW} \geq \text{bw}_\alpha; \quad \text{otherwise do not reject,} \tag{15.19}$$

where the constant bw_α is chosen to make the type I error probability equal to α. Thus the constant bw_α is the upper α percentile for the null ($\Delta = 0$) distribution of BW. Comment 10 explains how to obtain the critical value bw_α for set sizes k and q, cycle sizes c and d, and available values of α.

b. *One-Sided Lower-Tail Test.* To test

$$H_0 : \Delta = 0$$

versus

$$H_2 : \Delta < 0,$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } \text{BW} \leq (mn - \text{bw}_\alpha); \quad \text{otherwise do not reject.} \tag{15.20}$$

c. *Two-Sided Test.* To test

$$H_0 : \Delta = 0$$

versus

$$H_3 : \Delta \neq 0,$$

at the α level of significance,

$$\text{Reject } H_0 \text{ if } \text{BW} \geq \text{bw}_{\alpha/2} \text{ or } \text{BW} \leq (mn - \text{bw}_{\alpha/2}); \quad \text{otherwise do not reject.} \tag{15.21}$$

The two-sided procedure given by (15.21) is the two-sided symmetric test with $\alpha/2$ probability in each tail of the distribution.

Large-Sample Approximation

The large-sample approximation is based on the asymptotic normality of BW, suitably standardized. We consider here only the case where we have a common set size $k = q = 2$ and the same number of observations from each population as well (i.e., the same number of cycles $c = d$). For this setting, the standardized version of BW when the null hypothesis H_0 is true is given by

$$\text{BW}^* = \sqrt{\frac{9}{8c^3}}\big[\text{BW} - E_0(\text{BW})\big] = \sqrt{\frac{9}{8c^3}}\big[\text{BW} - 2c^2\big]. \tag{15.22}$$

When H_0 is true, BW^* has, as the common number of cycles c tends to infinity, an asymptotic $N(0, 1)$ distribution.

The normal theory approximation to procedure (15.19) is

$$\text{Reject } H_0 \text{ if } \text{BW}^* \geq z_\alpha; \quad \text{otherwise do not reject.} \tag{15.23}$$

The normal theory approximation to procedure (15.20) is

$$\text{Reject } H_0 \text{ if } \text{BW}^* \leq -z_\alpha; \quad \text{otherwise do not reject.} \tag{15.24}$$

The normal theory approximation to procedure (15.21) is

$$\text{Reject } H_0 \text{ if } |\text{BW}^*| \geq z_{\alpha/2}; \quad \text{otherwise do not reject.} \tag{15.25}$$

(Additional discussion about the form of the appropriate standardization for BW and the resulting large-sample approximation for general k, q, c, and d is provided in Comment 12.)

Ties

If there are ties between the $X_{[s]t}'s$ and the $Y_{[u]v}'s$, replace $\phi(X_{[s]t}, Y_{[u]v})$ in (15.17) by

$$\phi^*(X_{[s]t}, Y_{[u]v}) = \begin{cases} 1, & \text{if } X_{[s]t} < Y_{[u]v} \\ \frac{1}{2}, & \text{if } X_{[s]t} = Y_{[u]v} \\ 0, & \text{if } X_{[s]t} > Y_{[u]v} \end{cases} \tag{15.26}$$

in the computation of BW (15.18). Thus, we add a count of $1/2$ to the value of BW for every tie between an $X_{[s]t}$ and a $Y_{[u]v}$.

EXAMPLE 15.4 ***Body Mass Index.***

The NHANES III survey, 1988–1994, was conducted by the National Center for Health Statistics, Centers for Disease Control and Prevention. This survey was designed to obtain nationally representative information on the health and nutritional status of the population of the United States. The data set contains information for 33,994 persons aged 2 months and older who participated in the survey. Specifically, it contains various body measurements and information on other health-related variables for the respondents. The survey used a complex, multistage cluster sample of households. (Since we are going to treat a subset of the NHANES sample as our population for this example, we ignore the complex nature of the sample design.)

BMI, which is commonly used to classify an adult's weight status, will be the variable of interest in this example. It is calculated as the ratio of weight (kg) to height squared (m^2). For consideration here, we exclude those NHANES III subjects who were either younger than 21 years or who were pregnant at the time of their NHANES interviews. The remaining 13,267 NHANES III subjects are viewed as our population for this example. In Table 15.2 in the NSM Third Edition R package we provide the following NHANES III data for each of these 13,267 subjects: Gender, Age, BMI, Arm Circumference (ArmCir), Buttocks Circumference (ButtocksCir), and Thigh Circumference (ThighCir).

In this example we use buttocks circumference (ButtocksCir) as the auxiliary ranking variable to obtain RSSs to compare the BMI values for males and females in the NHANES III population under consideration. In particular, letting $X(Y)$ denote the male (female) BMI values, respectively, we will use our RSS data to test the null hypothesis $H_0 : \Delta = 0$ against the one-sided alternative $H_1 : \Delta > 0$ corresponding to larger BMI values for females in the NHANES III population.

For illustrative purposes, we take a common set size $k = q = 4$ and a common cycle size $c = d = 5$ for the males and females, so that our measured BMI data will consist of balanced RSSs of $m = n = 20$ observations from each of the male and female subsets of the NHANES III subjects. Following the approach detailed in Example 15.3, we apply the R command `RSS(4,5, NHANES.III $ButtocksCir[NHANES.III$Gender == 1])`, to the male subset of the NHANES III population to obtain the following ordered balanced RSS of 20 male BMI values:

$X_{[s]t}$ measurements:	18.0	20.5	21.3	21.3	22.3	23.8	23.8
	24.6	25.0	25.2	25.3	25.9	26.1	27.0
	27.4	27.4	28.4	29.4	29.6	32.8.	

Similarly, we apply the R command `RSS(4,5, NHANES.III $ButtocksCir [NHANES.III$Gender = = 2])`, to the female subset of the NHANES III population to obtain the following ordered balanced RSS of 20 female BMI values:

$Y_{[u]v}$ measurements:							
	17.2	17.8	19.9	20.0	21.7	22.0	22.3
	23.1	23.9	25.8	27.1	29.6	30.1	30.3
	30.7	31.1	35.2	35.6	38.1	42.5.	

For significance level $\alpha = .0515$, we use the R command `cBohnWolfe(.0515, k, q, c, d)` with $k = q = 4$ and $c = d = 5$ (see Comment 10) to obtain the Bohn–Wolfe critical value `cBohnWolfe(.0515,4,4,5,5)` = $\text{bw}_{.0515} = 239$ and the test procedure (15.19) becomes

$$\text{Reject } H_0 : \Delta = 0 \text{ in favor of } H_1 : \Delta > 0 \text{ if BW} \geq 239.$$

Applying the Bohn–Wolfe statistic to the female and male RSS data, we see that $BW = 226$. Since this value of BW is less than the critical value 239, we do not reject $H_0 : \Delta = 0$ in favor of $H_1 : \Delta > 0$, leading to the conclusion that females do not tend to have larger BMI values than males in the NHANES III population. In fact, from the observed value BW = 226 we see that $P_0(\text{BW} \geq 226) =$ `pBohnWolfe (x, y,4, 4, 5, 5)` = .14. Thus the lowest significance level at which we can reject $H_0 : \Delta = 0$ in favor of $H_1 : \Delta > 0$ with the observed value of the test statistic BW = 226 is the P-value $\underline{\alpha} = .14$.

Comments

8. *Motivation for the Test.* When Δ is greater than 0, the RSS Y-values will tend to be larger than the corresponding RSS X-values. Hence, there will tend to be more $\phi(X_{[s]t}, Y_{[u]v})$ counts equal to 1 and the resulting value of BW will be large. This suggests rejecting H_0 in favor of $\Delta > 0$ for large values of BW and motivates procedure (15.19). An analogous motivation leads to procedure (15.20).
9. *Testing Δ is Equal to Some Specified Nonzero Value.* Procedures (15.19), (15.20), and (15.21) and the corresponding large-sample approximations given by procedures (15.23), (15.24), and (15.25) are for testing if Δ is equal to 0. To test $\Delta = \Delta_0$, where Δ_0 is some specified nonzero number, subtract Δ_0 from each $Y_{[u]v}$ value to form a pseudo-RSS, namely, $Y^*_{[u]v} = Y_{[u]v} - \Delta_0$, for $u = 1, \ldots, q$ and $v = 1, \ldots, d$. Then compute BW via (15.18) applied to the $X_{[s]t}$'s and $Y^*_{[u]v}$'s. The procedures (15.19), (15.20), and (15.21), and their corresponding large-sample approximations given by displays (15.23), (15.24), and (15.25), can then be carried out as described earlier.
10. *Derivation of the Distribution of BW under H_0 (No-ties Case).* Assume that the underlying common X and Y distribution under H_0 is continuous so that ties have probability zero of occurring and, in addition, assume that the judgment rankings are perfect for both the X and Y RSSs. Bohn and Wolfe (1992) showed that, just as for the SRS setting, the RSS Mann–Whitney statistic BW (with perfect rankings) is distribution-free under H_0 over the class of all continuous distributions. However, there is a major difference in how the null distributions of the test statistics and associated critical values are obtained for the SRS and RSS settings.

    For the SRS setting, the $N = (m + n)$ combined sample X and Y observations are not only mutually independent but also identically distributed. Thus, in

the case of SRSs, it suffices to look at each of the $\binom{N}{m}$ distinct (i.e., unchanged by permutations within the X's and Y's separately) ordered arrangements of these combined sample observations, and moreover, they are all equally likely under H_0. This makes tabulation of the associated null distribution for the SRS Mann–Whitney U statistic (4.15) relatively straightforward. (For a more detailed discussion, see Comment 4.3.)

However, the equally likely nature of these arrangements under H_0 does not carry over to the RSS setting, due to the fact that the ranked set X's and Y's, while still mutually independent, are no longer identically distributed. For example, even in the case of perfect rankings, there is nothing to prevent the smallest ordered item selected in a given cycle from being larger than the largest ordered item selected in the same cycle when the null hypothesis H_0 is true. While this probability will generally be small, it will not be zero as in the case of SRS. This means that for RSS data it is no longer sufficient to look at the $\binom{N}{m} = \binom{kc+qd}{kc}$ distinct (i.e., unchanged by permutations within the X's and Y's separately) ordered arrangements of the combined sample observations. Instead we need to calculate the probability of each of the $N! = (kc + qd)!$ possible permutations separately (no longer equally likely) and then combine these probabilities to obtain the null distribution for BW.

Fortunately, the probabilities of these $N!$ permutations under RSS still do not depend on the form of the common, continuous $F \equiv G$ under H_0 and perfect rankings, although the tabulation of these probabilities can be tedious. We illustrate the necessary computations with a small example.

Example.

For a single X and Y cycle (i.e., $c = d = 1$) and common X and Y set size $k = q = 2$, we must obtain the null probabilities for the $4! = 24$ different permutations. Under the assumption of perfect judgment rankings, the RSS observations $X_{(1)1}, X_{(2)1}, Y_{(1)1}$, and $Y_{(2)1}$ are independent order statistics with joint pdf given by

$$\begin{aligned} &g_{\mathrm{RSS}}(x_{(1)}, x_{(2)}, y_{(1)}, y_{(2)}) \\ &\quad = \left\{ \prod_{s=1}^{2} \frac{2!}{(s-1)!(2-s)!} [F(x_{(s)})]^{s-1} [1 - F(x_{(s)})]^{2-s} f(x_{(s)}) \right\} \\ &\qquad \times \left\{ \prod_{u=1}^{2} \frac{2!}{(u-1)!(2-u)!} [F(y_{(u)})]^{u-1} [1 - F(y_{(u)})]^{2-u} f(y_{(u)}) \right\}, \end{aligned}$$

which simplifies to

$$\begin{aligned} &g_{\mathrm{RSS}}(x_{(1)}, x_{(2)}, y_{(1)}, y_{(2)}) \\ &\quad = 16[1 - F(x_{(1)})][F(x_{(2)})][1 - F(y_{(1)})][F(y_{(2)})] \prod_{s=1}^{2} f(x_{(s)}) \prod_{u=1}^{2} f(y_{(u)}). \end{aligned}$$

Using this expression for g_{RSS} and straightforward integration, the null probabilities for each of the $4! = 24$ permutations of $X_{(1)1}, X_{(2)1}, Y_{(1)1}$, and $Y_{(2)1}$ can then

be computed by integrating over the appropriate region. Thus, for example, the four permutations $\{X_{(1)1} < Y_{(1)1} < X_{(2)1} < Y_{(2)1}\}$, $\{X_{(1)1} < Y_{(1)1} < Y_{(2)1} < X_{(2)1}\}$, $\{Y_{(1)1} < X_{(1)1} < Y_{(2)1} < X_{(2)1}\}$, and $\{Y_{(1)1} < X_{(1)1} < X_{(2)1} < Y_{(2)1}\}$ all have the same null probability of occurrence, p, given by

$$p = \int_{-\infty}^{\infty} \int_{-\infty}^{y_{(2)}} \int_{-\infty}^{x_{(2)}} \int_{-\infty}^{y_{(1)}} g_{\text{RSS}}(x_{(1)}, x_{(2)}, y_{(1)}, y_{(2)}) dx_{(1)} dy_{(1)} dx_{(2)} dy_{(2)} = 41/280.$$

Proceeding in this fashion for all 24 permutations yields the set of null probabilities (independent of the form of the continuous common distribution F under H_0 and perfect rankings) and the associated values of BW given in the following table.

Null Probabilities under Perfect Rankings and Values of BW for the 24 Permutations in an RSS with $c = d = 1$ and $k = q = 2$.

Permutation	Null probability	Value of BW
$y_{(2)} < y_{(1)} < x_{(2)} < x_{(1)}$	17 / 2520	0
$y_{(2)} < y_{(1)} < x_{(1)} < x_{(2)}$	7 / 360	0
$y_{(1)} < y_{(2)} < x_{(1)} < x_{(2)}$	137 / 2520	0
$y_{(1)} < y_{(2)} < x_{(2)} < x_{(1)}$	7 / 360	0
$y_{(1)} < x_{(1)} < y_{(2)} < x_{(2)}$	41 / 280	1
$y_{(1)} < x_{(2)} < y_{(2)} < x_{(1)}$	7 / 360	1
$y_{(2)} < x_{(1)} < y_{(1)} < x_{(2)}$	7 / 360	1
$y_{(2)} < x_{(2)} < y_{(1)} < x_{(1)}$	1 / 280	1
$x_{(1)} < y_{(1)} < y_{(2)} < x_{(2)}$	41 / 280	2
$x_{(1)} < y_{(2)} < y_{(1)} < x_{(2)}$	137 / 2520	2
$x_{(2)} < y_{(1)} < y_{(2)} < x_{(1)}$	17 / 2520	2
$x_{(2)} < y_{(2)} < y_{(1)} < x_{(1)}$	1 / 280	2
$y_{(1)} < x_{(1)} < x_{(2)} < y_{(2)}$	41 / 280	2
$y_{(1)} < x_{(2)} < x_{(1)} < y_{(2)}$	137 / 2520	2
$y_{(2)} < x_{(1)} < x_{(2)} < y_{(1)}$	17 / 2520	2
$y_{(2)} < x_{(2)} < x_{(1)} < y_{(1)}$	1 / 280	2
$x_{(1)} < y_{(1)} < x_{(2)} < y_{(2)}$	41 / 280	3
$x_{(1)} < y_{(2)} < x_{(2)} < y_{(1)}$	7 / 360	3
$x_{(2)} < y_{(1)} < x_{(1)} < y_{(2)}$	7 / 360	3
$x_{(2)} < y_{(2)} < x_{(1)} < y_{(1)}$	1 / 280	3
$x_{(1)} < x_{(2)} < y_{(1)} < y_{(2)}$	137 / 2520	4
$x_{(1)} < x_{(2)} < y_{(2)} < y_{(1)}$	7 / 360	4
$x_{(2)} < x_{(1)} < y_{(1)} < y_{(2)}$	7 / 360	4
$x_{(2)} < x_{(1)} < y_{(2)} < y_{(1)}$	17 / 2520	4

Combining the null probabilities for the various permutations with the associated values for BW, we see that the null distribution of BW under perfect rankings for this setting (i.e., $c = d = 1$ and $k = q = 2$) is given by

$$P_0(\text{BW} = 0) = P_0(\text{BW} = 4) = 1/10,$$
$$P_0(\text{BW} = 1) = P_0(\text{BW} = 3) = 17/90$$
$$\text{and} \quad P_0(\text{BW} = 2) = 19/45.$$

Note that the null distribution is symmetric about its mean $E_0(\text{BW}) = kcqd/2 = 2$.

Observe that we have derived the null distribution of BW without specifying a form of the common underlying continuous distribution of the two populations under H_0. This is why the procedures based on BW are called *distribution-free procedures*. From the null distribution of BW we can determine the critical value bw_α and control the probability α of falsely rejecting H_0 when H_0 is true, and this error probability does not depend on the specific form of the common underlying continuous distribution, as long as the rankings used to obtain the RSS data are perfect. For the effect that imperfect rankings have on these type I error probabilities, see Comments 16 and 17.

For given set sizes k and q and cycle sizes c and d, the `R` command `cBohnWolfe`(α, k, q, c, d) can be used to find the available upper-tail critical values bw_α associated with the possible values of BW. For a given available significance level α, the critical value bw_α then corresponds to $P_0(\text{BW} \geq \text{bw}_\alpha) = \alpha$ and is given by `cBohnWolfe`$(\alpha, k, q, c, d) = \text{bw}_\alpha$. Thus, for example, for $k = 2$, $q = 3$, $c = 3$, and $d = 3$, we have $P_0(\text{BW} \geq 40) = .0303$, so that $\text{bw}_{.0303} =$ `cBohnWolfe(.0303, 2, 3, 3, 3)` $= 40$ for $k = 2$, $q = 3$, $c = 3$, and $d = 3$.

11. *Calculation of the Mean and Variance of BW*. Theoretical properties of the RSS Mann–Whitney statistic BW can be obtained by using standard results about the general class of U-statistics. (See Randles and Wolfe (1979) for a discussion of U-statistics.) Let

$$\gamma = \sum_{s=1}^{k} \sum_{u=1}^{q} P(X_{[s]1} < Y_{[u]1}). \tag{15.27}$$

Then γ is a two-sample, multivariate, estimable parameter of degree $(1, 1)$ and BW/cd is the multivariate U-statistic estimator for γ. Standard U-statistic arguments can be used to establish the following general (arbitrary distribution functions F and G) expressions for the expected value and variance of BW (see Hoeffding (1948a) and Bohn and Wolfe (1992) for more details):

$$E(\text{BW}) = cd\gamma \tag{15.28}$$

and

$$\sigma^2 = \text{Var}(\text{BW}) = cd[(d-1)\zeta_{1,0} + (c-1)\zeta_{0,1} + \zeta_{1,1}], \tag{15.29}$$

where

$$\zeta_{1,0} = \sum_{s=1}^{k}\sum_{u=1}^{q}\Big\{P(X_{[s]1} < \min(Y_{[u]1}, Y_{[u]2})) - [P(X_{[s]1} < Y_{[u]1})]^2\Big\}$$
$$+\sum_{s=1}^{k}\sum_{1\le u\ne v\le q}\sum\Big\{P(X_{[s]1} < \min(Y_{[u]1}, Y_{[v]1}))$$
$$-P(X_{[s]1} < Y_{[u]1})P(X_{[s]1} < Y_{[v]1})\Big\}, \tag{15.30}$$

$$\zeta_{0,1} = \sum_{s=1}^{k}\sum_{u=1}^{q}\Big\{P(\max(X_{[s]1}, X_{[s]2}) < Y_{[u]1}) - [P(X_{[s]1} < Y_{[u]1})]^2\Big\}$$
$$+\sum_{u=1}^{q}\sum_{1\le s\ne t\le k}\sum\Big\{P(\max(X_{[s]1}, X_{[t]1}) < Y_{[u]1})$$
$$-P(X_{[s]1} < Y_{[u]1})P(X_{[t]1} < Y_{[u]1})\Big\}, \tag{15.31}$$

and

$$\zeta_{1,1} = \sum_{s=1}^{k}\sum_{u=1}^{q}\Big\{P(X_{[s]1} < Y_{[u]1})[1 - P(X_{[s]1} < Y_{[u]1})]\Big\}$$
$$+\sum_{u=1}^{q}\sum_{1\le s\ne t\le k}\sum\Big\{P(\max(X_{[s]1}, X_{[t]1}) < Y_{[u]1})$$
$$-P(X_{[s]1} < Y_{[u]1})P(X_{[t]1} < Y_{[u]1})\Big\}$$
$$+\sum_{s=1}^{k}\sum_{1\le u\ne v\le q}\sum\Big\{P(X_{[s]1} < \min(Y_{[u]1}, Y_{[v]1}))$$
$$-P(X_{[s]1} < Y_{[u]1})P(X_{[s]1} < Y_{[v]1})\Big\}. \tag{15.32}$$

We point out that these general expressions for the expected value and variance of BW are valid even if the judgment rankings are not perfect. Of course, in that setting, we would need to know the probability distributions of the imperfect judgment order statistics in order to actually compute the relevant probabilities. When the judgment rankings are perfect, however, the usual distributional properties of order statistics from continuous distributions can be used to evaluate these expressions.

12. *Large-Sample Approximation.* The asymptotic normality of the standardized Bohn–Wolfe statistic $\mathrm{BW}^* = \frac{\mathrm{BW}-E(\mathrm{BW})}{\sqrt{\mathrm{Var}(\mathrm{BW})}}$ also follows from the fact that BW/cd is the multivariate U-statistic estimator for γ (see Bohn and Wolfe (1992) for more details). Let $N = c + d$ and set $\lambda = \lim_{N\to\infty}(c/N)$. If $0 < \lambda < 1$ and $\frac{\zeta_{1,0}}{\lambda} + \frac{\zeta_{0,1}}{1-\lambda} > 0$, then $\frac{\sqrt{N}}{cd}$ $(\mathrm{BW} - E[\mathrm{BW}])$ has an asymptotic $(N \to \infty)$ normal

distribution with mean 0 and finite variance σ_∞^2 given by

$$\sigma_\infty^2 = \frac{\zeta_{1,0}}{\lambda} + \frac{\zeta_{0,1}}{1-\lambda}, \tag{15.33}$$

where $\zeta_{1,0}$ and $\zeta_{0,1}$ are as given in (15.30) and (15.31), respectively.

Under the null hypothesis $H_0 : \Delta = 0$ and perfect rankings, we have $E_0[\mathrm{BW}] = cdkq/2$, and the null asymptotic variance, $\sigma_{0(\infty)}^2$, from (15.33), does not depend on the form of the common underlying continuous F. Thus, the standardized test statistic BW^* is asymptotically ($N \to \infty$) distribution-free under H_0 and perfect rankings. Bohn and Wolfe (1992) provided detailed expressions for computing these null values of $\zeta_{0,1}$ and $\zeta_{1,0}$, and thus the value of $\sigma_{0(\infty)}^2$.

For the special case when the set sizes are both equal to 2 (i.e., $k = q = 2$), Bohn and Wolfe (1992) showed that $E_0[\mathrm{BW}] = 2cd$ and $\zeta_{1,0} = \zeta_{0,1} = 4/9$ under H_0, so that $\sigma_{0(\infty)}^2 = \frac{4}{9\lambda(1-\lambda)}$. Thus, when $k = q = 2$, the asymptotic ($N \to \infty$) null distribution of $\mathrm{BW}^* = \frac{\sqrt{N}}{cd}\left(\frac{\mathrm{BW}-E_0[\mathrm{BW}]}{\sigma_{0(\infty)}}\right) = \frac{\sqrt{N}}{cd}\left(\frac{\mathrm{BW}-2cd}{\sqrt{\frac{4}{9\lambda(1-\lambda)}}}\right)$ is standard normal. Replacing λ and $1-\lambda$ by $\frac{c}{N}$ and $\frac{d}{N}$, respectively, it follows from Slutsky's theorem that the asymptotic ($N \to \infty$) null distribution of $\frac{3}{2}\sqrt{\frac{1}{cdN}}(\mathrm{BW} - 2cd)$ is standard normal. When the cycle sizes are also equal (i.e., $c = d$), this simplifies even further to the result that the asymptotic ($N \to \infty$) null distribution of $\sqrt{\frac{9}{8c^3}}(\mathrm{BW} - 2c^2)$ is standard normal, as noted previously in (15.22) in the Large-Sample Approximation section.

It follows from this result that $P_0\left\{\sqrt{\frac{9}{8c^3}}(\mathrm{BW} - 2c^2) \geq z_\alpha\right\} \approx \alpha$, where z_α is the upper αth percentile for the standard normal distribution, when $k = q = 2$ and $c = d$, so that the approximate upper αth percentile for the null distribution of BW is then $\mathrm{bw}_\alpha \approx \sqrt{\frac{8c^3}{9}}z_\alpha + 2c^2$ for this setting.

13. *Symmetry of the Distribution of BW under the Null Hypothesis and Perfect Rankings.* When H_0 is true and the judgment rankings are perfect, the distribution of BW is symmetric about its mean $cdkq/2$ for any (k, q, c, d) configuration. This implies that when H_0 is true,

$$P_0(\mathrm{BW} \leq x) = P_0(\mathrm{BW} \geq cdkq - x), \tag{15.34}$$

for $x = 0, 1, \ldots, cdkq$.

14. *Shift Parameter Estimator Associated with the Bohn–Wolfe Statistic.* For perfect judgment rankings, Bohn and Wolfe (1992) showed that the Hodges–Lehmann (1963) shift parameter estimator, $\hat{\Delta}_{\mathrm{BW}}$, associated with the BW statistic is given by

$$\hat{\Delta}_{\mathrm{BW}} = \mathrm{median}\{Y_{(u)v} - X_{(s)t} : u = 1, \ldots, q; v = 1, \ldots, d; \\ s = 1, \ldots, k; t = 1, \ldots, c\}. \tag{15.35}$$

Not surprisingly, $\hat{\Delta}_{\mathrm{BW}}$ has the same form as the shift parameter estimator $\hat{\Delta}$ (4.34) associated with the Mann–Whitney–Wilcoxon statistic W, the difference being that $\hat{\Delta}_{\mathrm{BW}}$ utilizes RSS data, whereas $\hat{\Delta}$ is computed with SRS observations. The usual small sample and asymptotic properties (see Randles and Wolfe

(1979)) for the Hodges–Lehmann estimator $\hat{\Delta}_{\mathrm{BW}}$ follow from the properties of the associated test statistic BW.

15. *Shift Parameter Confidence Intervals Associated with the Bohn–Wolfe Statistic.* To develop a $100(1-\delta)\%$ confidence interval under perfect rankings for the shift parameter Δ, let $D^*_{(1)} \le D^*_{(2)} \le \ldots \le D^*_{(cdkq)}$ be the $cdkq$ ordered differences $Y_{(u)v} - X_{(s)t}$, for $s = 1, \ldots, k$; $t = 1, \ldots, c$; $u = 1, \ldots, q$; and $v = 1, \ldots d$. Since we have a shift (Δ) model and $\mathrm{BW} = \sum_{i=1}^{cdkq} \Psi(D^*_{(i)})$, where $\Psi(w) = 1, 0$ as $w >, \le 0$, it follows from Randles and Wolfe (1979, Theorem 6.1.13) that$[D^*_{(cdkq+1-r)}, D^*_{(r)})$ is a $100(1-\delta)\%$ confidence interval for Δ, where r satisfies $P_0(\mathrm{BW} \ge r) = \delta/2$. The associated one-sided $100(1-\delta)\%$ upper {lower} confidence bound for Δ is given by $(-\infty, D^*_{(r^*)})$ $\{[D^*_{(cdkq+1-r^*)}, \infty)\}$, where $P_0(\mathrm{BW} \ge r^*) = \delta$. As with the shift estimator $\hat{\Delta}_{\mathrm{BW}}$ (Comment 14), these confidence intervals and bounds for Δ are direct RSS analogs of the SRS confidence intervals and bounds for Δ associated with the Mann–Whitney–Wilcoxon statistic W (see Section 4.3).

16. *Robustness of Level—Effect of Imperfect Rankings on BW.* All of the properties of the test procedures based on BW discussed in Section 15.4 are predicated on the assumption that the judgment rankings are perfect. This assumption leads directly to the distribution-free property for BW under the null hypothesis $H_0 : \Delta = 0$ and enables us to use standard distributional properties of order statistics to construct the necessary critical values for the test procedures. In practice, of course, it is likely that some ranking errors will occur in obtaining our RSS data. Bohn and Wolfe (1994) used expected spacings to develop an imperfect ranking model to address this issue. They found that small imperfections in the rankings did not seriously affect the overall performance of the tests based on BW. However, it was also clear from their study that significant ranking error could lead to substantial inflation of the true significance level over the nominal level set by using the null distribution of BW under perfect rankings. Fligner and MacEachern (2006) and Frey (2007a) studied this issue under different classes of imperfect ranking models. This level of inflation for the BW test procedures under imperfect rankings clearly emphasizes the importance of having a reliable ranking process for collecting the RSS data.

17. *Lack of Distribution-Freeness When the Rankings Are Imperfect.* As noted in Comment 16, the significance levels for the BW test procedures are inflated when there are serious ranking errors. This is an immediate consequence of the fact that the BW statistic is no longer distribution-free under the null hypothesis $H_0 : \Delta = 0$ in the presence of ranking error (see Bohn and Wolfe (1994)). In fact, the standardized statistic BW^* is not even asymptotically ($N \to \infty$) distribution-free when there are ranking errors. The null expected value $E_0[\mathrm{BW}] = 2cd$ is maintained (and not dependent on the common underlying distribution F under H_0) even in the presence of ranking errors (as long as the ranking process is consistent across the cycles). On the other hand, when there are ranking errors, the asymptotic variance σ^2_∞ (15.33) depends on the form of the common F under H_0. Thus, without some knowledge about the common F or the use of additional sample information (see Comment 18), even the asymptotic distribution cannot be used to set reliable approximate critical values for the test procedures based on BW in the presence of ranking errors. In fact, the true asymptotic variance for BW under H_0 in the presence of ranking errors is generally higher than the value

$\sigma^2_{0(\infty)}$ stipulated under perfect rankings. This is not surprising, as we would expect more variability in the BW procedures when unaccounted-for ranking errors are present in the data collection process. The net result is that using the perfect ranking expression $\sigma^2_{0(\infty)}$ for the asymptotic variance leads to approximate critical values that actually correspond to higher significance levels than stipulated.

18. *Modifications to Accommodate Imperfect Rankings.* As noted in Comments 16 and 17, the Bohn–Wolfe statistic BW is no longer distribution-free under the null hypothesis H_0 in the presence of ranking errors during the collection of the RSS data. This can lead to substantial inflation of the true significance level (over the nominal level under the assumption of perfect rankings) for the associated test procedures based on BW. While there is little that can be done to ameliorate these concerns in the presence of substantial ranking errors and small sample sizes, Özturk (2008, 2010) proposed a clever way to deal with them in the case of larger sample sizes. First, he employed a number of techniques, including minimizing a distance measure and nonparametric maximum likelihood, to estimate unknown parameters in the classes of imperfect ranking models developed by Bohn and Wolfe (1994) and Frey (2007a). He then used these fitted imperfect ranking models to provide an appropriate estimator, $\hat{\sigma}^2_{0(\infty)}$, of the asymptotic null variance that takes into account the additional variability associated with the presence of ranking errors. Finally, replacing the asymptotic null variance $\sigma^2_{0(\infty)}$ associated with perfect rankings, as discussed in Comment 12 and Bohn and Wolfe (1992), by the adjusted estimator $\hat{\sigma}^2_{0(\infty)}$ in the expression for BW* given in Comment 12 leads to large-sample test procedures that maintain their nominal approximate significance levels even in the presence of substantial ranking errors.

19. *Comparisons Only Within, but not Across, Judgment Ranks.* As discussed in Comments 17 and 18, the Bohn–Wolfe statistic BW is not necessarily distribution-free under the null hypothesis $H_0 : \Delta = 0$ when the RSS ranking process is not perfect and this leads to inflated significance levels for the associated test procedures in the presence of imperfect rankings. For the setting where the set size is the same for the X and Y RSSs (i.e., $k = q$), Fligner and MacEachern (2006) proposed a competitor statistic T that includes the Mann–Whitney comparisons between X's and Y's only for those (X, Y)'s that have the same within-set judgment ranks.

    Let

$$T_j = \sum_{t=1}^{c} \sum_{v=1}^{d} \phi(X_{[j]t}, Y_{[j]v}), \text{for} j = 1, \ldots, k, \tag{15.36}$$

    where

$$\phi(X_{[j]t}, Y_{[j]v}) = \begin{cases} 1, & \text{if } X_{[j]t} < Y_{[j]v} \\ 0, & \text{otherwise.} \end{cases}$$

    Thus, T_j is simply the Bohn–Wolfe statistic utilizing the Mann–Whitney counts only between those X and Y observations that have the same within-set judgment rank j, for $j = 1, \ldots, k$. The Fligner–MacEachern test statistic is then the sum of these common-rank Mann–Whitney statistics, namely,

$$\text{FM} = \sum_{j=1}^{k} T_j. \tag{15.37}$$

The statistic FM is distribution-free under $H_0 : \Delta = 0$ for any ranking mechanism (perfect or imperfect) that is the same for both the X and Y populations. In fact, the null distribution of each $T_j, j = 1, \ldots, k$, is precisely that of the usual two-sample Mann–Whitney statistic for c X observations and d Y observations (see Comment 4.7). Moreover, $T_1, \ldots, T_k$ are mutually independent since all $cdkq$ RSS observations are mutually independent. Thus, the null distribution for FM (15.37) can be obtained as the convolution of k independent Mann–Whitney null distributions, each for the same sample sizes of c X's and d Y's, and this is true whether the ranking process is perfect or imperfect in any fashion, including completely at random. MacEachern and Fligner compared the performance of tests based on the Bohn–Wolfe statistic BW with tests based on their statistic FM under perfect rankings and under a variety of imperfect ranking models. Since the BW statistic includes more individual comparisons between the two samples than does the FM statistic, it is not surprising that Fligner and MacEachern found the Bohn–Wolfe procedure to generally have higher power than the FM procedure when the rankings are perfect, although this edge in power for BW under perfect rankings is never overwhelming for the underlying distributions considered in their study. On the other hand, when the rankings are imperfect, they found that the FM procedure was generally superior to the BW procedure. This is also not surprising, given the FM procedure is truly distribution-free under H_0 so that it maintains its nominal significance level even when the rankings are imperfect, while the true significance level for the BW procedure can be considerably inflated over its nominal level in the presence of less than perfect rankings.

20. *Consistency of the BW Test.* Under Assumptions A1–A5, which includes the condition that the rankings are perfect, the consistency of the tests based on BW depends on the parameter

$$\gamma^* = \sum_{s=1}^{k} \sum_{u=1}^{q} P(X_{[s]1} < Y_{[u]1}) - \frac{kq}{2}. \tag{15.38}$$

Bohn and Wolfe (1992) showed that under Assumptions A1–A5 the test procedures defined by (15.19), (15.20), and (15.21) are consistent against the alternatives for which $\gamma^* >, <$, and $\neq 0$, respectively.

Properties

1. *Consistency.* Under perfect rankings and the location-shift model defined by (15.16), the tests defined by (15.19), (15.20), and (15.21) are consistent against the alternatives $\Delta >, <, \neq 0$, respectively. Also see Comment 20.
2. *Asymptotic Normality.* See Bohn and Wolfe (1992).
3. *Efficiency.* See Bohn and Wolfe (1992).

Problems

37. Consider the population of NHANES III data given in Table 15.2 and discussed in Example 15.4. Divide the subjects in the NHANES III population into two groups: (a) age 30 years or younger and (b) age 31 years or older. Using buttocks circumference as the auxiliary ranking

variable and common set size $k = q = 4$, select independent RSSs of size $m = n = 20$ each from these two age groups and obtain the BMI values for the individuals in the two RSSs. Find the P-value for the BW test of the conjecture that the average BMI value for individuals aged 31 years or older is greater than the average BMI value for individuals aged 30 years or younger in the NHANES III population.

38. In 1997, scientists at the Horticulture Research International–East Manning conducted an exploratory study to assess the potential impact of RSS. The results of this investigation were reported in Murray, Ridout, and Cross (2000). A portion of the study focused on a comparison of spray deposit coverage on tree leaves under two different sprayer configurations. Two similar plots of apple trees were sprayed with a fluorescent, water-soluble tracer Tinopal CBS-X at 2% concentration in water. One of the plots was sprayed at high volume with a coarse nozzle on the sprayer (coarse treatment) to provide large average droplet size. The second plot was sprayed at low volume with a fine nozzle on the sprayer (fine treatment) to provide small average droplet size. The investigators were interested in both the percentage of upper leaf surface covered by the spray and the total amount of spray deposited on the upper surfaces of the leaves. In this problem, we concentrate solely on the percentage of upper leaf surface covered by the spray and denote this variable by %Coverage (expressed as a decimal).

With the RSS methodology in mind, 25 sets of five leaves each (common set size $k = q = 5$) were collected from the central trees of each of these two plots. Leaves of similar size were selected throughout the tree canopies, with care to avoid any intentional bias, but without any formal randomization scheme. For our purposes in this problem, we will view these 250 collected leaves as 50 independent random samples (25 from each plot) of size 5 each.

The leaves were taken to the laboratory where a scientist ranked (without formal measurement) the %Coverage on the upper surfaces of each of the leaves within these 100 sets of five leaves based on visual appearance of the deposits on the upper leaf surfaces when viewed under ultraviolet light. Once these visual rankings were completed, the image analysis system Optimax V was used to formally estimate the percentage areas of the individual upper leaf surfaces that were covered with deposit.

The scientist's within-set rankings and the Optimax V estimated %Coverage data from this experiment are presented in Table 15.7. With common set size $k = q = 5$, randomly select five of the 25 sets of rankings for the coarse treatment data on which to utilize the Optimax V measurement for the smallest ranked leaf in each of the five sets. Then randomly select a second set of five of the remaining 20 sets of rankings for the coarse treatment data on which to utilize the Optimax V measurement for the second smallest ranked leaf in each of the five sets. Continue in this fashion through a third randomly selected set of five (from among the 15 remaining sets) where the Optimax V measurement for the third smallest ranked leaf is utilized, then a fourth randomly selected set of five (from among the 10 remaining sets) where the Optimax V measurement for the fourth smallest ranked leaf is utilized. Finally, the Optimax V measurement for the largest ranked leaf will be utilized from the remaining five sets. This yields an RSS of size $m = c(k) = 5(5) = 25$ from the coarse treatment data. We proceed in a similar fashion to obtain an independent RSS of size $n = d(q) = 5(5) = 25$ from the fine treatment data.

Find the P-value for the BW test of the conjecture that the upper leaf surface %Coverage is higher for the coarse treatment than for the fine treatment.

39. For a single X and Y cycle (i.e., $c = d = 1$), X set size $k = 2$ and Y set size $q = 3$, obtain the null probabilities under the assumption of perfect rankings for each of the $5! = 120$ distinct possible permutations of the judgment ordered X and Y observations. Compile the resulting null distribution of the test statistic BW for this setting.

40. For X set size $k = 4$, Y set size $q = 3$, $c = 4$ cycles for the X sample and $d = 3$ cycles for the Y sample, how many distinct possible permutations of the judgment ordered X and Y observations must be considered in order to compile the null distribution of the test statistic

Table 15.7 Percentage Upper Leaf Coverage by Spray

High volume coarse nozzle		
Ranked set	Observer ranking	Percentage cover
1	1	.003
1	2	.027
1	3	.029
1	4	.264
1	5	.347
2	1	.012
2	2	.028
2	3	.350
2	4	.378
2	5	.527
3	1	.052
3	2	.107
3	3	.244
3	4	.104
3	5	.194
4	1	.032
4	2	.087
4	3	.076
4	4	.057
4	5	.260
5	1	.095
5	2	.089
5	3	.115
5	4	.096
5	5	.143
6	1	.039
6	2	.137
6	3	.069
6	4	.141
6	5	.216
7	1	.225
7	2	.119
7	3	.212
7	4	.119
7	5	.252
8	1	.026
8	2	.096
8	3	.126
8	4	.100
8	5	.377
9	1	.007
9	2	.043
9	3	.083
9	4	.105
9	5	.153
10	1	.070
10	2	.128
10	3	.095
10	4	.191
10	5	.565
11	1	.034
11	2	.096
11	3	.204

(*continued*)

Table 15.7 (*Continued*)

High volume coarse nozzle		
Ranked set	Observer ranking	Percentage cover
11	4	.269
11	5	.223
12	1	.042
12	2	.118
12	3	.219
12	4	.236
12	5	.373
13	1	.004
13	2	.062
13	3	.130
13	4	.230
13	5	.360
14	1	.013
14	2	.133
14	3	.155
14	4	.218
14	5	.289
15	1	.012
15	2	.069
15	3	.046
15	4	.123
15	5	.296
16	1	.051
16	2	.102
16	3	.094
16	4	.241
16	5	.276
17	1	.098
17	2	.104
17	3	.226
17	4	.264
17	5	.622
18	1	.063
18	2	.059
18	3	.193
18	4	.204
18	5	.155
19	1	.046
19	2	.229
19	3	.224
19	4	.210
19	5	.194
20	1	.040
20	2	.126
20	3	.117
20	4	.285
20	5	.150
21	1	.032
21	2	.044
21	3	.100
21	4	.143
21	5	.273

Table 15.7 (*Continued*)

High volume coarse nozzle		
Ranked set	Observer ranking	Percentage cover
22	1	.108
22	2	.141
22	3	.158
22	4	.154
22	5	.100
23	1	.004
23	2	.143
23	3	.130
23	4	.137
23	5	.261
24	1	.028
24	2	.090
24	3	.292
24	4	.250
24	5	.271
25	1	.007
25	2	.057
25	3	.091
25	4	.103
25	5	.229
Low volume fine nozzle		
Ranked set	Observer ranking	Percentage cover
1	1	.036
1	2	.129
1	3	.090
1	4	.083
1	5	.360
2	1	.080
2	2	.137
2	3	.241
2	4	.363
2	5	.298
3	1	.091
3	2	.136
3	3	.183
3	4	.426
3	5	.564
4	1	.005
4	2	.074
4	3	.185
4	4	.270
4	5	.505
5	1	.310
5	2	.166
5	3	.337
5	4	.404
5	5	.487
6	1	.250
6	2	.287
6	3	.320
6	4	.457
6	5	.696

(*continued*)

Table 15.7 (*Continued*)

	Low volume fine nozzle	
Ranked set	Observer ranking	Percentage cover
7	1	.156
7	2	.181
7	3	.177
7	4	.167
7	5	.343
8	1	.171
8	2	.192
8	3	.290
8	4	.217
8	5	.217
9	1	.013
9	2	.037
9	3	.074
9	4	.328
9	5	.285
10	1	.094
10	2	.395
10	3	.544
10	4	.550
10	5	.715
11	1	.089
11	2	.186
11	3	.227
11	4	.436
11	5	.512
12	1	.043
12	2	.032
12	3	.042
12	4	.108
12	5	.237
13	1	.026
13	2	.044
13	3	.269
13	4	.336
13	5	.451
14	1	.137
14	2	.100
14	3	.379
14	4	.419
14	5	.518
15	1	.050
15	2	.086
15	3	.140
15	4	.186
15	5	.315
16	1	.180
16	2	.138
16	3	.133
16	4	.181
16	5	.211
17	1	.017
17	2	.111

Table 15.7 (*Continued*)

Low volume fine nozzle		
Ranked set	Observer ranking	Percentage cover
17	3	.105
17	4	.051
17	5	.086
18	1	.071
18	2	.067
18	3	.130
18	4	.140
18	5	.417
19	1	.044
19	2	.120
19	3	.293
19	4	.194
19	5	.342
20	1	.024
20	2	.102
20	3	.277
20	4	.464
20	5	.742
21	1	.100
21	2	.364
21	3	.438
21	4	.134
21	5	.333
22	1	.011
22	2	.009
22	3	.085
22	4	.052
22	5	.218
23	1	.111
23	2	.056
23	3	.184
23	4	.132
23	5	.322
24	1	.152
24	2	.199
24	3	.227
24	4	.277
24	5	.386
25	1	.111
25	2	.121
25	3	.170
25	4	.201
25	5	.122

Source: R. A. Murray, M. S. Ridout, and J. V. Cross (2000).

BW? Select 10 of these permutations and compute the value of BW and the null probability for each of them under the assumption of perfect rankings.

41. General expressions for E(BW) and Var(BW) are provided in (15.28) and (15.29), respectively, in Comment 11.

(a) Simplify these expressions for the setting of common set size $k = q = 2$ and the same number of observations from each population (i.e., common cycle size $c = d$).

(b) For the setting in part (a) under the assumption of perfect rankings, show that the null $H_0 : \Delta = 0$ mean and variance for BW are given by:

$$E_0(\text{BW}) = 2c^2 \qquad \text{and} \qquad \text{Var}_0(\text{BW}) = 8c^3/9.$$

42. Compute the Fligner–MacEachern test statistic FM (15.37) discussed in Comment 19 for the coarse spray and fine spray leaf %Coverage RSSs from Problem 38.

43. Compute the mean μ_{FM} and variance σ^2_{FM} under the null hypothesis $H_0 : \Delta = 0$ for the Fligner–MacEachern statistic FM (15.37) presented in Comment 19. What is the asymptotic distribution of the standardized statistic

$$\text{FM}^* = \frac{\text{FM} - \mu_{\text{FM}}}{\sigma_{\text{FM}}}$$

as both cycle sizes c and d become large? Provide arguments supporting this result.

15.5 OTHER IMPORTANT ISSUES FOR RANKED SET SAMPLING

McIntyre (1952) introduced the basic concept of RSS in his seminal paper 60 years ago and there were several bursts of related research activity over the next 30 years. However, it was not until the paper by Stokes and Sager (1988) that the true impact of this simple idea began to flourish. RSS has been an important aspect of statistical research for the past 20 years and continues to attract considerable attention even 60 years post-McIntyre. Part of this richness is due to the great flexibility provided by the ranked set paradigm. In this section we briefly discuss some aspects of this flexibility that provide excellent research opportunities as well as address complexities in applications.

Set Size

The set size plays a critical role in the performance of any RSS procedure. For given set size k, each measured RSS observation utilizes additional information obtained from its ranking relative to $k - 1$ other units from the population. With perfect rankings this additional information is clearly an increasing function of k. Thus, with perfect rankings, we would want to take our set size k to be as large as economically possible within available resources. However, it is also clear that the likelihood of errors in our rankings is an increasing function of the set size as well; that is, the larger k is, the more likely we are to experience errors in our rankings. Therefore, to select the set size k optimally, we need to be able to both model the probabilities for imperfect rankings and to assess their impact on our RSS statistical procedures. We discuss a number of approaches to this issue in the following section.

Imperfect Rankings

The effectiveness of RSS procedures depends directly on how well the within-set rankings to select the units for measurement can be accomplished. While perfect rankings are surely the goal of any RSS protocol, it is just as likely not to be feasible. Thus it is imperative in practice that we be able to assess the effect of imperfect rankings on our procedures, and the most appropriate way to do this is to develop statistical models to capture the uncertainty of the ranking process.

Dell and Clutter (1972) proposed the first class of models for this purpose. They view the ranks of the experimental units as being based on perceived values that are associated with the true measured values through an additive model. Taking a much different approach, Bohn and Wolfe (1994) considered the distributions of the judgment order statistics to be mixtures of distributions of the true order statistics and based their model on the expected spacings between order statistics. Presnell and Bohn (1999) pointed out some limitations with this approach. Frey (2007a) overcame the Presnell–Bohn concerns by producing a much larger class of models through a clever scheme of subsampling order statistics from the basic Bohn–Wolfe model. The most recent attempt by Fligner and MacEachern (2006) to understand the ranking process used the monotone likelihood ratio principle to develop a class of imperfect ranking models.

Bohn and Wolfe (1994) and Fligner and MacEachern (2006) studied the effect of imperfect rankings on the performance of the Bohn–Wolfe BW test procedure presented in Section 15.4. (More on this topic can be found in Comments 16–19.) Chen, Stasny, and Wolfe (2006a) used data from the NHANES III survey, 1988–1994, to provide an empirical assessment of the ranking accuracy in RSS.

Unbalanced Ranked Set Sampling

The emphasis in this chapter has been entirely on **balanced** RSS data of the form $X_{[i]j}, i = 1, \ldots, k$ and $j = 1, \ldots, m$, where k is the common set size and m is the number of cycles. Thus, in the case of balanced RSS data, we have the same number, m, of each of the judgment order statistics; that is, we have m mutually independent and identically distributed first judgment order statistics $X_{[1]1}, \ldots, X_{[1]m}$; m mutually independent and identically distributed second judgment order statistics $X_{[2]1}, \ldots, X_{[2]m}$; $\ldots$; and m mutually independent and identically distributed kth judgment order statistics $X_{[k]1}, \ldots, X_{[k]m}$. While balanced RSS is the most commonly occurring form of RSS data, there are situations where it is not optimal to collect the same number of measured observations for each of the judgment order statistics.

For example, consider an underlying distribution that is unimodal and symmetric about its median θ and suppose we are interested only in making inferences about θ using RSS data based on an odd set size k. Among all the order statistics for a random sample of size k, we know that the sample median $X_{(k+1)/2}$ contains the most information about θ. Thus, to estimate θ in this setting, it is natural to consider measuring the same judgment order statistic, namely, the judgment median $X_{[(k+1)/2]}$, in each set, so that it is measured all k times in each of the m cycles. The resulting RSS consists of mk measured observations, each of which is a judgment median from a set of size k. This would be the most efficient RSS for estimating the population median θ for a population that is both unimodal and symmetric about θ, and it is clearly as unbalanced as possible. A similar approach calls for a distinctly different unbalanced RSS for estimating the median of an asymmetric unimodal population. There are, of course, other considerations. While median judgment order statistics do provide an efficient estimator for the median of a symmetric population, they would not be an optimal choice if we also want to estimate the variance of the population—balanced RSS measurements would be preferable for this purpose. (See Özturk and Wolfe (2000) for more discussion of the pros and cons of balanced versus unbalanced RSS.)

Chen, Stasny, and Wolfe (2006b) and Chen et al. (2009) considered the use of unbalanced RSSs in estimation of a population proportion p. They used Neyman allocation

to decide on optimal representations of the various judgment order statistics in the formation of an RSS. This approach leads to the preferred use of balanced RSS for values of p near $1/2$, but the unbalanced nature of the optimal allocation grows dramatically as the value of p nears either 0 or 1.

Unequal Set Sizes

Sometimes the sets that arise naturally in RSS applications are of unequal sizes. For instance, commuters on different public buses in a large city or patients in a collection of doctors' waiting rooms represent naturally occurring sets of varying sizes. One alternative in such situations is to pare down the larger sets to agree in size with the smaller sets, but this can lead to the loss of valuable information that could have been obtained from the more comprehensive rankings within the larger sets. Gemayel, Stasny, and Wolfe (2010) proposed an estimator for the median of a symmetric population that combines the medians of RSSs of varying sizes. While not optimal for any specific symmetric distribution, they show that the estimator is robust over a wide class of symmetric distributions.

Cost Considerations

Even under perfect judgment rankings, the costs of the various components of RSS, namely, identifying sampling units, ranking of sets of sampling units, and eventual measurement of units selected for inclusion in an RSS, all affect the choice of an optimal set size k. For a basic discussion of these factors and their effect on optimal set size selection, the reader is referred to Nahhas, Wolfe, and Chen (2002).

Multiple Observations per Set

In all of the previous discussion of RSS in this chapter, we only consider measuring a single observation from each set. The rationale behind this approach is the fact that the correlation inherent in measuring more than one observation per set typically leads to a reduction in efficiency for RSS estimation. Wang, Chen, and Liu (2004), however, demonstrated that this is not necessarily the case when the cost involved in the ranking process itself is not small relative to the costs of unit selection and unit measurement. Under such conditions, they find that taking two or more observations from a set can lead to improved RSS estimation.

Problems

44. In Problem 14 we collected a balanced RSS of 96 subjects from the NHANES III population using a set size of $k = 6$ and arm circumference as the auxiliary variable to perform the judgment rankings with the goal of making inferences about the BMI for the population. Since this auxiliary variable does not provide perfect rankings for the BMI values, there will be ranking errors within the sets. Let

$$p_{ij} = P\ (\textit{true } i\text{th } \textit{order statistic in a set is ranked as the } j\text{th } \textit{judgment order statistic in that set}),$$

for $i = 1, \ldots, 6$ and $j = 1, \ldots, 6$. Thus, p_{ij}, $i \neq j$, is the probability that the ranking process in a given set incorrectly assigns the jth judgment rank to the BMI value that is actually the ith ordered BMI value in the set and $p_{ii}, i = 1, \ldots, 6$, is the probability that the ranking process correctly identifies the ith ordered BMI value in a set of size 6. Use the known BMI values for all 576 subjects involved in the ranking process leading to this RSS of size 96 to obtain sample estimates of the 36 probabilities p_{ij}, $i = 1, \ldots, 6$ and $j = 1, \ldots, 6$. Discuss the effectiveness of using arm circumference as the auxiliary ranking variable for BMI.

45. In Problem 15 we collected a balanced RSS of 96 subjects from the NHANES III population using a set size of $k = 8$ and arm circumference as the auxiliary variable to perform the judgment rankings with the goal of making inferences about the BMI for the population. Since this auxiliary variable does not provide perfect rankings for the BMI values, there will be ranking errors within the sets. Let

$$p_{ij} = P\ (\textit{true } i\text{th } \textit{order statistic in a set is ranked as the } j\text{th } \textit{judgment order statistic in that set}),$$

for $i = 1, \ldots, 8$ and $j = 1, \ldots, 8$. Thus, p_{ij}, $i \neq j$, is the probability that the ranking process in a given set incorrectly assigns the jth judgment rank to the BMI value that is actually the ith ordered BMI value in the set and $p_{ii}, i = 1, \ldots, 8$, is the probability that the ranking process correctly identifies the ith ordered BMI value in a set of size 8. Use the known BMI values for all 768 subjects involved in the ranking process leading to this RSS of size 96 to obtain sample estimates of the 64 probabilities p_{ij}, $i = 1, \ldots, 8$ and $j = 1, \ldots, 8$. Compare the imperfect ranking information obtained in this problem with that obtained in Problem 44 for set size $k = 6$.

46. In Problem 16 we collected a balanced RSS of 96 subjects from the NHANES III population using a set size of $k = 6$ and buttocks circumference as the auxiliary variable to perform the judgment rankings with the goal of making inferences about the BMI for the population. Since this auxiliary variable does not provide perfect rankings for the BMI values, there will be ranking errors within the sets. Let

$$p_{ij} = P\ (\textit{true } i\text{th } \textit{order statistic in a set is ranked as the } j\text{th } \textit{judgment order statistic in that set}),$$

for $i = 1, \ldots, 6$ and $j = 1, \ldots, 6$. Thus, p_{ij}, $i \neq j$, is the probability that the ranking process in a given set incorrectly assigns the jth judgment rank to the BMI value that is actually the ith ordered BMI value in the set and $p_{ii}, i = 1, \ldots, 6$, is the probability that the ranking process correctly identifies the ith ordered BMI value in a set of size 6. Use the known BMI values for all 576 subjects involved in the ranking process leading to this RSS of size 96 to obtain sample estimates of the 36 probabilities p_{ij}, $i = 1, \ldots, 6$ and $j = 1, \ldots, 6$. Compare the imperfect ranking information obtained in this problem with that obtained in Problem 44 when arm circumference was used as the auxiliary ranking variable. Discuss the implication of your findings.

47. Consider the NHANES III population data provided in Table 15.2 and discussed in Example 15.4. Collect an unbalanced RSS (see the "Unbalanced Ranked Set Sampling" discussion in this section) of size $n = 96$ from this population using set size $k = 5$ and buttocks circumference as the auxiliary variable for the ranking process by obtaining the BMI value for the subject with the median buttocks circumference in each of the 96 sets of size 5. Thus, the collected BMI data will be of the form $X_{[3]j} = j^{\text{th}}$ set sample median for each of the sets $j = 1, 2, \ldots, 96$. Use these data to estimate the average BMI for the entire NHANES III population. Compare this estimate with the four estimates of the same quantity using the balanced RSSs of 96 NHANES III subjects each obtained in Problems 14–17 and the estimate from the SRS of

size 96 obtained in Problem 26. How do these six estimates compare with the actual average BMI for the entire population?

48. Consider an underlying population that is unimodal and symmetric about its finite mean μ and suppose we are interested only in making inferences about μ using RSS data based on an odd set size k. Consider an unbalanced RSS of size n (see the "Unbalanced Ranked Set Sampling" discussion in this section) collected from this population, where the judgment sample median $X_{[(k+1)/2]}$ is measured in each of the n sets of size k each. Thus, the collected RSS data will be of the form $\{X_{[(k+1)/2]j} = j\text{th set sample median for each of the sets } j = 1, 2, \ldots, n\}$. Let $\hat{\mu} = \overline{X}_{\text{med}} = \frac{1}{n}\sum_{j=1}^{n} X_{[(k+1)/2]j}$ be the average of these n RSS set medians. Show that $\hat{\mu}$ is an unbiased estimator for μ when the ranking process is perfect. (See Özturk and Wolfe (2000) for more discussion on such median unbalanced RSSs.)

49. An individual is considered to be obese if his or her BMI is at least 30 (Kuczmarski et al., 1997). Consider the population of NHANES III data given in Table 15.2 (discussed in Example 15.4) and let p denote the proportion of individuals in this population who are obese.

(a) Using buttocks circumference as the auxiliary variable to perform the judgment rankings and set size $k = 5$, select an unbalanced RSS (see the "Unbalanced Ranked Set Sampling" discussion in this section) of size $n = 100$, where the BMI value $X_{[5]}$ is obtained for the largest judgment ordered subject in half (50) of these sets and the BMI value $X_{[4]}$ is obtained for the second largest judgment ordered subject in the other half (50) of these sets. Find the sample percentage, $\hat{p}_{\text{high}}$, of the individuals in this RSS who are obese.

(b) Using buttocks circumference as the auxiliary variable to perform the judgment rankings and set size $k = 5$, select an unbalanced RSS (see the "Unbalanced Ranked Set Sampling" discussion in this section) of size $n = 100$, where the BMI value $X_{[1]}$ is obtained for the smallest judgment ordered subject in half (50) of these sets and the BMI value $X_{[2]}$ is obtained for the second smallest judgment ordered subject in the other half (50) of these sets. Find the sample percentage, $\hat{p}_{\text{low}}$, of the individuals in this RSS who are obese.

(c) Using buttocks circumference as the auxiliary variable to perform the judgment rankings and set size $k = 5$, select an unbalanced RSS (see the "Unbalanced Ranked Set Sampling" discussion in this section) of size $n = 100$, where the BMI value $X_{[3]}$ is obtained for the median judgment ordered subject in all 100 of these sets. Find the sample percentage, $\hat{p}_{\text{middle}}$, of the individuals in this RSS who are obese.

(d) Compare the sample percentages obtained in parts (a)–(c) of this problem in conjunction with the true proportion of obese individuals in the NHANES III population. (See Chen, Stasny, and Wolfe (2006b) for more discussion about the use of unbalanced RSSs in the estimation of population proportions.)

50. An individual is considered to be overweight if his or her BMI is at least 25 (Kuczmarski et al., 1997). Consider the population of NHANES III data given in Table 15.2 (discussed in Example 15.4) and let q denote the proportion of individuals in this population who are overweight.

(a) Using the RSS sample obtained in part (a) of Problem 49, find the sample percentage, $\hat{q}_{\text{high}}$, of the individuals in this RSS who are overweight.

(b) Using the RSS sample obtained in part (b) of Problem 49, find the sample percentage, $\hat{q}_{\text{low}}$, of the individuals in this RSS who are overweight.

(c) Using the RSS sample obtained in part (c) of Problem 49, find the sample percentage, $\hat{q}_{\text{middle}}$, of the individuals in this RSS who are overweight.

(d) Discuss the sample percentages obtained in parts (a)–(c) of this problem in conjunction with the true proportion of overweight individuals in the NHANES III population.

(e) Compare the results obtained in this problem with those obtained in Problem 49. (See Chen, Stasny, and Wolfe (2006b) for more discussion about the use of unbalanced RSSs in estimation of population proportions.)

15.6 EXTENSIONS AND RELATED APPROACHES

In addition to the rapid development of the field of RSS over the past two decades, it has also provided a stimulus for the emergence of other important related approaches to statistical inference. In this section we discuss four such areas that have arisen directly from previous RSS considerations.

Judgment Post-Stratification

One of the features of RSS is that a researcher is required to judgment rank the potential units prior to obtaining any measurements; that is, the researcher must commit to the RSS approach from the onset of the experiment. MacEachern, Stasny, and Wolfe (2004) introduced a data collection method, called *judgment post-stratification* (*JPS*), that enables a researcher to collect an initial SRS in a standard fashion from the population of interest and then to post-stratify the SRS observations by ranking each of them among its own randomly chosen comparison sample. Thus the variable of interest is first measured on all of the original SRS units and only then is relative judgment ranking information obtained from the comparison samples to enable the JPS. This approach allows the researcher to utilize the measurements in the full SRS as well as the additional information obtained from the JPS process.

The JPS approach provides a mechanism for incorporating both imprecise rankings and information from multiple rankers via the JPS process. For additional work on this aspect of JPS, see Wang et al. (2006), Stokes, Wang, and Chen (2007), Wang, Lim, and Stokes (2008), and Frey and Özturk (2011).

Order Restricted Randomization

Özturk and MacEachern (2004, 2007) built on the general framework of RSS to develop order restricted randomized (ORR) designs that utilize subjective judgment ranking to enable restricted randomization in the comparison of two treatments (one of which could be a control). The units within a given set are assigned to different treatments and then instead of the typical RSS approach that selects a single unit from each ranked set for full measurement, the ORR designs allow for all of the units within a set to be fully measured. The positive dependence between the units within sets leads to contrast estimators and confidence intervals with smaller variability than those based on either completely randomized designs or purely RSS designs. An added feature of ORR estimation is that it does not rely on perfect judgment rankings.

Intentionally Representative Sampling

Frey (2007b) introduced a novel approach to data collection dubbed intentionally representative sampling (IRS) that allows a researcher more flexibility in the use of prior and auxiliary information than is possible with RSS. Once a target sample size n has been established, the IRS process requires that the researcher divide the population of interest into disjoint potential samples of size n, each of which is considered (based on prior and auxiliary information) to be at least roughly representative of the overall population with respect to the measurement of interest. In this way, the researcher can exclude from

the very beginning any potential samples that are considered to be unrepresentative of the population. To effectively implement the IRS approach, we must, of course, have reasonably good auxiliary information about **all** of the units in the population, not just the ranking subsets that are required for implementation of RSS procedures.

Sampling from Partially Rank-Ordered Sets

There are times when it is difficult to rank all of the experimental units in a set with high confidence, particularly when subjective information is utilized in the ranking process. Özturk (2011, 2012) and Gao and Özturk (2012) considered a judgment ordering process called *judgment subsetting* that allows a judgment ranker to use tied ranks when it is difficult to fully rank the experimental units in a set. They showed that this added flexibility leads to improved precision for RSS estimation procedures in settings where the full ranking cannot be done with high confidence.

Chapter 16

An Introduction to Bayesian Nonparametric Statistics via the Dirichlet Process

INTRODUCTION

Bayesian statistics incorporate prior information about the parameter of interest into the inferential method. The prior information is specified through the use of a prior distribution on the parameter of interest, thereby treating that parameter as a random quantity. The prior information can be obtained in many ways including pilot experiments and the opinions of experts. The parameter is chosen by the prior and then after the data are obtained, the posterior distribution is computed. The posterior distribution is the conditional distribution of the parameter, given the data. If more data are obtained, the posterior is used as the new prior and then a new posterior distribution is computed. This is called *Bayesian updating*.

It is easiest to do Bayesian statistics when the unknown parameter lies in a finite-dimensional space. For example, in the problem of estimating a success probability considered in Chapter 2, the typical prior used for the one-dimensional parameter p is a member of the Beta(r, s) family whose density function $f(x)$ is proportional to $x^{r-1}(1-x)^{s-1}$. Then after observing the outcomes of the n Bernoulli trials, say $X = (X_1, \ldots, X_n)$, where X_i is 1 if the *ith* trial results in a success and 0 if the *ith* trial results in a failure, the posterior distribution of p given the data is Beta$(r + B, s + n - B)$, where $B = \sum_{i=1}^{n} X_i$ is the number of successes in the n trials. The Bayes estimator of p for a squared-error loss function is the mean of the posterior distribution, namely, $E(p|X) = (r + B)/(r + s + n)$. As n gets large, the Bayes estimator approaches B/n, the frequentist estimator considered in Chapter 2.

Bayesian methods are more difficult in the nonparametric case because in the nonparametric setting, the unknown parameter may vary in an infinite-dimensional space. For example, if the parameter is the distribution function F itself, we need to estimate $F(x)$ for every value of x between $-\infty$ and ∞. A challenging problem for statisticians who wish to pursue a Bayesian nonparametric approach is to put a prior distribution on F and then compute the posterior distribution of F, given the data. Although there were several earlier attempts at doing this, Ferguson's (1973) elegant approach, creating what is now called *Ferguson's Dirichlet process prior*, or more succinctly, the

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

Dirichlet process, has proved to have staying power and is frequently used. It is surprisingly tractable because, in the simplest problems, the posterior distribution, given the data, is also a Dirichlet process prior. When the prior distribution, given the data, is also a member of the family, the family is said to be a conjugate family of distributions. The Dirichlet process priors are a conjugate family, and the posterior distribution, given the data, is readily obtained as we describe in Section 16.1. Section 16.2 considers Ferguson's (1973) Bayesian nonparametric estimator of the distribution function F. Section 16.3 treats a rank order estimation problem and the Bayesian nonparametric rank order estimator of Campbell and Hollander (1978). Section 16.4 treats the censored data counterpart to Section 16.2 and gives the Susarla–van Ryzin (1976) Bayesian nonparametric estimator of F for the case where the data are right censored. In Section 16.5 we discuss other Bayesian nonparametric approaches to estimation including Gibbs sampling.

16.1 FERGUSON'S DIRICHLET PROCESS

We consider probability measures P on the real line R_1. Ferguson's approach considers partitions $\mathcal{A} = (A_1, \ldots, A_k)$ of disjoint measurable sets whose union is R_1. For any such partition $\mathcal{A}$, the vector $P(\mathcal{A}) = (P(A_1), \ldots, P(A_k))$ is a k-vector of nonnegative numbers (probabilities) that sum to 1. The probability measure P is completely determined if $P(\mathcal{A})$ is known for all or most partitions $\mathcal{A}$. P must satisfy certain additivity and consistency conditions, which we will not present here, but see Ferguson (1973) for the mathematical rigor. It can be shown that a consistent family of probability measures for $P(\mathcal{A})$ as A ranges among all finite partitions will define a unique probability measure for P. This is the way Ferguson defines his process. Ferguson's Dirichlet process is arrived at by starting with finite-dimensional Dirichlet distributions. These are distributions in R_k concentrated on the subset $\mathcal{P}_k = \left\{p : (p_1, ..., p_k) : p_1 \geq 0, \ldots, p_k \geq 0, \sum_{i=1}^{k} p_i = 1\right\}$. A random vector $(Y_1, \ldots, Y_k) \epsilon \mathcal{P}_k$ is said to have a finite-dimensional Dirichlet distribution with parameters $(\beta_1, \ldots, \beta_k)$ if

$$Y_i = \frac{Z_i}{\sum_{j=1}^{k} Z_j}, \tag{16.1}$$

where $Z_1, \ldots, Z_k$ are independent Gamma random variables with scale parameter 1 and shape parameters $\beta_1, \ldots, \beta_k$. We write this distribution as $\mathcal{D}(\beta_1, \ldots, \beta_k)$. When $\beta_i > 0$, $i = 1, \ldots, k$, the density function of $(Y_1, \ldots, Y_k)$ is

$$f(y_1, \ldots, y_k) = \frac{\Gamma(\beta_0)}{\prod_{j=1}^{k} \Gamma(\beta_j)}, \tag{16.2}$$

where β_0 is defined as $\beta_0 = \sum_{j=1}^{k} \beta_j$ and $y_k = 1 - \sum_{j=1}^{k-1} y_j$. This is a $(k-1)$-dimensional distribution because the y's have the constraint $\sum_{j=1}^{k} y_j = 1$.

Ferguson's Dirichlet process produces a random probability measure P (say), taking values in $\mathcal{P}$, the space of all probability measures on R_1. The process is defined by having, for every finite partition $\mathcal{A}$, $P(\mathcal{A})$ follow the Dirichlet distribution with parameters $(\alpha\mu(A_1), \ldots, \alpha\mu(A_k))$, that is $P(\mathcal{A})$ has the distribution $\mathcal{D}(\alpha\mu(A_1), ..., \alpha\mu(A_k))$. In particular, for the partition (A, A^c), $P(A)$ follows the Beta distribution with parameters $(\alpha(A), \alpha(A^c))$. We denote the Dirichlet measure just described by $\mathcal{D}(\alpha, \mu)$.

The Bayesian works with this process as follows. The Bayesian chooses $\mathcal{D}(\alpha, \mu)$ as a prior distribution for P. The measure μ and the constant α need to be specified. The measure μ can be viewed as the Bayesian's guess of the unknown P and α is viewed as the prior sample size, a measure of confidence in the guess. The larger the values of α, the more confidence that is being expressed (see Comments 3 and 4). Ferguson showed (see Comment 5)

$$E(P(A)) = \mu(A) \tag{16.3}$$

$$\text{var}(P(A)) = \mu(A)\mu(A^c)/(\alpha + 1). \tag{16.4}$$

In particular with $A = (-\infty, x]$, we have

$$E(P(-\infty, x]) = E(F(x)) = \mu(-\infty, x] \tag{16.5}$$

$$\text{var}(P(-\infty, x]) = \text{var}(F(x)) = \mu(-\infty, x]\mu(x, \infty)/(\alpha + 1). \tag{16.6}$$

Suppose that $X_1, X_2, \ldots$ are random variables such that, given P, they are independent and identically distributed according to P. That is, P is first chosen by the Dirichlet process $\mathcal{D}(\alpha, \mu)$, and then given P, the X's are a sample from P. Then Ferguson (1973) showed that the posterior distribution of P, given $X_1, \ldots, X_n$, is $\mathcal{D}(\alpha + n, \mu_n)$ where

$$\mu_n = \left(\frac{\alpha}{\alpha + n}\right)\mu + \left(\frac{n}{\alpha + n}\right)F_n, \tag{16.7}$$

where

$$F_n(x) = \frac{\sum_{i=1}^{n} I(X_i \leq x)}{n} \tag{16.8}$$

is the empirical distribution of $X_1, \ldots, X_n$.

For a weighted squared-error loss function (see Comment 6), the Bayes estimator of any function $g(P)$ is its expectation with respect to the posterior distribution. In particular, the Bayes estimator of P is μ_n and the Bayes estimator of $F(x) = P((-\infty, x])$ is $\mu_n((-\infty, x])$ where μ_n is given by (16.7).

Comments

1. *The Dirichlet Process Produces Only Discrete Distributions.* The Dirichlet process assigns probability 1 to the class of all discrete distributions in R_1. This means the random distribution produced by the Dirichlet prior is always a discrete distribution. This result was first proved by Ferguson (1973) in his seminal paper. An alternative proof can be obtained using Sethuraman's (see Sethuraman and Tiwari (1982) and Sethuraman(1994)) constructive definition of a Dirichlet measure (see Comment 2).

   Figure 16.1 gives 10 randomly selected P's chosen by the Dirichlet measure $\mathcal{D}(\alpha, \mu)$, where μ is chosen to be the exponential distribution with scale parameter 3, and α is taken to be 4 yielding

$$\mu((0, x]) = 1 - \exp(-3x) \text{ for } x \geq 0, \ = 0 \text{ for } x < 0, \tag{16.9}$$

$$\alpha = 4.$$

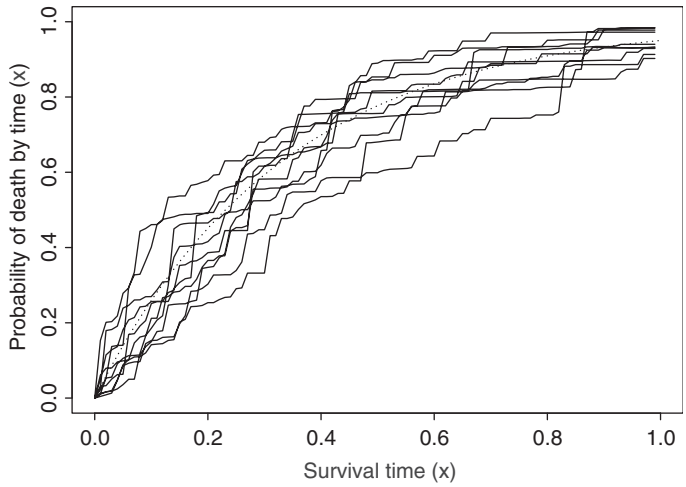

Figure 16.1 Ten randomly selected P's (dark curves) chosen by $\mathcal{D}(\alpha, \mu), \alpha = 4, \mu$ is exponential with scale parameter 3 (dotted curve).

The dark lines in Figure 16.1 are the randomly selected P's and the smooth curve is $\mu((0, x])$, which is the expected value of $F(x) = P((-\infty, x])$.

2. *Sethuraman's Constructive Definition of the Dirichlet Process.* Let $Y_1, Y_2, \ldots,$ $\theta_1, \theta_2, \ldots$ be independent random variables where the Y's are identically distributed with common distribution μ and the θ's are identically distributed according to a Beta distribution with parameters 1 and α. Set

$$p_1 = \theta_1, \, p_k = \theta_k \prod_{j=1}^{k-1} (1 - \theta_j), \tag{16.10}$$

$$P(A) = \sum_k p_k I(Y_k \epsilon A), \tag{16.11}$$

where for any event A, $I(A) = 1$ if A occurs and 0 otherwise. The random measure P is a random discrete probability measure putting mass p_k at Y_k. Sethuraman (1994) showed its distribution is the Dirichlet measure $\mathcal{D}(\alpha, \mu)$. The representation defined by (16.10) and (16.11) was announced in Sethuraman and Tiwari (1982) and studied in detail by Sethuraman (1994).

By its construction, the random measure P defined by (16.10) and (16.11) is a random discrete measure. It still had to be proved, however, that this process is the same as Ferguson's Dirichlet process. Sethuraman needed to show that finite-dimensional distributions are finite-dimensional Dirichlet. He did this using a fixed point theorem and he also used the same fixed point theorem to show that the posterior distribution is also Dirichlet. Since Sethuraman's representation uses only independent random variables, other Dirichlet results are relatively easy to prove using the representation. The representation also lends itself to computation. In particular, it can be used to obtain random samples from the Dirichlet process (see Sethuraman (1994), Doss (1994), and Hollander and Sethuraman (2001)).

3. *Relationship of the Bayes Estimator to the Empirical Distribution Function.* From (16.7), we see that the Bayes estimator μ_n is a linear combination of the prior guess μ and the empirical distribution F_n. Furthermore, as n gets large, the first

term in (16.7) approaches 0 and the second term approaches F_n. Thus, for large n, the influence of the prior information is diminished (the sample itself does most of the "talking"!) as is typically the case for Bayesian procedures.

4. *Interpretation of α as the Prior Sample Size.* It is a common practice to view α as the prior sample size partially because setting $\alpha = 0$ in (16.7) yields $\mu_n = F_n$. Sethuraman and Tiwari (1982) pointed out that there is a difficulty with this view. The Dirichlet process prior is not defined if $\alpha = 0$ and as α tends to 0 with the prior guess μ held fixed, it can be shown that the limiting prior distribution is a degenerate distribution that puts all of its mass at one random point Y, where Y has the distribution μ. This is too confining a prior opinion and this limiting distribution would typically not be an acceptable prior for use in Bayesian problems.
5. *Mean and Variance of $P(A)$.* Taking the partition (A, A^c) it follows directly from the definition of the Dirichlet process that if the Dirichlet prior is $\mathcal{D}(\alpha, \mu)$, then $P(A)$ has a Beta distribution with parameters $(\alpha\mu(A), \alpha\mu(A^c))$. It then follows that $E(P(A)) = \mu(A)$, $var(P(A)) = \mu(A)((1 - \mu(A))/(\alpha + 1)$.
6. *Bayes Estimator of P When the Loss Function Is Weighted Squared Error.* When estimating $F(x) = P((-\infty, x])$ by $\hat{F}(x)$, we will assume that the loss function is weighted squared error so that when we estimate P by $\hat{F}$ we "lose"

$$L(P, \hat{F}) = \int (F(x) - \hat{F}(x))^2 dW(x), \tag{16.12}$$

where W is a finite measure. The Bayes estimator is the $\hat{F}$ that minimizes the expected loss with respect to the posterior distribution. The minimizer is the expectation of $P((-\infty, x])$ with respect to $\mathcal{D}(\alpha + n, \mu_n)$ and this expectation is $\mu_n((-\infty, x])$.
7. *Unconditional Distribution of Sample Observations*. Ferguson (1973) showed that if $P \sim \mathcal{D}(\alpha, \mu)$ and X_1 is a sample of size 1 from P, then

$$Q(X_1 \in A) = \mu(A),$$

where Q denotes probability. He established the result using conditional expectations as follows.

$$Q(X_1 \in A) = \mathcal{E}Q(X_1 \in A | P(A)) = \mathcal{E}P(A) = \mu(A), \tag{16.13}$$

where the second equality uses the fact that given P, X_1 is a sample of size 1 from the distribution P. With $A = (-\infty, x]$, (16.13) yields the conditional distribution function of X_1, that is

$$Q(X_1 \leq x) = \mu((-\infty, x]).$$

Hollander and Korwar (1976) used moments of the Dirichlet distribution to generalize Ferguson's result to a sample of size m. Let $P \sim \mathcal{D}(\alpha, \mu)$ and let $X_1, \ldots, X_m$ be a sample of size m from P. Then

$$Q\{X_1 \leq x_1, \ldots, X_m \leq x_m\} = \frac{\prod_{j=1}^{m} \left(\alpha\mu\left(A_{x_{(j)}}\right) + j - 1\right)}{\prod_{j=1}^{m} (\alpha + j - 1)}, \tag{16.14}$$

where $x_{(1)} \leq \cdots \leq x_{(m)}$ are the order statistics of $x_1, \ldots, x_m$ and $A_x = (-\infty, x]$.

16.2 A BAYES ESTIMATOR OF THE DISTRIBUTION FUNCTION (FERGUSON)

Assumption

A1. $X_1, X_2, \ldots, X_n$ are a random sample from a distribution F with probability measure P, where $P((-\infty, x]) = F(x)$. Let the prior guess of the unknown P be the measure μ, and let α denote the degree of confidence in μ. The parameter α is often referred to as the *prior sample size*.

Procedure

For weighted squared-error loss (see Comment 6), the Bayes estimator of $F(x) = P((-\infty, x])$ is

$$\mu_n((-\infty, x]) = \left(\frac{\alpha}{\alpha + n}\right)\mu((-\infty, x]) + \left(\frac{n}{\alpha + n}\right)F_n((-\infty, x]), \tag{16.15}$$

where $\mu((-\infty, x]) = F_0(x)$ (say) is the prior guess at $F(x)$ and $F_n((-\infty, x]) = F_n(x) = \sum_{i=1}^{n} I(X_i \le x)/n$ is the empirical distribution function based on $X_1, \ldots, X_n$.

EXAMPLE 16.1 ***Framingham Heart Study.***

The Framingham Heart Study is a well-known ongoing longitudinal study of cardiovascular disease. The original study cohort consisted of a random sample of 5209 adults aged 28 through 62 years residing in Framingham, Massachusetts, between 1948 and 1951. The data in Table 16.1 consist of an extinct cohort of 12 men who were 67 and over at the fourth exam. The data are the times in days from the fourth exam to death.

McGee (2010) supplied the data of Table 16.1. He is an expert on the Framingham study. McGee postulated *a priori* an expected remaining life of about 8 years for men under those or similar situations in 1958 (about the time of the fourth exam). He felt that a two-parameter Gompertz distribution or a two-parameter Weibull distribution might be a good guess for μ but also an exponential with a mean of 8 (roughly 2292 days) would be a reasonable prior. Thus, partially for convenience of illustration, we set μ to be exponential with mean 2292, so that

$$\mu(x) = F_0(x) = 1 - e^{-\lambda x},\ \lambda = 1/2922,$$

and we take $\alpha = 4$. Figure 16.2 gives a plot of the empirical distribution F_n, the Bayes estimator, and the exponential prior μ.

Ferguson's estimator can be obtained from the R function `ferg.df(x, alpha, mu, npoints, ...)`, where x is a vector of length n, alpha is the prior sample size, mu is the prior guess for P, npoints is the number of estimated points returned, and "..." are all of the parameters needed for mu.

Table 16.1 Time from Fourth Exam to Death (Days)

i	1	2	3	4	5	6	7	8	9	10	11	12
X_i	2273	2710	141	4725	5010	6224	4991	458	1587	1435	2565	1863

Source: D. McGee (2010).

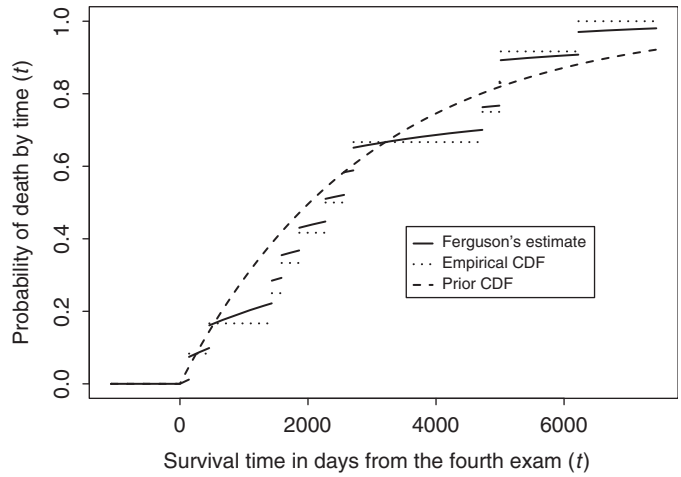

Figure 16.2 Ferguson's estimate, the empirical CDF, and the prior CDF for the fourth exam data of Table 16.1.

Comments

8. *Bayes Estimator.* With the Dirichlet process prior and weighted squared-error loss, the estimator μ_n given by (16.15) minimizes the Bayes risk among all estimators of F. That is, μ_n is the Bayes estimator for the Dirichlet Process prior.
9. *G-Invariant Dirichlet Process Priors.* Dalal (1979) introduced G-invariant Dirichlet process priors whose realizations are probability measures invariant under a finite group G of transformations. For the specific choice $G = \{e, g\}$ where $e(x) = x$ and $g(x) = 2\theta - x$, Dalal obtained the Bayes estimator against his prior for the problem of estimating a symmetric distribution with known center of symmetry θ. Dalal's estimator is a symmetrized version of Ferguson's estimator.

   Assume P follows a G-invariant process with $G = \{e, g\}$ and parameters $\mu\{(-\infty, x]\} = F_0$ (the prior guess at the symmetric F) and α. Then Dalal's estimator for a sample of size n is

$$\mu_n^*\{(-\infty, x]\} = p_n F_0(x) + (1 - p_n) \sum_{i=1}^{n} \left\{\delta_{X_i}(x) + \delta_{2\theta - X_i}(x)\right\} / 2n, \qquad (16.16)$$

   where $p_n = \alpha/(\alpha + n)$ and

$$\delta_X(t) = \begin{cases} 1 & \text{when } X \in \{(-\infty, t]\}, \\ 0 & \text{otherwise.} \end{cases}$$

   Hannum and Hollander (1983) used Dalal's G-invariant prior to study the robustness of Ferguson's estimator in the symmetric setting where Dalal's estimator is the Bayes estimator. Their paper gives information about the robustness of Ferguson's estimator against a prior for which it is not Bayes.

Properties

1. *Bayes optimality.* See Ferguson (1973) and Comment 8.
2. *Robustness.* See Hannum and Hollander (1983) and Comment 9.

Problems

1. The morning arrival times for 10 days of a school bus at the Tipperary street pickup stop are 8:05, 8:06, 8:09, 8:12, 8:03, 8:15, 8:11, 8:16, 8:00, and 8:14. The scheduled pickup time is 8:05 but past experience has led the students to believe there is a tendency for the bus to arrive late rather than be on time or early. For convenience we relabel so that 8:00 is represented by 0, 8:05 by 5, and so forth, hence we can consider the possible arrival interval to be $[0, 20]$. To utilize the students' experience, we choose the prior where the prior density $\mu' = f_0$ is

$$\mu'(x) = f_0(x) = \begin{cases} 1/40, & 0 \le x \le 10 \\ 3/40, & 10 \le x \le 20 \\ 0, & \text{otherwise,} \end{cases}$$

This choice makes arrivals a priori three times more likely to be in the interval $[10, 20]$ than in the interval $[0, 10]$. The corresponding prior distribution μ is

$$\mu(-\infty, x] = F_0(x) = \begin{cases} 0, & x < 0 \\ x/40, & 0 \le x \le 10 \\ \dfrac{1}{4} + \dfrac{3(x-10)}{40}, & 10 \le x \le 20 \\ 1, & x > 20. \end{cases}$$

For the choice $\alpha = 3$, find the Bayes estimator given by (16.15). What is the posterior probability that the bus will be more than 10 min late?

2. Consider the arrival times data of Problem 1 and again take $\alpha = 3$ but change the prior to be uniform, namely,

$$\mu'(x) = f_0(x) = \begin{cases} 1/20, & 0 \le x \le 20 \\ 0, & \text{otherwise,} \end{cases}$$

so that

$$\mu(-\infty, x] = F_0(x) = \begin{cases} 0, & x < 0 \\ x/20, & 0 \le x \le 20 \\ 1, & x > 20. \end{cases}$$

Find the Bayes estimator given by (16.15) and compare your answer to that of Problem 1. What is the posterior probability that the bus will be more than 10 min late? Is the unknown prior reasonable for a situation where *a priori* the bus has a tendency to be late?

3. Consider the arrival times of Problem 1, take $\alpha = 3$ and F_0 to be uniform on $[0, 20]$. Thus F_0 is symmetric with center of symmetry $\theta = 10$. Compute Dalal's estimator (see Comment 9) and compare it with Ferguson's estimator obtained in Problem 2.

4. Let P have the Dirichlet distribution $\mathcal{D}(\alpha, \mu)$. Show

(a) The expected value of $P(A)$ is $\mu(A)$.

(b) The variance of $P(A)$ is $\mu(A)\mu(A^c)/(\alpha + 1)$.

5. Show that Dalal's estimator μ_n^*, given by (16.16), of a symmetric distribution is a symmetrized version of Ferguson's Bayes estimator of an arbitrary distribution. That is, show

$$\mu_n^*((-\infty, x]) = \frac{1}{2}\left\{\mu_n(-\infty, x] + 1 - \mu_n(-\infty, [2\theta - x]^-)\right\},$$

where $F([t]^-)$ denotes $P(X < t)$ when X has the distribution F.

16.3 RANK ORDER ESTIMATION (CAMPBELL AND HOLLANDER)

Assumptions

A1 of Section 16.2

Rank Order Problem

We let $X_1, \ldots, X_n$ be a sample of size n from F. The problem is to estimate the rank order G of X_1 among $X_1, \ldots, X_n$ from the knowledge of r observed values $X_1, \ldots, X_r$ with $r < n$. Without loss of generality we can consider $X_1, \ldots, X_r$ to be the first r values in the unordered sample.

Some Possible Settings

(a) A pilot group of 10 astronauts are training for an important space mission. Each trainee earns a score based on his/her overall performance. Suppose X_1 represents astronaut BB's score. Based on $X_1, \ldots, X_{10}$, how do we estimate BB's rank in a total pool of 30 astronauts. In this scenario, $r = 10$ and $n = 30$. (The six best astronauts, as measured by X, will be chosen as the crew and the next best six will be chosen as the backup crew.)

(b) In 2008, there were five major hurricane (Category 3 strength or higher) with corresponding top wind speeds $X_1, \ldots, X_5$. How can we estimate, on the basis of $X_1, \ldots, X_5$, the rank of the storm represented by X_1 in the set of the five 2008 major hurricanes and the next 15 major hurricanes to occur? Here, $r = 5$ and $n = 20$.

Procedure

Let

$$\begin{aligned} K &= \text{ number of observations of } X_1, \ldots, X_n \text{ less than } X_1, \\ L &= \text{ number of observations of } X_1, \ldots, X_n \text{ equal to } X_1, \\ M &= \text{ number of observations of } X_1, \ldots, X_n \text{ greater than } X_1. \end{aligned}$$

Using average ranks when ties are present, the rank order G of X_1 among $X_1, ..., X_n$ is the average of the ranks that would be assigned to the L values tied at X_1, in a joint ranking from least to greatest, that is,

$$G = \{(K+1) + (K+2) + \cdots + (K+L)\}/L = K + \{(L+1)/2\}. \tag{16.17}$$

In a similar fashion, we define K', L', M' as the number of observations of $X_1, \ldots, X_r$ less than, equal to, and greater than X_1, respectively, so the rank order of X_1 among $X_1, \ldots, X_r$ is

$$G' = K' + (L' + 1)/2.$$

For squared-error loss, the mean of the posterior distribution of G, given $X_1, \ldots, X_r$, is the Bayes estimator $\hat{G}$ of G. Campbell and Hollander (1978) showed

$$\hat{G} = G' + (n - r)\left\{\alpha\mu(-\infty, X_1) + \sum_{i=1}^{r} \delta_{X_i}(-\infty, X_1) + \frac{1}{2}\alpha\mu(\{X_1\}) + \frac{1}{2}\sum_{i=1}^{r} \delta_{X_i}(\{X_1\})\right\} \Big/ (\alpha + n), \tag{16.18}$$

where the measure $\mu(\cdot)$ and the "prior sample size" α are the parameters of the Dirichlet, where $\delta_z(A) = 1$ if $z \in A$, 0 otherwise, and $\{X_1\}$ is the set consisting of the single point X_1. Campbell and Hollander showed that $\hat{G}$ can also be written as

$$\hat{G} = \{(n + \alpha)/(r + \alpha)\}\, G' - \frac{1}{2}\{(n - r)/(r + \alpha)\} + (n - r)\left[\alpha\mu(-\infty, X_1) + \frac{1}{2}\alpha\mu(\{X_1\})\right] \Big/ (r + \alpha). \tag{16.19}$$

The formula (16.19) for $\hat{G}$ is more computationally direct than (16.18).

We note that the estimator $\hat{G}$ depends on $X_1, \ldots, X_r$ only through X_1 and G' and of course $\hat{G}$ also depends on the prior guess μ and the prior sample size α, the parameters of the Dirichlet process.

EXAMPLE 16.2 ***Swimming the Womens' 50 yard Freestyle.***

A competitive Division I NCAA college swimming team is practicing for the 2009 conference championship. In the womens' 50 yard freestyle, the college team has decided to have two heats. There are six swimmers in the first heat and six swimmers in the second heat. A swimmer (WW, say) competes in the first heat and her time is X_1 seconds. The times $X_1, \ldots, X_6$ of the six swimmers in the first heat are given in Table 16.2. The team has decided that the six fastest swimmers in the two heats will be chosen to compete in the event at the conference championship. We wish to estimate WW's rank order among $X_1, \ldots, X_{12}$. Here $r = 6$ and $n = 12$.

An expert swimmer postulated, based on NCAA Division I data and her experience in the event, that a reasonable prior distribution is $N(22.52, .24)$, this is a normal distribution with mean 22.52 and standard deviation .24. Taking that as the prior parameter μ and

Table 16.2 Womens' 50 Yard Freestyle Times (seconds)

X_1	X_2	X_3	X_4	X_5	X_6
22.43	21.88	22.39	22.78	22.65	22.60

$\alpha = 10$ as the "prior sample size" parameter, we have from (16.19) with $r = 6$, $n = 12$, $G' = 3$, and $X_1 = 22.43$,

$$\hat{G} = \left(\frac{22}{16}\right)3 - \frac{1}{2}\left(\frac{6}{16}\right) + 60\left[P_{22.52,.24}(Y \leq 22.43) + \frac{1}{2}(0)\right] \Big/ 16,$$

where $P_{22.52,.24}(Y \leq 22.43)$ is the probability that a normal random variable Y with mean 22.52 and standard deviation .24 is less than or equal to 22.43, and $\mu(22.43) = 0$ because the normal distribution is continuous. Now, $P_{22.52,.24}(Y \leq 22.43) = P(Z \leq \frac{22.43-22.52}{.24}) = P(Z \leq -.375)$, where Z is an $N(0,1)$ random variable, and the probability is found to be .3538. Thus, $\hat{G} = \frac{66}{16} - \frac{3}{16} + \frac{60(.3538)}{16} = 5.264$, which may be rounded to 5.

The Campbell–Hollander estimator can be obtained from the R function `ch.ro(x, n, alpha, mu, ...)` where x is a vector of length r, n is the sample size, alpha is the prior sample size, mu is the prior guess for P, and ... are all the parameters needed for mu.

Note the role played by the prior's parameters μ and α. WW's time of 22.43 is slightly below the prior's mean of 22.52. Without the prior information we would have expected her rank in the 12 would be at the 50% point because her rank of 3 in the initial group of 6 put her at the 50% point of that group. The statistic $\hat{G}$ incorporates the prior information and adjusts WW's final ranking slightly downward.

Comments

10. *Bayes Optimality of* $\hat{G}$. With the Dirichlet process prior and squared-error loss, $\hat{G}$ minimizes the Bayes risk among all estimators of G (see Campbell and Hollander (1978)).
11. *Non-Bayesian Competitors of* $\hat{G}$. Johnson (1974) introduced non-Bayesian estimators of rank order for the rank order problem. His estimators are

$$G_F = G' + (n - r)F(X_1), \tag{16.20}$$

for the case when F is known and continuous, and

$$G_u = \{(n + 1)/(r + 1)\}\, G' \tag{16.21}$$

for the case when F is unknown.

For the case when F is known, Johnson showed that

$$\hat{\hat{T}} = \left[G' + (n - r + 1)F(X_1) - \{(n - r)/(n - 1)\}\right] \Big/ \left[1 + \left\{(2r - n - 1)/(n^2 - 1)\right\}\right] \tag{16.22}$$

is, conditional on $G = g$, an unbiased estimator of g. For the case when F is unknown, Johnson showed that for $r > 1$, the estimator

$$\tilde{T} = (r - 1)^{-1}(n - 1)(G' - 1) + 1 \tag{16.23}$$

is, conditional on $G = g$, an unbiased estimator of g.

12. *Optimal Properties of G_u and G_F in the Nonrandom Model.* In the nonrandom model (Model I), it is assumed that F is a nonrandom, continuous distribution from which a sample $X_1, \ldots, X_r, \ldots, X_n$ is taken. For this nonrandom model, Campbell and Hollander showed that in the class of linear rank order estimators of the form $aG' + c$, the estimator G_u minimizes the average mean-squared error. Campbell and Hollander also showed that in the nonrandom model, in the class of estimators of the form $aG' + bF(X_1) + c$, the estimator G_F minimizes the average mean square.
13. *Mean-Squared Error Comparisons.* Campbell and Hollander (1978) compared average mean-squared errors of the estimators $\hat{G}$, G_u, G_F, $\tilde{T}$, and $\hat{\tilde{T}}$ in the nonrandom model (Model I) and in the Dirichlet model (Model II). In Model II, $X_1, \ldots, X_n$ is a sample of size n from the Dirichlet process. Campbell and Hollander's Table 4.2 reflects the Bayesian optimality of $\hat{G}$.

Properties

1. *Bayes Optimality of $\hat{G}$.* See Campbell and Hollander(1978) and Comment 10.
2. *Average Mean-Squared Errors.* See Campbell and Hollander (1978) and Comment 13.

Problems

6. Suppose the prior used in Example 16.2 was changed to $\mu = N(22.35, .24)$ rather than $N(22.52, .24)$ and the prior sample size used remains $\alpha = 10$. How does this affect, if at all, WW's predicted rank? Compare the two results and, if a difference occurs, give an explanation for the difference.

7. Compute the estimators G_u and $\tilde{T}$ for the freestyle times of Table 16.2. Compare with the values of the estimator $\hat{G}$. Explain the differences, if any.

8. Show that formulas (16.18) and (16.19) are equivalent.

9. Describe three other possible scenarios for the rank order problem.

10. How would you obtain prior information for a problem involving California earthquakes and their severity as measured on the Richter scale?

11. Show that G_u can be obtained from G_F by replacing $F(X_1)$ with $G'/(r+1)$ in (16.20).

16.4 A BAYES ESTIMATOR OF THE DISTRIBUTION WHEN THE DATA ARE RIGHT-CENSORED (SUSARLA AND VAN RYZIN)

Assumptions

B1. We assume that given distribution functions F and G, $X_1, \ldots, X_n$ are independent and identically distributed according to the continuous life distribution F, and $Y_1, \ldots, Y_n$ are independent and identically distributed according to the continuous censoring distribution G. We observe

$$Z_i = \min\{X_i, Y_i\}$$

and δ_i where

$$\delta_i = \begin{cases} 1, & \text{if } X_i \le Y_i, \\ 0, & \text{if } X_i > Y_i. \end{cases}$$

Note that this is a change of notation from Section 11.6. We retain the usage of the X's earlier in this chapter.

B2. A prior distribution is needed for the pair (F, G). The prior used assumes that, under the prior, F and G are independent and that F is $\mathcal{D}(\alpha, \mu)$. As shown by Lo (1993), the distribution of G is irrelevant and the values $Y_1, \ldots, Y_n$ can be considered as constants.

Procedure

Rearrange the sample to take

$$Z_1 = X_1, \ldots, Z_r = X_r, Z_{r+1} = Y_{r+1}, \ldots, Z_n = Y_n,$$

where r is the number of uncensored X's.

The Bayes estimator of F, under the Dirichlet prior $\mathcal{D}(\alpha, \mu)$, was derived by Susarla and van Ryzin (1976). For squared-error loss, the estimator is given by (16.25). In (16.25), $Z_1^* < Z_2^* < \cdots < Z_m^*$ denote the distinct values among the censored observations $Z_{r+1}, \ldots, Z_n$, and

$$N_n(A) = \sum_{i=1}^{n} I(Z_i \in A), \tag{16.24}$$

where $I(Z_i \in A)$ is 1 if $Z_i \in A$, 0 otherwise. Note that $N_n(A)$ is a count among all the Z's, censored or not. The Susarla–van Ryzin estimator is

$$\hat{\bar{F}}_n(x) = \frac{\alpha\mu(x, \infty) + N_n(x, \infty)}{\alpha\mu(0, \infty) + n} \prod_{j=1}^{m} \frac{\alpha\mu([Z_j^*, \infty)) + N_n([Z_j^*, \infty))}{\alpha\mu([Z_j^*, \infty)) + N_n([Z_j^*, \infty)) - \lambda_j}, \tag{16.25}$$

where, in (16.25),

$$\lambda_k = \text{number of censored values equal to } Z_k^*, \ k = 1, \ldots, m,$$

and (16.18) holds for $Z_l^* \le x < Z_{l+1}^*$, $l = 0, \ldots, m$, Z_0^* is defined to be 0 and Z_{m+1}^* is defined to be ∞.

EXAMPLE 16.3 ***Hodgkin's Disease Data.***

We consider the "radiation of affected node" data given in Table 11.16. If one had available data from earlier studies in which the patients were given the affected node treatment, such data could be used to formulate the Dirichlet prior with corresponding specifications of the "guess" parameter $\mu(x) = F_0(x)$ and the "prior sample size" parameter α. We do not have such data but it is reasonable to specify an exponential for μ, namely, $\mu(x) = F_0(x) = 1 - e^{-\lambda x}$, an exponential with mean $1/\lambda$, and use the actual data of Table 11.16 to guide the choice of λ. That is, use that data to

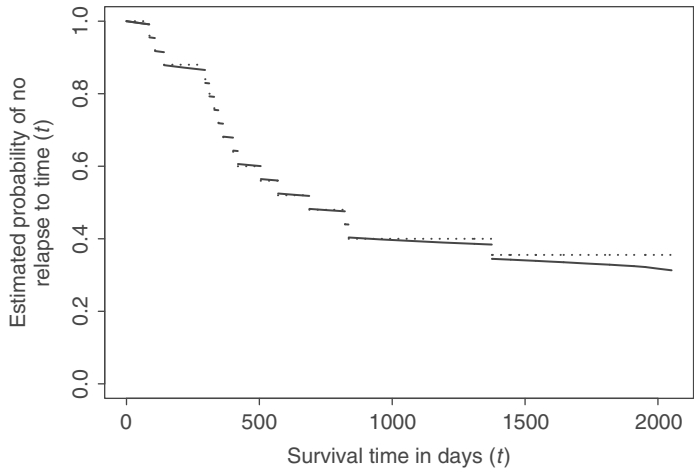

Figure 16.3 The Susarla–van Ryzin estimator (dark curve) and the Kaplan–Meier estimator (dotted curve) for the affected node data of Table 11.16.

obtain an estimate $\hat{\lambda}$ of λ. We could calculate the mean of the Kaplan–Meier estimator, which is $\int_0^\infty \bar{F}_{\mathrm{KM}}(x)dx$ and then set $\int_0^\infty \bar{F}_{\mathrm{KM}}(x) = 1/\hat{\lambda}$, and solve for $\hat{\lambda}$ to obtain $\hat{\lambda} = 1/(\int_0^\infty \bar{F}_{\mathrm{KM}}(x)dx)$. This would require choosing a convention to complete the tail of $\bar{F}_{\mathrm{KM}}$ because for the affected node data, the largest observation 2052 is a censored observation. It is more straightforward to use the median ($\hat{m}$ say) of $\bar{F}_{\mathrm{KM}}$. The median $\hat{m}$ is the smallest value x^* such that $\bar{F}_{\mathrm{KM}}(x^*) \leq 1/2$. From Table 11.16, we see that this value is $\hat{m} = 688$ because $\bar{F}_{KM}(688) = .480$ and $\bar{F}_{KM}(570) = .520$. The median m of $\bar{F}_0(x) = e^{-\lambda x}$ satisfies $e^{-\lambda m} = 1/2$, or equivalently, $m = \ln(2)/\lambda$. Therefore we set $\hat{m} = \ln(2)/\hat{\lambda}$, that is $688 = \ln(2)/\hat{\lambda}$, and find $\hat{\lambda} = \ln(2)/688 = .001$. Thus we use the "guess" $F_0(x) = 1 - e^{-(.001)x}$. The choice of α, the degree of faith in F_0, is somewhat arbitrary but here we choose $\alpha = 3$ and in Problem 13 ask you to investigate how other reasonable choices of α affect the calculation of $\hat{\bar{F}}_n$ given by (16.25).

Figure 16.3 is a plot of the Susarla–van Ryzin estimator and the Kaplan–Meier estimator.

The Susarla–van Ryzin estimator can be obtained from the R function `svr.df`.

Comments

14. *Bayes Optimality of* $\hat{\bar{F}}_n$. With the Dirichlet process prior and weighted squared-error loss, $\hat{\bar{F}}_n$ minimizes the Bayes risk among all estimators of $\bar{F}$.
15. *Calculation of the Bayes Estimator.* From (16.25) we see that both the uncensored and censored values have to be known to calculate the Bayes estimator. To calculate the Kaplan–Meier estimator we do not need to know the actual censored observations but only the number of censored observations between two uncensored observations.
16. *The Bayes Estimator is a Function of the Sufficient Statistic.* Susarla and van Ryzin pointed out that when μ is continuous, the data $(\underline{Z}, \underline{\delta}) = \{(Z_i, \delta_i), i = 1, \ldots, n\}$ can be recovered from the Bayes estimator. The Bayes estimator is thus a function of the sufficient statistic $(\underline{Z}, \underline{\delta})$. However, $(\underline{Z}, \underline{\delta})$ cannot be reclaimed

from the Kaplan–Meier estimator. Thus the Bayes estimator makes a fuller use of the data and in that sense is more informative.

17. *A Relationship between the Bayes and the Kaplan–Meier Estimators.* As α tends to 0, the Bayes estimator converges to the Kaplan–Meier estimator.
18. *Full Bayesian Analysis*. The Bayes estimator of F is the mean of the posterior distribution and is given by the Susarla–van Ryzin estimator of (16.25). The Susarla–van Ryzin development, however, does not provide a full Bayesian analysis. To do a full Bayesian analysis, the posterior distribution is needed. Furthermore, in the censored data framework, the posterior distribution is not a Dirichlet distribution. In Section 16.5, we show how Gibbs sampling approximates the posterior distribution.

Properties

1. *Bayes Optimality of* $\hat{\bar{F}}_n$. See Susarla and van Ryzin (1976) and Comment 14.
2. *Convergence to the Kaplan–Meier Estimator*. As α tends to 0, $\hat{\bar{F}}_n$ converges to $\bar{F}_{\mathrm{KM}}$. See Susarla and van Ryzin (1976) and Comment 17.

Problems

12. Recall Problem 37 of Chapter 11. The data corresponded to 211 patients with stage IV prostate cancer who were treated with estrogen. A random sample of size 14 from the 211 observations of survival times and withdrawal times (in months) from Table 11.19 yielded

Deaths:	31	33	97	3	32	17		
Censored:	66	3	5	13	93	13	59	38

As reported by Koziol and Green(1976), prior experience suggested that had the patients not been treated with estrogen, their survival distribution from cancer of the prostate would be exponential with mean 100. Compute the Susarla–van Ryzin estimator of F taking μ to be exponential with mean 100 and $\alpha = 4$.

13. Return to Example 16.3 and change the choice of α to $\alpha = 5$. How does this affect the Susarla–van Ryzin estimator? Now change $\alpha = 10$ and compute the Susarla–van Ryzin estimator. Compare with results with $\alpha = 3$ and $\alpha = 5$. What do you conclude?

14. Consider Table 11.18, times to first review of 1994 JASA Theory and Methods papers, and Table 11.23, times to first review of 1995 JASA Theory and Methods papers.

(a) Compute the Kaplan–Meier estimator of F for the 1994 data.

(b) Use the KM estimator obtained in part (a) as the prior guess μ to compute the Susarla–van Ryzin estimator of F for the 1995 data. Take $\alpha = 30$.

15. (a) Compute the Kaplan–Meier estimator of F for the 1995 data of Table 11.23.

(b) Use the KM estimator obtained in part (a) of Problem 15 as the prior guess μ for the 1994 JASA data (Table 11.18). Take $\alpha = 30$ and compute the Susarla–van Ryzin estimator for the 1994 data.

(c) Compare the results obtained in Problem 14(b) with those obtained in Problem 15(b).

16.5 OTHER BAYESIAN APPROACHES

In this section we briefly describe some other Bayesian approaches, including using different priors, Gibbs sampling, Gibbs sampling with the Dirichlet process, and the Sethuraman and Hollander partition-based (PB) priors developed in the context of repair models. For advanced Bayesian methods in survival analysis, see the January 2011 special issue of Lifetime Data Analysis. A brief summary of the papers contained in that issue is given by the issues' guest coeditors Chen and Gustafson (2011).

Other Priors

Mauldin, Sudderth, and Williams (2002) constructed a conjugate family of prior distributions from trees of Pólya urns. In particular, unlike the Dirichlet priors which assign probability one to the class of discrete, Pólya tree priors can assign probability one to the class of continuous distributions.

Dalal (1979) introduced G-invariant Dirichlet process priors whose realizations are probability measures invariant under a finite group of transformations (see Comment 9).

Dykstra and Laud (1981), using an extended gamma process, found posterior distributions of hazard rates for uncensored and censored data, Bayes estimators of hazard rates, and distribution functions for weighted squared-error loss.

Lo (1982) developed Bayesian nonparametric procedures for shock models using independent gamma-Dirichlet priors. He showed that these priors are conjugate when sampling from the shock models he considered. He obtained the Bayes estimator for the survival function under weighted squared-error loss.

One can find posterior distributions that are more general than Dirichlet distributions and more general than those mentioned earlier. Partition the whole space into two sets A_0, A_1, then each of these is partitioned into A_{00}, A_{01}, and A_{10}, A_{11}, respectively. Continue this partitioning. At the *nth* stage, the partitioning yields 2^n sets A_{t_n}, where t_n is an n-vector of 0's and 1's. The set A_{t_n} is partitioned as $A_{t_n,0}, A_{t_n,1}$, and there are 2^{n+1} sets at the $(n+1)$st stage. Under mild regularity, any probability measure P on R_1 provides the quantities $u_e = P(A_1)$, $u_0 = P(A_{01}|A_0)$, $u_1 = P(A_{11}|A_1), \ldots, u_{t_n} = P(A_{t_n 1}|A_{t_n}), \ldots$. These quantities are between 0 and 1 and conversely any sequence of numbers $(u_e, u_0, u_1, \ldots, u_{t_n}, \ldots)$ in $[0,1]$ yields a unique probability measure P on R_1, and the range of each u_{t_n} is unrestricted. Thus one can assign probability distributions to (u_{t_n}). Such probability distributions induce probability measures on the space of probability measures on R_1. Thus any such distribution can be used as a prior distribution. Blackwell (1973) showed that if we take the u's as independent, with u_{t_n} being a Beta distribution with parameters $(\alpha\mu(A_{t_n 1}), \alpha\mu(A_{t_n 0}))$, that yields the Dirichlet prior $D(\alpha, \mu)$ for P. Other choices for the distributions of the u's yield priors that concentrate on absolutely continuous distributions, symmetric distributions, unimodal distributions, distributions with increasing failure rate, and so forth.

Chen and Gustafson (2011) coedited a special issue of Lifetime Data Analysis devoted to Bayesian Methods in Survival Analysis.

Gibbs Sampling

Suppose we want to produce samples from a joint distribution K of a bivariate random variable (X, Y) or, for some function $t(x, y)$, desire to calculate $\int t(x, y) dK(x, y)$.

These calculations may be mathematically intractable directly, but Gibbs sampling provides a computational method for obtaining them. Suppose we can generate observations from the conditional distributions $Q(x|y) = P(X \le x|Y = y)$ and $R(y|x) = P(Y \le y|X = x)$. Then let (X_0, Y_0) be arbitrary. Generate X_n from the distribution $Q(\cdot|Y_{n-1})$ and Y_n from the distribution $R(\cdot|X_n)$, $n = 1, \ldots, N$. The sequence $(X_0, Y_0), (X_1, Y_1), \ldots,$ is a Markov chain with a transition function for which K is an invariant distribution. Under mild conditions on the transition function, the Markov chain converges in one of several senses to K. For large N, and a set A, $\frac{1}{N}\sum I((X_i, Y_i) \in A) \approx K(A)$ and $\frac{1}{N}\sum t(X_i, Y_i) \approx \int t(x, y)dK(x, y)$. If the convergence is stronger, (X_N, Y_N) approximately has distribution K.

The above-mentioned approach can be generalized to more than two variables and the variable need not be finite dimensional.

Gibbs Sampling with the Dirichlet

We describe Gibbs sampling with the Dirichlet in the context of the nonparametric censoring problem of Section 16.4. We follow the exposition in Hollander and Sethuraman (2001). Recall we observe $Z_i = \min(X_i, Y_i)$ and $\delta_i = I(X_i \le Y_i)$, $i = 1, \ldots, n$. Thus if $\delta_i = 1$ we observe the actual value of X_i but if $\delta_i = 0$ we only know that $X_i \in A_i$, where $A_i = (Y_i, \infty)$. We rearrange the samples so that $Z_1 = X_1, \ldots, Z_r = X_r, Z_{r+1} = Y_{r+1}, \ldots, Z_n = Y_n$ where r is the number of uncensored values and $m = n - r$ is the number of censored values. The Bayes posterior distribution is the conditional distribution of F given the data. We will use Gibbs sampling to simulate from this distribution. This will also produce a method of obtaining, via computational methods, the posterior estimate of the survival function. We will do this for the "radiation of affected node" data of Table 11.16. Then we will compare this estimate with the Susarla–van Ryzin estimate computed for those data.

Recall Ferguson's result (16.7) that the conditional distribution of F, given $X_1, \ldots, X_r$, is $\mathcal{D}(\alpha + r, \mu_r)$, where

$$\mu_r = \frac{\alpha}{\alpha + r}\mu + \frac{r}{\alpha + r}F.$$

With F having distribution $\mathcal{D}(\alpha + r, \mu_r)$, given F let $X_{r+1}, \ldots, X_n$ be independent and identically distributed according to F. The posterior distribution is the conditional distribution of F given $X_i \in A_i$, $i = r + 1, \ldots, n$. The conditional distribution $Q(F|X_r, \ldots, X_n, \text{data})$ of F is $\mathcal{D}(\alpha + n, \mu_n)$ and the conditional distribution $R(X_{r+1}, \ldots, X_n|F, \text{data})$ is that of independent random variables having distributions $F_{A_{r+1}}, \ldots, F_{A_n}$, where for any set A, F_A is the distribution F restricted to the set A. Hence, in the joint distribution of F and $(X_{r+1}, \ldots, X_n)$ given the data, we know the two conditional distributions to use Gibbs sampling. An observation $(F, X_{r+1}, \ldots, X_n)$ is simulated from the joint distribution of F and $(X_{r+1}, \ldots, X_n)$ given the data. The F can be considered an observation from the posterior distribution.

To generate a distribution from the Dirichlet prior, Sethuraman's (1994) representation of the Dirichlet process was used (see Comment 2 and (16.10), (16.11)). We start at stage 0 of Gibbs sampling. Let F^0, $X^0_{r+1}, \ldots, X^0_n$ be arbitrary or educated guesses at F and the unknown censored values. Define the measure

$$\mu^0_n = \frac{1}{n}\left[\alpha\mu + \sum_{i=1}^{r}\delta_{X_i} + \sum_{i=r+1}^{n}\delta_{X^0_i}\right],$$

where δ_x is the distribution that concentrates a mass of 1 at the point x, that is,

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

For a set A, say $A = (0, a)$,

$$\mu_n^0(A) = \frac{1}{n}\left[\alpha\mu(A) + \sum_{i=1}^{r} \delta_{X_i}(A) + \sum_{i=r+1}^{n} \delta_{X_i}(A)\right]. \tag{16.26}$$

Note that in (16.26), depending on the value of a, some of the $\delta_{X_i}((0,a))$ will be 1, others 0, and likewise for the $\delta_{X_i^0}((0,a))$. To generate F^1, take p_k (given by (16.10)) based on independent $\theta_1, \theta_2, \ldots$, which are iid Beta$(1, \alpha + n)$ and

$$F^1(x) = \sum_{k=1}^{\infty} p_k I(W_k \le x), \tag{16.27}$$

where $W_1, W_2, \ldots,$ are iid according to μ_n^0. The sum in (16.27) cannot be computed exactly because it is an infinite sum. We avoid generating the F^1 precisely, and go to the next step of generating $(X_{r+1}^1, ..., X_n^1)$ as iid according to this F^1 and satisfying, for $i = r+1, \ldots, n$, $X_i^1 \in A_i$. From Sethuraman's constructive definition of the Dirichlet, $X_{r+1}^1, \ldots, X_n^1$ can be taken to be $W_{I_{r+1}}, \ldots, W_{I_n}$, where $I_{r+1}, \ldots, I_n$ are iid according to the random discrete distribution $\{p_k,\ k = 1, \ldots,\}$. Thus first generate a sufficiently large vector $W_1, \ldots, W_N$, which are iid μ_n^0. The random index I_{r+1} is then generated as $\min\left\{j : \sum_{i=1}^{j} p_i \ge U\right\}$, where U is uniform on $[0, 1]$. Thus one needs to generate only a finite (through random) number of p_i's to generate I_{r+1}. If $X_{r+1}^1 = W_{I_{r+1}}$ is not in A_{r+1}, repeat this step till it happens. This yields X_{r+1}^1. Continue in this fashion till $X_{r+1}^1, \ldots, X_n^1$ are generated. Similarly, generate $(F^M, X_{r+1}^M, \ldots, X_n^M)$ for $M = 1, \ldots, K$. At each step, the infinite series for F^M does not have to be calculated. The average of μ_n^M over several such runs is the computational Bayes estimate of μ. See Doss (1994) where this computational method is described. Athreya, Doss and Sethuraman (1994) showed the Gibbs sampler converges. Table 16.3 computes the exact and computational Bayes estimates of F for the "radiation of affected node" data of Table 11.16.

Repair Models and Partition-Based Priors

Repair models are important in reliability because when a system (or item) fails in the field, cost and other considerations typically preclude replacement with a new item and instead the system is repaired, then put back in use. Hollander and Sethuraman (2002) discussed various repair models and described non-Bayesian nonparametric methods for such models.

One repair model that has been frequently considered in the literature is the age-dependent repair model of Block, Borges, and Savits (1985) (BBS, 1985). In this model, two types of repairs are possible at the time of the repair, namely, a perfect repair or a minimal repair. A perfect repair replaces the failed item with a new one and is performed with probability $p(t)$. A minimal repair restores the system to its state just before failure and is performed with probability $1 - p(t)$. The function $p(t)$ is allowed to depend on the age of the failed system.

Table 16.3 Exact (Susarla–van Ryzin) and Computational (Gibbs) Estimates of the Survival Function for the "Radiation of Affected Node" Data

x	Exact (SVR)	Computational (Gibbs)
86	.9555	.9555
107	.9177	.9177
141	.8788	.8788
296	.8297	.8297
312	.7927	.7927
330	.7556	.7556
346	.7187	.7187
364	.6816	.6816
401	.6432	.6432
419	.6062	.6062
505	.5647	.5647
570	.5249	.5249
688	.4824	.4824
822	.4400	.4399
836	.4036	.4036
1309	.3861	.3861
1375	.3447	.3485
1378	.3446	.3484
1446	.3424	.3467
1540	.3392	.3444
1645	.3353	.3302
1818	.3286	.3269
1910	.3245	.3254
1953	.3221	.3188
2052	.3130	.3173

Let F be the distribution of the time to first failure of the system. At first glance, it would seem reasonable, if one wanted to perform Bayesian nonparametric inference for F, to put a Dirichlet prior on F. Repair models such as the BBS model, however, introduce certain dependencies so that with a Dirichlet prior on F, the posterior distribution of F, given the data, is not Dirichlet. In other words, the Dirichlet process is not conjugate for these repair models. Sethuraman and Hollander (2009) introduced a new class of partition-based (PB) priors that are conjugate and they also defined a subclass of the PB class called *partition-based Dirichlet* (*PBD*) priors that also form a conjugate family. They considered the general framework of a sequence of dependent random variables $X_1, X_2, \ldots, X_n$, where X_1 is distributed according to a probability measure P(or equivalently, a distribution F), and the distribution of X_{i+1} given $X_1, \ldots, X_i$ and other covariates depends on P and the previous data, $i = 1, \ldots, n$. Sethuraman and Hollander (2009) developed nonparametric Bayesian methods to estimate P, which, unlike many of the non-Bayesian nonparametric methods described in Hollander and Sethuraman (2002), do not make any assumptions about when data collection is stopped.

Bibliography

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics* **6**, 701–726.

Abdushukurov, A. A. (1987). Nonparametric estimation in the proportional hazards model of random censorship. Akad Nauk UzSSR, Tashkent VINITI No. 3448-V.

Adichie, J. N. (1967). Estimates of regression parameters based on rank tests. *Annals of Mathematical Statistics* **38**, 894–904.

Adichie, J. N. (1974). Rank score comparison of several regression parameters. *Annals of Statistics* **2**, 396–402.

Adichie, J. N. (1976). Testing parallelism of regression lines against ordered alternatives. *Communications in Statistics: Theory and Methods* **5**, 985–997.

Adichie, J. N. (1984). In Krishnaiah and P. K. Sen (Eds), *Rank Tests in Linear Models*, Handbook in Statistics, Volume **4**, pp. 229–257. Amsterdam: Elsevier Science.

Agresti, A. (2013). *Categorical Data Analysis*, 3rd edn. New York: John Wiley and Sons, Inc.

Agresti, A., and B. Caffo (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician* **54**, 280–288.

Agresti, A., and B. A. Coull. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician* **52**, 119–126.

Agresti, A., C. R. Mehta, and N. R. Patel. (1990). Exact inference for contingency tables with ordered categories. *Journal of the American Statistical Association* **85**, 453–458.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Akritas, M. G. (1986). Empirical processes associated with V-statistics and a class of estimators under random censoring. *Annals of Statistics* **14**, 619–637.

Akritas, M. G. (1988). Pearson-type goodness-of-fit tests: The univariate case. *Journal of the American Statistical Association* **83**, 222–230.

Akritas, M. G. (2004). Nonparametric survival analysis. *Statistical Science* **19**, 615–623.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16** pp. 125–127.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician* **46**, 175–185.

Aly, E.-E. (1990). Tests for monotonicity properties of the mean residual life function. *Scandinavian Journal of Statistics* **17**, 189–200.

Anděl, J. (1967). Local asymptotic power and efficiency of tests of Kolmogorov-Smirnov type. *Annals of Mathematical Statistics* **38**, 1705–1725.

Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding. (1993). *Statistical Models Based on Counting processes*. New York: Springer-Verlag.

Anderson, J. D., L. Efron, and S. K. Wong. (1970). Martian mass and earth-moon mass ratio from coherent S-band tracking of Mariners 6 and 7. *Science* **167**, 277–279.

Anderson, V. L., and R. A. McLean. (1974). *Design of Experiments: A Realistic Approach*. New York: Dekker.

Andrews, B. (1995). Bodily shame as a mediator between abusive experiences and depression. *Journal of Abnormal Psychology* **104**, 277–285.

Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics* **16** pp. 523–531.

Andrews, D.F., and A.M. Herzberg. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. New York: Springer-Verlag.

Andrews, F. C. (1954). Asymptotic behavior of some rank tests for analysis of variance. *Annals of Mathematical Statistics* **25**, 724–736.

Ansari, A. R., and R. A. Bradley. (1960). Rank-sum tests for dispersions. *Annals of Mathematical Statistics* **31**, 1174–1189.

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

Anscombe, F. J. (1965). Comments on Kurtz-Link-Turkey-Wallace paper. *Technometrics* **7**, 167–168.

Aragon, T. J. (2012). epitools: Epidemiology Tools. R package version 0.5-7.

Arbuthnott, J. (1710). An argument for divine providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society* **27**, 186–190.

Archambault, W. A. T. Jr., G. A. Mack, and D. A. Wolfe. (1977). K-sample rank tests using pair-specific scoring functions. *Canadian Journal of Statistics* **5**, 195–207.

Arjas, E., and D. Gasbarra. (1994). Nonparametric Bayesian inference for right-censored survival data, using the Gibbs sampler. *Statistica Sinica* **2**, 505–524.

Arnold, B. C., N. Balakrishnan, and H. N. Nagaraja. (1992). *A First Course in Order Statistics*. New York: John Wiley and Sons, Inc.

Arnold, H. J. (1965). Small sample power for the one-sample Wilcoxon test for non-normal shift alternatives. *Annals of Mathematical Statistics* **36**, 1767–1778.

Astin, M. C., S. M. Ogland-Hand, E. M. Coleman, and D. W. Foy. (1995). Posttraumatic stress disorder and childhood abuse in battered women: Comparisons with maritally distressed women. *Journal of Consulting and Clinical Psychology* **63**, 308–312.

Athreya, K. B., H. Doss, and J. Sethuraman. (1994). On the convergence of the Markov chain simulation method. *Annals of Statistics* **24**, 69–100.

August, G. P., W. Hung, and J. C. Houck. (1974). The effects of growth hormone therapy on collagen metabolism in children. *Journal of Clinical Endocrinology and Metabolism* **39**, 1103–1109.

Barlow, R. E. (1968). Likelihood ratio tests for restricted families of probability distributions. *Annals of Mathematical Statistics* **39**, 547–560.

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. (1972). *Statistical Inference Under Order Restrictions*. New York: John Wiley and Sons, Inc.

Barlow, R. E., and R. Campo. (1975). Total time on test processes and applications to failure data analysis. In R. E. Barlow, J. Fursell, and N. Singpurwalla (Eds), *Reliability and Fault Tree Analysis*, Philadelphia: SIAM, pp. 451–481.

Barlow, R. E., and K. Doksum. (1972). Isotonic tests for convex orderings. *Proceedings of the 6th Berkeley Symposium* **1**, 293–323.

Barlow, R. E., and F. Proschan. (1966). Inequalities for linear combinations of order statistics from restricted families. *Annals of Mathematical Statistics* **37**, 1574–1592.

Barlow, R. E., and F. Proschan. (1969). A note on tests for monotone failure rate based on incomplete data. *Annals of Mathematical Statistics* **40**, 595–600.

Barlow, R. E., and F. Proschan. (1981). *Statistical Theory of Reliability and Life Testing*, Second Printing Publisher: To begin With, 1137 Hornell Drive, Silver Spring, MD 20904.

Barlow, R. E., and E. M. Scheuer. (1971). Estimation from accelerated life tests. *Technometrics* **13**, 145–159.

Barnard, G. A. (1945). A new test for 2×2 tables *Nature* **156**, 177.

Barnard, G. A. (1947). Significance tests for 2×2 tables. *Biometrics* **34**, 123–138.

Bauer, D. F. (1972). Constructing confidence sets using rank statistics. *Journal of the American Statistical Association* **67**, 687–690.

Beaton, A. E., and J. W. Tukey. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics* **16** pp. 147–185.

Benard, A., and P. van Elteren. (1953). A. generalization of the method of *m* rankings. *Indagationes Mathematicae* **15**, 358–369.

Benedetti, J. (1977). On the nonparametric estimation of regression functions. *Journal of the Royal Statistical Society Series B (Methodological)* **39**, 248–253.

Berger, J.O. (1985). Statistical Decision Theory and Bayesian Analysis, 2nd edn. New York: Springer-Verlag.

Bergman, B. (1977). Crossings in the total time on test plot. *Scandinavian Journal of Statistics* **4**, 171–177.

Bernoulli, J. (1713). *Ars Conjectandi*.

Bhattacharyya, G. K., R. A. Johnson, and H. R. Neave. (1970). Percentage points of some non-parametric tests for independence and empirical power comparisons. *Journal of the American Statistical Association* **65**, 976–983.

Bick, R. L., T. Adams, and W. R. Schmalhorst. (1976). Bleeding times, platelet adhesion, and aspirin. *Journal of Clinical Pathology* **65**, 69–72.

Bickel, P. J. (1965). On some robust estimates of location. *Annals of Mathematical Statistics* **36**, 847–858.

Bickel, P. J., and K. A. Doksum. (1969). Tests for monotone failure rate based on normalized spacings. *Annals of Mathematical Statistics* **40**, 1216–1235.

Bickel, P. J., and D. A. Freedman. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics* **9**, 1196–1217.

Bie, O., Ø. Borgan, and K. Liestøl. (1987). Confidence intervals and confidence bands for the cumulative hazard function and their small sample properties. *Scandinavian Journal of Statistics* **14**, 221–233.

Bierens, H. (1987). Kernel estimators of regression functions. *Advances in Econometrics: 5th World Congress of the Econometric Society*, vol. **1**.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: John Wiley and Sons, Inc.

Birch, M. W. (1964). The detection of partial correlation I: The 2×2 case. *Journal of the Royal Statistical Society, Series B* **26**, 313–324.

Birnbaum, Z. W. (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. *Journal of the American Statistical Association* **47**, 425–441.

Birnbaum, Z. W. (1956). On a use of the Mann-Whitney statistic. *Proceedings of the 3rd Berkeley Symposium*, **1**, 13–17.

Birnbaum, Z. W., and O. M. Klose. (1957). Bounds for the variance of the Mann-Whitney statistic. *Annals of Mathematical Statistics* **28**, 933–945.

Birnbaum, Z. W., and R. C. McCarty. (1958). A distribution-free upper confidence bound for $\Pr\{Y < X\}$, based on independent samples of X and Y. *Annals of Mathematical Statistics* **29**, 558–562.

Birnbaum, Z. W., and F. H. Tingey. (1951). One-sided confidence contours for probability distribution functions. *Annals of Mathematical Statistics* **22**, 592–596.

Bjerkedal, T. (1960). Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli. *American Journal of Hygiene* **72**, 130–148.

Blackwell, D. (1973). Discreteness of Ferguson selections. *Annals of Statistics* **1**, 356–358.

Bliss, C. I. (1944). A chart of the chi-square distribution. *Journal of the American Statistical Association* **39**, 246–248.

Block, H., W. Borges, and T. Savits. (1985). Age-dependent minimal repair. *Journal of Applied Probability* **22**, 370–385.

Blomqvist, N. (1950). On a measure of dependence between two random variables. *Annals of Mathematical Statistics* **21**, 593–600.

Blum, J. R., J. Kiefer, and M. Rosenblatt. (1961). Distribution free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics* **32**, 485–498.

Blyth, C. (1950). In E. L. Lehmann. *Notes on the Theory of Estimation, recorded from lectures*. Berkeley: University of California Press.

Bohn, L. L. (1996). A review of nonparametric ranked-set sampling methodology. *Communications in Statistics: Theory and Methods* **25**, 2675–2685.

Bohn, L. L., and D. A. Wolfe. (1992). Nonparametric two-sample procedures for ranked-set samples data. *Journal of the American Statistical Association* **87**, 552–561.

Bohn, L. L., and D. A. Wolfe. (1994). The effect of imperfect judgment rankings on properties of procedures based on the ranked-set samples analog of the Mann-Whitney-Wilcoxon statistic. *Journal of the American Statistical Association* **89**, 168–176.

Boos, D.B., and C. Brownie. (2004). Comparing variances and other measures of dispersion. *Statistical Science* **19**, 571–578.

Borgan, Ø., and K. Liestøl. (1990). A note on confidence intervals and bands for the survival curve based on transformations. *Scandinavian Journal of Statistics* **17**, 35–41.

Borges, W. S., F. Proschan, and J. Rodrigues. (1984). A simple test for new better than used in expectation. *Communications in Statistics: Theory and Methods* **13**, 3217–3223.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* **40**, 318–335.

Box, G. E. P., and S. L. Andersen. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society, Series B* **17**, 1–26.

Box, J. F. (1978). *R. A. Fisher: The Life of a Scientist*. Hoboken, New Jersey: John Wiley and Sons, Inc.

Boyles, R. A., and F. J. Samaniego. (1984). Estimating a survival curve when new is better than used. *Operations Research* **32**, 732–740.

Bradley, R. A. (1963). Some relationships among sensory difference tests. *Biometrics* **19**, 385–397.

Brady, J. P. (1969). Studies on the metronome effect on stuttering. *Behaviour Research and Therapy* **7**, 197–204.

Breslow, N. (1981). Odds ratio estimators when the data are sparse. *Biometrika* **68**, 73–84.

Breslow, N. E., and J. J. Crowley. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics* **2**, 437–453.

Breslow, N., and N. E. Day. (1980). *Statistical Methods in Cancer Research, 1. The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer.

Brinkman, N. D. (1981). Ethanol fuel-a single-cylinder engine study of efficiency and exhaust emissions. *SAE Transactions* **90** pp. 1410–1424.

Brown, B. W., Jr., and M. Hollander. (1977, 2008) *Statistics: A Biomedical Introduction*, A Volume in the Wiley Classics Library Series. New York: John Wiley and Sons, Inc.

Brown, B. W., Jr., M. Hollander, and R. M. Korwar. (1974). Nonparametric tests of independence, with applications to heart transplant studies. In F. Proschan and R. J. Serfling (Eds), *Reliability and Biometry: Statistical Analysis of Lifelength*, pp. 327–354. Philadelphia: SIAM.

Brown, G. W., B. Andrews, T. Haris, A. Alder, and L. Bridge. (1986). Social support, self-esteem and depression. *Psychological Medicine* **16**, 182–831.

Brown, L. D, T. T. Cai and A. DasGupta. (2001). Interval estimation for a binomial proportion (with comments by A. Agresti, B. A. Coull, G. Casella, C. Corcoran, C. Mehta, M. Ghosh, and T. J. Santner) *Statistical Science* **16**, 101–133.

Brown, M. (1984). On the choice of variance for the log rank test. *Biometrika* **71**, 65–74.

Brunden, M. N., and N. R. Mohberg. (1976). The Benard–van Elteren statistic and nonparametric computation. *Communications in Statistics: Simulation and Computation* **5**, 155–162.

Bryson, M. C., (1974). Heavy-tailed distributions: properties and tests. *Technometrics* **16**, 61–68.

Bryson, M. C., and M. M. Siddiqui. (1969). Some criteria for aging. *Journal of the American Statistical Association* **64**, 1472–1483.

Bugyi, H. L., E. Magnier, W. Joseph, and G. Frank. (1969). A method for measurement of sodium and potassium in erythrocytes and whole blood. *Clinical Chemistry* **15**, 712–719.

Burnett, W. C. Jr., and S. B. Jones, Jr. (1973). Differential feeding of the southern armyworm on Kentucky and Florida populations of pokeweed. *American Midland Naturalist* **90**, 231–234.

Burr, E. J. (1960). The distribution of Kendall's score S for a pair of tied rankings. *Biometrika* **47**, 151–171.

Byer, A. J., and D. Abrams. (1953). A comparison of the triangular and two-sample teste-test methods. *Food Technology* **7**, 185–187.

Cai, T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Annals of Statistics* **27**, 898–924.

Cai, T., and L. Brown. (1998). Wavelet shrinkage for nonequispaced samples. *Annals of Statistics* **26**, 1783–1799.

Cai, T., and L. D. Brown. (1999). Wavelet estimation for samples with random uniform design. *Statistics and Probability Letters* **42**, 313–321.

Cai, T., and B. Silverman. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhya-The Indian Journal of Statistics Series B* **63**, 127–148.

Cain, G. D., G. Mayer, and E. A. Jones. (1970). Augmentation of albumin but not fibrinogen synthesis by corticosteroids in patients with hepatocellular disease. *Journal of Clinical Investigation* **49**, 2198–2204.

Campbell, G., and M. Hollander. (1978). Rank order estimation with the Dirichlet prior. *Annals of Statistics* **6**, 142–153.

Campbell, J. A., and O. Pelletier. (1962). Determination of niacin (niacinamide) in cereal products. *Journal of the Association of Official Analytical Chemists* **45**, 449–453.

Capon, J. (1965). On the asymptotic efficiency of the Kolmogorov-Smirnov test. *Journal of the American Statistical Association* **60**, 843–853.

Casella, G. (1986). Refining binomial confidence intervals. *Canadian Journal of Statistics* **14**, 113–127.

Casella, G., and R. L. Berger. (2002). *Statistical Inference*. 2nd edn. Pacific Grove, CA: Duxbury.

Cencov, N. N. (1962). Evaluation of an unknown distribution density function from observations. *Doklady* **3**, 1559–1562.

Chambers, J. M., and T. J. Hastie. (1992). *Statistical Models in S*. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Chandra, M., N. D. Singpurwalla, and M. A. Stephens. (1981). Kolmogorov statistics for tests of fit for the extreme-value and Weibull distributions. *Journal of the American Statistical Association* **76**, 729–731.

Chang, T. (2004). Spatial statistics. *Statistical Science* **19**, 624–635.

Chen, H., E. A. Stasny, and D. A. Wolfe. (2005). Ranked set sampling for efficient estimation of a population proportion. *Statistics in Medicine* **24**, 3319–3329.

Chen, H., E. A. Stasny, and D. A. Wolfe. (2006a). An empirical assessment of ranking accuracy in ranked set sampling. *Computational Statistics & Data Analysis* **51**, 1411–1419.

Chen, H., E. A. Stasny, and D. A. Wolfe. (2006b). Unbalanced ranked set sampling for estimating a population proportion. *Biometrics* **62**, 150–158.

Chen, H., E. A. Stasny, and D. A. Wolfe. (2008). Ranked set sampling for ordered categorical variables. *Canadian Journal of Statistics* **36**, 179–181.

Chen, H., E. A. Stasny, D. A. Wolfe, and S. N. MacEachern. (2009). Unbalanced ranked set sampling for estimating a population proportion under imperfect rankings. *Communications in Statistics: Theory and Methods* **38**, 2116–2125.

Chen, M.-H., and P. Gustafson. (2011). Bayesian methods in survival analysis. *Lifetime Data Analysis* **17**, 1–2.

Chen, Y. I., and D. A. Wolfe. (1990a). Modifications of the Mack-Wolfe umbrella tests for a generalized Behrens-Fisher problem. *Canadian Journal of Statistics* **18**, 245–253.

Chen, Y. I., and D. A. Wolfe. (1990b). A study of distribution-free tests for umbrella alternatives. *Biometrical Journal* **32**, 47–57.

Chen, Y. I., and D. A. Wolfe. (1993). Nonparametric procedures for comparing umbrella pattern treatment effects with a control in a one-way layout. *Biometrics* **49**, 455–465.

Chen, Y. Y., M. Hollander, and N. A. Langberg. (1982). Small-sample results for the Kaplan-Meier estimator. *Journal of the American Statistical Association* **77**, 141–144.

Chen, Y. Y., M. Hollander, and N. A. Langberg. (1983a). Testing whether new is better than used with randomly censored data. *Annals of Statistics* **11**, 267–274.

Chen, Y. Y., M. Hollander, and N. A. Langberg. (1983b). Tests for monotone mean residual life using randomly censored data. *Biometrics* **39**, 119–127.

Chen, Z., Z. D. Bai, and B. K. Sinha. (2004). *Ranked Set Sampling: Theory and Applications*. New York: Springer-Verlag.

Cheng, P. E., and G. D. Lin. (1987). Maximum likelihood estimation of a survival function under the Koziol-Green proportional hazards model. *Statistics & Probability Letters* **5**, 75–80.

Chernoff, H., and I. R. Savage. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Annals of Mathematical Statistics* **29**, 972–994.

Chicken, E. (2003). Block thresholding and wavelet estimation for nonequispaced samples. *Journal of Statistical Planning and Inference* **116**, 113–129.

Chicken, E. (2005). Block-dependent thresholding in wavelet regression. *Journal of Nonparametric Statistics* **17**, 467–491.

Chicken, E. (2011). Nonparametric nonlinear profiles. In A. Saghaei, A. Amiri, and R. Noorossana eds., *Statistical Analysis of Profile Monitoring*, pp. 157–188. Hoboken, NJ: John Wiley and Sons.

Chicken, E., J. J. Pignatiello Jr., and J. R. Simpson. (2009). Statistical process monitoring of nonlinear profiles using wavelets. *Journal of Quality Technology* **41**(2), 198–212.

Clark, P. J., S. G. Vandenberg, and C. H. Proctor. (1961). On the relationship of scores on certain psychological tests with a number of anthropometric characters and birth order in twins. *Human Biology* **33**, 163–180.

Clarkson, D. B., Y. Fan, and H. Joe. (1993) A remark on algorithm 643: FEXACT: an algorithm for performing Fisher's exact test in `r x c` contingency tables. *ACM Transactions on Mathematical Software*, **19**, 484488.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Cleveland, W. S. (1994). *The Elements of Graphing Data.* Summit, NJ: Hobart. 2E.

Cleveland, W. S., and S. J. Devlin. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83** pp. 596–610.

Cleveland, W. S., and E. Grosse. (1991). Computational methods for local regression. *Statistics and Computing* **1**, 47–62.

Cleveland, W. S., E. Grosse, and W. M. Shyu. (1992). Local regression models. In J. M. Chambers and T. J. Hastie (Eds), *Statistical Models in S.* Pacific Grove, CA: Wadsworth, Chapter 8.

Clogg, C. C., and J. W. Shockey. (1988). Multivariate analysis of discrete data. In J. R. Nesselroade and R. B. Cattell (Eds), *Handbook of Multivariate Experimental Psychology*, pp. 337–365. New York: Plenum Press.

Clopper, C. J., and E. S. Pearson. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404–413.

Cochran, W. G. (1937). The efficiencies of the binomial series tests of significance of a mean and of correlation coefficient. *Journal Royal Statistical Society* **100**, 69–73.

Cochran, W. G. (1954). Some methods of strengthening the χ^2 common tests. *Biometrics* **10**, 317–451.

Cochran, W. G., and G. M. Cox. (1957). *Experimental Designs*. 2nd edn. New York: John Wiley and Sons, Inc.

Cohen, R. A., and B. Bloom. (2010). Access to and utilization of medical care for young adults aged 20-29 years: United States, 2008. *National Center for Health Statistics Data Brief No. 29* 1-8.

Coifman, R., and D. Donoho. (1995). Translation-invariant wavelet denoising. In A. Antoniadis and G. Oppenheim (Eds), *Wavelets and Statistics*, 125–150. New York: Springer-Verlag.

Cole, A. F. W., and M. Katz. (1966). Summer ozone concentrations in southern Ontario in relation to photochemical aspects and vegetation damage. *Journal of the Air Pollution Control Association* **16**, 201–206.

Comroe, J. H. Jr., S. Y. Botelho, and A. B. DuBois. (1959). Design of a body plethysmograph for studying cardiopulmonary physiology. *Journal of Applied Physiology* **14**, 439–444.

Conover, W. J. (1999). *Practical Nonparametric Statistics*, 3rd edn. New York: John Wiley and Sons, Inc.

Cooper, L. M., E. Schubot, S. A. Banford, and C. T. Tart. (1967). A further attempt to modify hypnotic susceptibility through repeated individualized experience. *International Journal of Clinical and Experimental Hypnosis* **15**, 118–124.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. *Proceedings of the 3rd Berkeley Symposium*, **4**: 135–138.

Cox, D. D. (1984). Multivariate smoothing spline functions. *SIAM Journal on Numerical Analysis* **21** pp. 789–813.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Cox, D. R., and D. Oakes. (1984). *Analysis of Survival Data*. London: Chapman & Hall.

Craig, A. T. (1932). On the distributions of certain statistics. *American Journal of Mathematics* **54**, 353–366.

Craigmile, P. F., and D. B. Percival. (2005). Asymptotic decorrelation of between-scale wavelet coefficients. *Information Theory, IEEE Transactions on*, **51**(3), 1039–1048.

Critchlow, D. E., and M. A. Fligner. (1991). On distribution-free multiple comparisons in the one-way analysis of variance. *Communications in Statistics: Theory and Methods* **20**, 127–139.

Crouse, C. F. (1966). Distribution free tests based on the sample distribution function. *Biometrika* **53**, 99–108.

Crowder, M. J., A. C. Kimber, R. L. Smith, and T. J. Sweeting. (1991). *Statistical Analysis of Reliability Data*. London: Chapman & Hall.

Cruess, D. F. (1989). Review of use of statistics in The American Journal of Tropical Medicine and Hygiene for January-December 1988. *American Journal of Tropical Medicine and Hygiene* **41**, 619–626.

Csörgõ, M., and R. Zitikis. (1996). Mean residual life processes. *Annals of Statistics* **24**, 1717–1739.

Csörgõ, S. (1988). Estimation in the proportional hazards model of random censorship. *Statistics* **19**, 437–463.

Csörgõ, S., and J. J. Faraway. (1998). The paradoxical nature of the proportional hazards model of random censorship. *Statistics* **31**, 61–78.

Csörgõ, S., and L. Horváth. (1986). Confidence bands from censored samples. *Canadian Journal of Statistics* **14**, 131–144.

Cuevas, J., and E. Chicken. (2012). A trimmed translation-invariant denoising estimator. *Journal of Statistical Computation and Simulation* **82**, 1299–1310.

Dalal, S. R. (1979). Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. *Stochastic Processes and their Applications* **9**, 99–107.

Dale, M. (1984). *Personal communication for report in Statistics 661*. Columbus: Ohio State University.

Dallal, G. E., and L. Wilkinson. (1986). An analytical approximation to the distribution of Lilliefors' test statistic. *American Statistician* **40**, 294–296.

Daly, D. A., and E. B. Cooper. (1967). Rate of stuttering adaptation under two electro-shock conditions. *Behaviour Research and Therapy* **5**, 49–54.

Damico, J. A., and D. A. Wolfe. (1987). Extended tables of the exact distribution of a rank statistic for all treatments multiple comparisons in one-way layout designs. *Communications in Statistics: Theory and Methods* **16**, 2343–2360.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM.

David, H. A. (1981). *Order Statistics,* 2nd edn. New York: John Wiley and Sons, Inc.

David, H. A., and H. N. Nagaraja. (2003). *Order Statistics*, 3rd ed. Hoboken, NJ: John Wiley and Sons.

Davis, C. E., and D. Quade. (1978). U-statistics for skewness or symmetry. *Communications in Statistics: Theory and Methods* **7**, 413–418.

Davison, A. C., and D. V. Hinkley. (1997). *Bootstrap Methods and Their Application*. New York: Cambridge University Press.

DeKroon, J., and P. Van der Laan. (1981). Distribution-free test procedures in two-way layouts; a concept of rank interaction. *Statistica Neerlandica* **35**, 189–213.

Dell, T. R., and J. L. Clutter. (1972). Ranked set sampling theory with order statistics background. *Biometrics* **28**, 545–555.

Delse, F. C., and B. W. Feather. (1968). The effect of augmented sensory feedback on the control of salivation. *Psychophysiology* **5**, 15–21.

Deshpande, J. V. (1983). A class of tests for exponentiality against increasing failure rate average alternatives. *Biometrika* **70**, 514–518.

Devore, J. L. (1991). *Probability and Statistics for Engineering and the Sciences* (3rd ed.). Belmont, CA: Wadsworth, Inc.

DiCiccio, T. J., and B. Efron. (1996). Bootstrap confidence intervals. *Statistical Science* **11**, 189–228.

Dickinson, M. B., F. E. Putz, and C. D. Canham. (1993). Canopy gap closure in thickets of the clonal shrub, Cornus racemosa, *Bulletin of the Torrey Botanical Club* **120**, 439–444.

Diehr, P., J. Yergan, J. Chu, P. Feigl, G. Glaefke, R. Moe, M. Bergner, and J. Rodenbaugh. (1989). Treatment modality and quality differences for black and white breast cancer patients treated in community hospitals. *Medical Care* **27**, 942–958.

Dietz, E. J. (1989). Teaching regression in a nonparametric statistics course. *American Statistician* **43**, 35–40.

Dixon V. Margolis. (1991). 56 FEP Cases (N. D. Illionois).

Doksum, K. (1967). Robust procedures for some linear models with one observation per cell. *Annals of Mathematical Statistics* **38**, 878–883.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *Annals of Statistics* **2**, 267–277.

Doksum, K. A., and G. L. Sievers. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika* **63**, 421–434.

Doksum, K. A., and B. S. Yandell. (1984). Tests for exponentiality. In P. R. Krishnaiah and P. K. Sen (Eds), *Handbook of Statistics*, Volume **4**, Nonparametric Methods, pp. 579–611. Amsterdam: North Holland.

Donner, A., and W. W. Hauck. (1986). The large-sample relative efficiency of the Mantel-Haenszel estimator in the fixed-strata case. *Biometrics* **42**, 537–545.

Donoho, D., and I. Johnstone. (1994). Ideal spatial adaptation via wavelet shrinkage. *Bionietrika* **81**, 425–455.

Donoho, D., and I. Johnstone. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200–1224.

Donoho, D. L., I. M. Johnstone, G. Kerkyacharian, and D. Picard. (1996). Density estimation by wavelet thresholding. *The Annals of Statistics* **24** pp. 508–539.

Doss, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Annals of Statistics* **22**, 1763–1786.

Doss, H., and R. D. Gill. (1992). An elementary approach to weak convergence for quantile processes, with applications to survival data. *Journal of the American Statistical Association* **87**, 869–877.

Dowdy, S., and S. Wearden. (1991), *2E. Statistics for Research*. New York: John Wiley and Sons, Inc.

Draper, D. (1988). Rank-based robust analysis of linear models. I. Exposition and review. *Statistical Science* **3**, 239–271.

DuBois, A. B., S. Y. Botelho, G. M. Bedell, R. Marshall, and J. H. Comroe Jr. (1956). A rapid plethysmographic method for measuring thoracic gas volume: A comparison with a nitrogen washout method for measuring functional residual capacity in normal subjects. *Journal of Clinical Investigation* **35**, 322–326.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics* **6**, 241–252.

Dunnett, C. W. (1964). New tables for multiple comparisons with a control. *Biometrics* **20**, 482–491.

Durbin, J. (1951). Incomplete blocks in ranking experiments. *British Journal of Mathematical and Statistical Psychology* **4**, 85–90.

Durbin, J. (1975). Kolmogorov-Smirnov tests when parameters are estimated with applications to tests of exponentiality and tests on spacings. *Biometrika* **62**, 5–22.

Dwass, M. (1960). Some k-sample rank-order tests. In I. Olkin, S. G. Ghurye, H. Hoeffding, W. G. Madow, and H. B. Mann (Eds), *Contributions to Probability and Statistics*, pp. 198–202. Stanford, CA: Stanford University Press.

Dykstra, R. L., and P. Laud. (1981). A Bayesian nonparametric approach to reliability. *Annals of Statistics* **9**, 356–367.

Easton, E. (2006). Personal communication.

Edwards, A. L. (1948). Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika* **13**, 185–187.

Edwards, A. W. F. (1963). The measure of association in a 2×2 table. *Journal of the Royal Statistical Society, Series A* **126**, 109–114.

Efron, B. (1967). The two-sample problem with censored data. *Proceedings of the 5th Berkeley Symposium* **4**, 831–854.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58**, 403–417.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Efron, B. (1982). *The Jackknife, the Bootstrap and other Resampling Plans*, Volume 38 of CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia, PA: SIAM.

Efron, B., and G. Gong. (1983). A leisurely look at the bootstrap, the jackknife and cross-validation. *American Statistician* **37**, 36–48.

Efron, B., and R. J. Tibshirani. (1994). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Ehlers, A. (1995). A 1-year prospective study of panic attacks: Clinical course and factors associated with maintenance. *Journal of Abnormal Psychology* **104**, 164–172.

Elmore, R.T., T.P. Hettmansperger and F. Xuan. (2004). The sign statistic, one-way layouts and mixture models. *Statistical Science* **19**, 579–587.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and its Applications* **14**, 153–158.

Epstein, B. (1960a). Tests for the validity of the assumption that the underlying distribution of life is exponential, I. *Technometrics* **2**, 83–101.

Epstein, B. (1960b). Tests for the validity of the assumption that the underlying distribution of life is exponential, II. *Technometrics* **2**, 167–183.

Eriksen, L., S. Björnstad, and K. G. Götestam. (1986). Social skills training in groups for alcoholics: One-year treatment outcome for groups and individuals. *Addictive Behaviors* **11**, 309–329.

Ernst, M.D. (2004). Permutation methods: A basis for exact inference. *Statistical Science* **19**, 676–685.

Eubank, R. L. (1999). *Nonparametric regression and Spline Smoothing 2E*. New York: Dekker.

Ezekiel, M. (1930). *Methods of Correlation Analysis*. Wiley, New York.

Fairbanks, D. J., and B. Rytting. (2001). Mendelian controversies: A botanical and historical review. *American Journal of Botany* **88**, 737–752.

Falkner, B., G. Onesti, T. Moshang, Jr., and D. T. Lowenthal. (1981). Growth hormone release in hypertensive adolescents treated with clonidine. *Journal of Clinical Pharmacology* **21**, 31–36.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American statistical Association* **87**, 998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics* **21**, 196–216.

Fan, J. (1996). Tests of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association* **91**, 674–688.

Fan, J., and J. S. Marron. (1993). Comment (discussion of Hastie and Loader). *Statistical Science* **8**, 129–134.

Faraway, J. J., and M. Jhun. (1990). Bootstrap choice of bandwidth for density estimation. *Journal of the American Statistical Association* **85** pp. 1119–1122.

Feather, B. W., and D. T. Wells. (1966). Effects of concurrent motor activity on the unconditioned salivary reflex. *Psychophysiology* **2**, 338–343.

Featherston, D. W. (1971). *Taenia hydatigena* II. Evagination of cysticerci and establishment in dogs. *Experimental Parasitology* **29**, 242–249.

Feller, W. (1948). On the Kolmogorov-Smirnov limit theorems for empirical distributions. *Annals of Mathematical Statistics* **19**, 177–189.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Finney, D. J., R. Latscha, B. M. Bennet, and P. Hsu. (1963). *Tables for Testing Significance in a.* 2×2 *Contingency table*. New York: Cambridge University Press.

Fisher, R. A. (1925), (1934), (1970). *Statistical Methods for Research Workers* (originally published 1925, 14th edition 1970). Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935). The logic of inductive inference (with discussion). *Journal of the Royal Statistical Society* **98**, 39–82.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Fisz, M. (1963). *Probability Theory and Mathematical Statistics,* 3rd edn. New York: John Wiley and Sons, Inc.

Fitzpatrick, S., and A. Scott. (1987). Quick simultaneous confidence intervals for multinomial proportions. *Journal of the American Statistical Association* **82**, 875–878.

Fleiss, J. L. (2003). *Statistical Methods for Rates and Proportions,* 3rd edn. New York: John Wiley and Sons, Inc.

Fleming, T. R., and D. P. Harrington. (1991). *Counting Processes and Survival Analysis*. New York: John Wiley and Sons, Inc.

Fleming, T. R., J. R. O'Fallon, P. C. O'Brien, and D. P. Harrington. (1980). Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics* **36**, 607–625.

Fligner, M. A. (1984). A note on two-sided distribution-free treatment versus control multiple comparisons. *Journal of the American Statistical Association* **79**, 208–211.

Fligner, M. A. (1985). Pairwise versus joint ranking: Another look at the Kruskal-Wallis statistic. *Biometrika* **72**, 705–709.

Fligner, M. A., and S. N. MacEachern. (2006). Nonparametric two-sample methods for ranked-set sample data. *Journal of the American Statistical Association* **101**, 1107–1108.

Fligner, M. A., and G. E. Policello II. (1981). Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association* **76**, 162–174.

Fligner, M. A., and S. W. Rust. (1983). On the independence problem and Kendall's tau. *Communications in Statistics: Theory and Methods* **12**, 1597–1607.

Fligner, M. A., and D. A. Wolfe. (1982). Distribution-free tests for comparing several treatments with a control. *Statistica Neerlandica* **36**, 119–127.

Flores, A. M., and L. R. Zohman. (1970). Energy cost of bedmaking to the cardiac patient and the nurse. *American Journal of Nursing* **70**, 1264–1267.

Forsman, A., and L. E. Lindell. (1993). The advantage of a big head: swallowing performance in adders, Vipera Berus. *Functional Ecology* **7**, 183–189.

Fox, J. R., and J. E. Randall. (1970). Relationship between forearm tremor and the biceps electromyogram. *Journal of Applied Physiology* **29**, 103–108.

Freedman, D., and P. Diaconis. (1981). On the histogram as a density estimator: L2 theory. *Probability Theory and Related Fields* **57**, 453–476.

Freund, R. J., and W. J. Wilson. (2010). In D. Mohr (Ed.), *Statistical Methods*, rev. ed. 2nd author. New York: Academic Press. 3E.

Frey, J. C. (2007a). New imperfect rankings models for ranked set sampling. *Journal of Statistical Planning and Inference* **137**, 1433–1445.

Frey, J. C. (2007b). Intentionally representative sampling for estimating a population mean. *Journal of Probability and Statistical Science* **5**, 103–112.

Frey, J., and Ö. Özturk. (2011). Constrained estimation using judgment post-stratification. *Annals of the Institute of Statistical Mathematics* **63**, 769–789.

Friedman, J. (1984). A variable span smoother. Tech. Rep. LCS-05, Stanford University Department of Statistics.

Friedman, J. (1991). Multivariate adaptive regression splines. *The Annals of Statistics* **19**, 1–67.

Friedman, J., and W. Stuetzle. (1981). Projection pursuit regression. *Journal of the American statistical Association* **76**, 817–823.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* **32**, 675–701.

Friedman, M., S. O. Byers, R. H. Rosenman, and R. Neuman. (1971). Coronary-prone individuals (Type A behavior pattern) growth hormone responses. *Journal of the American Medical Association* **217**, 929–932.

Friedman, M., and R. H. Rosenman. (1959). Association of specific overt behavior pattern with blood and cardiovascular findings: Blood cholesterol level, blood clotting time, incidence of arcus senilis, and clinical coronary artery disease. *Journal of the American Medical Association* **169**, 1286–1296.

Gabriel, K. R. (1969). Simultaneous test procedures—some theory of multiple comparisons. *Annals of Mathematical Statistics* **40**, 224–250.

Gail, M., and J. J. Gart. (1973). The determination of sample sizes for the use with the exact conditional test in 2×2 comparative trials. *Biometrics* **29**, 441–448.

Gamerman, D. (1991). Dynamic Bayesian models for survival data. *Applied Statistics* **40**, 63–79.

Gámiz, M.D., K.B. Kulasekera, N. Limnios and B.H. Lindqvist. (2011). *Applied Nonparametric Statistics in Reliability*. New York: Springer.

Gao, J., and Ö. Özturk. (2012). Two-sample distribution-free inference based on partially rank-ordered set samples. *Statistics & Probability Letters* **82**, 876–884.

Gart, J. J., and J. R. Zweiful. (1967). On the bias of various estimators of the logit and its variance

with applications to quantal bioassay. *Biometrika* **54**, 181–187.

Gasser, T., and H.-G. Müller. (1979). Kernel estimation of regression functions. In T. Gasser and M. Rosenblatt eds., *Smoothing Techniques for Curve Estimation, vol. 757 of Lecture Notes in Mathematics*, 23–68. Berlin: Springer.

Gastwirth, J. L., and H. Rubin. (1969). *The Behavior of Robust Estimators on Dependent Data*, Mimeo Ser. 197, Department of Statistics, Purdue University.

Gebhardt, A. (2009). *ash: David Scott's ASH routines*. R package version 1.0-12.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 203–223.

Gelfand, A. E., and A. F. M. Smith. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gemayel, N. M., E. A. Stasny, J. A. Tackett, and D. A. Wolfe. (2011). Ranked set sampling: an auditing application. *Review of Quantitative Finance and Accounting* Online First.

Gemayel, N. M., E. A. Stasny, and D. A. Wolfe. (2010). Optimal ranked set sampling estimation based on medians from multiple set sizes. *Journal of Nonparametric Statistics* **22**, 517–527.

Gentry, J., and J. Pike. (1970). An empirical study of the risk-return hypothesis using common stock portfolios of life insurance companies. *Journal of Financial and Quantitative Analysis* **5**, 179–185.

Gerstein, H. H. (1965). Lake Michigan pollution and Chicago's supply. *Journal American Water Works Association* **57**, 841–857.

Gerstein, H. H. (1997). *Nonparametric Methods for Quantitative Analysis,* 3rd edn. Syracuse, NY: American Sciences Press.

Gibbons, J. D., and S. Chakraborti. (2010). *Nonparametric Statistical Inference*, 5th edn. New York: Chapman and Hall.

Gilks, W. R., S. Richardson, and D. G. Spiegelhalter. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Gill, R. D. (1980). *Censoring and Stochastic Integrals*, Mathematical Centre Tracts 124. Amsterdam: Mathematisch Centrum.

Gill, R. D. (1983). Large sample behaviour of the product-limit estimator on the whole line. *Annals of Statistics* **11**, 49–58.

Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part 1). *Scandinavian Journal of Statistics* **16**, 97–128.

Gillespie, M. J., and L. Fisher. (1979). Confidence bands for the Kaplan-Meier survival curve estimate. *Annals of Statistics* **7**, 920–924.

Gleser, L. J. (1996). Comment on "Bootstrap confidence intervals" by T. J. DiCiccio and B. Efron. *Statistical Science* **11**, 219–221.

Goldsmith, J. R., and J. A. Nadel. (1969). Experimental exposure of human subjects to ozone. *Journal of the Air Pollution Control Association* **19**, 329–330.

Golub, G. H., M. Heath, and G. Wahba. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **11** pp. 215–223.

Goode, D. J., and H. Y. Meltzer. (1976). Effects of isometric exercise on serum creatine phosphokinase activity. *Archives of General Psychiatry* **33**, 1207–1211.

Goodman, L. A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics* **7**, 247–254.

Goodman, L. A., and W. H. Kruskal. (1963). Measures of association for cross classifications III. Approximate sampling theory. *Journal of the American Statistical Association* **58**, 310–364.

Gottlieb, G. (1965). Prenatal auditory sensitivity in chickens and ducks. *Science* **147**, 1596–1598.

Govindarajulu, Z. (1968). Distribution-free confidence bounds for $P(X < Y)$. *Annals of the Institute of Statistical Mathematics* **20**, 229–238.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **57**, 97–109.

Greenberg, V. L. (1966). Robust estimation in incomplete block designs. *Annals of Mathematical Statistics* **37**, 1331–1337.

Greenwood, M. (1926). The Natural duration of cancer. Reports on Public Health and Medical Subjects **33**, pp. 1–26. London: Her Majesty's Stationery Office.

Gregory, P. B. (1974). Personal communication.

Grenander, U. (1956). On the theory of mortality measure, Part II. *Skan Aktuarietidskr* **39**, 125–153.

Gripenberg, G. (1992). Confidence intervals for partial rank correlations. *Journal of the American Statistical Association* **87**, 546–551.

Gross, S. (1966). Nonparametric tests when nuisance parameters are present. PhD dissertation. Berkeley: University of California.

Guess, F. (1984). Testing whether mean residual life changes trend. PhD dissertation. Florida State University.

Guess, F., M. Hollander, and F. Proschan. (1986). Testing exponentiality versus a trend change in mean residual life. *Annals of Statistics* **14**, 1388–1398.

Gulati, S., and W. J. Padgett. (1996). Families of smooth confidence bands for the survival function under the general random censorship model. *Lifetime Data Analysis* **2**, 349–362.

Gupta, M. K. (1967). An asymptotically nonparametric test of symmetry. *Annals of Mathematical Statistics* **38**, 849–866.

Gurland, J., and J. Sethuraman. (1995). How pooling failure data may reverse increasing failure rates. *Journal of the American Statistical Association* **90**, 1416–1423.

Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen* **69**, 331–371.

Haber, M. (1986). An exact unconditional test for the 2×2 comparative trial. *Psychol. Bull.* **99**, 129–132.

Haber, M. (1987). A comparison of some conditional and unconditional exact tests for 2 by 2 contingency tables. *Communications in Statistics - Simulation and Computation* **16**, 999–1013.

Habib, M. G., and D. R. Thomas. (1986). Chi-squared goodness-of-fit tests for randomly censored data. *Annals of Statistics* **14**, 759–765.

Haciomeroglu, E., and E. Chicken. (2011). Investigating relations between ability, preference and calculus performance. In T. Lambert and L. Weist (Eds), *Proceedings of the 33rd annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Reno, NV: University of Nevada.

Hájek, J., and Z. Šidák. (1967). *Theory of Rank Tests*. New York: Academic Press.

Haldane, J. B. S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. *Ann. Human Genet.* **20**, 309–311.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Hall, P., and K.-H. Kang. (2001). Bootstrapping nonparametric density estimators with empirically chosen bandwidths. *The Annals of Statistics* **29** pp. 1443–1468.

Hall, P., G. Kerkyacharian, and D. Picard. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Annals of Statistics* **26**, 922–942.

Hall, P., G. Kerkyacharian, and D. Picard. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statistica Sinica* **9**, 33–50.

Hall, P., and M. Wand. (1996). On the accuracy of binned kernel density estimators. *Journal of Multivariate Analysis* **56**, 165 – 184.

Hall, W. J., and J. A. Wellner. (1979). Estimation of mean residual life. Technical Report, Department of Statistics. Rochester, New York: University of Rochester.

Hall, W. J., and J. A. Wellner. (1980). Confidence bands for a survival curve from censored data. *Biometrika* **67**, 133–143.

Hall, W. J., and J. A. Wellner. (1981). Mean residual life. In M. Csörgõ, D. A. Dawson, J. N. K. Rao, and A. K.Md.E. Saleh. *Statistics and Related Topics*, pp. 169–184. Amsterdam: North Holland.

Hallin, M., and D. Paindaveine. (2004). Multivariate signed-rank tests in vector autoregressive order identification. *Stat. Sci.* **19**, 697–711.

Halperin, M., P. R. Gilbert, and J. M. Lachin. (1987). Distribution-free confidence intervals for $\Pr(X_1 < X_2)$. *Biometrics* **43**, 71–80.

Hamilton. M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry* **23**, 56–62.

Hannum, R., and M. Hollander. (1983). Robustness of Ferguson's Bayes estimator of a distribution function. *Annals of Statistics* **11**, 632–639.

Härdle, W. (1992). *Applied Nonparametric Regression*. London: Cambridge University Press.

Härdle, W., and M. Müller. (2000). Multivariate and semiparametric kernel regression, In M. G. Schimek ed., *Smoothing and Regression: Approaches, Computation, and Application* 357–392. New York: John Wiley and Sons.

Harrell, F. E. Jr. (2012). Hmisc: Harrell Miscellaneous. R package version 3.9–3.

Harrington, D. P., and T. R. Fleming. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 553–566.

Harter, H. L. (1961). Expected values of normal order statistics. *Biometrika* **48**, 151–165.

Hartigan, J. A. (1969). Using subsample values as typical values. *Journal of the American Statistical Association* **64**, 1303–1317.

Hartigan, J. A. (1971). Error analysis by replaced samples. *Journal of the Royal Statistical Society, Series B* **33**, 98–110.

Hartigan, J. A. (1975). Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. *Annals of Statistics* **3**, 573–580.

Hastie, T., and C. Loader. (1993). Local regression: automatic kernel carpentry. *Statistical Science* **8**, 120–129 (discussion: 129–143).

Hastie, T., and R. Tibshirani. (1986). Generalized additive models. *Statistical Science* **1**, 297–310.

Hastie, T., and R. Tibshirani. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association* **82**, 371–386.

Hastie, T., and R. Tibshirani. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Hauck, W. W. (1979). The large-sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics* **35**, 817–819.

Hauck, W. W., and A. Donner. (1988). The asymptotic relative efficiency of the Mantel-Haenszel estimator in the increasing-number-of-strata case. *Biometrics* **44**, 379–384.

Hawkins, D. L., S. Kochar, and C. Loader. (1992). Testing exponentiality against IDMRL distributions with unknown change point. *Annals of Statistics* **20**, 280–290.

Hayfield, T., and J. S. Racine. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software* **27**(5), 1–32.

Hays, W. L. (1960). A note on average tau as a measure of concordance. *Journal of the American Statistical Association* **55**, 331–341.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparison procedure is conservative. *Annals of Statistics* **112**, 61–75.

Hayter, A. J. (2013). Simultaneous confidence intervals for several quantiles of an unknown distribution. (Submitted for publication).

Hayter, A. J., and G. Stone. (1991). Distribution free multiple comparisons for monotonically ordered treatment effects. *Australian Journal of Statistics* **33**, 335–346.

Hebb, D. O., and K. Williams. (1946). A method of rating animal intelligence. *Journal of General Psychology* **34**, 59–65.

Hettmansperger, T. P. (1984). *Statistical Inference Based on Ranks*. New York: John Wiley and Sons, Inc.

Hettmansperger. T. P., and J. W. McKean. (1977). A robust alternative based on ranks to least squares in analyzing linear models. *Technometrics* **19**, 275–284.

Hettmansperger, T. P., J. W. McKean, and S. J. Sheather. (1997). In G. S. Maddala and C. R. Rao (Eds), *Rank-based Analyses of Linear Models*, *Handbook of Statistics*, Volume **15**. Amsterdam: Elsevier Science.

Hettmansperger, T. P., and R. M. Norton. (1987). Tests for patterned alternatives in k-sample problems. *Journal of the American Statistical Association* **82**, 292–299.

Hilgard, E. R., L. W. Lauer, and A. H. Morgan. (1963). *Manual for Stanford Profile Scales of Hypnotic Susceptibility, Forms I and II*. Palo Alto, CA: Consulting Psychologist Press.

Hilton, J. F., and L. Gee. (1997a). The size and power of the exact bivariate symmetry test. *Computational Statistics & Data Analysis* **26**, 53–69.

Hilton, J. F., and L. Gee. (1997b). An exact Hollander bivariate symmetry test. Algorithm A5 321. *Journal of the Royal Statistical Society Series C* **46**, 533–540.

Hjort, N. L. (1990a). Goodness of fit tests in models for life history data based on cumulative hazard rates. *Annals of Statistics* **18**, 1221–1258.

Hjort, N. L. (1990b). Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics* **18**, 1259–1294.

Hochberg, Y., and A. C. Tamhane. (1987). *Multiple Comparison Procedures*. New York: John Wiley and Sons, Inc.

Hodges, J. L. Jr., and E. L. Lehmann. (1950). Some problems in minimax point estimation. *Annals of Mathematical Statistics* **21**, 182–197.

Hodges, J. L. Jr., and E. L. Lehmann. (1956). The efficiency of some nonparametric competitors of the t-test, *Annals of Mathematical Statistics* **27**, 324–335.

Hodges, J. L. Jr., and E. L. Lehmann. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics* **34**, 598–611.

Hodges, J. L. Jr., and E. L. Lehmann. (1967). On medians and quasimedians. *Journal of the American Statistical Association* **62**, 926–931.

Hodges, J. L. Jr., and E. L. Lehmann. (1970). Deficiency. *Annals of Mathematical Statistics* **41**, 783–801.

Hodges, J. L. Jr., and E. L. Lehmann. (1983). Hodges-Lehmann estimators. In S. Kotz, N. L. Johnson, and C. B. Read (Eds), *Encyclopedia of Statistical Sciences,* Volume **3**, pp. 463–465. New York: John Wiley and Sons, Inc.

Hoeffding, W. (1948a). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **19**, 293–325.

Hoeffding, W. (1948b). A non-parametric test of independence. *Annals of Mathematical Statistics* **19**, 546–557.

Hoeffding, W. (1951). "Optimum" nonparametric tests. *Proceedings of 2nd Berkeley Symposium*, Berkeley, pp. 83–92.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics* **23**, 169–192.

Hollander, M. (1966). An asymptotically distribution-free multiple comparison procedure—treatments versus control. *Annals of Mathematical Statistics* **37**, 735–738.

Hollander, M. (1967a). Rank tests for randomized blocks when the alternatives have an a priori ordering. *Annals of Mathematical Statistics* **38**, 867–877.

Hollander, M. (1967b). Asymptotic efficiency of two nonparametric competitors of Wilcoxon's two sample test. *Journal of the American Statistical Association* **62**, 939–949.

Hollander, M. (1971). A nonparametric test for bivariate symmetry. *Biometrika* **58**, 203–212.

Hollander, M. (1996). Personal communication.

Hollander, M., and R. M. Korwar. (1976). Nonparametric Bayes estimation of the probability that $X \leq Y$. *Communications in Statistics - Theory and Methods* **14**, 1369–1383.

Hollander, M., and R. M. Korwar. (1982). Nonparametric Bayesian estimation of the horizontal distance between two populations. In *Nonparametric Statistical Inference I*. New York: North Holland, pp. 409–415.

Hollander, M., I. W. McKeague, and J. Yang. (1997). Likelihood ratio-based confidence bands for survival functions. *Journal of the American Statistical Association* **92**, 215–226.

Hollander, M., and E. Peña. (1988). Nonparametric test under restricted treatment-assignment rules. *Journal of the American Statistical Association* **83**, 1144–1151.

Hollander, M., and E. Peña. (1989). Families of confidence bands for the survival function under the general random censorship model and the Koziol-Green model. *Canadian Journal of Statistics* **17**, 59–74.

Hollander, M., and E. Peña. (1992a). A chi-squared goodness-of-fit test for randomly censored data. *Journal of the American Statistical Association* **87**, 458–463.

Hollander, M., and E. Peña. (1992b). Classes of nonparametric goodness-of-fit tests for censored data: Simple null hypothesis case. In A. K. Md. E. Saleh (Eds), *Nonparametric Statistics and Related Topics*, pp. 97–118. Amsterdam: North-Holland.

Hollander, M., and E. A. Peña. (2004). Nonparametric methods in reliability. *Statistical Science* **19**, 644–651.

Hollander, M., G. Pledger, and P. Lin. (1974). Robustness of the Wilcoxon test to a certain dependency between samples. *Annals of Statistics* **2**, 177–181.

Hollander, M., and F. Proschan. (1972). Testing whether new is better than used. *Annals of Mathematical Statistics* **43**, 1136–1146.

Hollander, M., and F. Proschan. (1975). Tests for the mean residual life. *Biometrika* **62**, 585–593.

Hollander, M., and F. Proschan. (1979). Testing to determine the underlying distribution using randomly censored data. *Biometrics* **35**, 393–401.

Hollander, M., and F. Proschan. (1984). Nonparametric concepts and methods in reliability. In P. R. Krishnaiah and P. K. Sen (Eds), *Handbook of Statistics*, Nonparametric Methods, Volume **4**, pp. 613–655. Amsterdam: North-Holland.

Hollander, M., F. Proschan, and J. Sconing. (1985). Efficiency loss with the Kaplan-Meier estimator. Technical Report M707, Department of Statistics, Tallahassee: Florida State University.

Hollander, M., and J. Sethuraman. (2001). Nonparametric methods: Advanced computational approaches. In N. J. Smelser and P. B. Baltes (Eds), *International Encyclopedia of the Social & Behavioral Sciences*, pp. 10673–10680. Oxford: Permagon.

Hollander, M., and J. Sethuraman. (2002). Nonparametric inference for repair models. *Sankhya* **64**, 693–706.

Hotelling, H., and M. R. Pabst. (1936). Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics* **7**, 29–43.

Høyland, A. (1965). Robustness of the Hodges-Lehmann estimates for shift. *Annals of Mathematical Statistics* **36**, 174–197.

Høyland, A. (1968). Robustness of the Wilcoxon estimate of location against a certain dependence. *Annals of Mathematical Statistics* **39**, 1196–1201.

Hsieh, F. Y. (1987). A simple method of sample size calculation for unequal-sample-size designs that use the logrank or t-test. *Statistics in Medicine* **6**, 577–581.

Hsieh, F. Y. (1992). Comparing sample size formulae for trials with unbalanced allocation using the logrank test. *Statistics in Medicine* **11**, 1091–1098.

Huber, P. J., and E. M. Ronchetti. (2009). *Robust Statistics*. New York: John Wiley and Sons, Inc.

Hundal, P. S. (1969). Knowledge of performance as an incentive in repetitive industrial work. *Journal of Applied Psychology* **53**, 224–226.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **60**, 271–293.

Hurvich, C. M., and C.-L. Tsai. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Ijzermans, A. B. (1970). Pitting corrosion and intergranular attack of austenitic Cr-Ni stainless steels in Na SCN. *Corrosion Science* **10**, 607–615.

Iman, R. L. (1994). *A Data-Based Approach to Statistics*. Belmont, CA: Duxbury Press.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. *Annals of Mathematical Statistics* **43**, 1449–1458.

Jamison, H. H. (1971). Development of a gaseous oxygen impact testing method. *Materials Research and Standards* **11**, 22–27.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A* **186**, 453–461.

Jin, J., and J. Shi. (2001). Automatic feature extraction of waveform signals for in-process diagnostic performance improvement. *Journal of Intelligent Manufacturing* **12**(3), 257–268.

Joe, H. (1990). Multivariate concordance. *Journal of Multivariate Analysis* **35**, 12–30.

Johansen, S. (1978). The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics* **5**, 195–199.

Johnson, A. A, K. Mukherjee, S. Schlosser, and E. Raask. (1970). The behaviour of a cenosphere-resin composite under hydrostatic pressure. *Ocean Engineering* **2**, 45–48.

Johnson, B. (1984). *Personal communication for report in Statistics 661*. Columbus: Ohio State University.

Johnson, B. M. (1973). Decision making, faculty satisfaction, and the place of the School of Nursing in the university. *Nursing Research* **22**, 100–107.

Johnson, N. L. (1974). Estimation of rank order. Report 931. Univ. of North Carolina Institute of Statistics.

Johnson, R. A. (1988). Stress-strength models for reliability. In P. K. Krishnaiah and C. R. Rao (Eds), *Handbook of Statistics,* Volume **7**, pp. 27–54. New York: North Holland.

Johnson, S. K., and R. E. Johnson. (1972). Tonsillectomy history in Hodgkin's disease. *New England Journal of Medicine* **287**, 1122–1125.

Jonckheere, A. R. (1954a). A distribution-free k-sample test against ordered alternatives. *Biometrika* **41**, 133–145.

Jonckheere, A. R. (1954b). A test of significance for the relation between m rankings and k ranked categories. *British Journal of Statistical Psychology* **7**, 93–100.

Jones, M. P., T. W. O'Gorman, J. H. Lemke, and R. F. Woolson. (1989). A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. *Biometrics* **45**, 171–181.

Jung, D. H., and A. C. Parekh. (1970). A semimicromethod for the determination of serum iron and iron-binding capacity without deproteinization. *American Journal of Clinical Pathology* **54**, 813–817.

Kalbfleisch, J. D., and R. L. Prentice, (1980, 2002). *The Statistical Analysis of Failure Time Data*. New York: John Wiley and Sons, Inc.

Kaneto, A., K. Kosaka, and K. Nakao. (1967). Effects of stimulation of the Vagus nerve on insulin Secretion. *Endocrinology* **80**, 530–536.

Kaplan, E. L., and P. Meier. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

Kaplan, H. S., and S. A. Rosenberg. (1973). Personal communication.

Karpatkin, M., R. F. Porges, and S. Karpatkin. (1981). Platelet counts in infants of women with autoimmune thrombocytopenia. *New England Journal of Medicine* **305**, 936–939.

Kayle, K. A. (1984). *Personal communication for report in Statistics 661*. Columbus: Ohio State University.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* **30**, 81–93.

Kendall, M. G. (1962). *Rank Correlation Methods*, 3rd edn. London: Griffin.

Kendall, M. G., and J. D. Gibbons. (1990). *Rank Correlation Methods*, 5th edn. London: Arnold.

Kepner, J. L., and R. H. Randles. (1984). Comparison of test for bivariate symmetry versus location and/or scale alternatives. *Communications in Statistics: Theory and Methods* **13**, 915–930.

Kershenobich, D., F. J. Fierro, and M. Rojkind. (1970). The relationship between the free pool of proline and collagen content in human liver cirrhosis. *Journal of Clinical Investigation* **49**, 2246–2249.

Kiefer, J., and J. Wolfowitz. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Annals of Mathematical Statistics* **27**, 887–906.

Kim, D. H., and D. H. Lim. (1995). Rank tests for parallelism of regression lines against umbrella alternatives. *Journal of Nonparametric Statistics* **5**, 289–302.

Klefsjö, B. (1983). Some tests against aging based on the total time on test transform. *Communications in Statistics: Theory and Methods* **12**, 907–927.

Klein, J. P., and M. L. Moeschberger. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, 2nd edn. New York: Springer-Verlag.

Kloke, J., and J. McKean. (2011). Rfit: Rank Estimation for Linear Models. R package version 0.14.

Klotz, J. (1963). Small sample power and efficiency for the one-sample Wilcoxon and normal scores tests. *Annals of Mathematical Statistics* **34**, 624–632.

Klotz, J. (1964). On the normal scores two-sample rank test. *Journal of the American Statistical Association* **49**, 652–664.

Klotz, J. (1967). Asymptotic efficiency of the two sample Kolmogorov-Smirnov test. *Journal of the American Statistical Association* **62**, 932–938.

Koenker, R. (2011). *quantreg: Quantile Regression*. R package version 4.62.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari* **4**, 83–91.

Kolmogorov, A. N. (1941). Confidence limits for an unknown distribution function. *Annals of Mathematical Statistics* **12**, 461–483.

Konijn, H. S. (1956). On the power of certain tests for independence in bivariate populations. *Annals of Mathematical Statistics* **27**, 300–323. Correction: **29**(1958), 935–936.

Kontula, K., L. C. Andersson, T. Paavonen, G. Myllyla, L. Teerenhovi, and P. Vuopio. (1980). Glucocorticoid receptors and glucocorticoid sensitivity of human leukemic cells. *International Journal of Cancer* **26**, 177–183.

Kontula, K., T. Paavonen, P. Vuopio, and L. C. Andersson. (1982). Glucocorticoid receptors in hairy-cell leukemia. *International Journal of Cancer* **30**, 423–426.

Koul, H. L. (1977). A test for new is better than used. *Communications in Statistics: Theory and Methods* **6**, 563–573.

Koul, H. L. (1978a). A Class of tests for testing "new is better than used." *Canadian Journal of Statistics* **6**, 249–271.

Koul, H. L. (1978b). Testing for new is better than used in expectation. *Communications in Statistics: Theory and Methods* **7**, 685–701.

Koul, H. L., G. L. Sievers, and J. W. McKean. (1987). An estimator of the scale parameter for the rank analysis of linear models under general score functions. *Scandinavian Journal of Statistics* **14**, 131–141.

Kovac, A., and B. W. Silverman. (2000). Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of the American Statistical Association* **95**, 172–183.

Koziol, J. A. (1979). A test for bivariate symmetry based on the empirical distribution function. *Communications in Statistics: Theory and Methods* **8**, 207–221.

Koziol, J. A., and S. B. Green. (1976). A Cramër-von Mises statistic for randomly censored data. *Biometrika* **63**, 465–474.

Koziol, J. A., and S. B. Green. (1978). Personal communication.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics* **12**, 307–310.

Kramer, C. Y. (1957). Extension of multiple range tests to group correlated adjusted means. *Biometrics* **13**, 13–18.

Kronmal, R., and M. Tarter. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *Journal of the American Statistical Association* **63**, 925–952.

Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *Annals of Mathematical Statistics* **23**, 525–540.

Kruskal, W. H. (1957). Historical notes on the Wilcoxon unpaired two-sample test. *Journal of the American Statistical Association* **52**, 356–360.

Kruskal, W. H., and W. A. Wallis. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* **47**, 583–621.

Kuczmarski, R. J., M. D. Carol, K. M. Flegal, and R. P. Troiano. (1997). Varying body mass index cutoff points to describe overweight prevalence among U. S. adults: NHANES III (1988 to 1994). *Obesity Research* **5**, 542–548.

Kuehl, R. O. (1994). *Statistical Principles of Research Design and Analysis*. Belmont, CA: Duxbury Press.

Kurtz, T. E., R. F. Link, J. W. Tukey, and D. L. Wallace. (1965). Short-cut multiple comparisons for balanced single and double classification, Part I, results. *Technometrics* **7**, 95–161.

Kvam, P. H., and F. J. Samaniego. (1993). On the inadmissibility of empirical, averages as estimators in ranked set sampling. *Journal of Statistical Planning and Inference* **36**, 39–55.

Kvam, P. H., and F. J. Samaniego. (1994). Nonparametric maximum likelihood estimation based on ranked set samples. *Journal of the American Statistical Association* **89**, 526–537.

Lakatos, E., and G. Lan. (1992). A comparison of sample size methods for the logrank statistic. *Statistics in Medicine* **11**, 179–191.

Lamp, W. O. (1976). Statistical treatment of a study on the distribution of a stream insect by age. Master's thesis, Ohio State University.

Langenberg, P., and R. Srinivasan. (1979). Null distribution of the Hollander-Proschan statistic for decreasing mean residual life. *Biometrika* **66**, 679–680.

Laplace, P. S. (1812). Théorie Analytique des Probabilités. Paris: Courcier.

Latta, R. (1981). A Monte Carlo study of some two-sample rank tests with censored data. *Journal of the American Statistical Association* **76**, 713–729.

Leach, S. P. (1972). Personal communication.

Leach, S. P. (1979). *Introduction to Statistics. A Nonparametric Approach for the Social Sciences*. Chichester: John Wiley and Sons, Inc.

Leaf, D. A., W. E. Connor, R. Illingworth, S. P. Bacon, and G. Sexton. (1989). The hypolipidemic effects of gemfibrozil in type V hyperlipidemia. *Journal of the American Medical Association* **262**, 3154–3160.

LeCam, L., and J. Neyman (Eds). (1965). *Bernoulli Bayes Laplace Anniversary Volume*. New York: Springer-Verlag.

Lee, S. C. S., C. Locke, and J. D. Spurrier. (1980). On a class of tests of exponentiality. *Technometrics* **22**, 547–554.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *Annals of Mathematical Statistics* **22**, 165–179.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. New York: John Wiley and Sons, Inc.

Lehmann, E. L. (1963a). Robust estimation in analysis of variance. *Annals of Mathematical Statistics* **34**, 957–966.

Lehmann, E. L. (1963b). Asymptotically nonparametric inference: An alternative approach to linear models. *Annals of Mathematical Statistics* **34**, 1494–1506.

Lehmann, E. L. (1963c). Nonparametric confidence intervals for a shift parameter. *Annals of Mathematical Statistics* **34**, 1507–1512.

Lehmann, E. L. (1964). Asymptotically nonparametric inference in some linear models with one observation per cell. *Annals of Mathematical Statistics* **35**, 726–734.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco, CA: Holden-Day.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd edn. New York: John Wiley and Sons, Inc.

Lehmann, E. L., and H. J. M. D'Abrera. (2006). *Nonparametrics: Statistical Methods Based on Ranks,* 2nd edn. New York: Springer.

Lepage, Y. (1971). A combination of Wilcoxon's and Ansari-Bradley's statistics. *Biometrika* **58**, 213–217.

Lepage, Y. (1973). A table for a combined Wilcoxon Ansari-Bradley statistic. *Biometrika* **60**, 113–116.

Leurgans, S. (1983). Three classes of censored data rank tests: strengths and weaknesses under censoring. *Biometrika* **70**, 651–658.

Leurgans, S. (1984). Asymptotic behavior of two-sample rank tests in the presence of random censoring. *Annals of Statistics* **12**, 572–589.

Levin, A. (2011). No U.S. airline fatalities in 2010. January 21, 2011 USA TODAY.

Li, G., and H. Doss. (1993). Generalized Pearson-Fisher chi-square goodness-of-fit tests, with application to models with life history data. *Annals of Statistics* **21**, 772–797.

Li, G., M. Hollander, I. W. McKeague, and J. Yi. (1996). Nonparametric likelihood ratio confidence bands for quantile functions from incomplete survival data. *Annals of Statistics* **24**, 628–640.

Li, J., and R. Y. Liu. (2004). New nonparametric tests of multivariate locations and scales using data depth. *Statistical Science* **19**, 686–696.

Li, Q., and J. Racine. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica* **14**, 485–512.

Liang, K. Y., and S. G. Self. (1985). Tests for homogeneity of odds ratio when the data are sparse. *Biometrika* **72**, 353–358.

Lilliefors, H. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* **62**, 399–402.

Lilliefors, H. (1969). On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association* **64**, 387–389.

Lim, D. H., and D. A. Wolfe. (1997). Nonparametric comparisons of several regression lines with a control. *Far East Journal of Theoretical Statistics* **1**, 51–61.

Livesey, P. J. (1967). The Hebb-Williams elevated pathway test: A comparative study of rat, rabbit and cat performance. *Australian Journal of Psychology* **19**, 55–62.

Lloyd, S. J., K. D. Garlid, R. C. Reba, and A. E. Seeds. (1969). Permeability of different layers of the human placenta to isotopic water. *Journal of Applied Physiology* **26**, 274–276.

Lo, A. (1982). Bayesian nonparametric statistical inference for shock models. *Scandinavian Journal of Statistics* **8**, 237–242.

Lo, A. (1993). A Bayesian bootstrap for censored data. *Annals of Statistics* **21**, 100–123.

Loftsgaarden, D. O., and C. P. Quesenberry. (1965). A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics* **36**, 1049–1051.

Low, P. P., C. K. Luk, M. J. Dulfano, and P. J. P. Finch. (1984). Ciliary beat frequency of human respiratory tract by different sampling techniques. *American Review of Respiratory Disease* **130**, 497–498.

Lu, H. H. S., M. T. Wells, and R. C. Tiwari. (1994). Inference for shift functions in the two-sample problem with right-censored data: With applications. *Journal of the American Statistical Association* **89**, 1017–1026.

MacEachern, S. N., Ö. Özturk, G. V. Stark, and D. A. Wolfe. (2002). A new ranked set sample estimator of variance. *Journal of the Royal Statistical Society, Series B* **64**, 177–188.

MacEachern, S. N., E. A. Stasny, and D. A. Wolfe. (2004). Judgement post-stratification with imprecise rankings. *Biometrics* **60**, 207–215.

Mack, G. A. (1981). A quick and easy distribution-free test for main effects in a two-factor ANOVA. *Communications in Statistics: Simulation and Computation* **10**, 571–591.

Mack, G. A., and J. H. Skillings. (1980). A Friedman-type rank test for main effects in a two-factor ANOVA. *Journal of the American Statistical Association* **75**, 947–951.

Mack, G. A., and D. A. Wolfe. (1981). *K*-sample rank tests for umbrella alternatives. *Journal of the American Statistical Association* **76**, 175–181.

Maesono, Y. (1996). Higher order comparisons of jackknife variance estimators. *Journal of Nonparametric Statistics* **7**, 35–45.

Mallat, S. (1989a). Multiresolution approximations and wavelet orthonormal bases of $L^2(R)$. *Transactions of the American Mathematical Society* **315**, 69–89.

Mallat, S. (1989b). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674–693.

Mallat, S. (2009). *A Wavelet Tour of Signal Processing* (3rd ed.). Burlington, MA: Academic Press.

Manly, B. F. J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd edn. Boca Raton, FL: Chapman & Hall.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica* **13**, 245–259.

Mann, H. B., and D. R. Whitney. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals Of Mathematical Statistics* **18**, 50–60.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association* **58**, 690–700.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**, 113–170.

Mantel, N. (1985). Propriety of the Mantel-Haenszel variance for the log rank test. *Biometrika* **72**, 471–472.

Mantel, N., and W. Haenszel. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**, 719–748.

March, G. L., T. M. John, B. A. McKeown, L. Sileo, and J. C. George. (1976). The effects of lead poisoning on various plasma constituents in the Canada goose. *Journal of Wildlife Diseases* **12**, 14–19.

Marden, J. I. (2004). Positions and QQ plots. *Statistical Science* **19**, 606–614.

Marron, J. S. (1994). Visual understanding of higher-order kernels. *Journal of Computational and Graphical Statistics* **3**, 447–458.

Marshall, A. W., and F. Proschan. (1965). Maximum likelihood estimation for distributions with monotone failure rate. *Annals of Mathematical Statistics* **36**, 69–77.

Mather, M. (1984). *Personal communication for report in Statistics 661*. Columbus: Ohio State University.

Mauldin, R. D., W. D. Sudderth, and S. C. Williams. (1992). Pólya trees and random distributions. *Annals of Statistics* **20**, 1203–1221.

Maxson, S. J. (1977). Activity patterns of female ruffed grouse during the breeding season. *Wilson Bulletin* **89**, 439–455.

McClave, J. T., and G. Benson. (1978). *Statistics for Business and Economics*. San Francisco, CA: Dellen.

McDonald, B. J., and W. A. Thompson Jr. (1967). Rank sum multiple comparisons in one- and two-way classifications. *Biometrika* **54**, 487–497.

McGee, D. (2010). Personal communication.

McGinnity, K., E. Chicken, and J. J. Pignatiello Jr. (2013). Distribution-free changepoint detection for nonlinear profiles. In *Proceedings of the 2013 Industrial and Systems Engineering Research Conference*, A. Krishnamurthy and W. K. V. Chan eds., 3142–3151.

McIntyre, G. A. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research* **3**, 385–390.

McIntyre, G. A. (2005). A method for unbiased selective sampling, using ranked sets. *American Statistician* **59**, 230–232.

McKean, J.W. (2004). Robust analysis of linear models. *Statistical Science* **19**, 562–570.

McKean, J. W., and T. P. Hettmansperger. (1976). Tests of hypotheses of the general linear model based on ranks. *Communications in Statistics: Theory and Methods* **5**, 693–709.

McKean, J. W., and T. A. Ryan Jr. (1977). An algorithm for obtaining confidence intervals and point estimates based on ranks in the two-sample location problem. *Transactions on Mathematical Software* **3**, 183–185.

McKean, J. W., and S. J. Sheather. (1991). Small sample properties of robust analyses of linear models based on R-estimates: A survey. In W. Stahel and S. Weisberg (Eds), *Directions in Robust Statistics and Diagnostics, Part II,* pp. 1–19. New York: Springer-Verlag.

McLain, A. C., and S. K. Ghosh (2011). Nonparametric estimation of the conditional mean residual life function with censored data *Lifetime Data Analysis* **17**, 514–532.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157.

Mehta, C. R., and J. F. Hilton. (1993). Exact power of conditional and unconditional tests: going beyond the 2×2 contingency table. *American Statistician* **47**, 91–98.

Mehta, C. R., and N. R. Patel. (1986a). A hybrid algorithm for Fisher's exact test on unordered $r \times c$ contingency tables. *Communications in Statistics: Theory and Methods* **15**, 387–403.

Mehta, C. R., and N. R. Patel. (1986b). Algorithm 643: FEXACT: A Fortran subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. *ACM Transactions on Mathematical Software* **12**, 154–161.

Mehta, C. R., N. R. Patel, and R. Gray. (1985). Computing an exact confidence interval for the common odds ratio in several 2 by 2 contingency tables. *Journal of the American Statistical Association* **80**, 969–973.

Mehta, C. R., N. R. Patel, and L. J. Wei. (1988). Computing exact significance tests with restricted randomization rules. *Biometrika* **75**, 295–302.

Meier, P. (1975). Estimation of a distribution function from incomplete observations. In J. Gani (Ed.), *Perspectives in Probability and Statistics*, pp. 67–87. Sheffield: Applied Probability Trust.

Mendenhall, W. (1968). *Introduction to Linear Models and the Design and Analysis of Experiments*. Belmont, CA: Wadsworth.

Mendis, K. N., R. I. Ihalamulla, and P. H. David. (1988). Diversity of plasmodium vivax-induced antigens on the surface of infected human erythrocytes. *American Journal of Tropical Medicine and Hygiene* **38**, 42–46.

Merline, J. W. (1991). Will more money improve education? *Journal of Consumer Research* **74**, 26–27.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Mi, J. (1994). Estimation related to mean residual life. *Journal of Nonparametric Statistics* **4**, 179–190.

Miller, L. E. (1956). Table of percentage points of Kolmogorov statistics. *Journal of the American Statistical Association* **51**, 111–121.

Miller, R. G. Jr. (1964). A trustworthy jackknife. *Annals of Mathematical Statistics* **35**, 1594–1605.

Miller, R. G. Jr. (1966). *Simultaneous Statistical Inference*. New York: McGraw-Hill.

Miller, R. G. Jr. (1968). Jackknifing variances. *Annals of Mathematical Statistics* **38**, 567–582.

Miller, R. G. Jr. (1974). The jackknife: A review. *Biometrika* **61**, 1–15.

Miller, R. G. Jr. (1980). Combining 2×2 contingency tables. In R. G. Miller Jr., B. Efron, B. Wm. Brown Jr., and L. E. Moses (Eds), pp. 73–83. *Biostatistics Casebook*. New York: John Wiley and Sons, Inc.

Miller, R. G. Jr. (1981a). *Simultaneous Statistical Inference*, 2nd edn. New York: Springer-Verlag.

Miller, R. G. Jr. (1981b). *Survival Analysis*. New York: John Wiley and Sons, Inc.

Miller, R. G. Jr. (1983). What price Kaplan-Meier? *Biometrics* **39**, 1077–1082.

Miller, R. G. Jr. (1998). *Survival Analysis*. Wiley Classics Library. New York: John Wiley and Sons, Inc.

Milton, R. C. (1970). *Rank Order Probabilities*. New York: John Wiley and Sons, Inc.

Mittal, Y. (1991). Homogeneity of subpopulations and Simpson's paradox. *Journal of the American Statistical Association* **86**, 167–172.

Moeschberger, M. L., and J. P. Klein. (1985). A comparison of several methods of estimating the survival function when there is extreme censoring. *Biometrics* **41**, 253–259.

Molitor, F. T. (1989). Children's acceptance of others' fighting after viewing violent TV and playing cooperatively. MA thesis, California State University.

Moore, W., and C. I. Bliss. (1942). A method for determining insecticidal effectiveness using *Aphis rumicis* and certain organic compounds. *Journal of Economic Entomology* **35**, 544–553.

Mortazavi, B. (1997). Personal communication.

Morton, R. (1978). Regression analysis of life tables and related nonparametric tests. *Biometrika* **65**, 329–333.

Moses, L. E. (1963). Rank tests of dispersion. *Annals of Mathematical Statistics* **34**, 973–983.

Moses, L. E. (1964). One sample limits of some two-sample rank tests. *Journal of the American Statistical Association* **59**, 645–651.

Mosteller, F. (1952). Some statistical procedures in measuring the subjective response to drugs. *Biometrics* **8**, 220–226.

Müller, H. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *The Annals of Statistics* **12**, 766–774.

Müller, H. G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*, Lecture Notes in Statistics, Volume **46**. New York: Springer-Verlag.

Murphy, S. (1995). Likelihood ratio based confidence intervals in survival analysis. *Journal of the American Statistical Association* **90**, 1399–1405.

Murray, R. A., M. S. Ridout, and J. V. Cross. (2000). The use of ranked set sampling in spray deposit assessment. *Aspects of Applied Biology* **57**, 141–146.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability & its Applications* **9**, 141–142.

Nadaraya, E. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability & its Applications* **10**, 186–190.

Nahhas, R. W., D. A. Wolfe, and H. Chen. (2002). Ranked set sampling: Cost and optimal set size. *Biometrics* **58**, 964–971.

Nair, V. (1981). Plots and tests for goodness of fit with randomly censored data. *Biometrika* **68**, 99–103.

Nair, V. (1984). Confidence bands for survival functions with censored data: A comparative study. *Technometrics* **26**, 265–275.

Nason, G. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society, Series B-Methodological* **58**, 463–479.

Nason, G. (2008). *Wavelet Methods in Statistics with R*. New York: Springer-Verlag.

Nason, G. (2010). *wavethresh: Wavelets statistics and transforms*. R package version 4.5.

Nason, G., and B. Silverman. (1995). The stationary wavelet transform and some statistical applications. In A. Antoniadis and G. Oppenheim (Eds), *Wavelets and Statistics*, pp. 281–300. New York: Springer-Verlag.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology* **1**, 25–72.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics* **14**, 945–965.

Nemenyi, P. (1963). Distribution-free multiple comparisons. PhD dissertation, Princeton University.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London, Series A* **236**, 333–380.

Ng'andu, N. H. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's Model. *Statistics in Medicine* **16**, 611–626.

Nicholls, G. H., and D. Ling. (1982). Cued speech and the reception of spoken language. *Journal of Speech and Hearing Research* **25**, 262–269.

Noether, G. E. (1955). On a theorem of Pitman. *Annals of Mathematical Statistics* **26**, 64–68.

Noether, G. E. (1963). Note on the Kolmorogov statistic in the discrete case. *Metrika* **7**, 115–116.

Noether, G. E. (1967a). *Elements of Nonparametric Statistics*. New York: John Wiley and Sons, Inc.

Noether, G. E. (1967b). Wilcoxon confidence intervals for location parameters in the discrete case. *Journal of the American Statistical Association* **62**, 184–188.

Noether, G. E. (1987). Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association* **82**, 645–647.

Norman, R. D. (1966). A revised formula for the Wechsler adult intelligence scale. *Journal of Clinical Psychology* **22**, 287–294.

Nussbaum, B. D., and B. K. Sinha. (1997). Cost effective gasoline sampling using ranked set sampling. In *American Statistical Association 1997 Proceedings of the Section on Statistics and the Environment*, 83–87. American Statistical Association.

Obenchain, R. L. (1969). Rank tests invariant only under linear transformations, Mimeo Series 617, Institute of Statistics, University of North Carolina.

Odeh, R. E. (1977). Extended tables of the distributions of rank statistics for treatment versus control in randomized block designs. *Communications in Statistics: Simulation and Computation* **6**, 101–113.

Oja, H., and R.H. Randles. (2004). Multivariate nonparametric tests. *Statistical Science* **19**, 598–605.

Oppenheim, R. W. (1968). Light responsivity in chick and duck embryos just prior to hatching. *Animal Behaviour* **16**, 276–280.

Owen, D. B. (1962). *Handbook of Statistical Tables*. Reading, MA: Addison-Wesley.

Owen, D. B., K. J. Craswell, and D. L. Hanson. (1964). Nonparametric upper confidence bounds for $P\{Y < X\}$ and confidence limits for $P\{Y < X\}$ when X and Y are normal. *Journal of the American Statistical Association* **59**, 906–924.

Özturk, Ö. (2008). Inference in the presence of ranking error in ranked set sampling. *Canadian Journal of Statistics* **36**, 1–18.

Özturk, Ö. (2010). Nonparametric maximum-likelihood estimation of within-set ranking errors in ranked set sampling. *Journal of Nonparametric Statistics* **22**, 823–840.

Özturk, Ö. (2011). Sampling from partially rank-ordered sets. *Environmental and Ecological Statistics* **18**, 757–779.

Özturk, Ö. (2012). Quantile inference based on partially rank-ordered set samples. *Journal of Statistical Planning and Inference* **142**, 2116–2127.

Özturk, Ö., Ö. C. Bilgin, and D. A. Wolfe. (2005). Estimation of population mean and variance in flock management: a ranked set sampling approach in a finite population setting. *Journal of Statistical Computation and Simulation* **75**, 905–919.

Özturk, Ö., and S. N. MacEachern. (2004). Order restricted randomized designs for control versus treatment comparison. *Annals of the Institute of Statistical Mathematics* **56**, 701–720.

Özturk, Ö., and S. N. MacEachern. (2007). Order restricted randomized designs and two-sample inference. *Environmental and Ecological Statistics* **14**, 365–381.

Özturk, Ö., and D. A. Wolfe. (2000). Optimal allocation procedure in ranked set sampling for unimodal and multi-modal distributions. *Environmental and Ecological Statistics* **7**, 343–356.

Pagan, A., and A. Ullah. (1999). *Nonparametric Econometrics*. Cambridge: Cambridge University Press.

Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association* **58**, 216–230.

Pan, G. (1996). Distribution-free confidence procedure for umbrella orderings. *Australian Journal of Statistics* **38**, 161–172.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33**, 1065–1076.

Parzen, E. (2004). Quantile probability and statistical data modeling. *Statistical Science* **19**, 652–662.

Patil, G. P. (1995). Editorial: Ranked set sampling. *Environmental and Ecological Statistics* **2**, 271–285.

Patil, G. P., A. K. Sinha, and C. Taillie. (1995). Finite population corrections for ranked set sampling. *Annals of the Institute of Statistical Mathematics* **47**, 621–636.

Patil, G. P., A. K. Sinha, and C. Taillie. (1999). Ranked set sampling: a bibliography. *Environmental and Ecological Statistics* **6**, 91–98.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika* **23**, 114–133.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen

from random sampling, *Philosophical Magazine Series 5* **50**, 157–175.

Pearson, K. (1911). On the probability that two independent distributions of frequency are really samples from the same population. *Biometrika* **8**, 250–254.

Pensky, M., and B. Vidakovic. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *Annals of Statistics* **27**, 2033–2053.

Percival, D. B., and A. T. Walden. (2000). *Wavelet Methods for Time Series Analysis (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press.

Peterson, A. V. (1977). Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *Journal of the American Statistical Association* **90**, 1399–1405.

Peto, R., and J. Peto. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Association* **72**, 854–858.

Pettitt, A. N., and V. Siskind. (1981). Effect of within-sample dependence on the Mann-Whitney-Wilcoxon statistic. *Biometrika* **68**, 437–441.

Pettitt, A. N., and M. A. Stephens. (1977). The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics* **19**, 205–210.

Phadke, C. P., S. S. Wu, F. J. Thompson, and A. L. Behrman. (2006). Soleus H-reflex modulation in response to change in percentage of leg loading in standing after incomplete spinal cord injury. *Neuroscience Letters* **403**, 6–10.

Pires, A. M., and J. A. Branco. (2010). A statistical model to explain the Mendel-Fisher controversy. *Statistical Science* **25**, 545–565.

Pitman, E. J. G. (1948). Notes on non-parametric statistical inference. Columbia University (duplicated).

Platt, W. J., G. W. Evans, and S. L. Rathbun. (1988). The population dynamics of a long-lived conifer *(Pinus palustris)*. *The American Naturalist* **131**, 491–525.

Poland, A., D. Smith, R. Kuntzman, M. Jacobson, and A. H. Conney. (1970). Effect of intensive occupational exposure to DDT on phenylbutazone and cortisol metabolism in human subjects. *Clinical Pharmacology & Therapeutics* **11**, 724–732.

Portnoy, S., and R. Koenker. (1989). Adaptive l-estimation for linear models. *The Annals of Statistics* **17** pp. 362–381.

Potthoff, R. F. (1974). A non-parametric test of whether two simple regression lines are parallel. *Annals of Statistics* **2**, 295–310.

Pratt, J. W. (1959). Remarks on zeros and ties in the Wilcoxon signed rank procedures. *Journal of the American Statistical Association* **54**, 655–667.

Pratt, J. W. (1964). Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association* **59**, 665–680.

Prentice, M. J. (1979). On the problem of m incomplete rankings. *Biometrika* **66**, 167–170.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65**, 167–179.

Presnell, B., and L. L. Bohn. (1999). U-statistics and imperfect ranking in ranked set sampling. *Journal of Nonparametric Statistics* **10**, 111–126.

Priestley, M., and M. Chao. (1972). Non-parametric function fitting. *Journal of the Royal Statistical Society, Series B-Methodological* **34**, 385–392.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics* **5**, 375–383.

Puri, M. L. (1965). Some distribution-free k-sample rank tests of homogeneity against ordered alternatives. *Communications on Pure and Applied Mathematics* **18**, 51–63.

Puri, M. L., and P. K. Sen. (1968). On Chernoff-Savage tests for ordered alternatives in randomized blocks. *Annals of Mathematical Statistics* **39**, 967–972.

Puri, M. L., and P. K. Sen. (1971). *Nonparametric Methods in Multivariate Analysis*. New York: John Wiley and Sons, Inc.

Putt, M. E., and V. M. Chinchilli. (2004). Nonparametric approaches to the analysis of crossover studies. *Statistical Science* **19**, 712–719.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society, Series B* **11**, 68–84.

Racine, J., and Q. Li. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* **119**, 99–130.

Ramachandramurty, P. V. (1966a). On some nonparametric estimates for shift in the Behrens-Fisher situation. *Annals of Mathematical Statistics* **37**, 593–610.

Ramachandramurty, P. V. (1966b). On the Pitman efficiency of one-sided Kolmogorov and Smirnov tests for normal alternatives. *Annals of Mathematical Statistics* **37**, 940–944.

Ramsay, W. N. M. (1957). The determination of iron in blood plasma or serum. *Clinica Chimica Acta* **2**, 214–220.

Randles, R. H., M. A. Fligner, G. E. Policello II, and D. A. Wolfe. (1980). An asymptotically distribution-free test for symmetry versus asymmetry. *Journal of the American Statistical Association* **75**, 168–172.

Randles, R. H., T. P. Hettmansperger, and G. Casella. (2004). Introduction to the special issue: Nonparametric statistics. *Statistical Science* **19**, 561.

Randles, R. H., and R. V. Hogg. (1971). Certain uncorrelated and independent rank statistics. *Journal of the American Statistical Association* **66**, 569–574.

Randles, R. H., and D. A. Wolfe. (1979). *Introduction to the Theory of Nonparametric Statistics*. New York: John Wiley and Sons, Inc. (Randles and Wolfe (1979) was reprinted in 1991 by Krieger Publishing Company.)

Rao, K. S. M., and A. P. Gore. (1984). Testing concurrence and parallelism of several sample regressions against ordered alternatives. *Mathematische Operationsforschung und Statistik, Series Statistics* **15**, 43–50.

Rasekh, J., A. Kramer, and R. Finch. (1970). Objective evaluation of canned tuna sensory quality. *Journal of Food Science* **35**, 417–423.

Reed, O. M. (1973). *Papio Cynocephalus* age determinations. *American Journal of Physical Anthropology* **38**, 309–314.

Rice, J. A. (1998). *Mathematical Statistics and Data Analysis*. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis* (3rd edn.). Belmont, CA: Thomson Brooks/Cole.

Risen, J. L., and T. Gilovich. (2008). Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology* **95**, 293–307.

Robertson, T., F. T. Wright, and R. L. Dykstra. (1988). *Order Restricted Statistical Inference*. New York: John Wiley and Sons, Inc.

Robins, J., N. Breslow, and S. Greenland. (1986). Estimators of the Mantel-Haenszel variance consistent in both spare data and large-strata limiting models. *Biometrics* **42**, 311–323.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* **27**, 832–837.

Ruppert, D., and M. Wand. (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics* **22**, 1346–1370.

Rust, S. W., and M. A. Fligner. (1984). A modification of the Kruskal-Wallis statistic for the generalized Behrens-Fisher problem. *Communications in Statistics: Theory and Methods* **13**, 2013–2028.

Ryan, T. P. (2009). *Modern Regression Methods*. New York: John Wiley and Sons, Inc.

Salsburg, D. S. (1970). Personal communication (with the cooperation of Pfizer and Co., Groton, Conn.).

Samara, B., and R. H. Randles. (1988). A test for correlation based on Kendall's tau. *Communications in Statistics: Theory and Methods* **17**, 3191–3205.

Santner, T. J. (1988). Teaching large-sample binomial confidence intervals. *Teaching Statistics* **20**, 20–23.

Sauber, S. R. (1971). Approaches to precounseling and therapy training: An investigation of its potential influence on process outcome. PhD dissertation, Florida State University.

Savage, I. R. (1953). Bibliography of nonparametric statistics and related topics. *Journal of the American Statistical Association* **48**, 844–906, Correction: (1958), **53**, 1031.

Savage, I. R. (1956). Contributions to the theory of rank order statistics—the two-sample case. *Annals of Mathematical Statistics* **27**, 590–615.

Savage, I. R. (1962). *Bibliography of Nonparametric Statistics*. Cambridge: Harvard University Press.

Savur, S. R. (1937). The use of the median in tests of significance. *Proceedings of the Indian Academy of Sciences - Section A* **5** (6), 564–576.

Saxena, K. M. L. (1969). Use of sign statistic in problems concerning $P(Y < X)$. *Abstract in Annals of Mathematical Statistics **40**, 1154.*

Scheffé, H. (1943). Statistical inference in the non-parametric case. *Annals of Mathematical Statistics* **14**, 305–332.

Scheffé, H. (1959). *The Analysis of Variance*. New York: John Wiley and Sons, Inc.

Scheffé, H., and J. W. Tukey. (1945). Nonparametric estimation. I Validation of order statistics. *Annals of Mathematical Statistics* **16**, 187–192.

Schoenberg, I. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. *Quarterly Of Applied Mathematics* **4**, 45–99.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68**, 316–319.

Schonrock, H. (1996). Personal communication.

Schucany, W.R. (2004). Kernel smoothers: An overview of curve estimators for the first graduate course in nonparametric statistics. *Statistical Science* **19**, 663–675.

Schuster, E. (1974). On the rate of convergence of an estimate of a functional of a probability density. *Scandinavian Actuarial Journal* **1**, 103–107.

Schweder, T. (1975). Window estimation of the asymptotic variance of rank estimators of location. *Scandinavian Journal of Statistics* **2**, 113–126.

Schweder, T. (1981). Correction note. *Scandinavian Journal of Statistics* **8**, 55.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika* **66**, 605–610.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Hoboken, NJ: John Wiley and Sons, Inc.

Scott, D. W., and G. R. Terrell. (1981). *Sequential Nonparametrics: Invariance Principles and Statistical Inference*. New York: John Wiley and Sons, Inc.

Scott, D. W., and G. R. Terrell. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association* **82**, 1131–1146.

Sen, P. K. (1967). A note on asymptotically distribution-free confidence bounds for $Pr(X < Y)$, based on two independent samples. *Sankhya A* **29**, 95–102.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* **63**, 1379–1389.

Sen, P. K. (1969). On a class of rank order tests for the parallelism of several regression lines. *Annals of Mathematical Statistics* **40**, 1668–1683.

Senchaudhuri, P., C. R. Mehta, and N. R. Patel. (1995). Estimating exact p values by the method of control variates or Monte Carlo rescue. *Journal of the American Statistical Association* **90**, 640–648.

Sengupta, J. M., I. M. Chakravarti, and D. Sarkar. (1951). Experimental survey for the estimation of cinchona yield. *Bulletin of the International Statistical Institute* **33**, 313–331.

Serfling, R. J. (1968). The Wilcoxon two-sample statistic on strongly mixing processes. *Annals of Mathematical Statistics* **39**, 1202–1209.

Serfling, R. J. (1984). Generalized L-, M-, and R-statistics. *Annals of Statistics* **12**, 76–86.

Serfling, R. J. (1992). Nonparametric confidence intervals for generalized quantile parameters in multi-sample contexts. In A. K. Md. E. Saleh (Ed.), *Nonparametric Statistics and Related Topics*, pp. 121–139. New York: North Holland.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Sethuraman, J., and M. Hollander. (2009). Nonparametric Bayesian estimation in repair models. *Journal of Statistical Planning and Inference* **139**, 1722–1733.

Sethuraman, J., and R. C. Tiwari. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In S. S. Gupta and J. O. Berger (Eds), *Statistical Decision Theory and Related Topics III (2)*, pp. 305–315. New York: Academic Press.

Shao, J. (1988). Consistency of jackknife estimators of the variances of sample quantiles. *Communications in Statistics: Theory and Methods* **17**, 3017–3028.

Shao, J., and C. F. J. Wu. (1989). A general theory for jackknife variance estimation. *Annals of Statistics* **17**, 1176–1197.

Sheather, S.M. (2004). Density estimation. *Statistical Science* **19**, 588–597.

Sheather, S. J., and M. C. Jones. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B-Methodological* **53**, 683–690.

Shelp, W. D., F. H. Bach, W. A. Kisken, M. Newton, R. E. Rieselbach, and A. B. Weinstein. (1970). Long-term integrity of renal function in cadaver allografts. *Journal of the American Medical Association* **213**, 1443–1447.

Shen, S., G. M. Reaven, J. W. Farquhar, and R. H. Nakanishi. (1970). Comparison of impedance to insulin-mediated glucose uptake in normal subjects and in subjects with latent diabetes. *Journal of Clinical Investigation* **49**, 2151–2160.

Shen, Z. Q., X. H. Feng, Z. X. Qian, R. L. Liu, and C. R. Yang. (1988). Application of biotinadvin system, determination of circulating immune complexes, and evaluation of antibody response in different hydatidosis patients. *American Journal of Tropical Medicine and Hygiene* **39**, 93–96.

Sherman, E. (1965). A note on multiple comparisons using rank sums. *Technometrics* **7**, 255–256.

Shlafer, M., and A. M. Karow Jr. (1971). Ultrastructure-function correlative studies for cardiac cryopreservation II. Hearts frozen to various temperatures without a cryoprotectant. *Cryobiology* **8**, 350–360.

Shorack, G. R. (1965). Nonparametric tests and estimation of scale in the two sample problem, Technical Report 10, Department of Statistics, Stanford University.

Shorack, G. R. (1969). Testing and estimating ratios of scale parameters. *Journal of the American Statistical Association* **64**, 999–1013.

Shorack, G. R., and J. A. Wellner. (1986). *Empirical Processes*. New York: John Wiley and Sons, Inc.

Siddiqui, M. M., and E. A. Gehan. (1966). *Statistical Methodology for Survival Time Studies*. Communication of National Cancer Institute.

Siegel, A. F. (1982), Robust regression using repeated medians. *Bionietrika* **69**(1), 242–244.

Sillitto. G. P. (1947). The distribution of Kendall's τ coefficient of rank correlation in rankings containing ties. *Biometrika* **34**, 36–40.

Silver, H., N. F. Colovos, J. B. Holter, and H. H. Hayes. (1969). Fasting metabolism of white-tailed deer. *Journal of Wildlife Management* **33**, 263–274.

Silverman, B. W. (1982). Algorithm as 176: Kernel density estimation using the fast Fourier transform. *Journal of the Royal Statistical Society, Series C, (Applied Statistics)* **31**, 93–99.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Simpson, D. G., and B. H. Margolin. (1986). Recursive nonparametric testing for dose-response relationships subject to downturns at high doses. *Biometrika* **73**, 589–596.

Skaug, H. J., and D. Tjøstheim. (1993). A nonparametric test of serial independence based on the empirical distribution function. *Biometrika* **80**, 591–602.

Skillings, J. H. (1980). On the null distribution of Jonckheere's statistic used in two-way models for ordered alternatives. *Technometrics* **22**, 431–436.

Skillings, J. H., and G. A. Mack. (1981). On the use of a Friedman-type statistic in balanced and unbalanced block designs. *Technometrics* **23**, 171–177.

Skillings, J. H., and D. A. Wolfe. (1977). Testing for ordered alternatives by combining independent distribution-free block statistics. *Communications in Statistics: Theory and Methods* **6**, 1453–1463.

Skillings, J. H., and D. A. Wolfe. (1978). Distribution-free tests for ordered alternatives in a randomized block design. *Journal of the American Statistical Association* **73**, 427–431.

Smid, L. J. (1956). On the distribution of the test statistics of Kendall and Wilcoxon when ties are present. *Statistica Neerlandica* **10**, 205–214.

Smirnov, N. V. (1935). Ueber die verteilung des allgemeinen gliedes in der variationsreihe. *Metron* **12**, 59–81.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *(Russian) Bulletin of Moscow University* **2**, 3–16.

Smirnov, N. V. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics* **19**, 279–281.

Smith, A. F. M., and G. O. Roberts. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* **55**, 3–23.

Smith, E. J. (1967). Cloud seeding experiments in Australia. Proceedings of the 5th Berkeley Symposium, Volume **V**, pp. 161–176.

Smith, J. D. (1969). Geomorphology of a sand ridge. *Journal of Geology* **77**, 39–55.

Smith, P. L. (1979). Splines as a useful and convenient statistical tool. *American Statistician* **33**, 57–62.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology* **15**, 72–101.

Spjøtvoll, E. (1968). A note on robust estimation in analysis of variance. *Annals of Mathematical Statistics* **39**, 1486–1492.

Sposto, R., and M. Krailo. (1987). Use of unequal allocation in survival trials. *Statistics in Medicine* **6**, 119–125.

Spurrier, J. D. (1991). Improved bounds for the moments of some rank statistics. *Communications in Statistics: Theory and Methods* **20**, 2603–2608.

Stanton, J. M. (1969). Murderers on parole. *Crime and Delinquency* **15**, 149–155.

StatXact-9 (2010). Cytel Software Corporation, Cambridge, MA.

Steel, R. G. D. (1959). A multiple comparison rank sum test. Treatments versus control. *Biometrics* **15**, 560–572.

Steel, R. G. D. (1960). A rank sum test for comparing all pairs of treatments. *Technometrics* **2**, 197–207.

Steel, R. G. D. (1961). Some rank sum multiple comparisons tests. *Biometrics* **17**, 539–552.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics* **9**, 1135–1151.

Stephens, M. A. (1974). EDF Statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* **69**, 730–735.

Stephens, M. A. (1979). Tests of fit for the logistic distribution based on the empirical distribution function. *Biometrika* **66**, 591–595.

Stephens, M. A. (1983a). Kolmogorov-Smirnov statistics. In S. Kotz, N. L. Johnson, and C. B. Read (Eds), *Encyclopedia of Statistical Sciences*, **Volume 4**, pp. 393–396. New York: John Wiley and Sons, Inc.

Stephens, M. A. (1983b). Kolmogorov-Smirnov-type tests of fit. In S. Kotz, N. L. Johnson, and C. B. Read (Eds), *Encyclopedia of Statistical Sciences*, **Volume 4**, pp. 398–402. New York: John Wiley and Sons, Inc.

Sternhell, S. (1958). Chemistry of brown coals VI: Further aspects of the chemistry of hydroxyl groups in Victorian brown coals. *Australian Journal of Applied Science* **9**, 375–379.

Stigler, S. M. (1974). Linear functions of order statistics with smooth weight functions. *Annals of Statistics* **2**, 676–693.

Stitt, J. T., J. D. Hardy, and E. R. Nadel. (1971). Surface area of the squirrel monkey in relation to body weight. *Journal of Applied Physiology* **31**, 140–141.

Stokes, S. L. (1980). Estimation of variance using judgment ordered ranked set samples. *Biometrics* **36**, 35–42.

Stokes, S. L., and T. W. Sager. (1988). Characterization of a ranked-set sample with application to estimating distribution functions. *Journal of the American Statistical Association* **83**, 374–381.

Stokes, L., X. Wang, and M. Chen. (2007). Judgement post-stratification with multiple rankers. *Journal of Statistical Theory and Applications* **6**, 344–359.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B-Methodological* **36**, 111–147.

Stone, C. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Annals of Statistics* **22**, 118–184.

Storer, B. E., and C. Kim. (1990). Exact properties of some exact tests for comparing two binomial proportions. *Journal of the American Statistical Association* **85**, 146–155.

Strawderman, R. L. (1997). An asymptotic analysis of the logrank test. *Lifetime Data Analysis* **3**, 225–249.

Strawderman, R. L., and C. R. Mehta. (1992). On the validation of exact tests for nonparametric inference. *Computational Statistics & Data Analysis* **14**, 263–266.

Stuart, A. (1954). The asymptotic relative efficiencies of tests and the derivatives of their power functions. *Skandinavisk Aktuarietidskrift* **37**, 163–169.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association* **21**, 65–66.

Susarla, V., and J. van Ryzin. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association* **71**, 897–902.

Sussia, S., and J. J. Shuster. (1985). Exact unconditional sample sizes for the 2 by 2 binomial trial. *Journal of the Royal Statistical Society, Series A* **148**, 317–327.

Switzer, P. (1976). Confidence procedures for two-sample problems. *Biometrika* **63**, 13–25.

Sylvester. P. E. (1969). Pyramidal, lemniscal and parietal lobe status in cerebral palsy. *Journal of Mental Deficiency Research* **13**, 20–33.

Tackett, J.A. (2012). Personal communication.

Takahasi, K. (1970). Practical note on estimation of population means based on samples stratified by means of ordering. *Annals of the Institute of Statistical Mathematics* **22**, 421–428.

Takahasi, K., and K. Wakimoto. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics* **20**, 1–31.

Tanner, M., and W. Wong. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association* **82**, 528–550.

Tarone, R. E., J. J. Gart, and W. W. Hauck. (1983). On the asymptotic relative efficiency of certain noniterative estimators of a common relative risk or odds ratio. *Biometrika* **70**, 519–522.

Tarone, R. E., and J. Ware. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156–160.

Tate, M. W., W. L. Cullinan, and A. Ahlstrand. (1961). Measurement of adaptation in stuttering. *Journal of Speech and Hearing Research* **4**, 321–339.

Terpstra, J. T. (2004). On estimating a population proportion via ranked set sampling. *Biometrical Journal* **46**, 264–272.

Terpstra, J. T., and L. A. Liudahl. (2004). Concomitant-based rank set sampling proportion estimates. *Statistics in Medicine* **23**, 2061–2070.

Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae* **14**, 327–333.

Terrell, G. R., and D. W. Scott. (1992). Variable kernel density estimation. *The Annals of Statistics* **20**, 1236–1265.

Terry, M. E. (1952). Some rank order tests which are most powerful against specific parametric alternatives. *Annals of Mathematical Statistics* **23**, 346–366.

Theil, H. (1950a). A rank-invariant method of linear and polynomial regression analysis, I. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen Series A* **53**, 386–392.

Theil, H. (1950b). A rank-invariant method of linear and polynomial regression analysis, II. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen Series A* **53**, 521–525.

Theil, H. (1950c). A rank-invariant method of linear and polynomial regression analysis, III. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen Series A* **53**, 1397–1412.

Thomas, D. R., and G. L. Grunkemeier. (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association* **70**, 865–871.

Thomas, H. V., and E. Simmons. (1969). Histamine content in sputum from allergic and nonallergic individuals. *Journal of Applied Physiology* **26**, 793–797.

Thomson, M. L., and M. D. Short. (1969). Mucociliary function in health, chronic obstructive airway disease, and asbestosis. *Journal of Applied Physiology* **26**, 535–539.

Thompson, W. R. (1936). On confidence ranges for the median and other expectation distributions for populations of unknown distribution form. *Annals of Mathematical Statistics* **7**, 122–128.

Tocher, K. D. (1950). Extension of the Neyman-Pearson theory of tests to discontinuous variates. *Biometrika* **37**, 130–144.

Tukey, J. W. (1949). The simplest signed-rank tests, Memo Report 17, Statistical Research Group, Princeton University.

Tukey, J. W. (1953). The problem of multiple comparisons. Unpublished manuscript.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Abstract in Annals of Mathematical Statistics* **29**, 614.

Tukey, J. W. (1962). Data analysis and behavioral science. Unpublished manuscript.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Tuyl, F., R. Gerlach, and K. Mengersen. (2008). A comparison of Bayes-Laplace, Jeffreys, and other priors: The case of zero events. *American Statistician* **62**, 40–44.

Ujah, J. E., and K. B. Adeoye. (1984). Effects of shelterbelts in the Sudan Savanna zone of Nigeria on microclimate and yield of millet. *Agricultural and Forest Meteorology* **33**, 99–107.

van Belle, G. (1972). Personal communication.

van Dantzig, D. (1951). On the consistency and the power of Wilcoxon's two sample test. *Indagationes Mathematicae* **13**, 1–8.

van der Vaart, A. W. (1988). *Statistical Estimation for Large Parameter Spaces*, CWI Tracts 44. Amsterdam: Centre for Mathematics and Computer Science.

van der Vaart, A. W. (1991). Efficiency and Hadamard differentiability. *Scandinavian Journal of Statistics* **18**, 63–75.

van der Waerden, B. L., and E. Nievergelt. (1956). *Tafeln zum Vergleich Zweier Stichproben-mittels X-test und Zeichentest*. Berlin: Springer-Verlag, OHG.

van Elteren, P., and G. E. Noether. (1959). The asymptotic efficiency of the X_r^2-test for a balanced incomplete block design. *Biometrika* **46**, 475–477.

Veterans Administration Cooperative Urological Research Group. (1967). Treatment and survival of patients with cancer of the prostate. *Surgery Gynecology & Obstetrics* **124**, 1011–1017.

Vianna, N. J., P. Greenwald, and J. M. P. Davies. (1971). Tonsillectomy and Hodgkin's disease: the Lymphoid Tissue barrier. *Lancet* **1**, 431–432.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. New York: Wiley.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, **36**(3), 1–48.

Wahba, G. (1990). *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series, **Volume 59**. Philadelphia, PA: SIAM.

Walker, H. M., and J. Lev. (1953). *Statistical Inference,* 1st. edn. New York: Holt, Rinehart & Winston.

Walsh, J. E. (1949). Some significance tests for the median which are valid under very general conditions. *Annals of Mathematical Statistics* **20**, 64–81.

Walsh, J. E. (1963). Bounded probability properties of Kolmogorov-Smirnov and similar statistics for discrete data. *Annals of the Institute of Statistical Mathematics* **15**, 153–158.

Wang, J. G. (1987). A note on the uniform consistency of the Kaplan-Meier estimator. *Annals of Statistics* **15**, 1313–1316.

Wang, J. L. (1987). Estimators of a distribution function with increasing failure rate average. *Journal of Statistical Planning and Inference* **16**, 415–427.

Wang, X., J. Lim, and L. Stokes. (2008). A nonparametric mean estimator for judgment poststratified data. *Biometrics* **64**, 355–363.

Wang, X., L. Stokes, J. Lim, and M. Chen. (2006). Concomitants of multivariate order statistics with application to judgment poststratification. *Journal of the American Statistical Association* **101**, 1693–1704.

Wang, X.-F. (2010). *fANCOVA: Nonparametric Analysis of Covariance*. R package version 0.5-1.

Wang, Y. (2011). *Smoothing Splines: Methods and Applications*. Boca Raton, FL: CRC Press.

Wang, Y-G, Z. Chen, and J. Liu. (2004). General ranked set sampling with cost considerations. *Biometrics* **60**, 556–561.

Wardlaw, A. C., and P. J. Moloney. (1961). The assay of insulin with anti-insulin and mouse diaphragm. *Canadian Journal of Biochemistry and Physiology* **39**, 695–712.

Wardlaw, A. C., and G. van Belle. (1964). Statistical aspects of the mouse diaphragm test for insulin. *Diabetes* **13**, 622–633.

Wasserman, L. (2006). *All of Nonparametric Statistics*. New York: Springer.

Watson, G. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A* **29**, 359–372.

Wegman, E. J., and I. W. Wright. (1983). Splines in statistics. *Journal of the American Statistical Association* **78**, 351–365.

Weindling, S. (1977). Personal communication to J. A. Rice–statistics report: Math 80B.

Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350–362.

Welch, B. L. (1947). The generalization of "Student's" problem when several different populations are involved. *Biometrika* **34**, 28–35.

Wellner, J. A. (1982). Asymptotic optimality of the product limit estimator. *Annals of Statistics* **10**, 595–602.

Wellner, J. A. (1985). A heavy censoring limit theorem for the product limit estimator. *Annals of Statistics* **13**, 150–162.

Wells, J. M., and M. A. Wells. (1967). Note on project SCUD. *Proceedings of the 5th Berkley Symposium*, **Volume V**, pp. 357–369.

Wells, M. T., and R. C. Tiwari. (1989). Bayesian quantile plots and statistical inference for nonlinear models in the two-sample case with incomplete data. *Communications in Statistics: Theory and Methods* **18**, 2955–2964.

West, M. (1992). Modelling time-varying hazards and covariate effect. In J. P. Klein and P. K. Goel (Eds), *Survival Analysis: State of the Art*, pp. 47–62. Boston, MA: Kluwer.

Wheeler, B. (2009). SuppDists: Supplementary distributions. R package version 1.8.

Whitaker, L. R., and F. J. Samaniego. (1989). Estimating a survival curve when new is better than used in expectation. *Naval Research Logistics* **36**, 693–707.

Whitcher, B. (2010). *waveslim: Basic wavelet routines for one-, two- and three-dimensional signal processing*. R package version 1.6.4.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1**, 80–83.

Wilcoxon, F., and R. A. Wilcox. (1964). *Some Rapid Approximate Statistical Procedures*, 2nd edn. Pearl River, NY: American Cyanamid Co., Lederle Laboratories.

Wilding, G. E., and G. S. Mudholkar (2008). Empirical approximations for Hoeffdings test of bivariate independence using two Weibull extensions. *Statistical Methodology* **5**(2), 160–170.

Wilks, S. S. (1962). *Mathematical Statistics*. New York: John Wiley and Sons, Inc.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**, 209–212.

Windham, B. M. (1971). Personal communication.

Wolfe, B. E., and J. D. Maser. (1994). Treatment of panic disorder: Consensus statement. In B. E. Wolf and J. D. Maser (Eds), *Treatment of Panic Disorder: A Consensus Development Conference*, pp. 237–255. Washington, DC: American Psychiatric Press.

Wolfe, D. A. (1977). A Distribution-free test for related correlation coefficients. *Technometrics* **19**, 507–509.

Wolfe, D. A. (2004). Ranked set sampling: An approach to more efficient data collection. *Statistical Science* **19**, 636–643.

Wolfe, D. A., and R. V. Hogg. (1971). On constructing statistics and reporting data. *American Statistician* **25**, 27–30.

Woodward, W. F. (1970). A comparison of base running methods in baseball. MSc thesis, Florida State University.

Wu, C. F. J. (1990). On the asymptotic properties of the jackknife histogram. *Annals of Statistics* **18**, 1438–1452.

Yanagimoto, T. (1970). On measures of association and a related problem. *Annals of the Institute of Statistical Mathematics* **22**, 57–63.

Yang, G. L. (1978). Estimation of a biometric function. *Annals of Statistics* **6**, 112–116.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society* **1**, 217–235.

Ying, Z. (1989). A note on the asymptotic properties of the product-limit estimator on the whole line. *Statistics & Probability Letters* **7**, 311–314.

Yu, C. S. (1971). Pitman efficiencies of Kolmogorov-Smirnov tests. *Annals of Mathematical Statistics* **42**, 1595–1605.

Yule, G. U. (1900). On the association of attributes in statistics. *Philosophical Transactions of the Royal Society of London Series A* **194**, 257–319.

Yule, G. U. (1912). On the methods of measuring association between two attributes (with discussion). *Journal of the Royal Statistical Society* **75**, 579–642.

Zacks, S. (1992). *Introduction to Reliability Analysis: Probability Models and Statistical Methods*. New York: Springer-Verlag.

Zelen, M. (1971). The analysis of several 2×2 contingency tables. *Biometrika* **58**, 129–137.

Zheng, J. X. (1997). A consistent specification test of independence. *Journal of Nonparametric Statistics* **7**, 297–306.

R Program Index

The software R (freely available for download at http://www.r-project.org/) is utilized throughout this text. Some commonly used functions are included in R by default. Other functions are available in packages from sources independent of the authors. The remaining functions are defined in the **NSM3** package. The user needs only to install (once) and load (at the beginning of each R session) the **NSM3** package to have access to all of the functions used throughout the text. The independently maintained packages will be automatically included. Readers who are unfamiliar with R are referred to the "Introduction to R" available at http://cran.r-project.org/doc/manuals/R-intro.html.

Any bugs, suggestions, or other issues for the functions defined in the **NSM3** package should be sent to Grant Schneider at schneider.393@osu.edu. Issues relating to independent functions should be referred to their respective maintainers.

The following is a table of the R functions used throughout the text, a short description of each, and the package where each is defined. More extensive descriptions and examples can be found in the R documentation, which may be accessed by entering `?function.name` (with `function.name` replaced by the desired function) in the R console. General package information (including additional functions not referenced in the text) can be obtained by entering `help(package="NSM3")` in the R console.

Function Name	Description	Package
`akj`	Function to perform univariate adaptive kernel density estimation using Silverman's method	quantreg
`ash1`	Function to compute the univariate average shifted histogram	ash
`binom.confint`	Function to obtain a confidence interval for the Binomial parameter p	binom
`binom.test`	Function to test the null hypothesis about the probability of success in a Bernoulli experiment	stats
`cAnsBrad`	Quantile function for the Ansari-Bradley C distribution	NSM3
`cBohnWolfe`	Quantile function for the Bohn-Wolfe U distribution	NSM3
`cDurSkiMa`	Quantile function for the Durbin, Skillings-Mack D distribution	NSM3
`chisq.test`	Function to perform chi-square contigency table tests	stats
`cFligPoli`	Quantile function for the Fligner-Policello U distribution	NSM3
`cFrd`	Quantile function for the Friedman, Kendall-Babington Smith S distribution	NSM3

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

Function Name	Description	Package
`cHayStON`	Quantile function for the Hayter-Stone W distribution	NSM3
`cHayStonLSA`	Quantile function for the Hayter-Stone W asymptotic distribution	NSM3
`cHollBivSym`	Quantile function for the Hollander A distribution	NSM3
`ch.ro`	Function to compute the Campbell-Hollander estimator $\hat{G}$	NSM3
`cJCK`	Quantile function for the Jonckheere-Terpstra J distribution	NSM3
`cKolSmirn`	Quantile function for the Kolmogorov-Smirnov J distribution	NSM3
`cKW`	Quantile function for the Kruskal-Wallis H distribution	NSM3
`cLepage`	Quantile function for the Lepage D distribution	NSM3
`cMackSkil`	Quantile function for the Mack-Skillings MS distribution	NSM3
`cMaxCorrNor`	Quantile function for the maximum of k $N(0,1)$ random variables with common correlation ρ	NSM3
`cNDWol`	Quantile function for the Nemenyi, Damico-Wolfe Y distribution	NSM3
`cNWWM`	Quantile function for the Nemenyi, Wilcoxon-Wilcox, Miller R distribution	NSM3
`cor`	Function to compute correlation	stats
`cor.test`	Test for association between paired samples	stats
`CorrUpperBound`	Function to compute the upper bound for the null correlation between two overlapping signed rank statistics	NSM3
`coxph`	Function to fit a Cox proportional hazards regression model	survival
`cox.zph`	Function to test the proportional hazards assumption for a Cox regression model fit	survival
`cPage`	Quantile function for the Page L distribution	NSM3
`cRangeNor`	Quantile function for the range of k independent $N(0,1)$ random variables	NSM3
`cSDCFlig`	Quantile function for the Dwass, Steel, Critchlow-Fligner W distribution	NSM3
`cUmbrPK`	Quantile function for the Mack-Wolfe Peak Known A_p distribution	NSM3
`cUmbrPU`	Quantile function for the Mack-Wolfe Peak Unknown $A_{\hat{p}}$ distribution	NSM3
`cWNMT`	Quantile function for the Wilcoxon, Nemenyi, McDonald-Thompson R distribution	NSM3
`density`	Function to compute kernel density estimates	stats
`dmrl.mc`	Function to compute the Monte Carlo or asymptotic P-value for the observed Hollander-Proschan V' statistic	NSM3
`drop.test`	Function to perform a reduction in dispersion test	Rfit
`dwt`	Function to perform a level J decomposition of the input vector using the pyramid algorithm	waveslim
`ecdf`	Function to compute an empiricial cumulative distribution function	stats

(*continued*)

Function Name	Description	Package
`ecdf.ks.CI`	Function to compute and plot Kolmogorov's 95% confidence band for the distribution function $F(x)$	NSM3
`epstein`	Function to compute the P-value for the observed Epstein E statistic	NSM3
`e.mc`	Monte Carlo approximation to the `epstein` function	NSM3
`ferg.df`	Function to compute an approximation of Ferguson's estimator μ_n	NSM3
`fisher.test`	Function to perform Fisher's exact test for the null hypothesis of independence of rows and columns in a contingency table with fixed marginals	stats
`hist`	Function to compute a histogram of the given data values	graphics
`hoeffd`	Function to compute a matrix of Hoeffding's D statistic for all possible pairs of columns of a matrix	Hmisc
`HoeffD`	Function to compute Hoeffding's D statistic for small sample sizes, which is used in pHoeffD	NSM3
`HollBivSym`	Function to compute the Hollander A statistic for testing bivariate symmetry	NSM3
`hybrid.thresh`	Function to perform wavelet shrinkage using hybrid SURE thresholding	waveslim
`idwt`	Inverse of the `dwt` function, to reconstruct the original data	waveslim
`imodwt`	Inverse of the `modwt` function, to reconstruct the original data	waveslim
`kendall.ci`	Function to produce a confidence interval for Kendall's τ	NSM3
`klefsjo.ifr`	Function to compute the P-value for the observed Klefsjö's A^* statistic	NSM3
`klefsjo.ifra`	Function to compute the P-value for the observed Klefsjö's B^* statistic	NSM3
`klefsjo.ifr.mc`	Monte Carlo approximation to the `klefsjo.ifr` function	NSM3
`klefsjo.ifra.mc`	Monte Carlo approximation to the `klefsjo.ifra` function	NSM3
`km.ci`	Function to compute pointwise and simultaneous confidence intervals associated with the Kaplan-Meier estimator	km.ci
`kolmogorov`	Function to compute the asymptotic P-value for the observed Kolmogorov D statistic	NSM3
`lillie.test`	Function to perform the Lilliefors test for the composite hypothesis of normality	nortest
`loess`	Function to fit a polynomial surface determined by one or more numerical predictors, using local fitting	stats
`loess.as`	Function to fit a local polynomial regression with automatic smoothing parameter selection	fANCOVA

Function Name	Description	Package
`mantelhaen.test`	Function to perform a Cochran-Mantel-Haenszel chi-square test of the null hypothesis that two nominal variables are conditionally independent in each stratum	stats
`mblm`	Function to fit linear models based on Theil-Sen single median or Siegel repeated medians	mblm
`mcnemar.test`	Function to perform McNemar's chi-square test for symmetry of rows and columns in a two-dimensional contingency table	stats
`median`	Function to compute the sample median	stats
`MillerJack`	Function to compute the Miller Jackknife Q statistic	stats
`modwt`	Function to perform a level J decomposition of the input vector using the non-decimated discrete wavelet transform	waveslim
`mra`	Function to perform a level J additive decomposition of the input vector using the pyramid algorithm	waveslim
`mrl`	Function to return the mean residual life along with Hall and Wellner's upper and lower bounds	NSM3
`nb.mc`	Monte Carlo approximation to the `newbet` function	NSM3
`newbet`	Function to compute the asymptotic P-value for the observed Hollander-Proschan T^* statistic	NSM3
`npreg`	Function to perform kernel regression	np
`oddsratio`	Calculates odds ratio by one of several methods	epitools
`owa`	Function to compute the ordered Walsh averages and the value of the Hodges-Lehmann estimator	NSM3
`pAnsBrad`	Function to compute the P-value for the observed Ansari-Bradley C statistic	NSM3
`pbinom`	Distribution function for the binomial distribution	stats
`pBohnWolfe`	Function to compute the P-value for the observed Bohn-Wolfe U statistic	NSM3
`pchisq`	Distribution function for the chi-square distribution	stats
`pDurSkiMa`	Function to compute the P-value for the observed Durbin, Skillings-Mack D statistic	NSM3
`pf`	Distribution function for the F distribution	stats
`pFligPoli`	Function to compute the P-value for the observed Fligner-Policello U statistic	NSM3
`pFrd`	Function to compute the P-value for the observed Friedman, Kendall-Babington Smith S statistic	NSM3
`pHaySton`	Function to compute the P-value for the observed Hayter-Stone W statistic	NSM3
`pHayStonLSA`	Function to compute the upper-tail probability of the Hayter-Stone W asymptotic distribution for a given cutoff	NSM3
`pHoeff`	Function to approximate the distribution of Hoeffding's D statistic using a Monte Carlo sample	NSM3
`pHollBivSym`	Function to compute the P-value for the observed Hollander A statistic	NSM3
`pJCK`	Function to compute the P-value for the observed Jonckheere-Terpstra J statistic	NSM3

(*continued*)

Function Name	Description	Package
`pKendall`	Distribution function for the Kendall K statistic	SuppDists
`pKolSmirn`	Function to compute the P-value for the observed Kolmogorov-Smirnov J statistic	NSM3
`pKW`	Function to compute the P-value for the observed Kruskal-Wallis H statistic	NSM3
`pLepage`	Function to compute the P-value for the observed Lepage D statistic	NSM3
`pMackSkil`	Function to compute the P-value for the observed Mack-Skillings MS statistic	NSM3
`pMaxCorrNor`	Function to compute the upper-tail probability of the maximum of k $N(0,1)$ random variables with common correlation ρ for a given cutoff	NSM3
`pNDWol`	Function to compute the P-value for the observed Nemenyi, Damico-Wolfe Y statistic	NSM3
`pNWWM`	Function to compute the P-value for the observed Nemenyi, Wilcoxon-Wilcox, Miller R statistic	NSM3
`pnorm`	Distribution function for the normal distribution	stats
`pPage`	Function to compute the P-value for the observed Page L statistic	NSM3
`pPairedWilcoxon`	Function to extend `wilcox.test` to compute the (exact or Monte Carlo) P-value for paired Wilcoxon data in the presence of ties	NSM3
`pRangeNor`	Function to compute the upper-tail probability of the range of k independent $N(0,1)$ random variables for a given cutoff	NSM3
`pretty`	Function to compute equally spaced values that cover the range of values in the given data	base
`prop.test`	Function to test the null hypothesis that the proportions in several groups are the same or that they equal specified values	stats
`pSDCFlig`	Function to compute the P-value for the observed Dwass, Steel, Critchlow-Fligner W statistic	NSM3
`psignrank`	Distribution function for the Wilcoxon signed rank T^+ statistic	stats
`pSpearman`	Distribution function for the Spearman r_S statistic	SuppDists
`pUmbrPK`	Function to compute the P-value for the observed Mack-Wolfe Peak Known A_p statistic	NSM3
`pUmbrPU`	Function to compute the P-value for the observed Mack-Wolfe Peak Unknown $A_{\hat{p}}$ statistic	NSM3
`pwilcox`	Distribution function for the Wilcoxon rank sum W statistic	stats
`pWNMT`	Function to compute the P-value for the observed Wilcoxon, Nemenyi, McDonald-Thompson R statistic	NSM3
`qbinom`	Quantile function for the binomial distribution	stats
`qchisq`	Quantile function for the chi-square distribution	stats
`qf`	Quantile function for the F distribution	stats
`qKendall`	Quantile function for the Kendall K distribution	SuppDists
`qKolSmirnLSA`	Quantile function for the asymptotic distribution of the Kolmogorov-Smirnov J statistic	NSM3

Function Name	Description	Package
`qnorm`	Quantile function for the normal distribution	stats
`qSpearman`	Quantile function for the Spearman r_S distribution	SuppDists
`qt`	Quantile function for the Student t distribution	stats
`qwilcox`	Quantile function for the Wilcoxon rank sum W distribution	stats
`rexp`	Function for random generation from the exponential distribution	stats
`rfit`	Function to minimize Jaeckel's dispersion function to obtain a rank-based solution for linear models	Rfit
`RFPW`	Function to compute the P-value for the observed Randles-Fligner-Policello-Wolfe V statistic	NSM3
`rma.mh`	Function to fit a fixed-effects model via the Mantel-Haenszel method	metafor
`rnorm`	Function for random generation from the normal distribution	stats
`RSS`	Function to obtain a ranked-set sample of given set size and number of cycles based on a specified auxiliary variable	NSM3
`runif`	Function for random generation from the uniform distribution	stats
`sample`	Function to take a sample (with or without replacement) of the specified size from the input data	base
`sen.adichie`	Function to test for parallel lines	NSM3
`SIGN.test`	Function to test a hypothesis based on the sign statistic B and obtain confidence intervals for one-sample problems	BSDA
`supsmu`	Function to smooth values by Friedman's super smoother	stats
`sure.thresh`	Function to perform wavelet shrinkage using SURE thresholding	waveslim
`Surv`	Function to create a survival object, usually used as a response variable in a model formula	survival
`survdiff`	Function to test if there are differences between two or more survival curves	survival
`survfit`	Function to create survival curves from a formula or previously fitted model	survival
`svr.df`	Function to compute the Susarla-van Ryzin estimator $\hat{\hat{F}}_n$	NSM3
`tc`	Function to compute the asymptotic P-value for the observed Guess-Hollander-Proschan T_1 statistic	NSM3
`theil`	Function to estimate and perform tests on the slope and intercept of a simple linear model	NSM3
`universal.thresh`	Function to perform wavelet shrinkage using universal thresholding	waveslim
`unlist`	Function to, given a list, simplify the list to produce a vector containing the atomic components of that list	base

(*continued*)

Function Name	Description	Package
`waerden.test`	Function to compute the P-value for the observed van der Waerden c statistic	agricolae
`wilcox.test`	Function to perform one- and two-sample Wilcoxon (Mann-Whitney) tests	stats
`wilcox_test`	Function to compute the P-value for the observed two-sample Wilcoxon statistic based on the conditional distribution (when there are ties)	coin
`zelen.test`	Function to perform Zelen's test	NSM3

Author Index

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

Subject Index

Nonparametric Statistical Methods, Third Edition. Myles Hollander, Douglas A. Wolfe, Eric Chicken.

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

† ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data, *Second Edition*
AGRESTI · An Introduction to Categorical Data Analysis, *Second Edition*
AGRESTI · Categorical Data Analysis, *Third Edition*
ALTMAN, GILL, and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
AMARATUNGA, CABRERA, and SHKEDY . Exploration and Analysis of DNA Microarray and Other High-Dimensional Data, *Second Edition*
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Third Edition*
* ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
* ARTHANARI and DODGE · Mathematical Programming in Statistics
* BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BAJORSKI · Statistics for Imaging, Optics, and Photonics
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT · Environmental Statistics
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*
BARTHOLOMEW, KNOTT, and MOUSTAKI · Latent Variable Models and Factor Analysis: A Unified Approach, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference, *Second Edition*
BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications
BATES and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons
BEIRLANT, GOEGEBEUR, SEGERS, TEUGELS, and DE WAAL · Statistics of Extremes: Theory and Applications
BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
† BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Fourth Edition*
BERNARDO and SMITH · Bayesian Theory
BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*
BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN · Measurement Errors in Surveys
BILLINGSLEY · Convergence of Probability Measures, *Second Edition*
BILLINGSLEY · Probability and Measure, *Anniversary Edition*
BIRKES and DODGE · Alternative Methods of Regression
BISGAARD and KULAHCI · Time Series Analysis and Forecasting by Example
BISWAS, DATTA, FINE, and SEGAL · Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics
BLISCHKE and MURTHY (editors) · Case Studies in Reliability and Maintenance
BLISCHKE and MURTHY · Reliability: Modeling, Prediction, and Optimization
BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*
BOLLEN · Structural Equations with Latent Variables
BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective
BOROVKOV · Ergodicity and Stability of Stochastic Processes
BOSQ and BLANKE · Inference and Prediction in Large Dimensions
BOULEAU · Numerical Methods for Stochastic Processes
* BOX and TIAO · Bayesian Inference in Statistical Analysis
BOX · Improving Almost Anything, *Revised Edition*
* BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
BOX and DRAPER · Response Surfaces, Mixtures, and Ridge Analyses, *Second Edition*
BOX, HUNTER, and HUNTER · Statistics for Experimenters: Design, Innovation, and Discovery, *Second Editon*
BOX, JENKINS, and REINSEL · Time Series Analysis: Forcasting and Control, *Fourth Edition*
BOX, LUCEÑO, and PANIAGUA-QUIÑONES · Statistical Control by Monitoring and Adjustment, *Second Edition*
* BROWN and HOLLANDER · Statistics: A Biomedical Introduction
CAIROLI and DALANG · Sequential Stochastic Optimization
CASTILLO, HADI, BALAKRISHNAN, and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science
CHAN · Time Series: Applications to Finance with R and S-Plus®, *Second Edition*
CHARALAMBIDES · Combinatorial Methods in Discrete Distributions
CHATTERJEE and HADI · Regression Analysis by Example, *Fourth Edition*
CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
CHERNICK · Bootstrap Methods: A Guide for Practitioners and Researchers, *Second Edition*
CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty, *Second Edition*
CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, *Second Edition*
CLARKE · Linear Models: The Theory and Application of Analysis of Variance
CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
* COCHRAN and COX · Experimental Designs, *Second Edition*
COLLINS and LANZA · Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences
CONGDON · Applied Bayesian Modelling
CONGDON · Bayesian Models for Categorical Data
CONGDON · Bayesian Statistical Modelling, *Second Edition*
CONOVER · Practical Nonparametric Statistics, *Third Edition*
COOK · Regression Graphics
COOK and WEISBERG · An Introduction to Regression Graphics
COOK and WEISBERG · Applied Regression Including Computing and Graphics
CORNELL · A Primer on Experiments with Mixtures
CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
COX · A Handbook of Introductory Statistical Methods

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

CRESSIE · Statistics for Spatial Data, *Revised Edition*
CRESSIE and WIKLE · Statistics for Spatio-Temporal Data
CSÖRGŐ and HORVÁTH · Limit Theorems in Change Point Analysis
DAGPUNAR · Simulation and Monte Carlo: With Applications in Finance and MCMC
DANIEL · Applications of Statistics to Industrial Experimentation
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*
* DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
DAVID and NAGARAJA · Order Statistics, *Third Edition*
* DEGROOT, FIENBERG, and KADANE · Statistics and the Law
DEL CASTILLO · Statistical Process Adjustment for Quality Control
DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables
DEMIDENKO · Mixed Models: Theory and Applications with R, *Second Edition*
DENISON, HOLMES, MALLICK and SMITH · Bayesian Methods for Nonlinear Classification and Regression
DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
DEY and MUKERJEE · Fractional Factorial Plans
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
* DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
* DOOB · Stochastic Processes
DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*
DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
DRYDEN and MARDIA · Statistical Shape Analysis
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Fourth Edition*
DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
EDLER and KITSOS · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment
* ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ENDERS · Applied Econometric Time Series, *Third Edition*
† ETHIER and KURTZ · Markov Processes: Characterization and Convergence
EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
EVERITT, LANDAU, LEESE, and STAHL · Cluster Analysis, *Fifth Edition*
FEDERER and KING · Variations on Split Plot and Split Block Experiment Designs
FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition,* Revised; Volume II, *Second Edition*
FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis, *Second Edition*
* FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, *Third Edition*
† FLEMING and HARRINGTON · Counting Processes and Survival Analysis
FUJIKOSHI, ULYANOV, and SHIMIZU · Multivariate Statistics: High-Dimensional and Large-Sample Approximations
FULLER · Introduction to Statistical Time Series, *Second Edition*
† FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models
GEISSER · Modes of Parametric Statistical Inference
GELMAN and MENG · Applied Bayesian Modeling and Causal Inference from ncomplete-Data Perspectives
GEWEKE · Contemporary Bayesian Econometrics and Statistics
GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments
GIFI · Nonlinear Multivariate Analysis
GIVENS and HOETING · Computational Statistics
GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations,

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

Second Edition

GOLDSTEIN · Multilevel Statistical Models, *Fourth Edition*

GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues

GOLDSTEIN and WOOFF · Bayes Linear Statistics

GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing

GROSS, SHORTLE, THOMPSON, and HARRIS · Fundamentals of Queueing Theory, *Fourth Edition*

GROSS, SHORTLE, THOMPSON, and HARRIS · Solutions Manual to Accompany Fundamentals of Queueing Theory, *Fourth Edition*

* HAHN and SHAPIRO · Statistical Models in Engineering

HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners

HALD · A History of Probability and Statistics and their Applications Before 1750

† HAMPEL · Robust Statistics: The Approach Based on Influence Functions

HARTUNG, KNAPP, and SINHA · Statistical Meta-Analysis with Applications

HEIBERGER · Computation for the Analysis of Designed Experiments

HEDAYAT and SINHA · Design and Inference in Finite Population Sampling

HEDEKER and GIBBONS · Longitudinal Data Analysis

HELLER · MACSYMA for Statisticians

HERITIER, CANTONI, COPT, and VICTORIA-FESER · Robust Methods in Biostatistics

HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, *Second Edition*

HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 2: Advanced Experimental Design

HINKELMANN (editor) · Design and Analysis of Experiments, Volume 3: Special Designs and Applications

HOAGLIN, MOSTELLER, and TUKEY · Fundamentals of Exploratory Analysis of Variance

* HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes

* HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis

HOCHBERG and TAMHANE · Multiple Comparison Procedures

HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Third Edition*

HOEL · Introduction to Mathematical Statistics, *Fifth Edition*

HOGG and KLUGMAN · Loss Distributions

HOLLANDER, WOLFE, and CHICKEN · Nonparametric Statistical Methods, *Third Edition*

HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*

HOSMER, LEMESHOW, and MAY · Applied Survival Analysis: Regression Modeling of Time-to-Event Data, *Second Edition*

HUBER · Data Analysis: What Can Be Learned From the Past 50 Years

HUBER · Robust Statistics

† HUBER and RONCHETTI · Robust Statistics, *Second Edition*

HUBERTY · Applied Discriminant Analysis, *Second Edition*

HUBERTY and OLEJNIK · Applied MANOVA and Discriminant Analysis, *Second Edition*

HUITEMA · The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, *Second Edition*

HUNT and KENNEDY · Financial Derivatives in Theory and Practice, *Revised Edition*

HURD and MIAMEE · Periodically Correlated Random Sequences: Spectral Theory and Practice

HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek— with Commentary

HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data

JACKMAN · Bayesian Analysis for the Social Sciences

† JACKSON · A User's Guide to Principle Components

JOHN · Statistical Methods in Engineering and Quality Assurance

JOHNSON · Multivariate Statistical Simulation

JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz

JOHNSON, KEMP, and KOTZ · Univariate Discrete Distributions, *Third Edition*

JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present

JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1,

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

Second Edition

JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*

JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions

JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*

JUREK and MASON · Operator-Limit Distributions in Probability Theory

KADANE · Bayesian Methods and Ethics in a Clinical Trial Design

KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence

KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*

KARIYA and KURATA · Generalized Least Squares

KASS and VOS · Geometrical Foundations of Asymptotic Inference

† KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis

KEDEM and FOKIANOS · Regression Models for Time Series Analysis

KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory

KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*

KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models

* KISH · Statistical Design for Research

KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences

KLEMELÄ · Smoothing of Multivariate Data: Density Estimation and Visualization

KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions, *Third Edition*

KLUGMAN, PANJER, and WILLMOT · Loss Models: Further Topics

KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions, *Third Edition*

KOSKI and NOBLE · Bayesian Networks: An Introduction

KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*

KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index

KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume

KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1

KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2

KOWALSKI and TU · Modern Applied U-Statistics

KRISHNAMOORTHY and MATHEW · Statistical Tolerance Regions: Theory, Applications, and Computation

KROESE, TAIMRE, and BOTEV · Handbook of Monte Carlo Methods

KROONENBERG · Applied Multiway Data Analysis

KULINSKAYA, MORGENTHALER, and STAUDTE · Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence

KULKARNI and HARMAN · An Elementary Introduction to Statistical Learning Theory

KUROWICKA and COOKE · Uncertainty Analysis with High Dimensional Dependence Modelling

KVAM and VIDAKOVIC · Nonparametric Statistics with Applications to Science and Engineering

LACHIN · Biostatistical Methods: The Assessment of Relative Risks, *Second Edition*

LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction

LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*

LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*

LAWSON · Statistical Methods in Spatial Epidemiology, *Second Edition*

LE · Applied Categorical Data Analysis, *Second Edition*

LE · Applied Survival Analysis

LEE · Structural Equation Modeling: A Bayesian Approach

LEE and WANG · Statistical Methods for Survival Data Analysis, *Fourth Edition*

LEPAGE and BILLARD · Exploring the Limits of Bootstrap

LESSLER and KALSBEEK · Nonsampling Errors in Surveys

LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics

LIAO · Statistical Group Comparison

LIN · Introductory Stochastic Analysis for Finance and Insurance

LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*

LLOYD · The Statistical Analysis of Categorical Data

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

LOWEN and TEICH · Fractal-Based Point Processes
MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
MALLER and ZHOU · Survival Analysis with Long Term Survivors
MARCHETTE · Random Graphs for Statistical Pattern Recognition
MARDIA and JUPP · Directional Statistics
MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice
MARONNA, MARTIN and YOHAI · Robust Statistics: Theory and Methods
MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*
McCULLOCH, SEARLE, and NEUHAUS · Generalized, Linear, and Mixed Models, *Second Edition*
McFADDEN · Management of Data in Clinical Trials, *Second Edition*
* McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
McLACHLAN, DO, and AMBROISE · Analyzing Microarray Gene Expression Data
McLACHLAN and KRISHNAN · The EM Algorithm and Extensions, *Second Edition*
McLACHLAN and PEEL · Finite Mixture Models
McNEIL · Epidemiological Research Methods
MEEKER and ESCOBAR · Statistical Methods for Reliability Data
MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
MENGERSEN, ROBERT, and TITTERINGTON · Mixtures: Estimation and Applications
MICKEY, DUNN, and CLARK · Applied Statistics: Analysis of Variance and Regression, *Third Edition*
* MILLER · Survival Analysis, *Second Edition*
MONTGOMERY, JENNINGS, and KULAHCI · Introduction to Time Series Analysis and Forecasting
MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Fifth Edition*
MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
MUIRHEAD · Aspects of Multivariate Statistical Theory
MULLER and STOYAN · Comparison Methods for Stochastic Models and Risks
MURTHY, XIE, and JIANG · Weibull Models
MYERS, MONTGOMERY, and ANDERSON-COOK · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Third Edition*
MYERS, MONTGOMERY, VINING, and ROBINSON · Generalized Linear Models. With Applications in Engineering and the Sciences, *Second Edition*
NATVIG · Multistate Systems Reliability Theory With Applications
† NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
† NELSON · Applied Life Data Analysis
NEWMAN · Biostatistical Methods in Epidemiology
NG, TAIN, and TANG · Dirichlet Theory: Theory, Methods and Applications
OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
PALTA · Quantitative Methods in Population Health: Extensions of Ordinary Regressions
PANJER · Operational Risk: Modeling and Analytics
PANKRATZ · Forecasting with Dynamic Regression Models
PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
PARDOUX · Markov Processes and Applications: Algorithms, Networks, Genome and Finance
PARMIGIANI and INOUE · Decision Theory: Principles and Approaches
* PARZEN · Modern Probability Theory and Its Applications
PEÑA, TIAO, and TSAY · A Course in Time Series Analysis
PESARIN and SALMASO · Permutation Tests for Complex Data: Applications and Software
PIANTADOSI · Clinical Trials: A Methodologic Perspective, *Second Edition*
POURAHMADI · Foundations of Time Series Analysis and Prediction Theory
POURAHMADI · High-Dimensional Covariance Estimation
POWELL · Approximate Dynamic Programming: Solving the Curses of Dimensionality, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

POWELL and RYZHOV · Optimal Learning
PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*
PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach
PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
† PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
QIU · Image Processing and Jump Regression Analysis
* RAO · Linear Statistical Inference and Its Applications, *Second Edition*
RAO · Statistical Inference for Fractional Diffusion Processes
RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*
RAYNER, THAS, and BEST · Smooth Tests of Goodnes of Fit: Using R, *Second Edition*
RENCHER and SCHAALJE · Linear Models in Statistics, *Second Edition*
RENCHER and CHRISTENSEN · Methods of Multivariate Analysis, *Third Edition*
RENCHER · Multivariate Statistical Inference with Applications
RIGDON and BASU · Statistical Methods for the Reliability of Repairable Systems
* RIPLEY · Spatial Statistics
* RIPLEY · Stochastic Simulation
ROHATGI and SALEH · An Introduction to Probability and Statistics, *Second Edition*
ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance
ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
ROSSI, ALLENBY, and McCULLOCH · Bayesian Statistics and Marketing
† ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
ROYSTON and SAUERBREI · Multivariate Model Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables
* RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN and KROESE · Simulation and the Monte Carlo Method, *Second Edition*
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RUBINSTEIN, RIDDER, and VAISMAN · Fast Sequential Monte Carlo Methods for Counting and Optimization
RYAN · Modern Engineering Statistics
RYAN · Modern Experimental Design
RYAN · Sample Size Determination and Power
RYAN · Modern Regression Methods, *Second Edition*
RYAN · Statistical Methods for Quality Improvement, *Third Edition*
SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications
SALTELLI, CHAN, and SCOTT (editors) · Sensitivity Analysis
SCHERER · Batch Effects and Noise in Microarray Experiments: Sources and Solutions
* SCHEFFE · The Analysis of Variance
SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application
SCHOTT · Matrix Analysis for Statistics, *Second Edition*
SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives
SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
* SEARLE · Linear Models
† SEARLE · Linear Models for Unbalanced Data
† SEARLE · Matrix Algebra Useful for Statistics
† SEARLE, CASELLA, and McCULLOCH · Variance Components
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER · A Matrix Handbook For Statisticians
† SEBER · Multivariate Observations
SEBER and LEE · Linear Regression Analysis, *Second Edition*
† SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing Systems
* SERFLING · Approximation Theorems of Mathematical Statistics
SHAFER and VOVK · Probability and Finance: It's Only a Game!
SHERMAN · Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties
SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order, and Shape Restrictions

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

SINGPURWALLA · Reliability and Risk: A Bayesian Perspective
SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models, *Second Edition*
STAPLETON · Models for Probability and Statistical Inference: Theory and Applications
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYAN · Counterexamples in Probability, *Second Edition*
STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
STREET and BURGESS · The Construction of Optimal Stated Choice Experiments: Theory and Methods
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research
TAKEZAWA · Introduction to Nonparametric Regression
TAMHANE · Statistical Analysis of Designed Experiments: Theory and Applications
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building: Data, Models, and Reality, *Second Edition*
THOMPSON · Sampling, *Third Edition*
THOMPSON · Simulation: A Modeler's Approach
THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY · Analysis of Financial Time Series, *Third Edition*
TSAY · An Introduction to Analysis of Financial Data with R
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
† VAN BELLE · Statistical Rules of Thumb, *Second Edition*
VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, *Second Edition*
VESTRUP · The Theory of Measures and Integration
VIDAKOVIC · Statistical Modeling by Wavelets
VIERTL · Statistical Methods for Fuzzy Data
VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments
WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data
WEISBERG · Applied Linear Regression, *Third Edition*
WEISBERG · Bias and Causation: Models and Judgment for Valid Comparisons
WELSH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment
* WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WOODWORTH · Biostatistics: A Bayesian Introduction
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization, *Second Edition*
WU and ZHANG · Nonparametric Regression Methods for Longitudinal Data Analysis
YIN · Clinical Trial Design: Bayesian and Frequentist Adaptive Methods
YOUNG, VALERO-MORA, and FRIENDLY · Visual Statistics: Seeing Data with Dynamic Interactive Graphics
ZACKS · Stage-Wise Adaptive Designs
* ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZELTERMAN · Discrete Distributions—Applications in the Health Sciences
ZHOU, OBUCHOWSKI, and McCLISH · Statistical Methods in Diagnostic Medicine, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.