Statistical Methods
for the Social Sciences

FIFTH EDITION

ALAN AGRESTI

STATISTICAL METHODS FOR THE SOCIAL SCIENCES

Fifth Edition
Global Edition

Alan Agresti

University of Florida

Pearson

Attributions of third party content appear on page 549, which constitutes an extension of this copyright page.

To my Parents
Louis J. Agresti and Marjorie H. Agresti

This page intentionally left blank

Contents

This page intentionally left blank

Preface

When Barbara Finlay and I undertook the first edition of this book nearly four decades ago, our goal was to introduce statistical methods in a style that emphasized their concepts and their application to the social sciences rather than the mathematics and computational details behind them. We did this by focusing on how the methods are used and interpreted rather than their theoretical derivations.

This edition of the book continues the emphasis on concepts and applications, using examples and exercises with a variety of "real data." This edition increases its illustrations of statistical software for computations, and takes advantage of the outstanding applets now available on the Internet for explaining key concepts such as sampling distributions and for conducting basic data analyses. I continue to downplay mathematics, in particular probability, which is all too often a stumbling block for students. On the other hand, the text is not a cookbook. Reliance on an overly simplistic recipe-based approach to statistics is not the route to good statistical practice.

Changes in the Fifth Edition

Users of earlier editions will notice that the book no longer lists Barbara Finlay as a co-author. I am grateful to Barbara Finlay for her contributions to the first two editions of this text. Combining her sociology background with my statistics background, she very much helped me develop a book that is not only statistically sound but also relevant to the social sciences.

Since the first edition, the increase in computer power coupled with the continued improvement and accessibility of statistical software has had a major impact on the way social scientists analyze data. Because of this, this book does not cover the traditional shortcut hand-computational formulas and approximations. The presentation of computationally complex methods, such as regression, emphasizes interpretation of software output rather than the formulas for performing the analysis. The text contains numerous sample software outputs, both in chapter text and in homework exercises. In the appendix on using statistical software, this edition adds R and Stata to the material on SPSS and SAS.

Exposure to realistic but simple examples and to numerous homework exercises is vital to student learning. This edition has updated data in most of the exercises and replaced some exercises with new ones. Each chapter's homework set is divided into two parts, straightforward exercises on the text material in *Practicing the Basics* and exercises dealing with open-ended data analyses, understanding of concepts, and advanced material in *Concepts and Applications*. The data sets in the examples and exercises are available at www.pearsonglobaleditions.com/Agresti.

This edition contains some changes and additions in content, directed toward a more modern approach. The main changes are as follows:

- The text has greater integration of *statistical software*. Software output shown now uses R and Stata instead of only SAS and SPSS, although much output has a generic appearance. The text appendix provides instructions about basic use of these software packages.

- New examples and exercises in Chapters 4–6 ask students to use applets to help learn the fundamental concepts of sampling distributions, confidence

intervals, and significance tests. The text also now relies more on applets for finding tail probabilities from distributions such as the normal, t, and chi-squared. I strongly encourage instructors and students to look at the excellent applets cited at `www.pearsonglobaleditions.com/Agresti.` They were prepared by Prof. Bernhard Klingenberg for the fourth edition of the text *Statistics: The Art and Science of Learning from Data*, by Agresti, Franklin, and Klingenberg (Pearson, 2017).

- Chapter 5 has a new section introducing maximum likelihood estimation and the bootstrap method.

- Chapter 12 on ANOVA has been reorganized to put more emphasis on using regression models with dummy variables to handle categorical explanatory variables.

- Chapter 13 on regression modeling with both quantitative and categorical explanatory variables has a new section using case studies to illustrate how research studies commonly use regression with both types of explanatory variables. The chapter also has a new section introducing linear mixed models.

- Chapter 14 has a new section introducing robust regression standard errors and nonparametric regression.

- The text Web site `www.pearsonglobaleditions.com/Agresti` has the data sets analyzed in the text, in generic form to copy for input into statistical software. Special directories there also have data files in Stata format and in SPSS format, so they are ready for immediate use with those packages.

- Answers to Select Odd-Numbered Exercises are available at the text Website `www.pearsonglobaleditions.com/Agresti.`

Use of Text in Introductory Statistics Courses

Like the first four editions, this edition is appropriate for introductory statistics courses at either the undergraduate or beginning graduate level, and for either a single-term or a two-term sequence. Chapters 1–9 are the basis for a single-term course. If the instructor wishes to go further than Chapter 9 or wishes to cover some material in greater depth, sections that can easily be omitted without disturbing continuity include 2.4, 5.5, 6.6–6.7, 7.5–7.7, and 8.5. Also, Chapters 7–9 are self-contained, and the instructor could move directly into any of these after covering the fundamentals in Chapters 1–6. Three possible paths for a one-term course are as follows:

- Chapters 1–9 (possibly omitting sections noted above): Standard cross-section of methods, including basic descriptive and inferential statistics, two-sample procedures, contingency tables, and linear regression.

- Chapters 1–7, 9, and 11: Emphasis on regression.

- Chapters 1–7, and 9, and Sections 11.1–11.3 and 12.1–12.3: After two-group comparisons, introduction to regression and analysis of variance.

Regardless of the type of data, my belief is that a modeling paradigm emphasizing parameter estimation is more useful than the artificial hypothesis-testing approach of many statistics texts. Thus, the basic inference chapters (5–8) explain the advantages confidence intervals have over significance testing, and the second half of this text (starting in Chapter 9) is primarily concerned with model building. The modeling material forms the basis of a second course. Instructors who focus on

observational data rather than designed experiments may prefer to cover only the first section of Chapter 12 (ANOVA), to introduce dummy variables before moving to later chapters that incorporate both categorical and quantitative explanatory variables.

Some material appears in sections, subsections, or exercises marked by asterisks. This material is optional, having lesser importance for introductory courses. The text does not attempt to present every available method, since it is meant to be a teaching tool, not an encyclopedic cookbook. It does cover the most important methods for social science research, however, and it includes topics not usually discussed in introductory statistics texts, such as

- Methods for contingency tables that are more informative than chi-squared, such as cell residuals and analyses that utilize category orderings.

- Controlling for variables, and issues dealing with causation.

- The generalized linear modeling approach, encompassing ordinary regression, analysis of variance and covariance, gamma regression for nonnegative responses with standard deviation proportional to the mean, logistic regression for categorical responses, and loglinear association models for contingency tables.

- Relatively new methods that are increasingly used in research, such as the linear mixed model approach of using both fixed effects and random effects (and related multilevel models), and multiple imputation for dealing with missing data.

I believe that the student who works through this book successfully will acquire a solid foundation in applied statistical methodology.

Acknowledgments

Finally, extra special thanks to my wife, Jacki Levine, for assistance with editing and style in the third edition and with overall encouragement during the preparation of the fourth and fifth editions.

Alan Agresti
Gainesville, Florida and Brookline, Massachusetts

Global Edition Acknowledgments

Pearson would like to thank the following people for their work on the content of the Global Edition:

Contributors:
MaryJane Bock, Murdoch University Dubai
Gagari Chakrabarti, Presidency University
Raghvi Garg, doctoral student, Ashoka University
Pooja Thakur

Reviewers:
Dave Centeno, University of the Philippines
Chitrita Bhowmick Chakrabarti, Victoria Institution (College)
Gagari Chakrabarti, Presidency University
Samprit Chakrabarti, ICFAI Business School
Albert Lee Kai Chung, Nanyang Technological University
Eric Li, The University of Hong Kong
Francisco de los Reyes, University of the Philippines
Pooja Sengupta
Raymond Wong, The University of Hong Kong
Elizabeth Wright, Murdoch University Dubai

Chapter
1

INTRODUCTION

1.1 Introduction to Statistical Methodology

Recent years have seen a dramatic increase in the use of statistical methods by social scientists, whether they work in academia, government, or the private sector. Social scientists study their topics of interest, such as analyzing how well a program works or investigating the factors associated with beliefs and opinions of certain types, by analyzing quantitative evidence provided by data. The growth of the Internet and computing power has resulted in an increase in the amount of readily available quantitative information. At the same time, the evolution of new statistical methodology and software makes new methods available that can more realistically address the questions that social scientists seek to answer.

This chapter introduces "statistics" as a science that deals with describing data and making predictions that have a much wider scope than merely summarizing the collected data. So, why should knowledge of statistical science be important for a student who is studying to become a social scientist?

WHY STUDY STATISTICAL SCIENCE?

The increased use of statistical methods is evident in the changes in the content of articles published in social science research journals and reports prepared in government and industry. A quick glance through recent issues of journals such as *American Political Science Review* and *American Sociological Review* reveals the fundamental role of statistical methodology in research. For example, to learn about which factors have the greatest impact on student performance in school or to investigate what affects people's political beliefs or the quality of their health care or their decisions about work and home life, researchers collect information and analyze it using statistical methods. Because of this, more and more academic departments require that their majors take statistics courses.

These days, social scientists work in a wide variety of areas that use statistical methods, such as governmental agencies, business organizations, and health care facilities. Social scientists in government agencies dealing with human welfare or environmental issues or public health policy commonly need to read reports that contain statistical arguments, and perhaps use statistical methods themselves in preparing such reports. Some social scientists help managers to evaluate employee performance using quantitative benchmarks and to determine factors that help predict sales of products. Medical sociologists and physicians often must evaluate

recommendations from studies that contain statistical evaluations of new therapies or new ways of caring for the elderly. In fact, a recent issue of *The Journal of the American Medical Association* indicated that the Medical College Admissions Test has been revised to require more statistics, because doctors increasingly need to be able to evaluate quantitatively the factors that affect peoples' health.

In fact, increasingly many jobs for social scientists require a knowledge of statistical methods as a basic work tool. As the joke goes, "What did the sociologist who passed statistics say to the sociologist who failed it? 'I'll have a Big Mac, fries, and a Coke.'"

But an understanding of statistical science is important even if you never use statistical methods in your own career. Often you are exposed to communications containing statistical arguments, such as in advertising, news reporting, political campaigning, and surveys about opinions on controversial issues. Statistical science helps you to make sense of this information and evaluate which arguments are valid and which are invalid. You will find concepts from this text helpful in judging the information you encounter in your everyday life. Look at `www.youtube.com/user/ThisisStats` to view some short testimonials with the theme that "Statistics isn't just about data analysis or numbers; it is about understanding the world around us. The diverse face of statistics means you can use your education in statistics and apply it to nearly any area you are passionate about, such as the environment, health care, human rights, sports...."

We realize you are not reading this book in hopes of becoming a statistician. In addition, you may suffer from math phobia and feel fear at what lies ahead. Please be assured that you can read this book and learn the primary concepts and methods of statistics with little knowledge of mathematics. Just because you may have had difficulty in math courses before does not mean you will be at a disadvantage here. To understand this book, logical thinking and perseverance are more important than mathematics. In our experience, the most important factor in how well you do in a statistics course is how much time you spend on the course—attending class, doing homework, reading and re-reading this text, studying your class notes, working together with your fellow students, and getting help from your professor or teaching assistant—not your mathematical knowledge or your gender or your race or whether you now feel fear of statistics.

Please do not be frustrated if learning comes slowly and you need to read a chapter a few times before it makes sense. Just as you would not expect to take a single course in a foreign language and be able to speak that language fluently, the same is true with the language of statistical science. Once you have completed even a portion of this text, however, you will better understand how to make sense of statistical information.

DATA

Information gathering is at the heart of all sciences, providing the *observations* used in statistical analyses. The observations gathered on the characteristics of interest are collectively called ***data***.

For example, a study might conduct a survey of 1000 people to observe characteristics such as opinion about the legalization of same-sex marriage, political party affiliation, how often attend religious services, number of years of education, annual income, marital status, race, and gender. The data for a particular person would consist of observations such as (opinion = do not favor legalization, party = Republican, religiosity = once a week, education = 14 years, annual income in the range of 40–60 thousand dollars, marital status = married, race = white, gender = female). Looking

at the data in the right way helps us learn about how the characteristics are associated. We can then answer questions such as "Do people who attend church more often tend to be less favorable toward same-sex marriage?"

To generate data, the social sciences use a wide variety of methods, including surveys using questionnaires, experiments, and direct observation of behavior in natural settings. In addition, social scientists often analyze data already recorded for other purposes, such as police records, census materials, and hospital files. Existing archived collections of data are called **databases**. Many databases are now available on the Internet. An important database for social scientists contains results since 1972 of the *General Social Survey*.

Example 1.1

The General Social Survey   Every other year, the National Opinion Research Center at the University of Chicago conducts the General Social Survey (GSS). This survey of about 2000 adults provides data about opinions and behaviors of the American public. Social scientists use it to investigate how adult Americans answer a wide diversity of questions, such as "Do you believe in life after death?" "Would you be willing to pay higher prices in order to protect the environment?" and "Do you think a preschool child is likely to suffer if his or her mother works?" Similar surveys occur in other countries, such as the General Social Survey administered by Statistics Canada, the British Social Attitudes Survey, and the Eurobarometer survey and European Social Survey for nations in the European Union.

It is easy to get summaries of data from the GSS database. We'll demonstrate, using a question it asked in one survey, "About how many good friends do you have?"

- Go to the website `sda.berkeley.edu/GSS/` at the Survey Documentation and Analysis site at the University of California, Berkeley.

- Click on *GSS—with NO WEIGHT VARIABLES predefined*. You will then see a "variable selection" listing in the left margin dealing with issues addressed over the years, and a menu on the right for selecting particular characteristics of interest.

- The GSS name for the question about number of good friends is NUMFREND. Type NUMFREND in the *Row* box. Click on *Run the table*.

The GSS site will then generate a table that shows the possible values for "number of good friends" and the number of people and the percentage who made each possible response. The most common responses were 2 and 3 (about 16% made each of these responses). ∎

WHAT IS STATISTICAL SCIENCE?

You already have a sense of what the word *statistics* means. You hear statistics quoted about sports events (such as the number of points scored by each player on a basketball team), statistics about the economy (such as the median income or the unemployment rate), and statistics about opinions, beliefs, and behaviors (such as the percentage of students who indulge in binge drinking). In this sense, a statistic is merely a number calculated from data. But this book uses *statistics* in a much broader sense—as a science that gives us ways of obtaining and analyzing data.

Statistics

> *Statistics* consists of a body of methods for obtaining and analyzing data.

Specifically, statistical science provides methods for

1. ***Design***: Planning how to gather data for a research study to investigate questions of interest to us.
2. ***Description***: Summarizing the data obtained in the study.
3. ***Inference***: Making predictions based on the data, to help us deal with uncertainty in an objective manner.

Design refers to planning a study so that the data it yields are informative. For a survey, for example, the design specifies how to select the people to interview and constructs the questionnaire to administer to those people.

Description refers to summarizing data, to help understand the information the data provide. For example, an analysis of the number of good friends based on the GSS data might start with a list of the number reported for each person surveyed. The raw data are then a complete listing of observations, person by person. These are not easy to comprehend, however. We get bogged down in numbers. For presentation of results, instead of listing *all* observations, we could summarize the data with a graph or table showing the percentages reporting 1 good friend, 2 good friends, 3 good friends, and so on. Or we could report the average number of good friends, which was about 5, or the most common response, which was 2. Graphs, tables, and numerical summaries such as averages and percentages are called ***descriptive statistics***. We use descriptive statistics to reduce the data to a simpler and more understandable form without distorting or losing much information.

Inference refers to using the data to make predictions. For instance, for the GSS data on reported number of good friends, 6.1% reported having only 1 good friend. Can we use this information to predict the percentage of the 250 million adults in the United States who have only 1 good friend? A method presented in this book allows us to predict confidently that that percentage is no greater than 8%. Predictions made using data are called ***statistical inferences***.

Description and ***inference*** are the two types of ways of analyzing the data. Social scientists use descriptive and inferential statistics to answer questions about social phenomena. For instance, "Is having the death penalty available for punishment associated with a reduction in violent crime?" "Does student performance in schools depend on the amount of money spent per student, the size of the classes, or the teachers' salaries?"

1.2 Descriptive Statistics and Inferential Statistics

Section 1.1 explained that statistical science consists of methods for *designing* studies and *analyzing* data collected in the studies. A statistical analysis is classified as **descriptive** or **inferential**, according to whether its main purpose is to describe the data or to make predictions. To explain this distinction further, we next define the *population* and the *sample*.

POPULATIONS AND SAMPLES

The entities on which a study makes observations are called the sample ***subjects*** for the study. Usually the subjects are people, such as in the General Social Survey, but they need not be. For example, subjects in social research might be families, schools, or cities. Although we obtain data for the sample subjects, our ultimate interest is in the population that the sample represents.

Population and Sample

> The **population** is the total set of subjects of interest in a study. A **sample** is the subset of the population on which the study collects data.

In the 2014 General Social Survey, the sample was the 2538 adult Americans who participated in the survey. The population was all adult Americans at that time—about 250 million people. One person was sampled for about every 100,000 people in the population.

The goal of any study is to learn about populations. But it is almost always necessary, and more practical, to observe only samples from those populations. For example, survey organizations such as the Gallup Poll usually select samples of about 1000–2000 Americans to collect information about opinions and beliefs of the population of *all* Americans.

Descriptive Statistics

> **Descriptive statistics** summarize the information in a collection of data.

Descriptive statistics consist of graphs, tables, and numbers such as averages and percentages. Descriptive statistics reduce the data to simpler and more understandable form without distorting or losing much information.

Although data are usually available only for a sample, descriptive statistics are also useful when data are available for the entire population, such as in a census. By contrast, inferential statistics apply when data are available only for a sample, but we want to make a prediction about the entire population.

Inferential Statistics

> **Inferential statistics** provide predictions about a population, based on data from a sample of that population.

Example 1.2

How Many People Believe in Heaven?   In three of its surveys, the General Social Survey asked, "Do you believe in heaven?" The population of interest was the collection of all adults in the United States. In the most recent survey in which this was asked, 85% of the 1326 sampled subjects answered *yes*. This is a descriptive statistic. We would be interested, however, not only in those 1326 people but in the *entire population* of all adults in the United States.

Inferential statistics use the sample data to generate a prediction about the entire population. An inferential method presented in Chapter 5 predicts that the population percentage that believe in heaven falls between 83% and 87%. That is, the sample value of 85% has a "margin of error" of 2%. Even though the sample size was tiny compared to the population size, we can conclude that a large percentage of the population believed in heaven. ∎

Inferential statistical analyses can predict characteristics of populations well by selecting samples that are small relative to the population size. That's why many polls sample only about a thousand people, even if the population has millions of people. In this book, we'll see why this works.

In recent years, social scientists have increasingly recognized the power of inferential statistical methods. Presentation of these methods occupies a large portion of this textbook, beginning in Chapter 4.

PARAMETERS AND STATISTICS

A descriptive statistic is a numerical summary of the sample data. The corresponding numerical summary for the population is called a **parameter**.

Parameter	A ***parameter*** is a numerical summary of the population.

Example 1.2 estimated the percentage of Americans who believe in heaven. The parameter was the population percentage who believed in heaven. Its value was unknown. The inference about this parameter was based on a descriptive statistic—the percentage of the 1326 subjects interviewed in the survey who answered *yes*, namely, 85%.

In practice, our main interest is in the values of the *parameters*, not merely the values of the *statistics* for the particular sample selected. For example, in viewing results of a poll before an election, we're more interested in the *population* percentages favoring the various candidates than in the *sample* percentages for the people interviewed. The sample and statistics describing it are important only insofar as they help us make inferences about unknown population parameters.

An important aspect of statistical inference involves reporting the likely *precision* of the sample statistic that estimates the population parameter. For Example 1.2 on belief in heaven, an inferential statistical method predicted how close the *sample* value of 85% was likely to be to the unknown percentage of the *population* believing in heaven. The reported margin of error was 2%.

When data exist for an entire population, such as in a census, it's possible to find the values of the parameters of interest. Then, there is no need to use inferential statistical methods.

DEFINING POPULATIONS: ACTUAL AND CONCEPTUAL

Usually the population to which inferences apply is an actual set of subjects, such as all adult residents of the United States. Sometimes, though, the generalizations refer to a *conceptual* population—one that does not actually exist but is hypothetical.

For example, suppose a medical research team investigates a newly proposed drug for treating lung cancer by conducting a study at several medical centers. Such a medical study is called a *clinical trial*. The conditions compared in a clinical trial or other experiment are called *treatments*. Basic descriptive statistics compare lung cancer patients who are given the new treatment to other lung cancer patients who instead receive a standard treatment, using the percentages who respond positively to the two treatments. In applying inferential statistical methods, the researchers would ideally like inferences to refer to the conceptual population of *all* people suffering from lung cancer now or at some time in the future.

1.3 The Role of Computers and Software in Statistics

Over time, powerful and easy-to-use software has been developed for implementing statistical methods. This software provides an enormous boon to the use of statistics.

STATISTICAL SOFTWARE

Statistical software packages include R, SPSS,[1] SAS,[2] and Stata. Appendix A explains how to use them, organized by chapter. You can refer to Appendix A for the software used in your course as you read each chapter, to learn how to implement the analyses of that chapter. It is much easier to apply statistical methods using software

---

[1] Originally, this was an acronym for *Statistical Package for the Social Sciences.*
[2] Originally, this was an acronym for *Statistical Analysis System.*

than using hand calculation. Moreover, many methods presented in this text are too complex to do by hand or with hand calculators. Software relieves us of computational drudgery and helps us focus on the proper application and interpretation of the statistical methods.

Many examples in this text also show output obtained by using statistical software. One purpose of this textbook is to teach you what to look for in output and how to interpret it. Knowledge of computer programming is not necessary for using statistical software.

DATA FILES

Statistical software analyzes data organized in the spreadsheet form of a ***data file***:

- Any one row contains the observations for a particular subject (e.g., person) in the sample.
- Any one column contains the observations for a particular characteristic.

Figure 1.1 shows an example of a data file, in the form of a window for editing data using Stata software. It shows data for the first 10 subjects in a sample, for the characteristics sex, racial group, marital status, age, and annual income (in thousands of dollars). Some of the data are numerical, and some consist of labels. Chapter 2 introduces the types of data for data files.

FIGURE 1.1: Example of Part of a Stata Data File
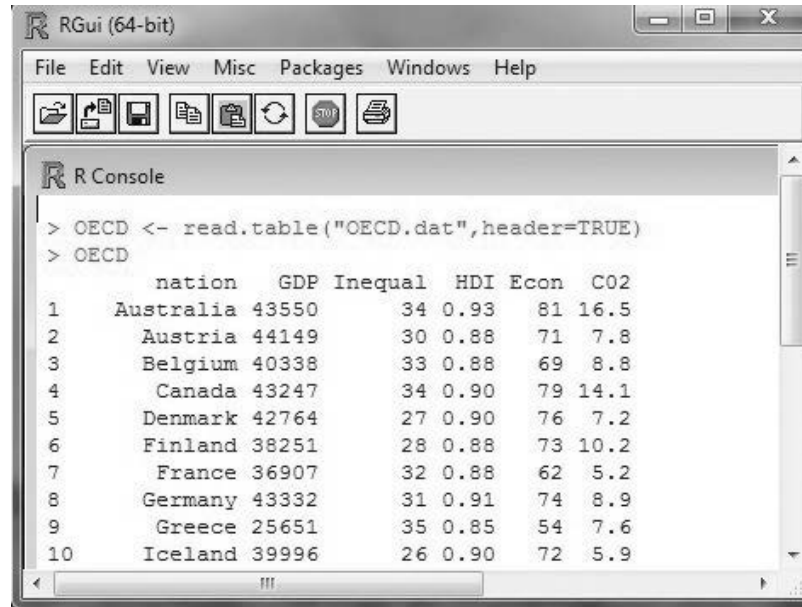
R is a software package that is increasingly popular, partly because it is available to download for free at `www.r-project.org`. Figure 1.2 shows part of an R session for loading a data file called `OECD.dat` from a PC directory and displaying it.

USES AND MISUSES OF STATISTICAL SOFTWARE

A note of caution: The easy access to statistical methods using software has dangers as well as benefits. It is simple to apply inappropriate methods. A computer performs the analysis requested whether or not the assumptions required for its proper use are satisfied.

Incorrect analyses result when researchers take insufficient time to understand the statistical method, the assumptions for its use, or its appropriateness for the

FIGURE 1.2: Example of Part of an R Session for Loading and Displaying Data. The full data file is in Table 3.13 on page 70.

specific problem. It is vital to understand the method before using it. Just knowing how to use statistical software does not guarantee a proper analysis. You'll need a good background in statistics to understand which method to select, which options to choose in that method, and how to make valid conclusions from the output. The purpose of this text is to give you this background.

1.4 Chapter Summary

The field of statistical science includes methods for

- designing research studies,
- describing the data (***descriptive statistics***), and
- making predictions using the data (***inferential statistics***).

Statistical methods apply to observations in a ***sample*** taken from a ***population***. ***Statistics*** summarize sample data, while ***parameters*** summarize entire populations.

- ***Descriptive statistics*** summarize sample or population data with numbers, tables, and graphs.
- ***Inferential statistics*** use sample data to make predictions about population parameters.

A ***data file*** has a separate row of data for each subject and a separate column for each characteristic. Software applies statistical methods to data files.

Exercises

Practicing the Basics

1.1. A medical university conducts an annual national survey of cancer patients who are in remission about their lifestyle habits. In 2016, 1764 patients were surveyed. Identify the **(a)** subject, **(b)** sample, and **(c)** population.

1.2. In 2015, a French national survey asked adults about marriage and divorce. Of the 1754 individuals surveyed, 43% reported that they were married. Of the entire adult French population, 41% were married.

(a) What was the population and what was the sample?

(b) Identify a statistic and a parameter.

1.3. A lecturer at an arts college wants to research social media usage among students. The lecturer sends out an online questionnaire to the entire student body of 2,234 students. Out of this, only 127 students respond. One question asked is, "How many hours per day do you spend on social media?"

(a) Identify the population of interest.

(b) For the 2,234 students, one characteristic of interest was the percentage who would respond "zero" to this question. This value is computed for the 127 students who responded. Is it a parameter or a statistic? Why?

1.4. In January 2016, an educational polling organization in the United Kingdom surveyed a sample of 1156 high school students asking, "Do you plan to attend university?" The response percentages were 42% ("Yes"), 27% ("No"), and 29% ("Undecided"). Are these values descriptive statistics or inferential statistics? Why?

1.5. A social survey agency in Norway conducted a study on a sample of 1520 subjects and asked whether eating a vegan diet reduces chances of suffering a stroke. Out of 1520 sampled subjects, 761 responded *probably true*, and 759 responded *probably not true*. The proportion responding *probably not true* was 759/1520 = 0.499.

(a) Describe the population that may not be of interest to this exercise.

(b) Which population parameter cannot be inferred from the sample statistic given?

(c) What is the proportion of people who responded *probably true*?

(d) Will the value of the population parameter be the same as the sample statistic? Explain.

1.6. Go to the General Social Survey website, `sda.berkeley.edu/GSS`. By entering TVHOURS in the *Row* box, find a summary of responses to the question "On a typical day, about how many hours do you personally watch television?"

(a) What was the most common response?

(b) Is your answer in (a) a descriptive statistic or an inferential statistic? Why?

1.7. At the General Social Survey website, `sda.berkeley.edu/GSS`, by entering HEAVEN in the *Row* box, you can find the percentages of people who said *yes, definitely*; *yes, probably*; *no, probably not*; and *no, definitely not* when asked whether they believed in heaven.

(a) Report the percentage who gave one of the *yes* responses.

(b) To obtain data for a particular year such as 2008, enter YEAR(2008) in the *Selection filter* option box before you click on *Run the table*. Do this for HEAVEN in 2008, and report the percentage who gave one of the *yes* responses. (The GSS asked this question only in 1991, 1998, and 2008.)

(c) Summarize opinions in 2008 about belief in hell (characteristic HELL in the GSS). Was the percentage of *yes* responses higher for HEAVEN or for HELL?

1.8. A private polling organization conducts a national study on physical exercise. The reported responses for 1,507 individuals surveyed were as follows: 21.7% exercised daily, 39.1% exercised two to three times per week, 21.4% exercised four to six times per week, and 17.8% reported being sedentary.

(a) Are these numbers statistics or parameters? Explain.

(b) A method from this text predicts that the percentage of the entire national population that is sedentary is at least 12% but no greater than 18%. What type of statistical method does this illustrate—descriptive or inferential? Why?

1.9. A multicountry poll in 2009 asked if people were willing to bear higher prices for energy and other goods to take steps to fight climate change. The percentages of people who responded by saying *yes* in various countries were 68% (China), 59% (Vietnam), 53% (Japan), 51% (Iran), and 51% (Mexico). The results indicate that majority of people in these five nations were willing to bear higher prices. The results of this poll are an example of (select one)

(a) inferential statistics.

(b) descriptive and inferential statistics.

(c) descriptive statistics.

(d) design of a study.

1.10. Using R (statistical software), create a data file to display the characteristics of five pet dogs. The characteristics measured were breed (Pomeranian, Apso, Dachshund, Labrador, Pug), color (White, Gray, Black, Brown, Gray), age (2, 2, 2, 3, 5), energy level (High, Moderate, Moderate, High, High), and Gender (Male, Male, Male, Female, Female).

Concepts and Applications

1.11. The `Students` data file at the text websites `www.pearsonglobaleditions.com/Agresti` shows responses of a class of social science graduate students at the University of Florida to a questionnaire that asked about *GENDER* (1 = female, 0 = male), *AGE*, *HSGPA* = high school GPA (on a four-point scale), *COGPA* = college GPA, *DHOME* = distance (in miles) of the campus from your home town, *DRES* = distance (in miles) of the classroom from your current residence, *TV* = average number of hours per week that you watch TV, *SPORT* = average number of hours per week that you participate in sports or have other physical exercise, *NEWS* = number of times a week you read a newspaper, *AIDS* = number of people you know who have died from AIDS or who are HIV+, *VEG* = whether you are a vegetarian (1 = yes, 0 = no), *AFFIL* = political affiliation (1 = Democrat, 2 = Republican, 3 = independent), *IDEOL* = political ideology (1 = very liberal, 2 = liberal, 3 = slightly

liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative), *RELIG* = how often you attend religious services (0 = never, 1 = occasionally, 2 = most weeks, 3 = every week), *ABOR* = opinion about whether abortion should be legal in the first three months of pregnancy (1 = yes, 0 = no), *AFFIRM* = support affirmative action (1 = yes, 0 = no), and *LIFE* = belief in life after death (1 = yes, 2 = no, 3 = undecided). You will use this data file for exercises in later chapters.

(a) Practice accessing a data file for statistical analysis with your software by going to this website and copying this data file. Print a copy of the data file. How many observations (rows) are there in the data file?

(b) Give an example of a question that could be addressed using these data with (i) descriptive statistics, (ii) inferential statistics.

1.12. Using statistical software or a spreadsheet program (such as Microsoft Excel), your instructor will help the class create a data file consisting of the values for class members of characteristics such as those in the previous exercise. One exercise in each chapter will use this data file.

(a) Copy the data file to your computer and print a copy.

(b) Give an example of a question that you could address by analyzing these data with (i) descriptive statistics, (ii) inferential statistics.

1.13. For the statistical software your instructor uses for your course, find out how to access the software, enter data, and print any data files that you create. Create a data file using the data in Figure 1.1 on page 19, and print it.

1.14. Illustrating with an example, explain

(a) the difference between a *sample* and a *population*.

(b) design of a research study.

1.15. A researcher in his study finds that a particular drug is very effective in alleviating symptoms of anxiety. The researcher would like to infer the results to the conceptual population. What is the conceptual population in this case?

1.16. A sociologist wants to estimate the average age at marriage for women in New England in the early eighteenth century. She finds within her state archives marriage records for a large Puritan village for the years 1700–1730. She then takes a sample of those records, noting the age of the bride for each. The average age in the sample is 24.1 years. Using a statistical method from Chapter 5, the sociologist estimates the average age of brides at marriage for the population to be between 23.5 and 24.7 years.

(a) What part of this example is descriptive?

(b) What part of this example is inferential?

(c) To what population does the inference refer?

1.17. In a recent public opinion poll conducted by a survey agency on the impact of globalization on the development of nations, a stark difference was found in opinion of various nations. The majority of nations that agreed on a positive impact of globalization included Bangladesh (59%), Vietnam (76%), Indonesia (54%), and Pakistan (61%). The nations that didn't see globalization as favorable included Italy (87%), Germany (82%), China (66%), and South Korea (61%). Of the 2370 subjects interviewed in Bangladesh, 59% agreed. It was predicted that for all the citizens of Bangladesh, the percentage who agreed falls between 56% and 62%. Identify in this discussion **(a)** a statistic, **(b)** a parameter, **(c)** a descriptive analysis, and **(d)** an inferential analysis.

1.18. At the homepage www.gallup.com for the Gallup Poll, from the information listed or linked, give an example of **(a)** sample in a study, and **(b)** population of interest.

1.19. Browse through electronic journals of your choice, such as, *Journal of Experimental Psychology*, *Journal of Educational and Behavioral Statistics*, *Journal of Social Psychology*, etc. Pick an article that interests you. Make a presentation to explain the gist of the article in terms of the design of the study, sample, population, statistical methods used, and inferences drawn at the end.

SAMPLING AND MEASUREMENT

To analyze social phenomena with a statistical analysis, *descriptive* methods summarize the data and *inferential* methods use sample data to make predictions about populations. In gathering data, we must decide which subjects to sample. (Recall that the *subjects* of a population to be sampled could be individuals, families, schools, cities, hospitals, records of reported crimes, and so on.) Selecting a sample that is representative of the population is a primary topic of this chapter.

For our sample, we must convert our ideas about social phenomena into data by deciding what to measure and how to measure it. Developing ways to measure abstract concepts such as performance, achievement, intelligence, and prejudice is one of the most challenging aspects of social research. A measure should have *validity*, describing what it is intended to measure and accurately reflecting the concept. It should also have *reliability*, being consistent in the sense that a subject will give the same response when asked again. Invalid or unreliable data-gathering instruments render statistical analyses of the data meaningless and even possibly misleading.

The first section of this chapter introduces definitions pertaining to measurement, such as types of data. The other sections discuss ways, good and bad, of selecting the sample.

2.1 Variables and Their Measurement

Statistical methods help us determine the factors that explain *variability* among subjects. For instance, variation occurs from student to student in their college grade point average (GPA). What is responsible for that variability? The way those students vary in how much they study per week? How much they watch TV per day? Their IQ? Their college board scores? Their high school GPA?

VARIABLES

Any characteristic that we can measure for each subject is called a ***variable***. The name reflects that values of the characteristic *vary* among subjects.

Variable   A ***variable*** is a characteristic that can vary in value among subjects in a sample or population.

Different subjects may have different values of a variable. Examples of variables are income last year, number of siblings, whether employed, and gender. The values the variable can take form the ***measurement scale***. For gender, for instance, the measurement scale consists of the two labels, (female, male). For number of siblings, it is (0, 1, 2, 3, 4, and so on).

The valid statistical methods for a variable depend on its measurement scale. We treat a numerical-valued variable such as annual income differently from a variable with a measurement scale consisting of categories, such as (yes, no) for whether

employed. We next present ways to classify variables. The first type refers to whether the measurement scale consists of categories or numbers. Another type refers to the number of levels in that scale.

QUANTITATIVE VARIABLES AND CATEGORICAL VARIABLES

A variable is called **quantitative** when the measurement scale has numerical values that represent different magnitudes of the variable. Examples of quantitative variables are a subject's annual income, number of siblings, age, and number of years of education completed.

A variable is called **categorical** when the measurement scale is a set of categories. For example, marital status, with categories (single, married, divorced, widowed), is categorical. For Canadians, the province of residence is categorical, with the categories (Alberta, British Columbia, and so on). Other categorical variables are whether employed (yes, no), primary clothes shopping destination (local mall, local downtown, Internet, other), favorite type of music (classical, country, folk, jazz, rock), religious affiliation (Christianity, Islam, Hinduism, Buddhism, Jewish, other, none), and political party preference.

For categorical variables, distinct categories differ in quality, not in numerical magnitude. Categorical variables are often called **qualitative**. We distinguish between categorical and quantitative variables because different statistical methods apply to each type. For example, the *average* is a statistical summary for quantitative variables, because it uses numerical values. It's possible to find the average for a quantitative variable such as income, but not for a categorical variable such as favorite type of music.

NOMINAL, ORDINAL, AND INTERVAL SCALES OF MEASUREMENT

For a quantitative variable, the possible numerical values are said to form an **interval** scale, because they have a numerical distance or *interval* between each pair of levels. For annual income, for instance, the interval between $40,000 and $30,000 equals $10,000. We can compare outcomes in terms of how much larger or how much smaller one is than the other.

Categorical variables have two types of scales. For the categorical variables mentioned in the previous subsection, such as religious affiliation, the categories are *unordered*. The scale does not have a "high" or "low" end. The categories are then said to form a **nominal scale**. For another example, a variable measuring primary mode of transportation to work might use the nominal scale (automobile, bus, subway, bicycle, walking). For a nominal variable, no category is greater than or smaller than any other category. Labels such as "automobile" and "bus" for mode of transportation identify the categories but do not represent different magnitudes. By contrast, each possible value of a quantitative variable is *greater than* or *less than* any other possible value.

A third type of scale falls, in a sense, between nominal and interval. It consists of categorical scales having a natural *ordering* of values. The categories form an **ordinal scale**. Examples are social class (upper, middle, lower), political philosophy (very liberal, slightly liberal, moderate, slightly conservative, very conservative), government spending on the environment (too little, about right, too much), and frequency of religious activity (never, less than once a month, about 1–3 times a month, every week, more than once a week). These scales are not nominal, because the categories are ordered. They are not interval, because there is no defined distance between

levels. For example, a person categorized as very conservative is *more* conservative than a person categorized as slightly conservative, but there is no numerical value for *how much more* conservative that person is.

The scales refer to the actual measurement and not to the phenomena themselves. *Place of residence* may indicate a geographic place name such as a county (nominal), the distance of that place from a point on the globe (interval), the size of the place (interval or ordinal), or other kinds of variables.

QUANTITATIVE ASPECTS OF ORDINAL DATA

Levels of nominal scales are qualitative, varying in quality, not in quantity. Levels of interval scales are quantitative, varying in magnitude. The position of ordinal scales on the quantitative–qualitative classification is fuzzy. Because their scale is a set of categories, they are often analyzed using the same methods as nominal scales. But in many respects, ordinal scales more closely resemble interval scales. They possess an important quantitative feature: Each level has a *greater* or *smaller* magnitude than another level.

Some statistical methods apply specifically to ordinal variables. Often, though, it's helpful to analyze ordinal scales by assigning numerical scores to categories. By treating ordinal variables as interval scale rather than nominal scale, we can use the more powerful methods available for quantitative variables. For example, course grades (such as A, B, C, D, E) are ordinal. But, we treat them as interval when we assign numbers to the grades (such as 4, 3, 2, 1, 0) to compute a grade point average.

DISCRETE AND CONTINUOUS VARIABLES

One other way to classify a variable also helps determine which statistical methods are appropriate for it. This classification refers to the *number* of values in the measurement scale.

Discrete and Continuous Variables

> A variable is ***discrete*** if its possible values form a set of separate numbers, such as (0, 1, 2, 3, …). It is ***continuous*** if it can take an infinite continuum of possible real number values.

An example of a discrete variable is the number of siblings. Any variable phrased as "the number of …" is discrete, because it is possible to list its possible values {0, 1, 2, 3, 4, …}.

Examples of continuous variables are height, weight, and the amount of time it takes to read a passage of a book. It is impossible to write down all the distinct potential values, since they form an interval of infinitely many values. The amount of time needed to read a book, for example, could take the value 8.62944 … hours.

Discrete variables have a basic unit of measurement that cannot be subdivided. For example, 2 and 3 are possible values for the number of siblings, but 2.571 is not. For a continuous variable, by contrast, between any two possible values there is always another possible value. For example, age is continuous in the sense that an individual does not age in discrete jumps. At some well-defined point during the year in which you age from 21 to 22, you are 21.385 years old, and similarly for every other real number between 21 and 22. A continuous, infinite collection of age values occurs between 21 and 22 alone.

Any variable with a finite number of possible values is discrete. Categorical variables, nominal or ordinal, are discrete, having a finite set of categories. Quantitative variables can be discrete or continuous; age is continuous, and number of siblings is discrete.

For quantitative variables, the distinction between discrete and continuous variables can be blurry, because of how variables are actually measured. In practice, we round continuous variables when measuring them, so the measurement is actually discrete. We say that an individual is 21 years old whenever that person's age is somewhere between 21 and 22. On the other hand, some variables, although discrete, have a very large number of possible values. In measuring annual family income in dollars, the potential values are (0, 1, 2, 3, . . .), up to some very large value in many millions.

What's the implication of this? Statistical methods for discrete variables are mainly used for quantitative variables that take relatively few values, such as the number of times a person has been married. Statistical methods for continuous variables are used for quantitative variables that can take lots of values, regardless of whether they are theoretically continuous or discrete. For example, statisticians treat variables such as age, income, and IQ as continuous.

In summary,

- Variables are either *quantitative* (numerical-valued) or *categorical*. Quantitative variables are measured on an *interval* scale. Categorical variables with unordered categories have a *nominal* scale, and categorical variables with ordered categories have an *ordinal* scale.

- Categorical variables (nominal or ordinal) are *discrete*. Quantitative variables can be either discrete or continuous. In practice, quantitative variables that can take lots of values are treated as *continuous*.

Figure 2.1 summarizes the types of variables, in terms of the (quantitative, categorical), (nominal, ordinal, interval), and (continuous, discrete) classifications.

FIGURE 2.1: Summary of Quantitative–Categorical, Nominal–Ordinal–Interval, and Continuous–Discrete Classifications
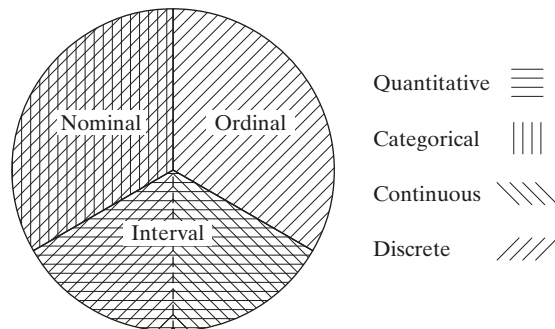
Note: Ordinal data are treated sometimes as categorical and sometimes as quantitative

2.2 Randomization

Inferential statistical methods use sample statistics to make predictions about values of population parameters. The quality of the inferences depends on how well the sample represents the population. This section introduces ***randomization***, the mechanism for achieving good sample representation.

In this section and throughout the text, we let n denote the number of subjects in the sample. This is called the ***sample size***.

SIMPLE RANDOM SAMPLING

Simple random sampling is a method of sampling for which every possible sample of size n has equal chance of selection.

> A **simple random sample** of n subjects from a population is one in which each possible sample of that size has the same probability (chance) of being selected.

For instance, suppose you want to select a simple random sample of a student from a class of 60 students. For a simple random sample of $n = 1$ student, each of the 60 students is equally likely to be selected. You could select one by numbering the students from 01 to 60, placing the 60 numbers on 60 identical ballots, and selecting one blindly from a hat. For a simple random sample of $n = 2$ students from the class, each possible sample of size 2 is equally likely. The potential samples are (01, 02), (01, 03), (01, 04), ..., (59, 60). To select the sample, you blindly select two ballots from the hat. But this is unwieldy if the population size is large, and these days we can easily select the sample using a *random number generator* with software.

A simple random sample is often just called a **random sample**. The *simple* adjective is used to distinguish this type of sampling from more complex sampling schemes presented in Section 2.4 that also have elements of randomization.

Why is it a good idea to use random sampling? Because everyone has the same chance of inclusion in the sample, so it provides fairness. This reduces the chance that the sample is seriously biased in some way, leading to inaccurate inferences about the population. Most inferential statistical methods assume randomization of the sort provided by random sampling.

HOW TO SELECT A SIMPLE RANDOM SAMPLE?

To select a random sample, we need a list of all subjects in the population. This list is called the **sampling frame**. Suppose you plan to sample students at your school. The population is all students at the school. One possible sampling frame is the student directory.

The most common method for selecting a random sample is to (1) number the subjects in the sampling frame, (2) generate a set of these numbers randomly, and (3) sample the subjects whose numbers were generated. Using *random numbers* to select the sample ensures that each subject has an equal chance of selection.

> **Random numbers** are numbers that are computer generated according to a scheme whereby each digit is equally likely to be any of the integers $0, 1, 2, \ldots, 9$ and does not depend on the other digits generated.

Table 2.1 shows a table containing random numbers, in sets of size 5. The numbers fluctuate according to no set pattern. Any particular number has the same chance of being a 0, 1, 2, ..., or 9. The numbers are chosen independently, so any one digit chosen has no influence on any other selection. If the first digit in a row of the table is a 9, for instance, the next digit is still just as likely to be a 9 as a 0 or 1 or any other number.

TABLE 2.1: A Table of Random Numbers

Line/Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	90826	68432	36255	32536	92103	76082	82293	78852
2	77714	33924	86688	94720	45943	83064	68007	10523
3	34371	53100	81078	34696	92393	92799	72281	62696

Source: Constructed using `sample` function in R.

Although random numbers are available in published tables, we can easily generate them with software and many statistical calculators. For example, suppose you want to randomly select $n = 4$ students out of a class of size 60. After assigning the numbers $(01, 02, \ldots, 60)$ to the class members, you can use software to generate four random numbers between 01 and 60. R is a software package that can do this. It is available to download for free at `www.r-project.org`. In R, the `sample` function performs simple random sampling from a numbered population list. Here is how to select a sample of size 4 from a population of size 60 (the $>$ is the R system prompt, and you type in *sample(1:60, 4)* and press the *enter* key on your keyboard):

```
> sample(1:60, 4) # put comments on command line after the # symbol
[1] 22 47 38 44   # these are the four numbers randomly generated
```

The sample of size 4 selects the students numbered 22, 47, 38, and 44.

COLLECTING DATA WITH SAMPLE SURVEYS

Many studies select a sample of people from a population and interview them. This method of data collection is called a ***sample survey***. The interview could be a personal interview, telephone interview, or self-administered questionnaire.

The General Social Survey (GSS) is an example of a sample survey. It gathers information using personal interviews of a random sample of subjects from the U.S. adult population to provide a snapshot of that population. (They do not use *simple* random sampling but rather a method discussed later in the chapter that incorporates multiple stages and clustering but is designed to give each family the same chance of inclusion.) National polls such as the Gallup Poll are also sample surveys. They usually use telephone interviews. Since it is often difficult to obtain a sampling frame, especially since many people now have cell phones but not landline phones, many telephone interviews obtain the sample with *random digit dialing*.

COLLECTING DATA WITH AN EXPERIMENT

In some studies, data result from a planned ***experiment***. The purpose of most experiments is to compare responses of subjects on some outcome measure, under different conditions. Those conditions are levels of a variable that can influence the outcome. The scientist has the experimental control of being able to assign subjects to the conditions.

The conditions in an experiment are called ***treatments***. For instance, the treatments might be different drugs for treating some illness. To conduct the experiment, the researcher needs a plan for assigning subjects to the treatments. These plans are called *experimental designs*. Good experimental designs use randomization to determine which treatment a subject receives. This reduces bias and allows us to use statistical inference to make predictions.

In the late 1980s, the Physicians' Health Study Research Group at Harvard Medical School designed an experiment to analyze whether regular intake of aspirin reduces mortality from heart disease. Of about 22,000 male physicians, half were randomly chosen to take an aspirin every other day. The remaining half took a placebo, which had no active agent. After five years, rates of heart attack were compared. By using randomization to determine who received which treatment, the researchers knew the groups would roughly balance on factors that could affect heart attack rates, such as age and quality of health. If the physicians could decide on their own

which treatment to take, the groups might have been out of balance on some important factor. Suppose, for instance, younger physicians were more likely to select aspirin. Then, a lower heart attack rate among the aspirin group could occur merely because younger subjects are less likely to suffer heart attacks.

In medicine, experiments using randomization (so-called *randomized clinical trials*) have been the gold standard for many years. But randomized experiments are also increasingly used in the social sciences. For example, researchers use randomized experiments to evaluate programs for addressing poverty in the developing world. For many examples, see the websites

`www.povertyactionlab.org/methodology` and `www.nature.com/news`,

at the latter site searching for the article "Can randomized trials eliminate global poverty?" (by J. Tollefson, August 12, 2015).

COLLECTING DATA WITH AN OBSERVATIONAL STUDY

In social research, it is often not feasible to conduct experiments. It's usually not possible to randomly assign subjects to the groups we want to compare, such as levels of gender or race or educational level or annual income. Many studies, such as sample surveys, merely *observe* the outcomes for available subjects on the variables without any experimental manipulation of the subjects. Such studies are called **observational studies**. The researcher measures subjects' responses on the variables of interest but has no experimental control over the subjects.

With observational studies, comparing groups is difficult because the groups may be imbalanced on variables that affect the outcome. This is true even with random sampling. For instance, suppose we plan to compare black students, Hispanic students, and white students on some standardized exam. If white students have a higher average score, a variety of variables might account for that difference, such as parents' education or parents' income or quality of school attended. This makes it difficult to compare groups with observational studies, especially when some key variables may not have been measured in the study.

Establishing cause and effect is central to science. But it is not possible to establish cause and effect definitively with a nonexperimental study, whether it be an observational study with an available sample or a sample survey using random sampling. An observational study always has the possibility that some unmeasured variable could be responsible for patterns observed in the data. By contrast, with an experiment that randomly assigns subjects to treatments, those treatments should roughly balance on any unmeasured variables. For example, in the aspirin and heart attack study mentioned above, the doctors taking aspirin would not tend to be younger or of better health than the doctors taking the placebo. Because a randomized experiment balances the groups being compared on other factors, we can use it to study cause and effect.

2.3 Sampling Variability and Potential Bias

Even if a study wisely uses randomization, the results of the study still depend on which subjects are sampled. Two researchers who separately select random samples from some population may have little overlap, if any, between the two sample memberships. Therefore, the values of sample statistics will differ for the two samples, and the results of analyses based on these samples may differ.

SAMPLING ERROR

Suppose the Gallup, Harris, Ipsos-Reid, and Pew polling organizations each randomly sample 1000 adult Canadians, in order to estimate the percentage of Canadians who give the prime minister's performance in office a favorable rating. Based on the samples they select, perhaps Gallup reports an approval rating of 53%, Harris reports 58%, Ipsos-Reid 55%, and Pew 54%. These differences could reflect slightly different question wording. But even if the questions are worded exactly the same, the percentages would probably differ somewhat because the samples are different.

For conclusions based on statistical inference to be worthwhile, we should know the potential *sampling error*—how much the statistic differs from the parameter it predicts because of the way results naturally exhibit variation from sample to sample.

Sampling Error

> The *sampling error* of a statistic is the error that occurs when we use a statistic based on a sample to predict the value of a population parameter.

Suppose that the percentage of the population of adult Canadians who give the prime minister a favorable rating is 56%. Then the Gallup organization, which predicted 53%, had a sampling error of 53% − 56% = −3%. The Harris organization, which predicted 58%, had a sampling error of 58% − 56% = 2%. In practice, the sampling error is unknown, because the values of population parameters are unknown.

Random sampling protects against bias, in the sense that the sampling error tends to fluctuate about 0, sometimes being positive (as in the Harris Poll) and sometimes being negative (as in the Gallup Poll). Random sampling also allows us to predict the likely size of the sampling error. For sample sizes of about 1000, we'll see that the sampling error for estimating percentages is usually no greater than plus or minus 3%. This bound is the *margin of error*. Variability also occurs in the values of sample statistics with nonrandom sampling, but the extent of the sampling error is not predictable as it is with random sampling.

SAMPLING BIAS: NONPROBABILITY SAMPLING

Other factors besides sampling error can cause results to vary from sample to sample. These factors can also possibly cause bias. We next discuss three types of bias. The first is called *sampling bias*.

For simple random sampling, each possible sample of *n* subjects has the same probability of selection. This is a type of *probability sampling* method, meaning that the probability any particular sample will be selected is known. Inferential statistical methods assume probability sampling. *Nonprobability sampling* methods are ones for which it is not possible to determine the probabilities of the possible samples. Inferences using such samples have unknown reliability and result in *sampling bias*.

The most common nonprobability sampling method is *volunteer sampling*. As the name implies, subjects volunteer for the sample. But the sample may poorly represent the population and yield misleading conclusions. Examples of volunteer sampling are visible daily on Internet sites and television news programs. Viewers register their opinions on an issue by voting over the Internet. The viewers who respond are unlikely to be a representative cross section, but will be those who can easily access the Internet and who feel strongly enough to respond. Individuals having a particular opinion might be much more likely to respond than individuals having a different opinion. For example, one night the ABC TV program *Nightline* asked viewers whether the United Nations should continue to be located in the United

States. Of more than 186,000 respondents, 67% wanted the United Nations out of the United States. At the same time, a poll using a random sample of about 500 respondents estimated the population percentage to be about 28%. Even though the random sample had a much smaller size, it is far more trustworthy.

A large sample does not help with volunteer sampling—the bias remains. In 1936, the newsweekly *Literary Digest* sent over 10 million questionnaires in the mail to predict the outcome of the presidential election. The questionnaires went to a relatively wealthy segment of society (those having autos or telephones), and fewer than 25% were returned. The journal used these to predict an overwhelming victory by Alfred Landon over Franklin Roosevelt. The opposite result was predicted by George Gallup with a much smaller sample in the first scientific poll taken for this purpose. In fact, Roosevelt won in a landslide.

The sampling bias inherent in volunteer sampling is also called ***selection bias***. It is problematic to evaluate policies and programs when individuals can choose whether or not to participate in them. For example, if we were evaluating a program such as Head Start in which participation is partly based on a parental decision, we would need to consider how family background variables (such as mother's educational level) could play a role in that decision and in the outcome evaluated.

Unfortunately, volunteer sampling is sometimes unavoidable, especially in medical studies. Suppose a study plans to investigate how well a new drug performs compared to a standard drug, for subjects who suffer from high blood pressure. The researchers are not going to be able to find a sampling frame of all who suffer from high blood pressure and take a simple random sample of them. They may, however, be able to sample such subjects at certain medical centers or using volunteers. Even then, randomization should be used wherever possible. For the study patients, the researchers can randomly select who receives the new drug and who receives the standard one.

Even with random sampling, sampling bias can occur. One case is when the sampling frame suffers from ***undercoverage***: It lacks representation from some groups in the population. A telephone survey will not reach prison inmates or homeless people, whereas families that have many phones will tend to be over-represented. Responses by those not having a telephone might tend to be quite different from those actually sampled, leading to biased results. About 21% of adults are under age 30, yet only 5% of the population having a landline phone are under age 30, so substantial bias could occur if we sampled only landlines.[1] Likewise there would be bias if we sampled only cell phones, because adults who have only a cell phone tend to be younger, poorer, more likely to be renters, to live with unrelated adults, and to be Hispanic than those who also have a landline phone.

RESPONSE BIAS

In a survey, the way a question is worded or asked can have a large impact on the results. For example, when a New York Times/CBS News poll asked whether the interviewee would be in favor of a new gasoline tax, only 12% said yes. When the tax was presented as reducing U.S. dependence on foreign oil, 55% said yes, and when asked about a gas tax that would help reduce global warming, 59% said yes.[2]

Poorly worded or confusing questions result in ***response bias***. Even the order in which questions are asked can influence the results dramatically. During the Cold War, a study asked, "Do you think the U.S. should let Russian newspaper reporters come here and send back whatever they want?" and "Do you think Russia should let American newspaper reporters come in and send back whatever they want?" The

---

[1] See `http://magazine.amstat.org/blog/2014/10/01/prescolumnoct14`.
[2] Column by T. Friedman, *New York Times*, March 2, 2006.

percentage of yes responses to the first question was 36% when it was asked first and 73% when it was asked second.[3]

In an interview, characteristics of the interviewer may result in response bias. Respondents might lie if they think their belief is socially unacceptable. They may be more likely to give the answer that they think the interviewer prefers. In a study on the effect of the interviewer's race, following a phone interview, respondents were asked whether they thought the interviewer was black or white (all were actually black). Perceiving a white interviewer resulted in more conservative opinions. For example, 14% agreed that "American society is fair to everyone" when they thought the interviewer was black, but 31% agreed to this when they thought the interviewer was white.[4]

NONRESPONSE BIAS: MISSING DATA

Some subjects who are selected for the sample may refuse to participate, or it may not be possible to reach them. This results in **nonresponse bias**. If only half the intended sample was actually observed, we should worry about whether the half not observed differ from those observed in a way that causes biased results. Even if we select the sample randomly, the results are questionable if there is substantial nonresponse, say, over 20%.

For her book *Women in Love*, author Shere Hite surveyed women in the United States. One of her conclusions was that 70% of women who had been married at least five years have extramarital affairs. She based this conclusion on responses to questionnaires returned by 4500 women. This sounds like an impressively large sample. However, the questionnaire was mailed to about 100,000 women. We cannot know whether the 4.5% of the women who responded were representative of the 100,000 who received the questionnaire, much less the entire population of American women. This makes it dangerous to make an inference to the larger population.

A problem in many studies is **missing data**: Some subjects do not provide responses for some of the variables measured. This problem is especially common in studies that observe people over time (called *longitudinal studies*), as some people may drop out of the study for various reasons. Even in censuses, which are designed to observe everyone in a country, some people are not observed or fail to cooperate. A statistical analysis that ignores cases for which some observations are missing wastes information and has possible bias.

SUMMARY OF TYPES OF BIAS

In summary, sample surveys have potential sources of bias:

- **Sampling bias** occurs from using nonprobability samples, such as the *selection bias* inherent in volunteer samples.

- **Response bias** occurs when the subject gives an incorrect response (perhaps lying), or the question wording or the way the interviewer asks the questions is confusing or misleading.

- **Nonresponse bias** occurs when some sampled subjects cannot be reached or refuse to participate or fail to answer some questions.

These sources of bias can also occur in observational studies other than sample surveys and even in experiments. In any study, carefully assess the scope of conclusions. Evaluate critically the conclusions by noting the makeup of the sample. How

---

[3] See Crossen (1994).
[4] *Washington Post*, June 26, 1995.

was the sample selected? How large was it? How were the questions worded or the variables measured? Who sponsored and conducted the research? The less information that is available, the less you should trust it.

Finally, be wary of any study that makes inferences to a broader population than is justified by the sample chosen. Suppose a psychologist performs an experiment using a random sample of students from an introductory psychology course. With statistical inference, the sample results generalize to the population of all students in the class. For the results to be of wider interest, the psychologist might claim that the conclusions extend to *all* college students, to all young adults, or even to all adults. These generalizations may well be wrong, because the sample may differ from those populations in fundamental ways, such as in average age or socioeconomic status.

2.4 Other Probability Sampling Methods*

Section 2.2 introduced **simple random sampling** and explained its importance to statistical inference. In practice, other probability sampling methods that utilize randomization can be simpler to obtain.

SYSTEMATIC RANDOM SAMPLING

Systematic random sampling selects a subject near the beginning of the sampling frame list, skips names and selects another subject, skips names and selects the next subject, and so forth. The number of names skipped at each stage depends on the chosen sample size. Here's how it is done:

Systematic Random Sample

> Denote the sample size by n and the population size by N. Let $k = N/n$, the population size divided by the sample size. A **systematic random sample** (1) selects a subject at random from the first k names in the sampling frame, and (2) selects every kth subject listed after that one. The number k is called the *skip number*.

Suppose you want a systematic random sample of 100 students from a population of 30,000 students listed in a campus directory. Then, $n = 100$ and $N = 30,000$, so $k = 30,000/100 = 300$. The population size is 300 times the sample size, so you need to select one of every 300 students. You select one student at random, using random numbers, from the first 300 students in the directory. Then you select every 300th student after the one selected randomly. This produces a sample of size 100. For example, suppose the random number you choose between 001 and 300 is 104. Then, the numbers of the students selected are 104, 104 + 300 = 404, 404 + 300 = 704, 704 + 300 = 1004, 1004 + 300 = 1304, and so on. The 100th student selected is listed in the last 300 names in the directory.

Systematic random sampling typically provides as good a representation of the population as simple random sampling, because for alphabetic listings such as directories of names, values of most variables fluctuate randomly through the list. With this method, statistical formulas based on simple random sampling are usually valid.

A systematic random sample is not a simple random sample, because all samples of size n are not equally likely. For instance, unlike in a simple random sample, two subjects listed next to each other on the sampling frame list cannot both appear in the sample.

STRATIFIED RANDOM SAMPLING

Another probability sampling method, useful in social science research for studies comparing groups, is **stratified sampling**.

Stratified Random Sample	A ***stratified random sample*** divides the population into separate groups, called ***strata***, and then selects a simple random sample from each stratum.

Suppose a study in Cambridge, Massachusetts, plans to compare the opinions of registered Democrats and registered Republicans about whether government should guarantee health care to all citizens. Stratifying according to political party registration, the study selects a random sample of Democrats and another random sample of Republicans.

Stratified random sampling is called ***proportional*** if the sampled strata proportions are the same as those in the entire population. For example, in the study of opinions about health care, if 90% of registered voters in Cambridge are Democrats and 10% are Republicans, then the sampling is proportional if the sample size for Democrats is nine times the sample size for Republicans.

Stratified random sampling is called ***disproportional*** if the sampled strata proportions differ from the population proportions. This is useful when the population size for a stratum is relatively small. A group that comprises a small part of the population may not have enough representation in a simple random sample to allow precise inferences. It is not possible to compare accurately Republicans to Democrats, for example, if only 10 people in a sample of size 100 are Republican. By contrast, a disproportional stratified sample of size 100 might randomly sample 50 of each party.

To implement stratification, we must know the stratum into which each subject in the sampling frame belongs. This usually restricts the variables that can be used for forming the strata. The variables must have strata that are easily identifiable. For example, it would be easy to select a stratified sample of a school population using grade level as the stratification variable, but it would be difficult to prepare an adequate sampling frame of city households stratified by household income.

CLUSTER SAMPLING

Simple, systematic, and stratified random sampling are often difficult to implement, because they require a complete sampling frame. Such lists are easy to obtain for sampling cities or hospitals or schools, but more difficult for sampling individuals or families. ***Cluster sampling*** is useful when a complete listing of the population is not available.

Cluster Random Sample	Divide the population into a large number of ***clusters***, such as city blocks. Select a simple random sample of the clusters. Use the subjects in those clusters as the sample.

For example, a study might sample about 1% of the families in a city, using city blocks as clusters. Using a map to identify city blocks, it could select a simple random sample of 1% of the blocks and then sample every family on each block. A study of patient care in mental hospitals in Ontario could first sample mental hospitals (the clusters) in that province and then collect data for patients within those hospitals.

What's the difference between a stratified sample and a cluster sample? A stratified sample uses *every* stratum. The strata are usually groups we want to compare. By contrast, a cluster sample uses a *sample* of the clusters, rather than all of them. In cluster sampling, clusters are merely ways of easily identifying groups of subjects. The goal is not to compare the clusters but rather to use them to obtain a sample. Most clusters are not represented in the eventual sample.

Figure 2.2 illustrates the distinction among sampling subjects (simple random sample), sampling clusters of subjects (cluster random sample), and sampling

subjects from within strata (stratified random sample). The figure depicts ways to survey 40 students at a school, to make comparisons among Freshmen, Sophomores, Juniors, and Seniors.

FIGURE 2.2: Ways of Randomly Sampling 40 Students. The figure is a schematic for a simple random sample, a cluster random sample of 8 clusters of students who live together, and a stratified random sample of 10 students from each class (Fresh., Soph., Jnr., Snr.).
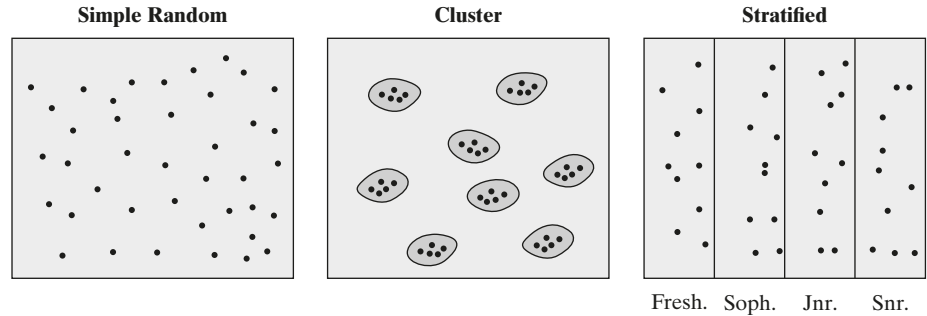
MULTISTAGE SAMPLING

When conducting a survey for predicting elections, the Gallup organization often identifies election districts as clusters and takes a simple random sample of them. But then it also takes a simple random sample of households within each selected election district. This is more feasible than sampling *every* household in the chosen districts. This is an example of **_multistage sampling_**, which uses combinations of sampling methods.

Here is an example of a multistage sample:

- Treat counties (or census tracts) as clusters and select a random sample of a certain number of them.

- Within each county selected, take a cluster random sample of square-block regions.

- Within each region selected, take a systematic random sample of every 10th house.

- Within each house selected, select one adult at random for the sample.

Multistage samples are common in social science research. They are simpler to implement than simple random sampling but provide a broader sampling of the population than a single method such as cluster sampling.

For statistical inference, stratified samples, cluster samples, and multistage samples use different formulas from the ones in this book. Cluster sampling requires a larger sample to achieve as much inferential precision as simple random sampling. Observations within clusters tend to be similar, because of the tendency of subjects living near one another to have similar values on opinion issues and on economic and demographic variables such as age, income, race, and occupation. So, we need more data to obtain a representative cross section. By contrast, the results for stratified sampling may be more precise than those stated in this textbook for simple random sampling. Books specializing in sampling methodology provide further details (Lohr, 2009; Scheaffer et al., 2011; Thompson, 2012).

2.5 Chapter Summary

Statistical methods analyze data on **_variables_**, which are characteristics that vary among subjects. The statistical methods we can use depend on the type of variable:

- Numerically measured variables, such as family income and number of children in a family, are ***quantitative***. They are measured on an *interval scale*.

- Variables taking in a set of categories are ***categorical***. Those measured with unordered categories, such as religious affiliation and province of residence, have a *nominal scale*. Those measured with ordered categories, such as social class and political ideology, have an *ordinal scale* of measurement.

- Variables are also classified as ***discrete***, having possible values that are a set of separate numbers (such as 0, 1, 2, …), or ***continuous***, having a continuous, infinite set of possible values. Categorical variables, whether nominal or ordinal, are discrete. Quantitative variables can be of either type, but in practice are treated as continuous if they can take a large number of values.

Inferential statistical methods require ***probability samples***, which incorporate randomization in some way. Random sampling allows control over the amount of ***sampling error***, which describes how results can vary from sample to sample. Random samples are much more likely to be representative of the population than are nonprobability samples such as volunteer samples.

- For a ***simple random sample***, every possible sample has the same chance of selection.

- Here are other types of probability sampling: ***Systematic*** random sampling takes every *k*th subject in the sampling frame list. ***Stratified*** random sampling divides the population into groups (strata) and takes a random sample from each stratum. ***Cluster*** random sampling takes a random sample of clusters of subjects (such as city blocks) and uses subjects in those clusters as the sample. ***Multistage*** sampling uses combinations of these methods.

Some social science research studies are ***experimental***, with subjects randomly assigned to different treatments that we want to compare. Most studies, such as ***sample surveys***, are ***observational***. They use available subjects in a sample to observe variables of interest, without any experimental control for randomly assigning subjects to groups we want to compare. We need to be very cautious in making causal conclusions based on inferential analyses with data from observational studies.

Chapter 3 introduces statistics for describing samples and corresponding parameters for describing populations. Hence, its focus is on *descriptive statistics*.

Exercises

Practicing the Basics

2.1. Explain the difference between

(a) quantitative and qualitative variables.

(b) interval scale and ordinal scale.

(c) experiment and survey.

2.2. Identify each variable as categorical or quantitative:

(a) Level of education (high school, bachelor's, master's).

(b) Number of times married.

(c) Hair color.

(d) Preferred type of exercise.

(e) Type of residence (rented apartment, owned apartment, rented house, owned house).

(f) Number of household members.

(g) Height.

(h) Field of study (humanities, social sciences, natural sciences).

(i) Number of close friends.

2.3. Which scale of measurement (nominal, ordinal, or interval) is most appropriate for

(a) City of residence?

(b) Number of people you have known who have had lung cancer (0, 1, 2, 3, or more)?

(c) Temperature (Celsius)?

(d) Eye color (brown, black, blue, hazel, etc.)?

(e) Attitude toward paid paternity leave (favor, neutral, oppose)?

(f) Satisfaction with salary (very satisfied, somewhat satisfied, somewhat unsatisfied, very unsatisfied)?

(g) Home ownership (yes/no)?

(h) Rating of cell phone service (awful, poor, neutral, good, great)?

(i) Marital status (married, divorced, single, widowed)?

(j) Number of pets (0, 1, 2, or more)?

(k) Industry of employment (healthcare, technology, education, any other)?

2.4. Which scale of measurement is most appropriate for

(a) Subjects in a course (French, Spanish, German, . . .)?

(b) Relationship status (single, married, divorced, widowed, . . .)?

(c) Time taken per trial (between 0 and 10 seconds)?

(d) Statewide female birth rate (per 1000 population)?

(e) Headcount of employees in a company?

(f) Number of books read in past one year?

(g) Attitude toward climate change (unconcerned, slightly concerned, very concerned)?

(h) College dropout rate (in percentage)?

(i) Mental health status (poor, average, good)?

(j) Score on a job involvement questionnaire (scores between 20 and 100)?

2.5. Which scale of measurement is most appropriate for "personal characteristics of children" measured as

(a) Height and weight of child?

(b) Developmental milestone (rolling over, sitting with support, crawling, . . .)?

(c) Color of eyes (brown, green, blue, . . .)?

2.6. Identify the following variables as categorical or quantitative, and as continuous or discrete.

(a) Gender (male, female)

(b) Number of siblings

(c) Age

(d) IQ

(e) Favorite film genre

(f) Employment status (yes, no)

(g) Annual income

2.7. A study conducted by YouGov (yougov.com) asked people from different countries from around the world, "Which of the following best describes the way you feel about your country?" The response choices were *best in the world, better than most other countries, as good as most other countries, not as good as most other countries*, and *worst in the world.*

(a) What scale of measurement has been used in the study?

(b) What method of data collection has been used in the study?

(c) Thirty-five percent of the Australian sample rated their country as the best in the world as compared to the French sample, where only 5% rated their country as best in the world. Are these values discrete or continuous?

2.8. A researcher decides to use the experimental method to see the effectiveness of a new teaching method on the achievement scores of subjects. Two groups are formed, that is, the experimental group (those who receive training using the new teaching method) and the control group (those who receive training using the usual method). After six months, an achievement test is administered on the subjects in the two groups. The scores on the test range from 0 to 100. What scale of measurement is used to measure the achievement scores of the subjects? Why?

2.9. Identify which of the following variables are discrete and which would be measured on a continuous scale.

(a) Level of education

(b) Favorite type of music

(c) Annual income

(d) Religious affiliation

(e) Number of family members

(f) Finish time for a marathon

(g) Distance traveled to work

2.10. Which of the following variables are both discrete and quantitative?

(a) Number of records sold

(b) Time to swim 200 meters

(c) Occupation

(d) Number of years employed

(e) Age when first married

(f) Student population

(g) Cups of coffee consumed per day

2.11. Select the telephone numbers of at least 30 people you know. Number these telephone numbers from 1 to 30. Using a statistical software, create a table of random numbers and draw a random sample of five telephone numbers from the total list. What were the numbers selected?

2.12. Make a list of the total number of states/districts in your country. Suppose you have to draw a sample of three states randomly. What are the steps involved in drawing a random sample without a statistical software?

2.13. Explain whether you would conduct an experiment or an observational study for the following scenarios. Justify your reasoning.

(a) Whether or not enrollment in a preparatory program improves student retention rates.

(b) Whether public transport or carpooling provides a shorter commute time.

(c) Whether or not a new drug helps individuals manage anxiety.

(d) Whether or not students who participate in music education have higher GPAs.

2.14. A study is planned about whether playing violent video games before the age of 10 has an impact on antisocial behavior.

(a) One possible study would take a sample of children under 10 and randomly select half of them to play violent video games, while the other half would not be exposed. Then 10 years later the researchers would observe whether each child had developed antisocial behavior. Would this study be experimental or observational? Why?

(b) What do you see as possible problems with the study in (a)? How would you redesign the study to make it feasible?

2.15. Table 2.2 shows the result of the 2012 Presidential election and the predictions of several organizations in the days before the election. The sample sizes were typically about 2000. (The percentages for each poll do not sum to 100 because of voters reporting as undecided or favoring another candidate.)

(a) What factors cause the results to vary somewhat among organizations?

(b) Identify the sampling error for the Gallup Poll.

TABLE 2.2

Poll	Predicted Vote	
	Obama	Romney
Gallup	49	50
CNN/Opinion Research	49	49
ABC/Washington Post	50	47
Rasmussen Reports	48	49
NBC/Wall Street Journal	48	47
Pew Research	50	47
Actual vote	51.1	47.2

Source: www.realclearpolitics.com.

2.16. Identify the type of sampling bias in each of the following studies. Explain.

(a) A graduate student is researching TV viewing habits. She posts a link to her questionnaire on her social media sites and receives 46 responses.

(b) Educational researchers want to examine the effectiveness of an after-school enrichment program. The program is optional, with parents deciding whether or not to enroll their children. At the end of the program, researchers test the children participating in the program.

2.17. A national polling organization surveyed 1751 individuals to investigate their opinion on public transportation services. One question in the survey was: "How happy or unhappy are you with current national highway tolls and the national railway service?" When calculating the results, the polling organization had to eliminate the question. Use this example to explain the concept of response bias.

2.18. A farmhouse has a population of 5067 animals. A researcher is keen to study the effect of a drug on the milk production of cows.

(a) How will the researcher select a sample of 10 cows from the farmhouse with a population of 560 cows? Explain the steps using systematic random sampling.

(b) Suppose in the population of 560 cows, 280 cows are black and 280 are white. How will the researcher draw a sample of 100 cows using proportional stratified random sampling?

2.19. You plan to sample from the 6000 students at a college to compare the proportions of men and women who believe that capital punishment should be abolished. Explain how you would proceed if you want a systematic random sample of 200 students.

2.20. You plan to sample from the 15,000 students enrolled at your university to compare the proportions of medical students and law students who believe that social media has a positive impact on their lives.

(a) Suppose that you use random numbers to select students but you stop selecting law students as soon as you have 30, and you stop selecting biology students as soon as you have 30. Is the resulting sample a simple random sample? Why or why not?

(b) What type of sample is the sample in (a)? What is the advantage of using this type of sampling compared to other sampling methods?

2.21. In stratified random sampling, what is the difference between a proportional and a disproportional stratified random sample?

(a) Which sampling method combines different sampling techniques?

(b) Which type of sampling is best when a complete list of population is not available?

Concepts and Applications

2.22. Refer to the Students data file introduced in Exercise 1.11 (page 21). For each variable in the data set, indicate whether it is

(a) Categorical or quantitative.

(b) Nominal, ordinal, or interval.

2.23. Repeat the previous exercise for the data file created in Exercise 1.12 (page 22).

2.24. You are directing a research to study different aspects of the refugee crisis.

(a) Give at least two research problems that you would like to study.

(b) What will be the characteristics of the sample that you wish to study? How will you measure the variables?

(c) What will be your choice of method to draw a sample for your study? Why?

2.25. A researcher wants to study the opinions of people about social media. For this, the researcher takes into consideration all the states/districts of his/her country and calls them clusters. Further, a simple random sample is drawn from the clusters listed. What could be the possible advantage of using this kind of sampling technique? How is this method different from stratified random sampling?

2.26. A social psychology website conducts a survey online using the volunteer sampling method. The survey collects information about a newly launched clothing brand. It asks the subjects, "Would you rate this brand as the best in the market?" Explain briefly what volunteer sampling is. Discuss selection bias and undercoverage as the two outcomes of volunteer sampling.

2.27. In each of the following situations, evaluate whether the method of sample selection is appropriate for obtaining information about the population of interest. How would you improve the sample design?

(a) A TV talent show with a viewership of 2 million wants to find out whether viewers prefer to have a one-hour final episode or a two-hour final episode. Viewers are invited to call in and vote for their preference. A total of 20000 viewers voluntarily call in and 86% indicate a preference for a two-hour finale. Based on the viewers' input, the producers arrange to have a two-hour finale.

(b) A voluntary suggestion box at a museum receives 374 entries. In the suggestion box, one out of every four entries is a complaint about the museum's no-camera policy. The museum's administrators determine that 75% of museum attendees find the no-camera policy acceptable.

(c) A university student council conducts a student satisfaction survey about catering services. Two representatives from the council survey students as they are entering the cafeteria. After surveying 100 students from a total population of 3000, they find that 76% students are satisfied or very satisfied with cafeteria services.

(d) A graduate student e-mails a survey about dieting to the entire student body (3000). She receives 121 responses. After compiling her data, she reports that her data represents a "simple random sample of the student body."

(e) A school district wants to examine the impact of sedentary behavior on academic performance. The total number of students in the district is spread across five schools. The research team lists all of the grade levels, assigning a number to each. Then using a random number table, they select a grade at random and survey every student in that grade about their level of activity and compare this level to their academic performance.

2.28. A researcher wants to analyze the relation between months in a year and the rate of depression among people. The study aims to see whether certain months report more number of cases of depression. The sampling frame consists of number of depression cases for each month for the past five years. What can be a potential problem in using systematic sampling with a skip number 12 or a multiple of 12?

2.29. Why might a researcher choose to use a systematic random sample? Explain.

2.30. With a total sample size of 150, you want to compare the study techniques of students majoring in engineering to other students at your university. What would be the best way to sample your study? Why?

2.31. Explain with an example the steps involved in multi-stage sampling. How is it different from cluster sampling?

2.32. Look for websites on social surveys on the Internet like the General Social Survey (GSS), World Values Survey, World Economic Forum, etc. Find three studies from different domains and report the following:

(a) The variables used in the studies and their scale of measurement.

(b) The sampling method used.

(c) The result of the study (descriptive or inferential?).

2.33. To sample residents of registered nursing homes in Yorkshire, UK, I construct a list of all nursing homes in the county, which I number from 1 to 110. Beginning randomly, I choose every 10th home on the list, ending up with 11 homes. I then obtain lists of residents from those 11 homes, and I select a simple random sample from each list. What kinds of sampling have I used?

For multiple-choice questions 2.34–2.37, select the best response.

2.34. A simple random sample of size n is one in which

(a) Every nth member is selected from the population.

(b) Each possible sample of size n has the same chance of being selected.

(c) There must be exactly the same proportion of women in the sample as is in the population.

(d) You keep sampling until you have a fixed number of people having various characteristics (e.g., males, females).

(e) A particular minority group member of the population is less likely to be chosen than a particular majority group member.

(f) All of the above.

(g) None of the above.

2.35. If we use random numbers to take a simple random sample of 50 students from the 3500 undergraduate students at the University of Rochester,

(a) It is impossible to get the random number 1111, because it is not a random sequence.

(b) If we get 2001 for the first random number, for the second random number that number is less likely to occur than the other possible four-digit random numbers.

(c) The draw 1234 is no more or less likely than the draw 1111.

(d) Since the sample is random, it is *impossible* that it will be non representative, such as having only females in the sample.

2.36. An analysis[5] of published medical studies involving treatments for heart attacks noted that in the studies having randomization and strong controls for bias, the new therapy provided improved treatment 9% of the time. In studies without randomization or other controls for bias, the new therapy provided improved treatment 58% of the time. For each of the following conclusions, state *true* or *false*.

(a) This result suggests it is better not to use randomization in medical studies, because it is harder to show that new ideas are beneficial.

(b) Many newspaper articles that suggest that a particular food, drug, or environmental agent is harmful or beneficial should be viewed skeptically, unless we learn more about the statistical design and analysis for the study.

(c) This result suggests that you should be skeptical about published results of medical studies that are not randomized, controlled studies.

(d) Controlling for biases, both suspected and unsuspected, is necessary in medical research but not in social research, because the social sciences deal in subjective rather than objective truth.

2.37. Keeping in mind the importance of survey design and wording, you want to investigate the percentage of students who have cheated in exams and assignments. From your random sample of 1478 students from a total student population of 10000, the percentage that admits to cheating is 28%. Why might this percentage be unrepresentative?

(a) Responses can depend greatly on the wording of the questions.

(b) Surveys sample only a small part of the population and can never be trusted.

(c) The sample may not have been randomly selected.

(d) The sample may have had problems with bias resulting from subjects not telling the truth.

2.38. Explain how voluntary surveys may not represent a random sample.

2.39. A national news network asks viewers to call in to participate in a poll about whether or not viewers believe crime has increased over the past five years. About 10000 viewers call in, with 72% reporting that they believe crime has increased. True or false: Since everyone had a chance to call, this was a simple random sample of the national population. Explain. What could make the sample random?

2.40.[*] An interval scale for which ratios are valid is called a *ratio scale*. Such scales have a well-defined 0 point, so, for instance, one can regard the value 20 as twice the quantity of the value 10. Explain why annual income is measured on a ratio scale, but temperature (in Fahrenheit or Centigrade) is not. Is IQ, as a measure of intelligence, a ratio-scale variable? Explain.

---

[5] *Source*: Crossen (1994, p. 168).
[*] Exercises marked with an asterisk are of greater difficulty or else introduce new and optional material.

Descriptive Statistics

W e've seen that statistical methods are *descriptive* or *inferential*. The purpose of descriptive statistics is to summarize data, to make it easier to assimilate the information. This chapter presents basic methods of descriptive statistics.

We first present tables and graphs that describe the data by showing the number of times various outcomes occurred. Quantitative variables also have two key features to describe numerically:

- The **center** of the data—a typical observation.

- The **variability** of the data—the spread around the center.

Most importantly, the **mean** describes the center and the **standard deviation** describes the variability.

The final section introduces descriptive statistics that investigate, for a pair of variables, their **association**—how certain values for one variable may tend to go with certain values of the other. For quantitative variables, the **correlation** describes the strength of the association, and **regression analysis** predicts the value of one variable from a value of the other variable.

3.1 Describing Data with Tables and Graphs

Tables and graphs are useful for all types of data. We'll begin with categorical variables.

RELATIVE FREQUENCIES: CATEGORICAL DATA

For categorical variables, we list the categories and show the number of observations in each category. To make it easier to compare different categories, we also report proportions or percentages in the categories, also called *relative frequencies*. The *proportion* equals the number of observations in a category divided by the total number of observations. It is a number between 0 and 1 that expresses the share of the observations in that category. The *percentage* is the proportion multiplied by 100. The sum of the proportions equals 1.00. The sum of the percentages equals 100.

Example 3.1

Household Structure in the United States   Table 3.1 lists the different types of households in the United States in 2015. Of 116.3 million households, for example, 23.3 million were a married couple with children, for a proportion of $23.3/116.3 = 0.20$.

A percentage is the proportion multiplied by 100. That is, the decimal place is moved two positions to the right. For example, since 0.20 is the proportion of families that are married couples with children, the percentage is $100(0.20) = 20\%$. Table 3.1

TABLE 3.1: U.S. Household Structure, 2015

Type of Family	Number (millions)	Proportion	Percentage (1970)
Married couple with children	23.3	0.20	20 (40)
Married couple, no children	33.7	0.29	29 (30)
Women living alone	17.4	0.15	15 (11)
Men living alone	14.0	0.12	12 (6)
Other family households	20.9	0.18	18 (11)
Other nonfamily households	7.0	0.06	6 (2)
Total	116.3	1.00	100 (100)

Source: U.S. Census Bureau; percentages from 1970 in parentheses.

also shows the percentages (in parentheses) from the year 1970. We see a substantial drop since 1970 in the relative number of married couples with children. ∎

FREQUENCY DISTRIBUTIONS AND BAR GRAPHS: CATEGORICAL DATA

A table, such as Table 3.1, that lists the categories and their numbers of observations is called a *frequency distribution*.

Frequency Distribution

> A *frequency distribution* is a listing of possible values for a variable, together with the number of observations at each value.

When the table shows the proportions or percentages instead of the numbers, it is called a *relative frequency distribution*.

   To more easily get a feel for the data, it is helpful to look at a graph of the frequency distribution. A *bar graph* has a rectangular bar drawn over each category. The height of the bar shows the frequency or relative frequency in that category. Figure 3.1 is a bar graph for the data in Table 3.1. The bars are separated to emphasize that the variable is categorical rather than quantitative. Since household structure is a nominal variable, there is no particular natural order for the bars. The order of presentation for an ordinal variable is the natural ordering of the categories.

FIGURE 3.1: Bar Graph of Relative Frequency Distribution of U.S. Household Types

Another type of graph, the *pie chart*, is a circle having a "slice of the pie" for each category. The size of a slice represents the percentage of observations in the category. A bar graph is more precise than a pie chart for visual comparison of categories with similar relative frequencies.

FREQUENCY DISTRIBUTIONS: QUANTITATIVE DATA

Frequency distributions and graphs also are useful for quantitative variables. The next example illustrates this.

Example 3.2

Statewide Violent Crime Rates   Table 3.2 lists all 50 states in the United States and their 2015 violent crime rates. This rate measures the number of violent crimes in that state per 10,000 population. For instance, if a state had 12,000 violent crimes and a population size of 2,300,000, its violent crime rate was (12,000/2,300,000) × 10,000 = 52. Tables, graphs, and numerical measures help us absorb the information in these data.

TABLE 3.2: List of States with 2015 Violent Crime Rates Measured as Number of Violent Crimes per 10,000 Population

Alabama	43	Louisiana	52	Ohio	29
Alaska	64	Maine	13	Oklahoma	44
Arizona	42	Maryland	47	Oregon	25
Arkansas	46	Massachusetts	41	Pennsylvania	34
California	40	Michigan	45	Rhode Island	26
Colorado	31	Minnesota	23	South Carolina	51
Connecticut	26	Mississippi	27	South Dakota	32
Delaware	49	Missouri	43	Tennessee	59
Florida	47	Montana	25	Texas	41
Georgia	37	Nebraska	26	Utah	22
Hawaii	25	Nevada	60	Vermont	12
Idaho	22	New Hampshire	22	Virginia	20
Illinois	38	New Jersey	29	Washington	29
Indiana	36	New Mexico	61	West Virginia	30
Iowa	27	New York	39	Wisconsin	28
Kansas	34	North Carolina	34	Wyoming	21
Kentucky	21	North Dakota	27		

Source: `www.fbi.gov`; data are in `Crime` data file at text website.

To summarize the data with a frequency distribution, we divide the measurement scale for violent crime rate into a set of intervals and count the number of observations in each interval. Here, we use the intervals {0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69}. Table 3.3 (page 44) shows that considerable variability exists in the violent crime rates.

Table 3.3 also shows the relative frequencies, using proportions and percentages. As with any summary method, we lose some information as the cost of achieving some clarity. The frequency distribution does not show the exact violent crime rates or identify which states have low or high rates. ∎

The intervals of values in frequency distributions are usually of equal width. The width equals 10 in Table 3.3. The intervals should include all possible values of the variable. In addition, any possible value must fit into one and only one interval; that is, they should be ***mutually exclusive***.

TABLE 3.3: Frequency Distribution and Relative Frequency Distribution for Violent Crime Rates			
Violent Crime Rate	Frequency	Proportion	Percentage
0–9	0	0.00	0
10–19	2	0.04	4
20–29	20	0.40	40
30–39	10	0.20	20
40–49	12	0.24	24
50–59	3	0.06	6
60–69	3	0.06	6
Total	50	1.00	100.0

HISTOGRAMS

A graph of a frequency distribution for a quantitative variable is called a ***histogram***. Each interval has a bar over it, with height representing the number of observations in that interval. Figure 3.2 is a histogram for the violent crime rates, as constructed by R software.

FIGURE 3.2: Histogram of Frequencies for Violent Crime Rates
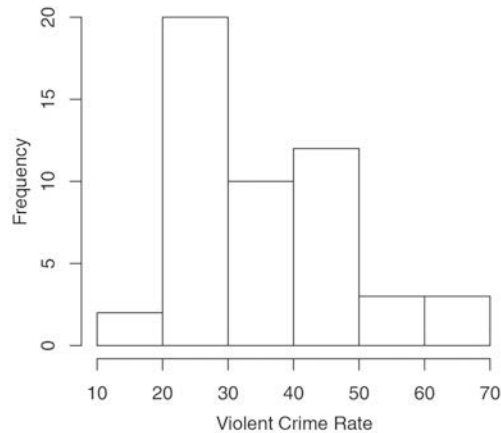
Choosing intervals for frequency distributions and histograms is primarily a matter of common sense. If too few intervals are used, too much information is lost. If too many intervals are used, they are so narrow that the information presented is difficult to digest, and the histogram may be irregular and the overall pattern of the results may be obscured. Ideally, two observations in the same interval should be similar in a practical sense. To summarize annual income, for example, if a difference of $5000 in income is not considered practically important, but a difference of $15,000 is notable, we might choose intervals of width less than $15,000, such as $0–$9999, $10,000–$19,999, $20,000–$29,999, and so forth.

For a discrete variable with relatively few values, a histogram has a separate bar for each possible value. For a continuous variable or a discrete variable with many possible values, you need to divide the possible values into intervals, as we did with the violent crime rates. Statistical software can automatically choose intervals for us and construct frequency distributions and histograms.

STEM-AND-LEAF PLOTS

Figure 3.3 shows an alternative graphical representation of the violent crime rate data. This figure, called a ***stem-and-leaf plot***, represents each observation by its leading digit(s) (the *stem*) and by its final digit (the *leaf*). Each stem is a number to the left of the vertical bar and a leaf is a number to the right of it. For instance, on the first line, the stem of 1 and the leaves of 2 and 3 represent the violent crime rates 12 and 13. The plot arranges the leaves in order on each line, from smallest to largest.

FIGURE 3.3:
Stem-and-Leaf Plot for
Violent Crime Rate Data in
Table 3.2

Stem	Leaf
1	2 3
2	0 1 1 2 2 2 3 5 5 5 6 6 6 7 7 7 8 9 9 9
3	0 1 2 4 4 4 6 7 8 9
4	0 1 1 2 3 3 4 5 6 7 7 9
5	1 2
6	0 1 4

A stem-and-leaf plot conveys information similar to a histogram. Turned on its side, it has the same shape as the histogram. In fact, since the stem-and-leaf plot shows each observation, it displays information that is lost with a histogram. From Figure 3.3, the largest violent crime rate was 64, and the smallest was 12. It is not possible to determine these exact values from the histogram in Figure 3.2.

Stem-and-leaf plots are useful for quick portrayals of small data sets. As the sample size increases, you can accommodate the increase in leaves by splitting the stems. For instance, you can list each stem twice, putting leaves of 0 to 4 on one line and leaves of 5 to 9 on another. When a number has several digits, it is simplest for graphical portrayal to drop the last digit or two. For instance, for a stem-and-leaf plot of annual income in thousands of dollars, a value of $27.1 thousand has a stem of 2 and a leaf of 7 and a value of $106.4 thousand has a stem of 10 and a leaf of 6.

POPULATION DISTRIBUTION AND SAMPLE DATA DISTRIBUTION

Frequency distributions and histograms apply both to a population and to samples from that population. The first type is called the ***population distribution***, and the second type is called a ***sample data distribution***. In a sense, the sample data distribution is a blurry photo of the population distribution. As the sample size increases, the sample proportion in any interval gets closer to the true population proportion. Thus, the sample data distribution looks more like the population distribution.

For a continuous variable, imagine the sample size increasing indefinitely, with the number of intervals simultaneously increasing, so their width narrows. Then, the shape of the sample histogram gradually approaches a smooth curve. This text uses such curves to represent population distributions. Figure 3.4 shows two sample

FIGURE 3.4: Histograms for a Continuous Variable. We use smooth curves to represent population distributions for continuous variables.

a) 100 measurements   b) 500 measurements   c) Population
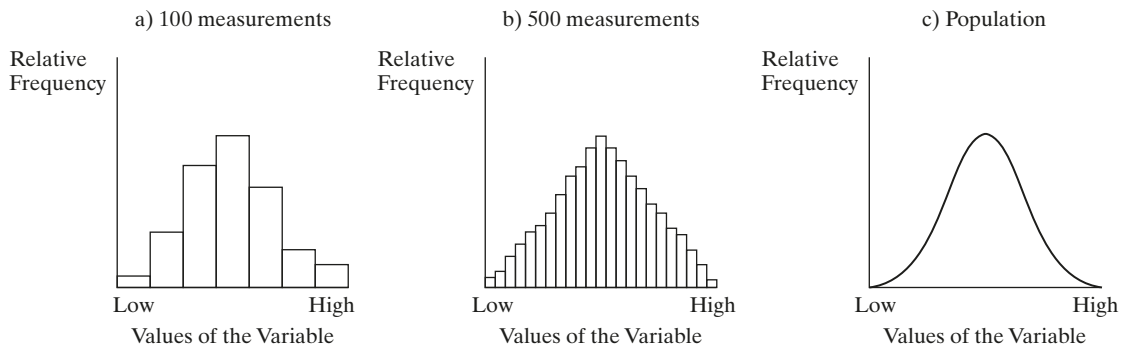
histograms, one for a sample of size 100 and the second for a sample of size 500, and also a smooth curve representing the population distribution. Even if a variable is discrete, a smooth curve often approximates well the population distribution, especially when the number of possible values of the variable is large.

THE SHAPE OF A DISTRIBUTION

Another way to describe a sample or a population distribution is by its shape. A group for which the distribution is bell shaped is fundamentally different from a group for which the distribution is U-shaped, for example. See Figure 3.5. In the U-shaped distribution, the highest points (representing the largest frequencies) are at the lowest and highest scores, whereas in the bell-shaped distribution, the highest point is near the middle value. A U-shaped distribution indicates a polarization on the variable between two sets of subjects. A bell-shaped distribution indicates that most subjects tend to fall near a central value.

FIGURE 3.5: U-Shaped and Bell-Shaped Frequency Distributions
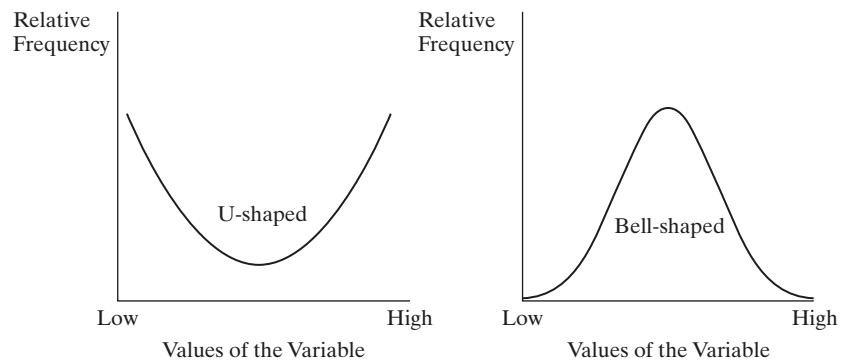
The distributions in Figure 3.5 are **symmetric**: The side of the distribution below a central value is a mirror image of the side above that central value. Most distributions encountered in the social sciences are not symmetric. Figure 3.6 illustrates this. The parts of the curve for the lowest values and the highest values are called the **tails** of the distribution. Often, as in Figure 3.6, one tail is much longer than the other. A distribution is said to be **skewed to the right** or **skewed to the left**, according to which tail is longer.

FIGURE 3.6: Skewed Frequency Distributions. The longer tail indicates the direction of skew.
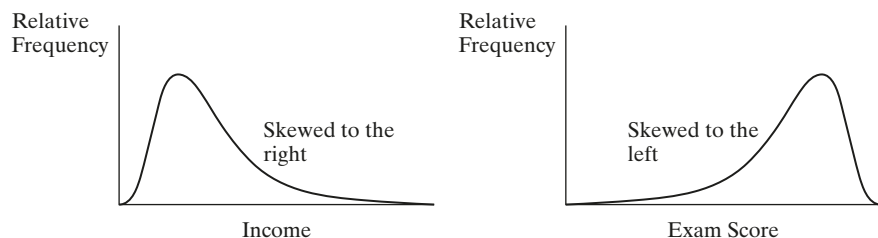
To compare frequency distributions or histograms for two groups, you can give verbal descriptions using characteristics such as skew. It is also helpful to make numerical comparisons such as "On the average, the violent crime rate for Southern states is 5.4 above the violent crime rate for Western states." We next present numerical descriptive statistics.

3.2 Describing the Center of the Data

This section presents statistics that describe the center of a frequency distribution for a quantitative variable. The statistics show what a *typical* observation is like.

THE MEAN

The best known and most commonly used measure of the center is the ***mean***.

Mean

> The ***mean*** is the sum of the observations divided by the number of observations.

The mean is often called the ***average***.

Example 3.3

Female Economic Activity in Europe and Middle East   Table 3.4 shows an index of female economic activity in 2014 for the 10 largest countries (in population) of Western Europe and of the Middle East. The number specifies female employment as a percentage of male employment. In Italy, for instance, the number of females in the work force was 66% of the number of males in the work force.

TABLE 3.4: Female Employment, as a Percentage of Male Employment, in Western Europe and the Middle East

Western Europe		Middle East	
Country	Employment	Country	Employment
Belgium	79	Egypt	29
France	79	Iran	42
Germany	78	Iraq	19
Greece	68	Israel	81
Italy	66	Jordan	39
Netherlands	82	Saudi Arabia	4
Portugal	78	Syria	38
Spain	71	Turkey	34
Sweden	85	United Arab Emirates	49
UK	81	Yemen	40

Source: www.socialwatch.org.

For the 10 observations for Western Europe, the sum equals

$$79 + 79 + 78 + 68 + 66 + 82 + 78 + 71 + 85 + 81 = 767.$$

The mean equals $767/10 = 76.7$. By comparison, you can check that the mean for the 10 Middle Eastern countries equals $375/10 = 37.5$. Female economic activity tends to be considerably lower in the Middle East. ∎

NOTATION FOR OBSERVATIONS, MEAN, AND SUMMATIONS

We use the following notation in formulas for the mean and statistics that use the mean:

Notation for Observations and Sample Mean

> The sample size is symbolized by n. For a variable denoted by y, its observations are denoted by $y_1, y_2, \ldots, y_n$. The sample mean is denoted by $\bar{y}$.

Throughout the text, letters near the end of the alphabet denote variables. The n sample observations on a variable y are denoted by y_1 for the first observation, y_2 for the second, and so forth. For example, for female economic activity in Western Europe, $n = 10$, and the observations are $y_1 = 79, y_2 = 79, \ldots, y_{10} = 81$. The symbol $\bar{y}$ for the sample mean is read as "y-bar." A bar over a letter represents the sample mean for that variable. For instance, $\bar{x}$ represents the sample mean for a variable denoted by x.

The definition of the sample mean says that

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}.$$

The symbol $\sum$ (upper case Greek letter sigma) represents the process of summing. For instance, $\sum y_i$ represents the sum $y_1 + y_2 + \cdots + y_n$. This symbol[1] stands for the sum of the y-values, where the index i represents a typical value in the range 1 to n. To illustrate, for the Western European data,

$$\sum y_i = y_1 + y_2 + \cdots + y_{10} = 79 + 79 + \cdots + 81 = 767.$$

Using this summation symbol, we have the shortened expression for the sample mean of n observations,

$$\bar{y} = \frac{\sum y_i}{n}.$$

The summation operation is sometimes even further abbreviated as $\sum y$.

PROPERTIES OF THE MEAN

Here are some properties of the mean:

- The formula for the mean uses numerical values for the observations. So, the mean is appropriate only for quantitative variables. It is not sensible to compute the mean for observations on a nominal scale. For instance, for religion measured with categories such as (Protestant, Catholic, Muslim, Jewish, Other), the mean religion does not make sense, even though for convenience these levels may be coded in a data file by numbers.

- The mean can be highly influenced by an observation that falls well above or well below the bulk of the data, called an ***outlier***.

  Here is an example illustrating an outlier: The owner of Leonardo's Pizza reports that the mean annual income of full-time employees in the business is $45,900. In fact, the annual incomes of the seven employees are $15,400, $15,600, $15,900, $16,400, $16,400, $16,600, and $225,000. The $225,000 income is the salary of the owner's son, who happens to be an employee. The value $225,000 is an outlier. The mean computed for the other six observations alone equals $16,050, quite different from the mean of $45,900 including the outlier.

- The mean is pulled in the direction of the longer tail of a skewed distribution, relative to most of the data.

  In the Leonardo's Pizza example, the large observation $225,000 results in an extreme skewness to the right of the income distribution. This skewness pulls the mean above six of the seven observations. This example shows that the mean is not always typical of the observations in the sample. The more highly skewed the distribution, the less typical the mean is of the data.

---

[1] You can also formally present the range of observations in the symbol, using $\sum_{i=1}^{n} y_i$ to represent summing y_i while letting i go from 1 to n.

- The mean is the point of balance on the number line when an equal weight is at each observation point.

  For example, Figure 3.7 shows that if we place an equal weight at each Middle Eastern observation on female economic activity from Table 3.4, then the line balances by placing a fulcrum at the point 37.5. The mean is the *center of gravity* (balance point) of the observations: The sum of the distances to the mean from the observations *above* the mean equals the sum of the distances to the mean from the observations *below* the mean.

FIGURE 3.7: The Mean as the Center of Gravity, for Middle Eastern Data from Table 3.4. The line balances with a fulcrum at 37.5.
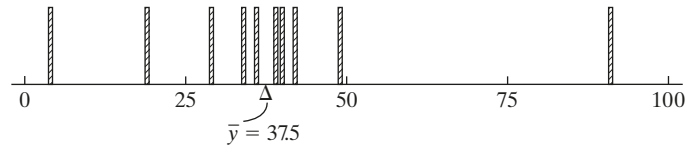
- Denote the sample means for two sets of data with sample sizes n_1 and n_2 by $\bar{y}_1$ and $\bar{y}_2$. The overall sample mean for the combined set of $(n_1 + n_2)$ observations is the **weighted average**

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}.$$

The numerator $n_1\bar{y}_1 + n_2\bar{y}_2$ is the sum of all the observations, since $n\bar{y} = \sum y$ for each set of observations. The denominator is the total sample size.

To illustrate, for the female economic activity data in Table 3.4, the Western European observations have $n_1 = 10$ and $\bar{y}_1 = 76.70$. Canada, the United States, and Mexico have $n_2 = 3$ and values (83, 69, 56), for which $\bar{y}_2 = 69.33$. The overall mean economic activity for the 13 nations equals

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} = \frac{10(76.70) + 3(69.33)}{10 + 3} = \frac{(767 + 208)}{13} = \frac{975}{13} = 75.0.$$

The weighted average of 75.0 is closer to 76.7, the value for Western Europe, than to 69.3, the value for the three North American nations. This happens because more observations are from Western Europe.

THE MEDIAN

The mean is a simple measure of the center. But other measures are also informative and sometimes more appropriate. Most important is the *median*. It splits the sample into two parts with equal numbers of observations, when they are ordered from lowest to highest or from highest to lowest.

Median

> The **median** is the observation that falls in the middle of the ordered sample. When the sample size n is odd, a single observation occurs in the middle. When the sample size is even, two middle observations occur, and the median is the midpoint between the two.

To illustrate, the ordered income observations for the seven employees of Leonardo's Pizza are

$15,400, $15,600, $15,900, $16,400, $16,400, $16,600, $225,000.

The median is the middle observation, $16,400. This is a more typical value for this sample than the sample mean of $45,900. When a distribution is highly skewed, the median describes a typical value better than the mean.

In Table 3.4, the ordered economic activity values for the Western European nations are

$$66, 68, 71, 78, 78, 79, 79, 81, 82, 85.$$

Since $n = 10$ is even, the median is the midpoint between the two middle values, 78 and 79, which is $(78 + 79)/2 = 78.5$. This is close to the sample mean of 76.7, because this data set has no outliers.

The middle observation has the index $(n+1)/2$. That is, the median is the value of observation $(n+1)/2$ in the ordered sample. When $n = 7$, $(n+1)/2 = (7+1)/2 = 4$, so the median is the fourth smallest, or equivalently fourth largest, observation. When n is even, $(n + 1)/2$ falls halfway between two numbers, and the median is the midpoint of the observations with those indices. For example, when $n = 10$, then $(n + 1)/2 = 5.5$, so the median is the midpoint between the fifth and sixth smallest observations.

Example 3.4

Median for Grouped or Ordinal Data Table 3.5 summarizes the distribution of the highest degree completed in the U.S. population of age 25 years and over, as estimated from the 2014 American Community Survey taken by the U.S. Bureau of the Census. The possible responses form an ordinal scale. The population size was $n = 209$ (in millions). The median score is the $(n + 1)/2 = (209 + 1)/2 = 105$th lowest. Now, 24 responses fall in the first category, $(24 + 62) = 86$ in the first two, $(24 + 62 + 35) = 121$ in the first three, and so forth. The 87th to 121st lowest scores fall in category 3, which therefore contains the 105th lowest, which is the median. The median response is "Some college, no degree." Equivalently, from the percentages in the last column of the table, $(11.5\% + 29.7\%) = 41.2\%$ fall in the first two categories and $(11.5\% + 29.7\% + 16.7\%) = 57.9\%$ fall in the first three, so the 50% point falls in the third category. ∎

TABLE 3.5: Highest Degree Completed, for a Sample of Americans		
Highest Degree	Frequency (millions)	Percentage
Not a high school graduate	24	11.5
High school only	62	29.7
Some college, no degree	35	16.7
Associate's degree	21	10.0
Bachelor's degree	42	20.1
Master's degree	18	8.6
Doctorate or professional	7	3.3

PROPERTIES OF THE MEDIAN

- The median, like the mean, is appropriate for quantitative variables. Since it requires only ordered observations to compute it, it is also valid for ordinal-scale data, as the previous example showed. It is not appropriate for nominal-scale data, since the observations cannot be ordered.

- For symmetric distributions, such as in Figure 3.5, the median and the mean are identical. To illustrate, the sample of observations 4, 5, 7, 9, and 10 is symmetric about 7; 5 and 9 fall equally distant from it in opposite directions, as do 4 and 10. Thus, 7 is both the median and the mean.

- For skewed distributions, the mean lies toward the longer tail relative to the median. See Figure 3.8.

FIGURE 3.8: The Mean and the Median for Skewed Distributions. The mean is pulled in the direction of the longer tail.
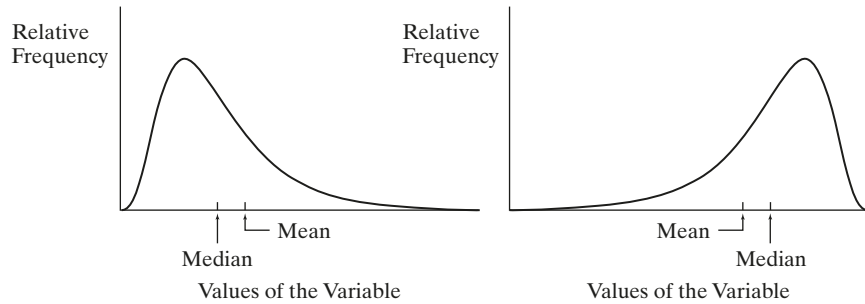
The mean is larger than the median for distributions that are skewed to the right. For example, income distributions are often skewed to the right. Household income in the United States in 2015 had a mean of about $73,000 and a median of about $52,000 (U.S. Bureau of the Census).

The mean is smaller than the median for distributions that are skewed to the left. The distribution of grades on an exam may be skewed to the left when some students perform considerably poorer than the others. For example, suppose that an exam scored on a scale of 0 to 100 has a median of 88 and a mean of 80. Then most students performed quite well (half being over 88), but apparently some scores were very much lower in order to bring the mean down to 80.

- The median is insensitive to the distances of the observations from the middle, since it uses only the ordinal characteristics of the data. For example, the following four sets of observations all have medians of 10:

Set 1:	8,	9,	10,	11,	12
Set 2:	8,	9,	10,	11,	100
Set 3:	0,	9,	10,	10,	10
Set 4:	8,	9,	10,	100,	100

- The median is not affected by outliers. For instance, the incomes of the seven Leonardo's Pizza employees have a median of $16,400 whether the largest observation is $20,000, $225,000, or $2,000,000.

MEDIAN COMPARED TO MEAN

The median is usually more appropriate than the mean when the distribution is very highly skewed, as we observed with the Leonardo's Pizza employee incomes. The mean can be greatly affected by outliers, whereas the median is not.

For the mean we need quantitative (interval-scale) data. The median also applies for ordinal scales. To use the mean for ordinal data, we must assign scores to the categories. In Table 3.5, if we assign scores 10, 12, 13, 14, 16, 18, and 20 to the categories of highest degree, representing approximate number of years of education, we get a sample mean of 13.7.

The median has its own disadvantages. For discrete data that take relatively few values, quite different patterns of data can have the same median. For instance, Table 3.6, from the 2014 General Social Survey, summarizes the responses of the 53 females of age 18–22 to the question "How many sex partners have you had in the last 12 months?" Only six distinct responses occur, and 50.9% of those are 1. The median response is 1. For the sample mean, to sum the 52 observations we multiply each possible value by the frequency of its occurrence, and then add. That is,

$$\sum y_i = 11(0) + 27(1) + 6(2) + 5(3) + 3(4) + 1(5) = 71.$$

The sample mean response is

$$\bar{y} = \frac{\sum y_i}{n} = \frac{71}{53} = 1.34.$$

If the distribution of the 53 observations among these categories were (0, 27, 6, 5, 3, 12) (i.e., we shift the 11 responses from 0 to 5), then the median would still be 1, but the mean would shift to 2.38. The mean uses the numerical values of the observations, not just their ordering.

TABLE 3.6: Number of Sex Partners Last Year, for Female Respondents in GSS of Age 18–22		
Response	Frequency	Percentage
0	11	20.8
1	27	50.9
2	6	11.3
3	5	9.4
4	3	5.7
5	1	1.9

The most extreme form of this problem occurs for **binary data**, which can take only two values, such as 0 and 1. The median equals the more common outcome, but gives no information about the relative number of observations at the two levels. For instance, consider a sample of size 5 for the variable, number of times married. The observations (1, 1, 1, 1, 1) and the observations (0, 0, 1, 1, 1) both have a median of 1. The mean is 1 for (1, 1, 1, 1, 1) and 3/5 for (0, 0, 1, 1, 1).

For binary (0, 1) data, proportion = mean

> When observations take values of only 0 or 1, the mean equals the proportion of observations that equal 1.

Generally, for highly discrete data, the mean is more informative than the median.

In summary,

- If a distribution is highly skewed, the median is better than the mean in representing what is typical.

- If the distribution is close to symmetric or only mildly skewed or if it is discrete with few distinct values, the mean is usually preferred over the median, because it uses the numerical values of all the observations.

THE MODE

Another measure, the *mode*, states the most frequent outcome.

Mode

> The *mode* is the value that occurs most frequently.

The mode is most commonly used with highly discrete variables, such as with categorical data. In Table 3.6 on the number of sex partners in the last year, for instance, the mode is 1, since the frequency for that outcome is higher than the frequency for any other outcome. Here are some properties of the mode:

- The mode is appropriate for all types of data. For example, we might measure the mode for religion in Australia (nominal scale), for the grade given by a teacher (ordinal scale), or for the number of years of education completed by Hispanic Americans (interval scale).

- A frequency distribution is called ***bimodal*** if two distinct mounds occur in the distribution. Bimodal distributions often occur with attitudinal variables when populations are polarized, with responses tending to be strongly in one direction or another. For instance, Figure 3.9 shows the relative frequency distribution of responses in a General Social Survey to the question "Do you personally think it is wrong or not wrong for a woman to have an abortion if the family has a very low income and cannot afford any more children?" The frequencies in the two extreme categories are much higher than those in the middle categories.

FIGURE 3.9: Bimodal Distribution for Opinion about Whether Abortion Is Wrong
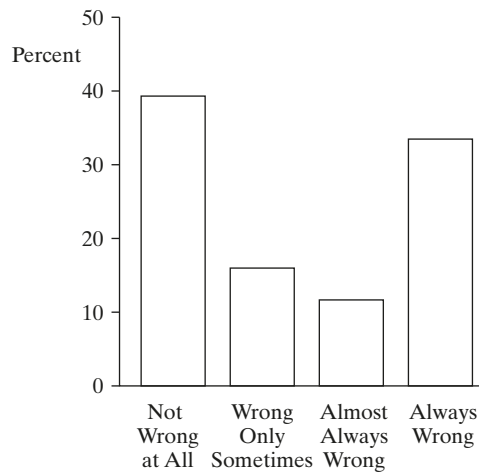
- The mean, median, and mode are identical for a unimodal, symmetric distribution, such as a bell-shaped distribution.

The mean, median, and mode are complementary measures. They describe different aspects of the data. In any particular example, some or all their values may be useful. Be on the lookout for misleading statistical analyses, such as using one statistic when another would be more informative. People who present statistical conclusions often choose the statistic giving the impression they wish to convey. Recall the Leonardo's Pizza employees, with the extreme outlying income observation. Be wary of the mean when the distribution may be highly skewed.

3.3 Describing Variability of the Data

A measure of center alone is not adequate for numerically describing data for a quantitative variable. It describes a typical value, but not the spread of the data about that typical value. The two distributions in Figure 3.10 illustrate this. The citizens of nation A and the citizens of nation B have the same mean annual income ($25,000). The distributions of those incomes differ fundamentally, however, nation B being much less variable. An income of $30,000 is extremely large for nation B, but not especially large for nation A. This section introduces statistics that describe the *variability* of a data set.

FIGURE 3.10:
Distributions with the
Same Mean but Different
Variability
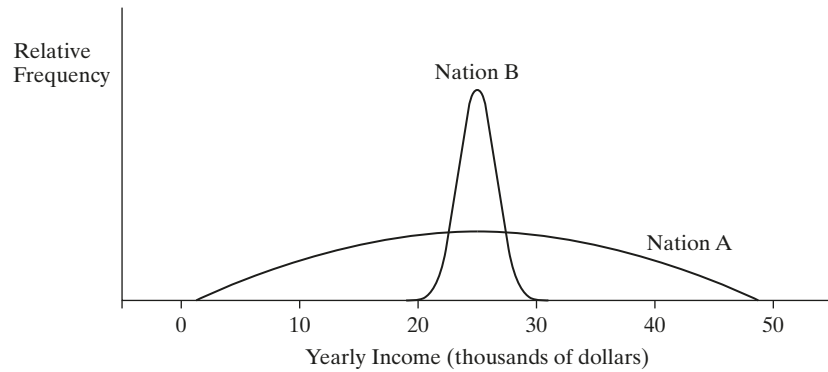
THE RANGE

The difference between the largest and smallest observations is the simplest way to describe variability.

Range

> The *range* is the difference between the largest and smallest observations.

For nation A, from Figure 3.10, the range of income values is about $50,000 − $0 = $50,000. For nation B, the range is about $30,000 − $20,000 = $10,000. Nation A has greater variability of incomes.

The range is not, however, sensitive to other characteristics of data variability. The three distributions in Figure 3.11 all have the same mean ($25,000) and range ($50,000), but they differ in variability about the center. In terms of distances of observations from the mean, nation A has the most variability, and nation B the least. The incomes in nation A tend to be farthest from the mean, and the incomes in nation B tend to be closest.

FIGURE 3.11:
Distributions with the
Same Mean and Range, but
Different Variability about
the Mean
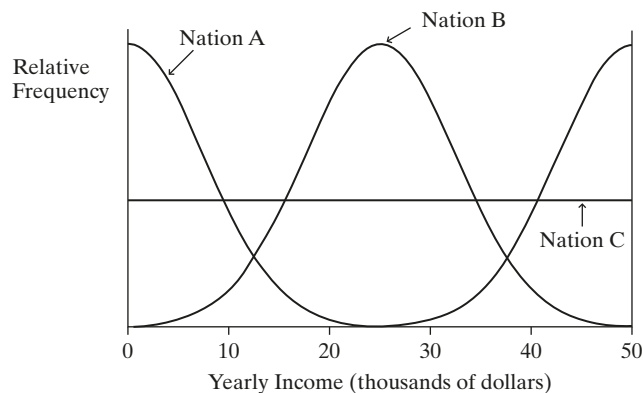
STANDARD DEVIATION

The most useful measure of variability is based on the *deviations* of the data from their mean.

Deviation

> The *deviation* of an observation y_i from the sample mean $\bar{y}$ is $(y_i - \bar{y})$, the difference between them.

Each observation has a deviation. The deviation is *positive* when the observation falls *above* the mean. The deviation is *negative* when the observation falls *below* the mean. The interpretation of $\bar{y}$ as the center of gravity of the data implies that the sum of the positive deviations equals the negative of the sum of negative deviations. Thus, the sum of all the deviations about the mean, $\sum(y_i - \bar{y})$, equals 0. Because of this, measures of variability use either the absolute values or the squares of the deviations. The most popular measure uses the squares.

Standard Deviation

> The **standard deviation** *s* of *n* observations is
>
> $$s = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{\text{sum of squared deviations}}{\text{sample size } - 1}}.$$
>
> This is the positive square root of the **variance** s^2, which is
>
> $$s^2 = \frac{\sum(y_i - \bar{y})^2}{n - 1} = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n - 1}.$$

The *variance* is approximately the average of the squared deviations. The units of measurement are the squares of those for the original data, since it uses squared deviations. This makes the variance difficult to interpret. It is why we use instead its square root, the *standard deviation*.

The expression $\sum(y_i - \bar{y})^2$ in these formulas is called a **sum of squares**. It represents squaring each deviation and then adding those squares. The larger the deviations, the larger the sum of squares and the larger *s* tends to be.

Although its formula looks complicated, the most basic interpretation of the standard deviation *s* is simple: *s* is a sort of *typical distance* of an observation from the mean. So, *the larger the standard deviation, the greater the spread of the data*.

Example 3.5

Comparing Variability of Quiz Scores Each of the following sets of quiz scores for two small samples of students has a mean of 5 and a range of 10:

$$\text{Sample 1:} \quad 0, 4, 4, 5, 7, 10$$
$$\text{Sample 2:} \quad 0, 0, 1, 9, 10, 10$$

By inspection, sample 1 shows less variability about the mean than sample 2. Most scores in sample 1 are near the mean of 5, whereas all the scores in sample 2 are quite far from 5.

For sample 1,

$$\sum(y_i - \bar{y})^2 = (0 - 5)^2 + (4 - 5)^2 + (4 - 5)^2 + (5 - 5)^2 + (7 - 5)^2 + (10 - 5)^2 = 56.$$

So, the variance is

$$s^2 = \frac{\sum(y_i - \bar{y})^2}{n - 1} = \frac{56}{6 - 1} = \frac{56}{5} = 11.2,$$

and the standard deviation is $s = \sqrt{11.2} = 3.3$. For sample 2, you can verify that $s^2 = 26.4$ and $s = \sqrt{26.4} = 5.1$. Since $3.3 < 5.1$, the standard deviations tell us that sample 1 is less variable than sample 2. ∎

Statistical software and many hand calculators can find the standard deviation. For example, for sample 2 the free software R finds

```
> quiz2 <- c(0, 0, 1, 9, 10, 10) # c COMBINES values listed
> sd(quiz2)                # sd is standard deviation function
[1] 5.138093
```

You should do the calculation yourself for a couple of small data sets to get a feel for what s represents. The answer you get may differ slightly from the value reported by software, depending on how much you round off in performing the calculation.

PROPERTIES OF THE STANDARD DEVIATION

- $s \geq 0$.

- $s = 0$ only when all observations have the same value. For instance, if the ages for a sample of five students are 19, 19, 19, 19, and 19, then the sample mean equals 19, each of the five deviations equals 0, and $s = 0$. This is the minimum possible variability.

- The greater the variability about the mean, the larger is the value of s.

- The reason for using $(n - 1)$, rather than n, in the denominator of s is technical. In Chapter 5, we'll see that doing this provides a better estimate of a corresponding parameter for the population. When we have data for an entire population, we replace $(n - 1)$ by the actual population size; the population variance is then precisely the mean of the squared deviations about the population mean.

- If the data are rescaled, the standard deviation is also rescaled. For instance, if we change annual incomes from dollars (such as 34,000) to thousands of dollars (such as 34.0), the standard deviation also changes by a factor of 1000 (such as from 11,800 to 11.8).

INTERPRETING THE MAGNITUDE OF s: THE EMPIRICAL RULE

A distribution with $s = 5.1$ has greater variability than one with $s = 3.3$, but how do we interpret *how large* $s = 5.1$ is? We've seen that a rough answer is that s is a typical distance of an observation from the mean. To illustrate, suppose the first exam in your course, graded on a scale of 0 to 100, has a sample mean of 77. A value of $s = 0$ in unlikely, since every student must then score 77. A value such as $s = 50$ seems implausibly large for a typical distance from the mean. Values of s such as 8 or 12 seem much more realistic.

More precise ways to interpret s require further knowledge of the *shape* of the frequency distribution. The following rule is applicable for many data sets.

Empirical Rule

If the histogram of the data is approximately bell shaped, then

1. About 68% of the observations fall between $\bar{y} - s$ and $\bar{y} + s$.
2. About 95% of the observations fall between $\bar{y} - 2s$ and $\bar{y} + 2s$.
3. All or nearly all observations fall between $\bar{y} - 3s$ and $\bar{y} + 3s$.

The rule is called the Empirical Rule because many frequency distributions seen in practice (i.e., *empirically*) are approximately bell shaped. Figure 3.12 is a graphical portrayal of the rule.

FIGURE 3.12: Empirical Rule: For Bell-Shaped Frequency Distributions, the Empirical Rule Specifies Approximate Percentages of Data within 1, 2, and 3 Standard Deviations of the Mean

About 68% of measurements

$\bar{y} - 3s$ $\bar{y} - 2s$ $\bar{y} - s$ $\bar{y}$ $\bar{y} + s$ $\bar{y} + 2s$ $\bar{y} + 3s$

About 95% of measurements

All or nearly all measurements
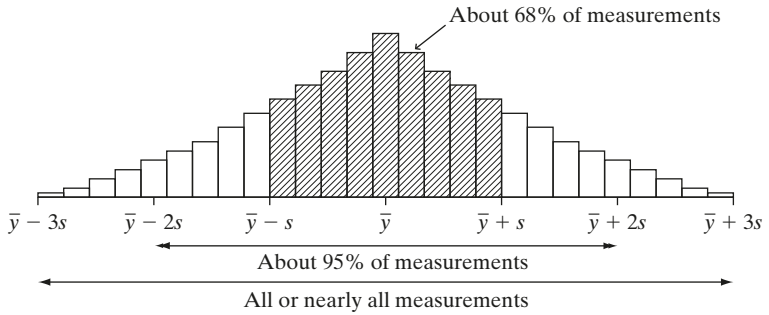
Example 3.6

Describing a Distribution of SAT Scores   The Scholastic Aptitude Test (SAT, see www.collegeboard.com) has three portions: critical reading, mathematics, and writing. For each portion, the distribution of scores is approximately bell shaped with mean about 500 and standard deviation about 100. Figure 3.13 portrays this. By the Empirical Rule, for each portion, about 68% of the scores fall between 400 and 600, because 400 and 600 are the numbers that are *one* standard deviation below and above the mean of 500. About 95% of the scores fall between 300 and 700, the numbers that are *two* standard deviations from the mean. The remaining 5% fall either below 300 or above 700. The distribution is roughly symmetric about 500, so about 2.5% of the scores fall above 700 and about 2.5% fall below 300.   ■

FIGURE 3.13: A Bell-Shaped Distribution of Scores for a Portion of the SAT, with Mean 500 and Standard Deviation 100

68% of scores

2.5% of scores

2.5% of scores

300   400   500   600   700

95% of scores
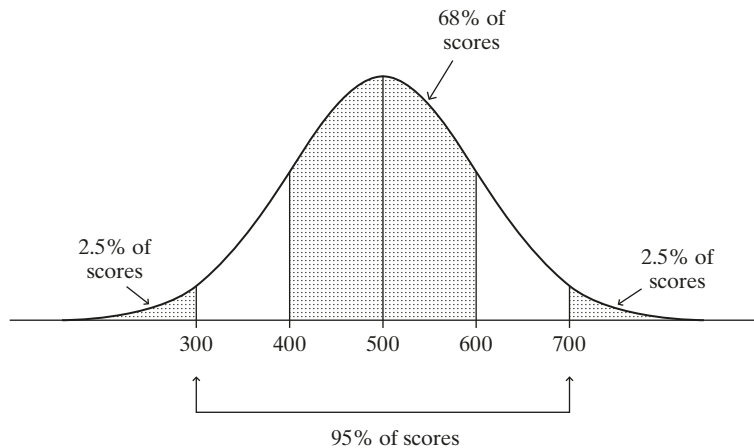
The Empirical Rule applies only to distributions that are approximately bell shaped. For other shapes, the percentage falling within two standard deviations of the mean need not be near 95%. It could be as low as 75% or as high as 100%. The Empirical Rule does not apply if the distribution is highly skewed or if it is highly discrete, with the variable taking few values. The exact percentages depend on the form of the distribution, as the next example demonstrates.

Example 3.7

Familiarity with AIDS Victims   A General Social Survey asked, "How many people have you known personally, either living or dead, who came down with AIDS?" Table 3.7 shows part of some software output for summarizing the 1598 responses on this variable. It indicates that 76% of the responses were 0.
    The mean and standard deviation are $\bar{y} = 0.47$ and $s = 1.09$. The values 0 and 1 both fall within one standard deviation of the mean. Now, 88.8% of the distribution falls at these two points, or within $\bar{y} \pm s$. This is considerably larger

than the 68% that the Empirical Rule states. The Empirical Rule does not apply to this distribution, because it is not even approximately bell shaped. Instead, it is highly skewed to the right, as you can check by sketching a histogram. The smallest value in the distribution (0) is less than one standard deviation below the mean; the largest value in the distribution (8) is nearly seven standard deviations above the mean. ∎

TABLE 3.7: Frequency Distribution of the Number of People Known Personally with AIDS		
AIDS	Frequency	Percent
0	1214	76.0
1	204	12.8
2	85	5.3
3	49	3.1
4	19	1.2
5	13	0.8
6	5	0.3
7	8	0.5
8	1	0.1
n = 1598	Mean = 0.47	Std Dev = 1.09

Whenever the smallest or largest observation is less than a standard deviation from the mean, this is evidence of severe skew. Suppose that the first exam in your course, having potential scores between 0 and 100, has $\bar{y} = 86$ and $s = 15$. The upper bound of 100 is less than one standard deviation above the mean. The distribution is likely highly skewed to the left.

The standard deviation, like the mean, can be greatly affected by an outlier, especially for small data sets. For instance, for the incomes of the seven Leonardo's Pizza employees shown on page 48, $\bar{y} = \$45,900$ and $s = \$78,977$. When we remove the outlier, $\bar{y} = \$16,050$ and $s = \$489$.

3.4 Measures of Position

Another way to describe a distribution is with a measure of *position*. This tells us the point at which a given percentage of the data fall below (or above) that point. As special cases, some measures of position describe center and some describe variability.

QUARTILES AND OTHER PERCENTILES

The range uses two measures of position, the maximum value and the minimum value. The median is a measure of position, with half the data falling below it and half above it. The median is a special case of a set of measures of position called *percentiles*.

<table>
<tr><td align="right">**Percentiles**</td><td>The ***pth percentile*** is the point such that p% of the observations fall below or at that point and $(100 - p)$% fall above it.</td></tr>
</table>

Substituting $p = 50$ in this definition gives the 50th percentile. This is the *median*. The median is larger than 50% of the observations and smaller than the other $(100 - 50) = 50$%. In proportion terms, a percentile is called a **quantile**. The 50th percentile is the 0.50 quantile.

Two other commonly used percentiles are the *lower quartile* and the *upper quartile*.

<table>
<tr><td align="right">**Lower and Upper Quartiles**</td><td>The 25th percentile is called the ***lower quartile***. The 75th percentile is called the ***upper quartile***. One quarter of the data fall below the lower quartile. One quarter fall above the upper quartile.</td></tr>
</table>

The quartiles result from the percentile definition when we set $p = 25$ and $p = 75$. The quartiles together with the median split the distribution into four parts, each containing one-fourth of the observations. See Figure 3.14. The lower quartile is the median for the observations that fall below the median, that is, for the bottom half of the data. The upper quartile is the median for the observations that fall above the median, that is, for the upper half of the data.

FIGURE 3.14: The Quartiles and the Median Split a Distribution into Four Equal Parts. The interquartile range describes the spread of the middle half of the distribution.
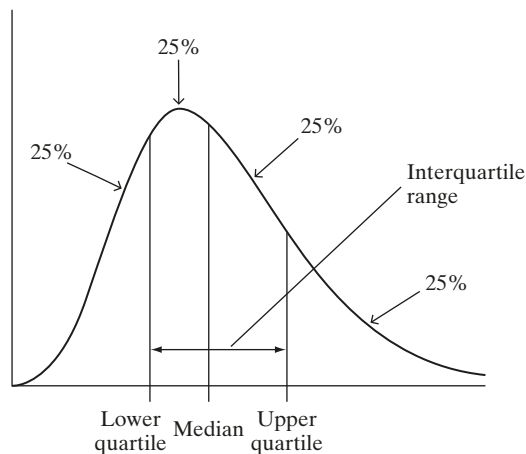
The median, the quartiles, and the maximum and minimum are five positions often used as a set to describe center and spread. Software can easily find these values as well as other percentiles. For instance, using R software we find $\bar{y}$ and s and then the five-number summary for the violent crime rates of Table 3.2, which the variable *violent* lists in the data file `Crime` at the text website:

```
> mean(violent); sd(violent)
[1] 34.9
[1] 12.43637
> summary(violent)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   12.0    26.0    33.0    34.9    43.0    64.0
```

The lower and upper quartiles are labeled as "1st Qu." and "3rd Qu." In Stata, we use the `summarize` command to get $\bar{y}$, s, and the min and max.

```
. summarize violent
  Variable |   Obs      Mean    Std. Dev    Min     Max
   violent |    50      34.9    12.43637     12      64
```

We can also find the quartiles:

```
. tabstat violent, stats(p25 p50 p75)

    variable |       p25        p50        p75
     violent |        26         33         43
```

In summary, about a quarter of the states had violent crime rates (i) below 26, (ii) between 26 and 33, (iii) between 33 and 43, and (iv) above 43. The distance between the upper quartile and the median, $43 - 33 = 10$, exceeds the distance $33 - 26 = 7$ between the lower quartile and the median. This commonly happens when the distribution is skewed to the right.

MEASURING VARIABILITY: INTERQUARTILE RANGE

The difference between the upper and lower quartiles is called the ***interquartile range***, denoted by IQR. This measure describes the spread of the middle half of the observations. For the U.S. violent crime rates just summarized by the five-number summary, the interquartile range $IQR = 43 - 26 = 17$. The middle half of the rates fall within a range of 17, whereas all rates fall within a range of $64 - 12 = 52$. Like the range and standard deviation, the IQR increases as the variability increases, and it is useful for comparing variability of different groups. For example, in 1990 the violent crime rates had quartiles of 33 and 77, giving an IQR of $77 - 33 = 44$. This indicates quite a bit more variability than in 2015, when $IQR = 17$.

An advantage of the IQR over the ordinary range or the standard deviation is that it is not sensitive to outliers. The violent crime rates ranged from 12 to 64, so the range was 52. When we include the observation for D.C., which was 130, the IQR changes only from 17 to 18. By contrast, the range changes from 52 to 118.

For bell-shaped distributions, the distance from the mean to either quartile is about two-thirds of a standard deviation. Then, IQR equals approximately $(4/3)s$.

BOX PLOTS: GRAPHING THE FIVE-NUMBER SUMMARY OF POSITIONS

The five-number summary consisting of (minimum, lower quartile, median, upper quartile, maximum) is the basis of a graphical display called[2] the ***box plot*** that summarizes center and variability. The *box* of a box plot contains the central 50% of the distribution, from the lower quartile to the upper quartile. The median is marked by a line drawn within the box. The lines extending from the box are called *whiskers*.

---

[2] Stem-and-leaf plots and box plots are relatively recent innovations, introduced by the statistician John Tukey (see Tukey, 1977), who also introduced the terminology "software."

These extend to the maximum and minimum, except for outliers, which are marked separately.

Figure 3.15 shows the box plot for the violent crime rates, including D.C., in the format provided with R software. The upper whisker and upper half of the central box are a bit longer than the lower ones. This indicates that the right tail of the distribution, which corresponds to the relatively large values, is longer than the left tail. The plot reflects the skewness to the right of violent crime rates.

FIGURE 3.15: Box Plot of Violent Crime Rates of U.S. States. The outlier is the observation for D.C.

Violent Crime Rates
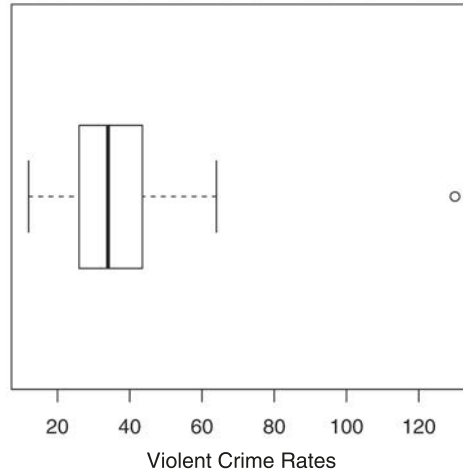
COMPARING GROUPS

Many studies compare different groups on some variable. Relative frequency distributions, histograms, and side-by-side box plots are useful for making comparisons.

Example 3.8

Comparing Canadian and U.S. Murder Rates   Figure 3.16 (page 62) shows side-by-side box plots of murder rates (measured as the number of murders per 100,000 population) in a recent year for the 50 states in the United States and for the provinces of Canada. From this figure, it is clear that the murder rates tended to be much lower in Canada, varying between 0.7 (Prince Edward Island) and 2.9 (Manitoba) whereas those in the United States varied between 1.6 (Maine) and 20.3 (Louisiana). These side-by-side box plots reveal that the murder rates in the United States tend to be much higher and have much greater variability.   ∎

OUTLIERS

Box plots identify outliers separately. To explain this, we now present a formal definition of an outlier.

Outlier

> An observation is an **_outlier_** if it falls more than 1.5(IQR) above the upper quartile or more than 1.5(IQR) below the lower quartile.

In box plots, the whiskers extend to the smallest and largest observations only if those values are not outliers, that is, if they are no more than 1.5(IQR) beyond the quartiles. Otherwise, the whiskers extend to the most extreme observations within 1.5(IQR), and the outliers are marked separately.
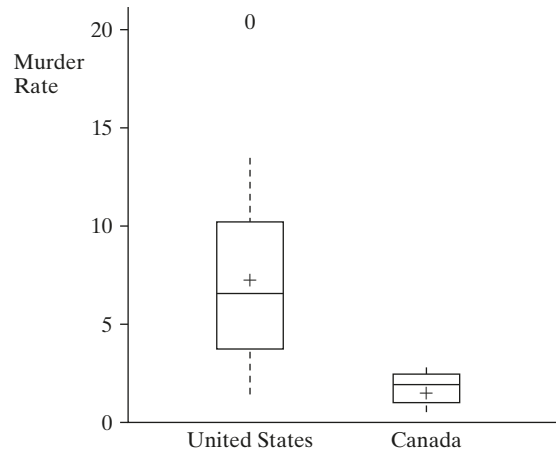
Figure 3.16 shows one outlier for the United States with a very high murder rate. This is the murder rate of 20.3 (for Louisiana). For these data, lower quartile = 3.9 and upper quartile = 10.3, so IQR = 10.3 − 3.9 = 6.4. Thus,

$$\text{Upper quartile} + 1.5(\text{IQR}) = 10.3 + 1.5(6.4) = 19.9.$$

Since 20.3 > 19.9, the box plot highlights the observation of 20.3 as an outlier.

Why highlight outliers? It can be informative to investigate them. Was the observation perhaps incorrectly recorded? Was that subject fundamentally different from the others in some way? Often it makes sense to repeat a statistical analysis without an outlier, to make sure the conclusions are not overly sensitive to a single observation. Another reason to show outliers separately in a box plot is that they do not provide much information about the shape of the distribution, especially for large data sets.

In practice, the 1.5(IQR) criterion for an outlier is somewhat arbitrary. It is better to regard an observation satisfying this criterion as a *potential* outlier rather than a definite outlier. When a distribution has a long right tail, some observations may fall more than 1.5(IQR) above the upper quartile even if they are not separated far from the bulk of the data.

HOW MANY STANDARD DEVIATIONS FROM THE MEAN? THE z-SCORE

Another way to measure position is by the number of standard deviations that a value falls from the mean. For example, the U.S. murder rates shown in the box plot in Figure 3.16 have a mean of 7.3 and a standard deviation of 4.0. The value of 20.3 for Louisiana falls 20.3 − 7.3 = 13.0 above the mean. Now, 13.0 is 13.0/4.0 = 3.25 standard deviations. The Louisiana murder rate is 3.25 standard deviations above the mean.

The number of standard deviations that an observation falls from the mean is called its ***z-score***. For the murder rates of Figure 3.16, Louisiana has a z-score of

$$z = \frac{20.3 - 7.3}{4.0} = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} = 3.25.$$

By the Empirical Rule, for a bell-shaped distribution it is very unusual for an observation to fall more than three standard deviations from the mean. An alternative criterion regards an observation as an outlier if it has a z-score larger than 3 in absolute value. By this criterion, the murder rate for Louisiana is an outlier.

3.5 Bivariate Descriptive Statistics

In this chapter, we've learned how to summarize categorical and quantitative variables graphically and numerically. In the next three chapters, we'll learn about statistical inference for a categorical or quantitative variable. Most studies have more than one variable, however, and Chapters 7–15 present methods that can handle two or more variables at a time.

ASSOCIATION BETWEEN RESPONSE AND EXPLANATORY VARIABLES

With multivariable analyses, the main focus is on studying *associations* among the variables. An association exists between two variables if certain values of one variable tend to go with certain values of the other.

For example, consider "religious affiliation," with categories (Protestant, Catholic, Jewish, Muslim, Hindu, Other), and "ethnic group," with categories (Anglo-American, African-American, Hispanic). In the United States, Anglo-Americans are more likely to be Protestant than are Hispanics, who are overwhelmingly Catholic. African-Americans are even more likely to be Protestant. An association exists between religious affiliation and ethnic group, because the proportion of people having a particular religious affiliation changes as the ethnic group changes.

An analysis of association between two variables is called a *bivariate* analysis, because there are two variables. Usually one is an outcome variable on which comparisons are made at levels of the other variable. The outcome variable is called the *response variable*. The variable that defines the groups is called the *explanatory variable*. The analysis studies how the outcome on the response variable *depends on* or is *explained by* the value of the explanatory variable. For example, when we describe how religious affiliation depends on ethnic group, religious affiliation is the response variable and ethnic group is the explanatory variable. In a comparison of men and women on income, income is the response variable and gender is the explanatory variable. Income may depend on gender, not gender on income.

Often, the response variable is called the *dependent variable* and the explanatory variable is called the *independent variable*. The terminology *dependent variable* refers to the goal of investigating the degree to which the response on that variable *depends on* the value of the other variable. We prefer not to use these terms, since *independent* and *dependent* are used for many other things in statistical science.

COMPARING TWO GROUPS: BIVARIATE CATEGORICAL AND QUANTITATIVE DATA

Chapter 7 presents descriptive and inferential methods for comparing two groups. For example, suppose we'd like to know whether men or women have more good friends, on the average. A General Social Survey reports that the mean number of good friends is 7.0 for men ($s = 8.4$) and 5.9 for women ($s = 6.0$). The two distributions have similar appearance, both being highly skewed to the right and with a median of 4.

Here, this is an analysis of two variables—number of good friends and gender. The response variable, number of good friends, is quantitative. The explanatory variable, gender, is categorical. In this case, it's common to compare categories of the categorical variable on measures of the center (such as the mean and median) for the response variable. Graphs are also useful, such as side-by-side box plots.

BIVARIATE CATEGORICAL DATA

Chapter 8 presents methods for analyzing association between two categorical variables. Table 3.8 is an example of such data. This table results from answers to two questions on the 2014 General Social Survey. One asked whether homosexual relations are wrong. The other asked about the fundamentalism/liberalism of the respondent's religion. A table of this kind, called a ***contingency table***, displays the number of subjects observed at combinations of possible outcomes for the two variables. It displays how outcomes of a response variable are *contingent* on the category of the explanatory variable.

TABLE 3.8: Contingency Table for Religion and Opinion about Homosexual Relations

| | Opinion about Homosexual Relations | | | | |
Religion	Always Wrong	Almost Always Wrong	Sometimes Wrong	Not Wrong at All	Total
Fundamentalist	262	10	19	87	378
Liberal	122	16	43	360	541

Table 3.8 has eight possible combinations of responses. (Another possible outcome, *moderate* for the religion variable, is not shown here.) We could list the categories in a frequency distribution or construct a bar graph. It's most informative to do this for the categories of the response variable, separately for each category of the explanatory variable. For example, if we treat opinion about homosexual relations as the response variable, we could report the percentages in the four categories for homosexual relations, separately for each religious category.

Consider the *always wrong* category. For fundamentalists, since 262/378 = 0.69, 69% believe homosexual relations are always wrong. For those who report being liberal, since 122/541 = 0.23, 23% believe homosexual relations are always wrong. Likewise, you can check that the percentages responding *not wrong at all* were 23% for fundamentalists and 67% for liberals. There seems to be an appreciable association between opinion about homosexuality and religious beliefs, with religious fundamentalists being more negative about homosexuality. (For comparison, in the 1974 GSS the percentages in the *always wrong* category were 84% for fundamentalists and 47% for liberals, so the change in views over time has been considerable.) Chapter 8 shows many other ways of analyzing bivariate categorical data.

BIVARIATE QUANTITATIVE DATA

To illustrate methods that are useful when both variables are quantitative, we use the UN data file at the text website, partly shown in Table 3.9. The file has United Nations data from 2014 for 42 nations on per capita gross domestic product (GDP, in thousands of dollars), a human development index (HDI, which has components referring to life expectancy at birth, educational attainment, and income per capita), a gender inequality index (GII, a composite measure reflecting inequality in achievement between women and men in reproductive health, empowerment, and the labor market), fertility rate (number of births per woman), carbon dioxide emissions per capita (metric tons), a homicide rate (number of homicides per 100,000 people), prison population (per 100,000 people), and percent using the Internet.

TABLE 3.9: National Data from UN Data File at Text Website

Nation	GDP	HDI	GII	Fertility	CO2	Homicide	Prison	Internet
Algeria	12.8	0.72	0.42	2.8	3.2	0.8	162	17
Argentina	14.7	0.81	0.38	2.2	4.7	5.5	147	60
Australia	42.3	0.93	0.11	1.9	16.5	1.1	130	83
Austria	43.1	0.88	0.06	1.4	7.8	0.8	98	81
Belgium	39.5	0.88	0.07	1.8	8.8	1.8	108	82
Brazil	14.3	0.74	0.44	1.8	2.2	21.8	274	52
Canada	40.6	0.90	0.14	1.6	14.1	1.5	118	86
...								
UK	34.7	0.89	0.19	1.9	7.1	1.2	148	90
US	50.9	0.91	0.26	1.9	17.0	4.7	716	84
Vietnam	4.9	0.64	0.32	1.7	2.0	1.6	145	44

Source: `http://hdr.undp.org/en/data` and `http://data.worldbank.org`; complete data file UN ($n = 42$) is at text website.

Figure 3.17 is an example of a type of graphical plot, called a *scatterplot*, that portrays bivariate relations between quantitative variables. It plots data on percent using the Internet and gross domestic product. Here, values of GDP are plotted on the horizontal axis, called the *x-axis*, and values of Internet use are plotted on the vertical axis, called the *y-axis*. The values of the two variables for any particular observation form a point relative to these axes. The figure plots the 42 observations as 42 points. For example, the point at the highest level on GDP represents Norway, which had a GDP of 62.9 and Internet use of 95 percent. The scatterplot shows a tendency for nations with higher GDP to have higher levels of Internet use.

FIGURE 3.17: Scatterplots for GDP as Predictor of Internet Use and of GII, for 42 Nations
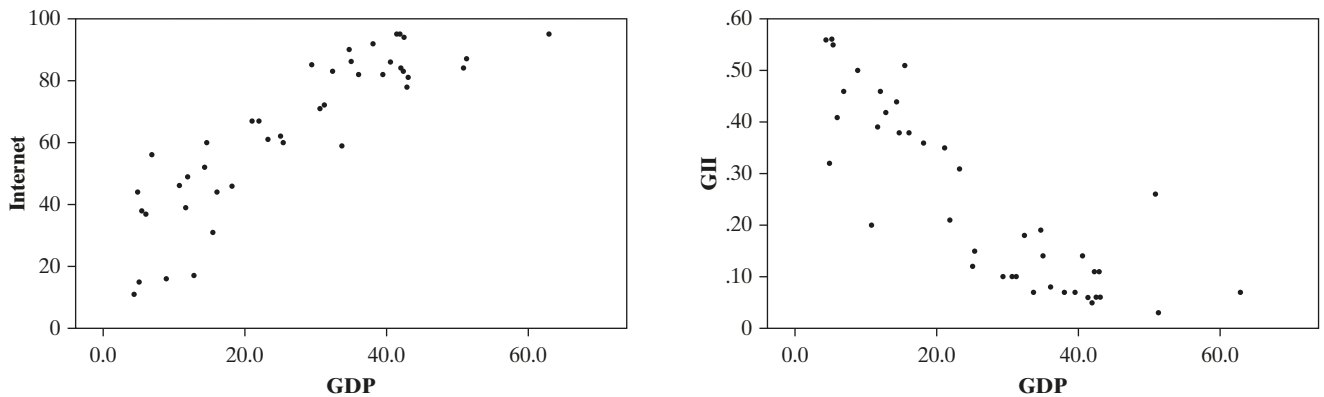
In Chapter 9, we'll learn about two ways to describe such a trend. One way to describe the trend, called the *correlation*, describes how strong the association is, in terms of how closely the data follow a *straight-line trend*. For Figure 3.17, the correlation is 0.88. The positive value means that Internet use tends to go *up* as GDP goes *up*. By contrast, Figure 3.17 also shows a scatterplot for GDP and GII. Those variables have a negative correlation of -0.85. As GDP goes up, GII tends to go down. The correlation takes values between -1 and $+1$. The larger it is in absolute value, that is, the farther from 0, the stronger the association. For example, GDP is more strongly associated with Internet use and with GII than it is with fertility, because correlations of 0.88 and -0.85 are larger in absolute value than the correlation of -0.49 between GDP and fertility.

The second useful tool for describing the trend is ***regression analysis***. This method treats one variable, usually denoted by *y*, as the response variable, and the other variable, usually denoted by *x*, as the explanatory variable. It provides a straight-line formula for predicting the value of *y* from a given value of *x*. For the data from Table 3.9 on *y* = fertility rate and *x* = GDP, this equation is

$$\text{Predicted fertility} = 2.714 - 0.025(\text{GDP}).$$

For a country with GDP = 4.4 (the lowest value in this sample), the predicted fertility rate is $2.714 - 0.025(4.4) = 2.6$ births per woman. For a country with GDP = 62.9 (the highest value in this sample), the predicted fertility rate is $2.714 - 0.025(62.9) = 1.1$ births per woman.

Chapter 9 shows how to find the correlation and the regression line. It is simple with software, as shown in Table 3.10 using R with variables from the data file UN at the text website. Later chapters show how to extend the analysis to handle categorical as well as quantitative variables.

TABLE 3.10: Using R Software for a Scatterplot, Correlation, and Regression Line

```
> UN <- read.table("http://www.stat.ufl.edu/~aa/smss/data/UN.dat",
+                   header=TRUE)
> attach(UN)
> plot(GDP, Fertility)  # requests scatterplot
> cor(GDP, Fertility); cor(GDP, Internet); cor(GDP, GII)
[1] -0.4861589
[1]  0.8771987
[1] -0.8506693

> lm(Fertility ~ GDP)  # lm is short for "linear model"
Coefficients:
(Intercept)          GDP
    2.71401      -0.02519
```

ANALYZING MORE THAN TWO VARIABLES

This section has introduced analyzing associations between two variables. One important lesson from later in the text is that *just because two variables have an association does not mean there is a causal connection*. For example, the correlation for Table 3.9 between the Internet use and the fertility rate is −0.48. But having more people using the Internet need not be the reason the fertility rate tends to be lower (e.g., because people are on the Internet rather than doing what causes babies). Perhaps high values on Internet use and low values on fertility are both a by-product of a nation being more economically advanced.

Most studies have *several* variables. The second half of this book (Chapters 10–15) shows how to conduct *multivariate* analyses. For example, to study what is associated with the number of good friends, we might want to simultaneously consider gender, age, whether married, educational level, whether attend religious services regularly, and whether live in urban or rural setting.

3.6 Sample Statistics and Population Parameters

Of the measures introduced in this chapter, the mean $\bar{y}$ is the most commonly used measure of center and the standard deviation s is the most common measure of spread. We'll use them frequently in the rest of the text.

Since the values $\bar{y}$ and s depend on the sample selected, they vary in value from sample to sample. In this sense, they are variables. Their values are unknown before the sample is chosen. Once the sample is selected and they are computed, they become known sample statistics.

With inferential statistics, we distinguish between sample statistics and the corresponding measures for the population. Section 1.2 introduced the term *parameter* for a summary measure of the population. A statistic describes a sample, while a parameter describes the population from which the sample was taken. In this text, lower case Greek letters usually denote population parameters and Roman letters denote the sample statistics.

Notation for Mean and Standard Deviation Parameters

> Greek letters denote parameters. For example, μ (mu) and σ (sigma) denote the population mean and standard deviation of a variable.

We call μ and σ the **population mean** and **population standard deviation**, respectively. The population mean is the average of the observations for the entire population. The population standard deviation describes the variability of those observations about the population mean.

Whereas the statistics $\bar{y}$ and s are variables, with values depending on the sample chosen, the parameters μ and σ are constants. This is because μ and σ refer to just one particular group of observations, namely, the observations for the entire population. The parameter values are usually unknown, which is the reason for sampling and computing sample statistics to estimate their values. Much of the rest of this text deals with ways of making inferences about parameters (such as μ) using sample statistics (such as $\bar{y}$). Before studying these inferential methods, though, you need to learn some basic ideas of *probability*, which serves as the foundation for the methods. Probability is the subject of the next chapter.

3.7 Chapter Summary

This chapter introduced **descriptive statistics**—ways of *describing* data to summarize key characteristics of the data.

OVERVIEW OF TABLES AND GRAPHS

- A **frequency distribution** summarizes numbers of observations for possible values or intervals of values of a variable.

- For a quantitative variable, a **histogram** uses bars over possible values or intervals of values to portray a frequency distribution. It shows shape—such as whether the distribution is approximately bell shaped or skewed to the right (longer tail pointing to the right) or to the left.

- The **box plot** portrays the quartiles, the extreme values, and any outliers.

Cook (2014) and Tufte (2001) showed other innovative ways to present data graphically.

OVERVIEW OF MEASURES OF CENTER

Measures of center describe the center of the data, in terms of a typical observation.

- The *mean* is the sum of the observations divided by the sample size. It is the center of gravity of the data.

- The *median* divides the ordered data set into two parts of equal numbers of observations, half below and half above that point.

- The lower quarter of the observations fall below the *lower quartile*, and the upper quarter fall above the *upper quartile*. These are the 25th and 75th *percentiles*. The median is the 50th percentile. The quartiles and median split the data into four equal parts. These *measures of position*, portrayed with extreme values in *box plots*, are less affected than the mean by outliers or extreme skew.

OVERVIEW OF MEASURES OF VARIABILITY

Measures of variability describe the spread of the data.

- The *range* is the difference between the largest and smallest observations. The *interquartile range* is the range of the middle half of the data between the upper and lower quartiles. It is less affected by outliers.

- The *variance* averages the squared deviations about the mean. Its square root, the *standard deviation*, is easier to interpret, describing a typical distance from the mean.

- The *Empirical Rule* states that for a bell-shaped distribution, about 68% of the observations fall within one standard deviation of the mean, about 95% fall within two standard deviations of the mean, and nearly all, if not all, fall within three standard deviations of the mean.

Table 3.11 summarizes measures of center and variability. A *statistic* summarizes a sample. A *parameter* summarizes a population. *Statistical inference* uses statistics to make predictions about parameters.

TABLE 3.11: Summary of Measures of Center and Variability

Measure	Definition	Interpretation
Center		
Mean	$\bar{y} = \sum y_i / n$	Center of gravity
Median	Middle observation of ordered sample	50th percentile, splits sample into two equal parts
Mode	Most frequently occurring value	Most likely outcome, valid for all types of data
Variability		
Standard deviation	$s = \sqrt{\sum (y_i - \bar{y})^2 / (n - 1)}$	Empirical Rule: If bell shaped, 68%, 95% within s, $2s$ of $\bar{y}$
Range	Largest − smallest observation	Greater with more variability
Interquartile range	Upper quartile (75th percentile) − lower quartile (25th percentile)	Encompasses middle half of data

OVERVIEW OF BIVARIATE DESCRIPTIVE STATISTICS

Bivariate statistics summarize data on two variables together, to analyze the *association* between them.

- Many studies analyze how the outcome on a *response variable* depends on the value of an *explanatory variable*.

- For categorical variables, a *contingency table* shows the number of observations at the combinations of possible outcomes for the two variables.

- For quantitative variables, a *scatterplot* graphs the observations. It shows a point for each observation, plotting the response variable on the *y*-axis and the explanatory variable on the *x*-axis.

- For quantitative variables, the *correlation* describes the strength of straight-line association. It falls between -1 and $+1$ and indicates whether the response variable tends to increase (positive correlation) or decrease (negative correlation) as the explanatory variable increases. A *regression line* is a straight-line formula for predicting the response variable using the explanatory variable.

Exercises

Practicing the Basics

3.1. Table 3.12 shows the number (in millions) of the foreign-born population of the United States, by place of birth.

(a) Construct a relative frequency distribution.

(b) Sketch the data in a bar graph.

(c) Is "place of birth" quantitative, or categorical?

(d) Use whichever of the following measures is relevant for these data: mean, median, mode.

TABLE 3.12

Place of Birth	Number
Europe	4.5
Asia	10.1
Caribbean	3.6
Central America	14.4
South America	2.4
Other	2.6
Total	37.6

Source: *Statistical Abstract of the United States, 2012.*

3.2. According to a survey done by a large technology firm, the number of Internet users was 630 million in China, 400 million in Europe, and 290 million in the United States.

(a) Construct a relative frequency distribution.

(b) Sketch a bar graph.

(c) Can you find a mean, median, or mode for these data? If yes, do so and interpret.

3.3. A teacher shows her class the scores on the midterm exam in the stem-and-leaf plot:

```
6 | 5 8 8
7 | 0 1 1 3 6 7 7 9
8 | 1 2 2 3 3 3 4 6 7 7 7 8 9
9 | 0 1 1 2 3 4 4 5 8
```

(a) Identify the number of students and the minimum and maximum scores.

(b) Sketch a corresponding histogram with four intervals.

3.4. A countrywide demographical survey of a population revealed that 43.2 million people had black-colored eyes, 12.1 million people had hazel-colored eyes, 26.7 million people had brown-colored eyes, 17.4 million people had green-colored eyes, and 10.0 million people had gray-colored eyes.

(a) Make a relative frequency distribution to display the above data.

(b) Make a histogram.

(c) What is the mode in this case?

3.5. Create a data file with your software for the `Crime` data file from the text website. Use the variable *murder*, which is the murder rate (per 100,000 population). Using software,

(a) Construct a relative frequency distribution.

(b) Construct a histogram. How would you describe the shape of the distribution?

(c) Construct a stem-and-leaf plot. How does this plot compare to the histogram in (b)?

TABLE 3.13

nation	GDP	Gini	HDI	Econ	CO2	Prison
Australia	43550	34	0.93	81	16.5	130
Austria	44149	30	0.88	71	7.8	98
Belgium	40338	33	0.88	69	8.8	108
Canada	43247	34	0.90	79	14.1	118
...						
UK	36197	38	0.89	76	7.1	148
US	53143	41	0.91	76	17.0	716

Source: `stats.oecd.org`, `hdr.undp.org/en/data`, and `www.pewresearch.org`; complete data file is at text website.

3.6. The OECD (Organization for Economic Cooperation and Development) consists of advanced, industrialized countries that accept the principles of representative democracy and a free market economy. Table 3.13 shows part of the `OECD` data file at the text website that has data on several variables for the 24 nations that made up the OECD before its recent enlargement to include nations that have recently undergone market economy reforms. The variables are gross domestic product (GDP, per capita in U.S. dollars), the Gini measure of inequality, a human development index (HDI, which has components referring to life expectancy at birth, educational attainment, and income per capita), an index of economic freedom, carbon dioxide emissions (CO_2, per capita, in metric tons), and prison population (per 100,000 people). Using the complete data file from the text website:

(a) Construct a stem-and-leaf plot of the GDP values, by rounding and reporting the values in thousands of dollars (e.g., replacing $43,550 by 44).

(b) Construct a histogram. Interpret.

(c) Identify the outlier in each plot.

3.7. Refer to the prison values in the previous exercise.

(a) Find the mean and the median.

(b) Based on a histogram or box plot for these data, why would you expect the mean to be larger than the median?

(c) Identify an outlier. Investigate how it affects the mean and the median by recalculating them without this observation.

3.8. A study conducted by a world opinion website on the number of Internet users across different countries revealed the following data: 11.6 million in Singapore, 69.8 million in Japan, 22.1 million in France, 14.2 million in Belgium, 14.1 million in Austria, and 12.7 million in Bangladesh.

(a) Find the mean and median for the above data.

(b) Which value in the above data would be considered an outlier? What will be the value of the mean if the outlier was not taken into consideration?

3.9. In order to develop a new line of hair dye, a cosmetics company poll asked, "What color best describes your natural hair?" (black, brown, blonde, red), the percentage of responses were 31%, 54%, 7% and 3%, respectively.

(a) Which response is the mode?

(b) Can you compute a mean or a median for these data? If so, do so; if not, explain why not.

3.10. A researcher in an alcoholism treatment center, to study the length of stay in the center for first-time patients, randomly selects 10 records of individuals institutionalized within the previous two years. The lengths of stay, in days, were 11, 6, 20, 9, 13, 4, 39, 13, 44, and 7. For a similar study 25 years ago, lengths of stay, in days, for 10 sampled individuals were 32, 18, 55, 17, 24, 31, 20, 40, 24, and 15. Software shows results:

```
------------------------------------------
Variable | Obs  Mean  Std. Dev.  Min  Max
 stay_new |  10  16.6  13.91402     4   44
 stay_old |  10  27.6  12.39355    15   55
------------------------------------------
```

(a) Summarize results in the two studies, using measures of center and variability. Interpret any differences you find.

(b) Actually, the new study also selected one other record. That patient is still institutionalized after 40 days. Thus, that patient's length of stay is at least 40 days, but the actual value is unknown. Can you calculate the mean or median for the complete sample of size 11 including this partial observation? Explain. (This observation is said to be *censored*, meaning that the observed value is "cut short" of its true, unknown value.)

3.11. Access the GSS at `sda.berkeley.edu/GSS`. Entering TVHOURS for the variable and year(2014) in the selection filter, you obtain data on hours per day of TV watching in the United States in 2014.

(a) Construct the relative frequency distribution for the values 0, 1, 2, 3, 4, 5, 6, 7 or more.

(b) How would you describe the shape of the distribution?

(c) Explain why the median is 2.

(d) The mean is larger than 2. Why do you think this is?

3.12. Table 3.14 shows 2012 female economic activity (FEA) for countries in Eastern Europe. Construct plots and find summary statistics to compare these values with those from the Middle East in Table 3.4. Interpret.

TABLE 3.14

Country	FEA	Country	FEA	Country	FEA	Country	FEA
Bosnia/Herz.	68	Estonia	80	Moldova	87	Slovakia	75
Bulgaria	81	Hungary	82	Poland	75	Slovenia	79
Croatia	79	Latvia	81	Romania	80	Ukraine	67
Czech Rep.	74	Lithuania	83	Serbia	75		

Source: www.socialwatch.org.

3.13. Calculate the weighted average for two sets of sample where $n_1 = 12$, $n_2 = 5$, $\bar{y}_1 = 360$, and $\bar{y}_2 = 211$. The weighted average is close to which sample mean?

3.14. Table 3.15 summarizes responses of 2223 subjects in the 2014 GSS to the question "About how often did you have sex during the last 12 months?"

TABLE 3.15

How Often Had Sex	Frequency
Not at all	571
Once or twice	220
About once a month	255
2 or 3 times a month	357
About once a week	333
2 or 3 times a week	365
More than 3 times a week	122

(a) Report the median and the mode. Interpret.

(b) Treat this scale in a quantitative manner by assigning the scores 0, 0.1, 1.0, 2.5, 4.3, 10.8, and 17 to the categories, for approximate monthly frequency. Find the sample mean, and interpret.

3.15. A survey conducted on a sample of employees from an organization asked, "How would you rate yourself on a scale of happiness?" The responses ranged from extremely happy, very happy, somewhat happy, neutral, somewhat unhappy, very unhappy, to extremely unhappy. The counts in each category were 234, 267, 387, 459, 350, 301, and 231.

(a) Make a relative frequency distribution and plot a histogram for the data.

(b) What are the mode and median for the above data?

3.16. According to a study done by an educational research organization, the median tuition cost for a private university was $32,500 per year and $17,350 per year for a public university, whereas the mean was $34,000 for private universities and $19,500 for public universities.

(a) Does this suggest that the distribution of tuition for public and private universities is symmetric, or skewed to the right, or skewed to the left? Explain.

(b) The results refer to 25 private universities and 23 public universities. Find the overall mean.

3.17. A national polling organization conducts a research study on the median amount of money spent on grooming products. Women spent $2,350 in the past year and men spent $866 in the past year.

(a) Identify the response variable and the explanatory variable for this analysis.

(b) Is enough information given to find the median when the data from both groups are combined? Why or why not?

(c) If the reported values were means, what else would you need to know to find the overall mean?

3.18. The General Social Survey has asked, "During the past 12 months, how many people have you known personally that were victims of homicide?" Table 3.16 shows software output from analyzing responses.

(a) Is the distribution bell shaped, skewed to the right, or skewed to the left?

TABLE 3.16

VICTIMS	Frequency	Percent
0	1244	90.8
1	81	5.9
2	27	2.0
3	11	0.8
4	4	0.3
5	2	0.1
6	1	0.1

n	Mean	Std Dev	Min	1st Qu.	Med	3rd Qu.	Max
1370	0.146	0.546	0	0	0	0	6

(b) Does the Empirical Rule apply to this distribution. Why or why not?

(c) Report the median. If 500 observations shift from 0 to 6, how does the median change? What property does this illustrate for the median?

3.19. According to an article reported by `en.wikipedia.org` on the top oil producers in 2011, the oil production (measured in metric ton, mt) was 517 for Saudi Arabia, 510 for Russia, 215 for Iran, 203 for China, 139 for Nigeria, 144 for the UAE, and 144 for Mexico.

(a) Calculate the mean and the median.

(b) What is the range for the given distribution?

(c) What is the standard deviation for the data and what does it tell us about the spread of the data?

3.20. *National Geographic Traveler* magazine recently presented data on the annual number of vacation days averaged by residents of eight different countries. They reported 42 days for Italy, 37 for France, 35 for Germany, 34 for Brazil, 28 for Britain, 26 for Canada, 25 for Japan, and 13 for the United States. (The number of days mandated by these governments differs, varying between 0 for the United States and 30 for France.)

(a) Find the mean and standard deviation. Interpret.

(b) Report the five-number summary. (*Hint*: You can find the lower quartile by finding the median of the four values below the median.)

3.21. The Human Development Index (HDI) is an index the United Nations uses to give a summary rating for each nation based on life expectancy at birth, educational attainment, and income. In 2014, the 10 nations (in order) with the highest HDI rating, followed by the percentage of seats in their parliament held by women (which is a measure of gender empowerment), were Norway 40, Australia 31, Switzerland 28, Netherlands 38, United States 18, Germany 32, New Zealand 32, Canada 28, Singapore 24, and Denmark 39. For these data, $\bar{y} = 31$ and $s = 7$. Would s increase, or decrease, **(a)** if the United States were removed from the data set? **(b)** if Australia were removed? Explain.

3.22. A report published by a survey agency on the number of female workers (in millions) in each country revealed the following data. For Western Europe, the values were as follows: Germany, 86; Norway, 89; France, 72; Ireland, 77; Finland, 81; Greece, 85; the United Kingdom, 89; Belgium, 73; Italy, 79; and Denmark, 83. For Africa, the values reported were as follows: Congo, 32; Sudan, 12; Botswana, 11; Kenya, 36; Ghana, 22; Zimbabwe, 27; Madagascar, 12; and Zambia, 38.

(a) Calculate the mean for the two sets of nations. Which set of nations has a higher number of female workers?

(b) Find the standard deviation for the two sets of nations. Which set has a larger spread of scores and why?

3.23. A report indicates that teacher's total annual pay (including bonuses) in Toronto, Ontario, has a mean of $61,000 and standard deviation of $10,000 (Canadian dollars). Suppose the distribution has approximately a bell shape.

(a) Give an interval of values that contains about (i) 68%, (ii) 95%, (iii) all or nearly all salaries.

(b) Would a salary of $100,000 be unusual? Why?

3.24. Excluding the United States, the national mean number of holiday and vacation days in a year for OECD nations (see Exercise 3.6) is approximately bell shaped with a mean of 35 days and standard deviation of 3 days.[3]

(a) Use the Empirical Rule to describe the variability.

(b) The observation for the United States is 19. If this is included with the other observations, will the (i) mean increase, or decrease? (ii) standard deviation increase, or decrease?

(c) Using the mean and standard deviation for the other countries, how many standard deviations is the U.S. observation from the mean?

3.25. For GSS data on "the number of people you know who have committed suicide," 88.8% of the responses were 0, 8.8% were 1, and the other responses took higher values. The mean equals 0.145, and the standard deviation equals 0.457.

(a) What percentage of observations fall within one standard deviation of the mean?

(b) Is the Empirical Rule appropriate for the distribution of this variable? Why or why not?

3.26. A sample of 100 voters completed a questionnaire regarding their level of satisfaction with the current government. The scores on the questionnaire ranged between 25 and 150. The mean of the sample was 92. Which of the following value can be a possible standard deviation for this sample: 0, −92, 12, 64? Why?

3.27. Why is standard deviation preferred in place of variance to explain variability?

3.28. According to an online survey, the average number of hours spent by a person on the Internet is 12. Which of the following values seems the most realistic in terms of standard deviation:
(i) 5, (ii) 0, (iii) −5, (iv) 32, (v) 64? Why?

3.29. For all homes in Gainesville, Florida, the annual residential electrical consumption[4] recently had a mean of 10,449 and a standard deviation of 7489 kilowatt-hours (kWh). The maximum usage was 336,240 kWh.

---

[3] *Source*: `www.stateofworkingamerica.org`.
[4] Data supplied by Todd Kamhoot, Gainesville Regional Utilities.

(a) What shape do you expect this distribution to have? Why?

(b) Do you expect this distribution to have any outliers? Explain.

3.30. A recent study[5] of the effect of work hours and commuting time on political participation estimated that for those engaged in paid work in the United States, the time it takes on a typical day to get to work has a mean of 19.8 minutes and standard deviation of 13.6 minutes. What shape do you expect this distribution to have? Why?

3.31. According to *Statistical Abstract of the United States, 2015*, the mean salary (in dollars) of secondary school teachers in the United States varied among states with a five-number summary of

```
------------------------------------------
  100% Max       68,800 (New York)
   75% Q3        54,700
   50% Med       45,500
   25% Q1        43,100
    0% Min       37,700 (South Dakota)
------------------------------------------
```

(a) Find and interpret the range.

(b) Find and interpret the interquartile range.

3.32. Refer to the previous exercise.

(a) Sketch a box plot.

(b) Based on (a), predict the direction of skew for this distribution. Explain.

(c) If the distribution, although skewed, is approximately bell shaped, which value is most plausible for the standard deviation:
(i) 100, (ii) 1000, (iii) 7000, (iv) 25,000? Explain.

3.33. Table 3.17 shows part of software output for analyzing the murder rates (per 100,000) in the `Crime2` data

TABLE 3.17

Variable = MURDER			
n	51	n	50
Mean	5.6	Mean	4.8
Std Dev	6.05	Std Dev	2.57
Quartiles		Quartiles	
100% Max	44	100% Max	13
75% Q3	6	75% Q3	6
50% Med	5	50% Med	5
25% Q1	3	25% Q1	3
0% Min	1	0% Min	1
Range	43	Range	12
Q3-Q1	3	Q3-Q1	3

file at the text website (to be analyzed in Chapter 9). The first column refers to the entire data set, and the second column deletes the observation for D.C. For each statistic reported, evaluate the effect of including the outlying observation for D.C.

3.34. The text website has a data file `Houses` that lists recent selling prices of 100 homes in Gainesville, Florida. Software reports $\bar{y} = \$155,331$, $s = \$101,262$, and a five-number summary of minimum $= \$21,000$, Q1 $= \$91,875$, median $= \$132,600$, Q3 $= \$173,875$, and maximum $= \$587,000$.

(a) Does the Empirical Rule apply to this distribution? Why?

(b) What do these values suggest about the shape of the distribution? Why?

(c) Use the 1.5(IQR) criterion to determine if any outliers are present.

3.35. For the following examples, discuss the shape of the distribution by giving a valid explanation.

(a) The mean score on a mathematics test was 67 and the median was 44.

(b) In a sample of 100 infants, the average number of hours slept was 18 and the median was 22.

(c) The most frequently occurring scores in a distribution are at the lowest and the highest points.

(d) In an experiment, the average time taken to complete a trial is 30 seconds. The median is also 30 seconds.

(e) The average calories consumed by an adult in a day are 2,300 and the median is 1,300.

3.36. For each of the following variables, indicate whether you would expect its relative frequency histogram to be bell shaped, U-shaped, skewed to the right, or skewed to the left.

(a) Exam score of easy exam (with $\bar{y} = 88$, $s = 10$, minimum $= 65$, Q1 $= 77$, median $= 85$, Q3 $= 91$, and maximum $= 100$).

(b) IQ for the general population.

(c) Number of times arrested in past year.

(d) Time needed to complete difficult exam (maximum time is 1 hour).

(e) Age at death.

(f) Weekly church contribution (median is $10 and mean is $17).

(g) Attitude toward legalization of abortion.

3.37. For parts (a), (b), and (f) of the previous exercise, sketch box plots that would be plausible for the variable.

3.38. The January 2014 unemployment rates of adults of age 24 or less in the 28 countries in the European Union

---

[5] B. Newman, J. Johnson, and P. Lown, *Am. Politics Res.*, vol. 42 (2014), pp. 141–170.

ranged from 7.9 (Germany) to 57.3 (Greece), with lower quartile = 18.9, median = 23.7, upper quartile = 33.5, mean = 26.0, and standard deviation = 13.0. Sketch a box plot, labeling which of these values are used in the plot.

3.39. For the number of times a week reading a newspaper, from the `Students` data file referred to in Exercise 1.11, Figure 3.18 shows software output (rather crude) of the stem-and-leaf plot and the box plot.

(a) From the box plot, identify the minimum, lower quartile, median, upper quartile, and maximum.

(b) Identify this five-number summary using the stem-and-leaf plot.

(c) Do the data appear to contain any outliers? If so, identify.

(d) The standard deviation is one of the following values—0.3, 3, 13, 23. Which do you think it is, and why?

FIGURE 3.18

```
Stem Leaf                      Boxplot
  14 00                           0
  13
  12 0                            0
  11
  10
   9
   8
   7 000000000                    |
   6 000                          |
   5 000000000                 +-----+
   4 0000                      |     |
   3 0000000000000             *-----*
   2 000000000                 +-----+
   1 000000                       |
   0 0000                         |
```
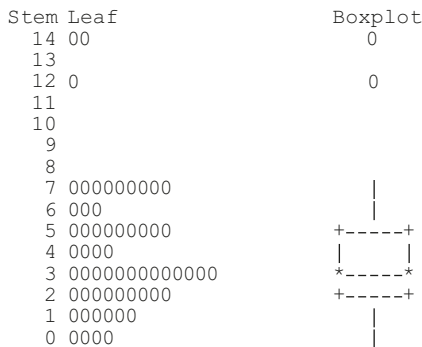
3.40. Infant mortality rates (number of infant deaths, per 1000 live births) are reported by the World Bank. In their report for 2010–2014, the five-number summaries (min, Q1, median, Q2, max) were (19, 55, 74, 86, 107) for Africa and (2, 3, 4, 4, 6) for Europe. Sketch side-by-side box plots, and use them to describe differences between the distributions. (The plot for Europe shows that the quartiles, like the median, are less useful when the data are highly discrete.)

3.41. In 2013, the five-number summary for the U.S. statewide percentage of people without health insurance had minimum = 4% (Massachusetts), Q1 = 10%, Med = 12%, Q3 = 15%, and maximum = 20% (Texas).

(a) Sketch a box plot.

(b) Do you think that the distribution is symmetric, skewed to the right, or skewed to the left? Why?

3.42. For a national graduation examination scored out of 110, the 2016 results were reported as minimum = 30,

lower quartile = 41, median = 66, upper quartile = 74, and maximum = 93.

(a) Report and interpret the range and interquartile range.

(b) Calculate the outliers low observations and the high observations.

3.43. Using software, analyze the murder rates from the `Crime` data file at the text website.

(a) Find the five-number summary.

(b) Construct a box plot, and interpret.

(c) Repeat the analyses, adding the D.C. murder rate of 15.9 to the data file, and compare results.

3.44. A report by the OECD[6] indicated that annual water consumption for nations in the OECD (see Exercise 3.6) was skewed to the right, with values (in cubic meters per capita) having a median of about 500 and ranging from about 200 in Denmark to 1700 in the United States Consider the possible values for the IQR: $-10, 0, 10, 350, 1500$. Which is the most realistic value? Why?

3.45. According to values from Table 3.9, for the nations in the European Union (EU) excluding Luxembourg, carbon dioxide emissions (metric tons per capita) had a mean of 7.4 and standard deviation of 1.7.

(a) How many standard deviations above the mean was the value of 20.9 for Luxembourg?

(b) Sweden's observation was 5.5. How many standard deviations below the mean was it?

(c) The carbon dioxide emissions were 14.1 for Canada and 17.0 for the United States Relative to the distribution for the European Union, find and interpret the z-score for (i) Canada, (ii) the United States.

3.46. The United Nations publication *Energy Statistics Yearbook* lists consumption of energy (`unstats. un.org/unsd/energy`). For the 27 nations that made up the European Union (EU) in 2011, the energy values (in kilowatt-hours per capita) had a mean of 5963 and a standard deviation of 2292.

(a) France had a value of 7946. How many standard deviations from the mean was it?

(b) The value for the United States was 13,930. Relative to the distribution for the European Union, find its z-score. Interpret.

(c) If the distribution of EU energy values were bell shaped, would a value of 13,930 be unusually high? Why?

3.47. A study compares three types of personalities (Extrovert, Ambivalent, Introvert) on their level of sociability, that is, very social, social, asocial.

(a) In this study, identify the explanatory variable and the response variable.

---

[6] *OECD Key Environmental Indicators.*

(b) How will you display this data in a contingency table?

3.48. Table 3.18 shows reported happiness for those subjects in the 2014 GSS who attend religious services rarely and for those who attend frequently.

(a) Identify the response variable and the explanatory variable.

(b) At each level of religious attendance, find the percentage who reported being very happy.

(c) Does there seem to be an association between these variables? Why?

TABLE 3.18

Religious	Happiness			
Attendance	Very Happy	Pretty Happy	Not Too Happy	Total
At least once a week	247	287	67	601
No more than once a year	298	723	165	1186

3.49. For 2014 World Bank data (`data.worldbank.org`) for several nations, a prediction equation relating fertility (the mean number of children per adult woman) and percentage of people using the Internet is

$$\text{Predicted fertility} = 3.20 - 0.02(\text{Internet use}).$$

(a) Compare the predicted fertility of a nation with 87% use of the Internet (Canada) to a nation with 0% use (North Korea).

(b) The correlation is -0.55. Explain what the negative value represents.

(c) The correlation for these nations between fertility and percentage of people using contraceptive methods is -0.89. Which variable seems to be more strongly associated with fertility—Internet use, or contraceptive use? Why?

3.50. Refer to the previous exercise. Using regression, the prediction equation relating GDP (in thousands of dollars per capita) to carbon dioxide emissions (in metric tons per capita) is

$$\text{Predicted } CO_2 = 1.93 + 0.178(\text{GDP}).$$

(a) What type of pattern would you expect for the points in a scatterplot for these data?

(b) In this data set, the highest CO_2 value was 17.0 metric tons, for the United States. Its GDP was 53.1. Find its predicted CO_2 use, according to the regression analysis.

(c) In this data set, GDP ranged from a low of 4.4 to a high of 62.9. Find the range of predicted CO_2 values.

3.51. For the data for OECD nations in Table 3.13 in Exercise 3.6 and in the UN data file, use software to construct a scatterplot relating $x =$ carbon dioxide emissions (CO_2) and $y =$ fertility rate.

(a) Based on this plot, would you expect the correlation between these variables to be positive, or negative? Why?

(b) Do you see an observation that falls apart from the others? Identify the nation.

(c) The correlation with CO_2 is 0.67 for GDP and -0.55 for the gender inequality index. Which variable is more strongly associated with CO_2? Why?

3.52. Using national surveys, the Pew Research Center has estimated the percent of people who say that religion plays a very important role in their lives. Values for OECD nations (with GDP values in parentheses from the OECD data file) include Australia 21% (43,550), Canada 25% (43,247), France 14% (36,907), Germany 21% (43,332), Greece 35% (25,651), Italy 32% (34,303), Japan 10% (36,315), Spain 22% (32,103), Turkey 71% (18,975), the United Kingdom 18% (36,197), and the United States 59% (53,143). Construct a scatterplot of these values (as y) against GDP. Summarize what you learn, highlighting any apparent outliers.

3.53. What is the difference between the descriptive measures symbolized by **(a)** $\bar{y}$ and μ? **(b)** s and σ?

Concepts and Applications

3.54. For the `Students` data file at the text website (see Exercise 1.11 on page 21), use software to conduct graphical and numerical summaries for **(a)** distance from home town, **(b)** weekly hours of TV watching. Describe the shapes of the distributions, and summarize your findings.

3.55. Refer to the data file your class created for Exercise 1.12 (page 22). For variables chosen by your instructor, conduct descriptive statistical analyses. In your report, give an example of a research question that could be addressed using your analyses, identifying response and explanatory variables. Summarize and interpret your findings.

3.56. Table 3.19, the `Guns` data file at the text website, shows annual homicide rates (including homicide, suicide,

TABLE 3.19

Nation	Gun Deaths	Nation	Gun Deaths	Nation	Gun Deaths
Australia	1.1 (15)	Greece	5.9 (22)	Norway	0.4 (31)
Austria	1.8 (30)	Iceland	3.2 (30)	Portugal	4.8 (8)
Belgium	2.9 (17)	Israel	9.4 (7)	Spain	1.5 (10)
Canada	5.1 (31)	Italy	8.1 (12)	Sweden	1.9 (32)
Denmark	2.2 (12)	Japan	0.4 (1)	Switzerland	2.3 (46)
Finland	2.6 (45)	Luxembourg	6.0 (15)	United Kingdom	0.4 (6)
France	2.2 (31)	Netherlands	2.0 (4)	United States	28.3 (89)
Germany	2.0 (30)	New Zealand	2.6 (23)		

Source: `www.smallarmssurvey.org`. Values in parentheses are number of firearms (per 100 people).

and accidental deaths) per million population in advanced industrialized nations. The values in parentheses are the nation's number of firearms, per 100 people. Prepare a report in which you summarize the data using graphical and numerical methods from this chapter.

3.57. In 2014, UNICEF reported child poverty rates[7] for many nations in 2012 and in 2008. In 2012, values in Scandinavia were Norway 5.3%, Finland 8.8%, Denmark 10.2%, and Sweden 12.1%, and values in North America were Canada 20.8%, United States 32.2%, and Mexico 34.3%.

(a) Use descriptive statistical methods to summarize and compare Scandinavia and North America.

(b) Compare 2012 and 2008 poverty rates for the full data set of 41 nations in the Poverty data file at the text website. Overall, did the distribution change much? Also, take differences between 2012 and 2008 for each nation and analyze the changes in poverty rates.

3.58. For Table 3.9, pose a research question for two variables relating to the direction of their association, identifying the response variable and explanatory variable. Using software, construct a scatterplot and find the correlation. Interpret, and indicate what they suggest about the research question.

3.59. Zagat restaurant guides publish ratings of restaurants for many large cities around the world (see www.zagat.com). The review for each restaurant gives a verbal summary as well as a 0-to-30-point rating of the quality of food, decor, service, and the cost of a dinner with one drink and tip. Figure 3.19 shows side-by-side box plots of the cost for Italian restaurants in Boston, London, and New York (Little Italy and Greenwich Village neighborhoods). Summarize what you learn from these plots.

3.60. Use software to load the UN data file shown in Table 3.9.

(a) Conduct a descriptive statistical analysis of the prison rates. Summarize your conclusions, highlighting any unusual observations.

(b) Use scatterplots and correlations to investigate the association between prison rate and the other variables in the table. Summarize the results for one of those associations.

3.61. For Table 3.19, construct a scatterplot to investigate the association between gun deaths and number of firearms. Identify the effect of any outlier.

3.62. The Internet site www.pearsonglobaleditions .com/Agresti has useful applets for illustrating data analyses and properties of statistical methods.

(a) Using the *Explore Quantitative Data* applet, construct a sample of 20 observations on y = number of hours of physical exercise in the past week having $\bar{y} < s$. What aspect of the shape of the distribution causes this to happen?

FIGURE 3.19

---
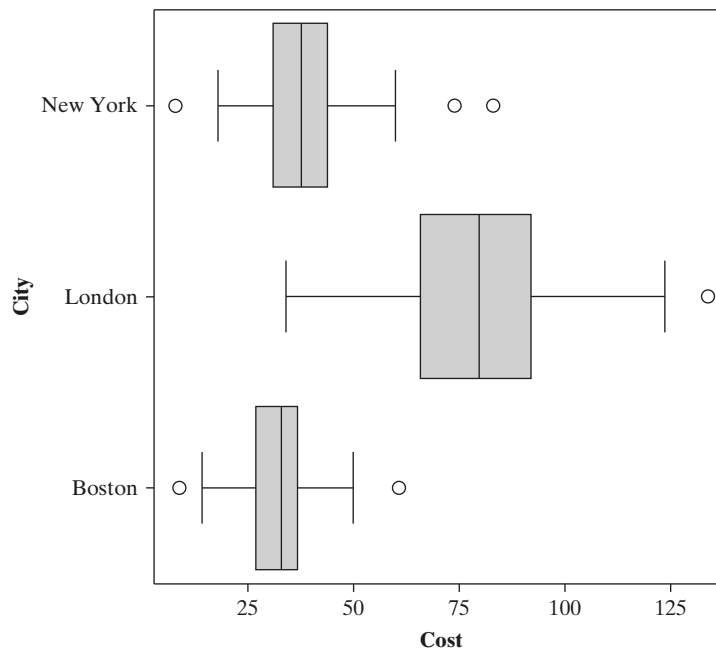
[7] See www.unicef-irc.org/publications/pdf/rc12-eng-web.pdf.

(b) Using the *Explore Linear Regression* applet with the *Draw Own* option, create 20 data points that are plausible for $x = $ number of hours of exercise last week and $y = $ number of hours of exercise this week. (Note that you can adjust the ranges of the axes in the applet.) Describe your data by the correlation and by the linear regression line, and interpret them.

3.63. When a distribution is highly skewed, which describes the typical value better, the median or the mean? Which one is more likely to be impacted by outliers, the median or the mean?

3.64. What is the size of the mean compared to the median when a distribution skews to the left? What about when it skews to the right?

3.65. The fertility rate (mean number of children per adult woman) varies in European countries between a low of 1.2 (Poland and Portugal) and a high of 1.9 (Ireland and France). For each woman, the number of children is a whole number, such as 0 or 1 or 2. Explain why it makes sense to measure a mean number of children per adult woman (which is not a whole number), for example, to compare these rates among European countries or with Canada (1.6), the United States (1.9), and Mexico (2.2).

3.66. According to a national weather service, during the month of June 2016, 4% of days were 20° Celsius or lower, 10% were 21° Celsius or lower, 26% were 22° Celsius or lower, 48% were 24° Celsius or lower, 67% were 25° Celsius or lower, and 96% were 27° Celsius or lower. These are called cumulative percentages.

(a) Find the median temperature for the month of June.

(b) Nearly all the daily temperatures fall between 26 and 19° Celsius, with less than 1% falling outside that range. If the temperatures are approximately bell shaped, give a rough approximation for the standard deviation. Explain your reasoning.

3.67. Give an example of a variable for which the mode applies, but not the mean or median.

3.68. Give an example of a variable having a distribution that you expect to be **(a)** approximately symmetric, **(b)** skewed to the right, **(c)** skewed to the left, **(d)** bimodal, **(e)** skewed to the right, with a mode and median of 0 but a positive mean.

3.69. To measure center, why is the **(a)** median sometimes preferred over the mean? **(b)** mean sometimes preferred over the median? In each case, give an example to illustrate your answer.

3.70. To measure variability, why is

(a) The standard deviation s usually preferred over the range?

(b) The IQR sometimes preferred to s?

3.71. Answer true or false to the following:

(a) The mean, median, and mode can never all be the same.

(b) The mean is always one of the data points.

(c) The median is the same as the second quartile and the 50th percentile.

(d) For 67 sentences for murder recently imposed using U.S. Sentencing Commission guidelines, the median length was 160 months and the mean was 251 months. This distribution is probably skewed to the right.

For multiple-choice problems 3.72–3.74, select the best response.

3.72. A national polling organization reported a recent survey of levels of education (high school, some college, college, master's, doctorate/professional). For level of education, the frequencies were (48%, 11%, 24%, 9%, 7%), respectively.

(a) The median level of education is doctorate/professional.

(b) Only 3% of the subjects fall within one standard deviation of the mean.

(c) The mode is high school.

(d) The master's response is an outlier.

3.73. A national survey asked whether capital punishment is (always wrong, almost always wrong, wrong only sometimes, not wrong at all). The response counts in these four categories were (621, 172, 112, 275). This distribution is

(a) Skewed to the right.

(b) Approximately bell shaped.

(c) Somewhat bimodal, but with overall mode "always wrong."

(d) Shape does not make sense, since the variable is nominal.

3.74. In a study of graduate students who took the Graduate Record Exam (GRE), the Educational Testing Service recently reported that for the quantitative exam, U.S. citizens had a mean of 529 and standard deviation of 127, whereas the non-U.S. citizens had a mean of 649 and standard deviation of 129.

(a) Both groups had about the same amount of variability in their scores, but non-U.S. citizens performed better, on the average, than U.S. citizens.

(b) If the distribution of scores was approximately bell shaped, then almost no U.S. citizens scored below 400.

(c) If the scores range between 200 and 800, then probably the scores for non-U.S. citizens were symmetric and bell shaped.

(d) A non-U.S. citizen who scored three standard deviations below the mean had a score of 200.

3.75. A teacher summarizes grades on the midterm exam by

Min = 26, Q1 = 67, Median = 80, Q3 = 87, Max = 100,
Mean = 76, Mode = 100, Standard dev. = 76, IQR = 20.

She incorrectly recorded one of these. Which one do you think it was? Why?

3.76. Ten people are randomly selected in Florida and another 10 people are randomly selected in Alabama. Table 3.20 provides summary information on mean income. The mean is higher in Alabama both in rural areas and in urban areas. Which state has the larger overall mean income? (The reason for this apparent paradox is that mean urban incomes are larger than mean rural incomes for both states and the Florida sample has a higher proportion of urban residents.)

TABLE 3.20

State	Rural		Urban	
Florida	$26,000	($n = 3$)	$39,000	($n = 7$)
Alabama	$27,000	($n = 8$)	$40,000	($n = 2$)

3.77. For Table 3.2 (page 43), explain why the mean of these 50 observations is not necessarily the same as the violent crime rate for the entire U.S. population.

3.78. For a sample with mean $\bar{y}$, adding a constant c to each observation changes the mean to $\bar{y} + c$, and the standard deviation s is unchanged. Multiplying each observation by c changes the mean to $c\bar{y}$ and the standard deviation to $|c|s$.

(a) Scores on a difficult exam have a mean of 57 and a standard deviation of 20. The teacher boosts all the scores by 20 points before awarding grades. Report the mean and standard deviation of the boosted scores.

(b) Suppose that annual income of Canadian lawyers has a mean of $100,000 and a standard deviation of $30,000. Values are converted to British pounds for presentation to a British audience. If one British pound equals $2.00, report the mean and standard deviation in British currency.

(c) Observations from a survey that asks about the number of miles traveled each day on mass transit are to be converted to kilometer units (1 mile = 1.6 kilometers). Explain how to find the mean and standard deviation of the converted observations.

3.79. * Show that $\sum(y_i - \bar{y})$ must equal 0 for any collection of observations $y_1, y_2, \ldots, y_n$.

3.80. * The Russian mathematician Tchebysheff proved that for any $k > 1$, the proportion of observations that fall more than k standard deviations from the mean can be no greater than $1/k^2$. This holds for *any* distribution, not just bell-shaped ones.

(a) Find the upper bound for the proportion of observations falling (i) more than 2 standard deviations from the mean, (ii) more than 3 standard deviations from the mean, (iii) more than 10 standard deviations from the mean.

(b) Compare the upper bound for $k = 2$ to the approximate proportion falling more than 2 standard deviations from the mean in a bell-shaped distribution. Why is there a difference?

3.81. * The *least squares* property of the mean states that the data fall closer to $\bar{y}$ than to any other number c, in the sense that the sum of squares of deviations about the mean is smaller than the sum of squares of deviations about c. That is,

$$\sum(y_i - \bar{y})^2 < \sum(y_i - c)^2.$$

If you have studied calculus, prove this property by treating $f(c) = \sum(y_i - c)^2$ as a function of c and deriving the value of c that provides a minimum. (*Hint*: Take the derivative of $f(c)$ with respect to c and set it equal to zero.)

PROBABILITY DISTRIBUTIONS

Compared to most mathematical sciences, statistical science is young. Methods of statistical inference were developed within the past century. By contrast, **probability**, the subject of this chapter, has a long history. For instance, mathematicians used probability in France in the seventeenth century to evaluate various gambling strategies. Probability is a highly developed subject, but this chapter limits attention to the basics that we'll need for statistical inference.

Following an introduction to probability, we introduce **probability distributions**, which provide probabilities for all the possible outcomes of a variable. The **normal distribution**, described by a bell-shaped curve, is the most important probability distribution for statistical inference. The **sampling distribution** is a fundamentally important type of probability distribution that we need to conduct statistical inference. It enables us to predict how close a sample mean falls to the population mean. The main reason for the importance of the normal distribution is the remarkable result that sampling distributions are usually bell shaped.

4.1 Introduction to Probability

In Chapter 2, we learned that randomness is a key component of good ways to gather data. For each observation in a random sample or randomized experiment, the possible outcomes are known, but it's uncertain which will occur.

PROBABILITY AS A LONG-RUN RELATIVE FREQUENCY

For a particular possible outcome for a random phenomenon, the **probability** of that outcome is the proportion of times that the outcome would occur in a very long sequence of observations.

Probability

> With a random sample or randomized experiment, the **probability** that an observation has a particular outcome is the proportion of times that outcome would occur in a very long sequence of like observations.

Later in this chapter, we'll analyze data for the 2014 California gubernatorial election, for which the winner was the Democratic party candidate, Jerry Brown. We'll use an exit poll that interviewed a random sample of voters in that election and asked whom they voted for. Suppose that the population proportion who voted for Brown is 0.60. Then, the probability that a randomly selected person voted for Brown is 0.60.

Why does probability refer to the *long run*? Because when you do not already know or assume some value for a probability, you need a large number of

observations to accurately assess it. If you sample only 10 people and they are all right-handed, you can't conclude that the probability of being right-handed equals 1.0.

This book defines a probability as a proportion, so it is a number between 0 and 1. In practice, probabilities are often expressed also as percentages, then falling between 0 and 100. For example, if a weather forecaster says that the probability of rain today is 70%, this means that in a long series of days with atmospheric conditions like those today, rain occurs on 70% of the days.

This *long-run* approach is the standard way to define probability. This definition is not always applicable, however. It is not meaningful, for instance, for the probability that human beings have a life after death, or the probability that intelligent life exists elsewhere in the universe. If you start a new business, you will not have a long run of trials with which to estimate the probability that the business is successful. You must then rely on *subjective* information rather than solely on *objective* data. In the subjective approach, the probability of an outcome is defined to be your degree of belief that the outcome will occur, based on the available information, such as data that may be available from experiences of others. A branch of statistical science uses subjective probability as its foundation. It is called *Bayesian statistics*, in honor of an eighteenth-century British clergyman (Thomas Bayes) who discovered a probability rule on which it is based.

BASIC PROBABILITY RULES

Next, we'll present four rules for finding probabilities. We won't try to explain them with precise, mathematical reasoning, because for our purposes it suffices to have an intuitive feel for what each rule says.

Let $P(A)$ denote the probability of a particular possible outcome denoted by the letter A. Then,

- **$P(\text{not } A) = 1 - P(A)$.**

    If you know the probability a particular outcome occurs, then the probability it does *not* occur is 1 minus that probability. Suppose A represents the outcome that a randomly selected person favors legalization of same-sex marriage. If $P(A) = 0.66$, then $1 - 0.66 = 0.34$ is the probability that a randomly selected person does *not* favor legalization of same-sex marriage.

- **If A and B are distinct possible outcomes (with no overlap), then $P(A \text{ or } B) = P(A) + P(B)$.**

    In a survey to estimate the population proportion of people who favor legalization of marijuana, let A represent the sample proportion estimate being much too low, say more than 0.10 *below* the population proportion. Let B represent the sample proportion estimate being much too high—at least 0.10 *above* the population proportion. These are two distinct possible outcomes. From methods in this chapter, perhaps $P(A) = P(B) = 0.03$. Then, the overall probability the sample proportion is in error by more than 0.10 (without specifying the direction of error) is

$$P(A \text{ or } B) = P(A) + P(B) = 0.03 + 0.03 = 0.06.$$

- **If A and B are possible outcomes, then $P(A \text{ and } B) = P(A) \times P(B \text{ given } A)$.**

    From U.S. Census data, the probability that a randomly selected American adult is married equals 0.56. Of those who are married, General Social Surveys estimate that the probability a person reports being *very happy* when asked to

choose among (very happy, pretty happy, not too happy) is 0.40; that is, given you are married, the probability of being very happy is 0.40. So,

$$P(\text{married and very happy}) =$$

$$P(\text{married}) \times P(\text{very happy given married}) = 0.56 \times 0.40 = 0.22.$$

About 22% of the adult population is both married *and* very happy. The probability $P(B$ given $A)$ is called a ***conditional probability*** and is often denoted by $P(B \mid A)$.

In some cases, A and B are "independent," in the sense that whether one occurs does not depend on whether the other does. That is, $P(B$ given $A) = P(B)$, so the previous rule simplifies:

- **If A and B are independent, then $P(A$ and $B) = P(A) \times P(B)$.**

For example, suppose that 60% of a population supports a carbon tax to diminish impacts of carbon dioxide levels on global warming. In random sampling from that population, let A denote the probability that the first person sampled supports the carbon tax and let B denote the probability that the second person sampled supports it. Then $P(A) = 0.60$ and $P(B) = 0.60$. With random sampling, successive observations are independent, so the probability that *both* people support a carbon tax is

$$P(A \text{ and } B) = P(A) \times P(B) = 0.60 \times 0.60 = 0.36.$$

This extends to multiple independent events. For 10 randomly sampled people, the probability that all 10 support a carbon tax is $0.60 \times 0.60 \times \cdots \times 0.60 = (0.60)^{10} = 0.006$.

4.2 Probability Distributions for Discrete and Continuous Variables

A variable can take at least two different values. For a random sample or randomized experiment, each possible outcome has a probability that it occurs. The variable itself is sometimes then referred to as a ***random variable***. This terminology emphasizes that the outcome varies from observation to observation according to random variation that can be summarized by probabilities. For simplicity, we'll continue to use the "variable" terminology regardless of whether the variation has a random aspect.

Recall (from Section 2.1) that a variable is *discrete* if the possible outcomes are a set of separate values, such as a variable expressed as "the number of ..." with possible values 0, 1, 2, ... It is *continuous* if the possible outcomes are an infinite continuum, such as all the real numbers between 0 and 1. A ***probability distribution*** lists the possible outcomes and their probabilities.

PROBABILITY DISTRIBUTIONS FOR DISCRETE VARIABLES

The probability distribution of a *discrete* variable assigns a probability to each possible value of the variable. Each probability is a number between 0 and 1. The sum of the probabilities of all possible values equals 1.

Let $P(y)$ denote the probability of a possible outcome for a variable y. Then,

$$0 \le P(y) \le 1 \ \text{ and } \ \sum_{\text{all } y} P(y) = 1,$$

where the sum is over all the possible values of the variable.

Example
4.1

Ideal Number of Children for a Family   Let y denote the response to the question "What do you think is the ideal number of children for a family to have?" This is a discrete variable, taking the possible values 0, 1, 2, 3, and so forth. According to recent General Social Surveys, for a randomly chosen person in the United States the probability distribution of y is approximately as Table 4.1 shows. The table displays the recorded y-values and their probabilities. For instance, $P(4)$, the probability that $y = 4$ children is regarded as ideal, equals 0.12. Each probability in Table 4.1 is between 0 and 1, and the sum of the probabilities equals 1.   ∎

TABLE 4.1: Probability Distribution of y = Ideal Number of Children for a Family	
y	$P(y)$
0	0.01
1	0.03
2	0.60
3	0.23
4	0.12
5	0.01
Total	1.00

A *histogram* can portray the probability distribution. The rectangular bar over a possible value of the variable has height equal to the probability of that value. Figure 4.1 is a histogram for the probability distribution of the ideal number of children, from Table 4.1. The bar over the value 4 has height 0.12, the probability of the outcome 4.

FIGURE 4.1: Histogram for the Probability Distribution of the Ideal Number of Children for a Family
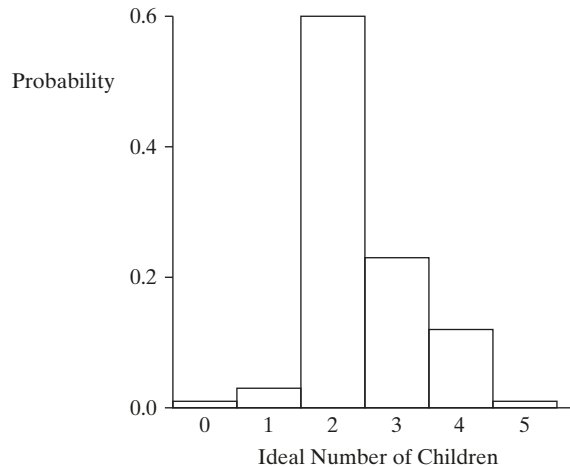
PROBABILITY DISTRIBUTIONS FOR CONTINUOUS VARIABLES

Continuous variables have an infinite continuum of possible values. Probability distributions of continuous variables assign probabilities to *intervals* of numbers. The probability that a variable falls in any particular interval is between 0 and 1, and the probability of the interval containing all the possible values equals 1.

A graph of the probability distribution of a continuous variable is a smooth, continuous curve. The *area* under the curve[1] for an interval of values represents the probability that the variable takes a value in that interval.

Commuting Time to Work   A recent U.S. Census Bureau study about commuting time for workers in the United States who commute to work[2] measured $y =$ travel time, in minutes. The probability distribution of y provides probabilities such as $P(y < 15)$, the probability that travel time is less than 15 minutes, or $P(30 < y < 60)$, the probability that travel time is between 30 and 60 minutes.

Figure 4.2 portrays the probability distribution of y. The shaded area in the figure refers to the region of values higher than 45. This area equals 15% of the total area under the curve, representing the probability of 0.15 that commuting time is more than 45 minutes. Those regions in which the curve has relatively high height have the values most likely to be observed. ∎

FIGURE 4.2: Probability Distribution of Commuting Time to Work. The area under the curve between two points represents the probability of that interval of values.
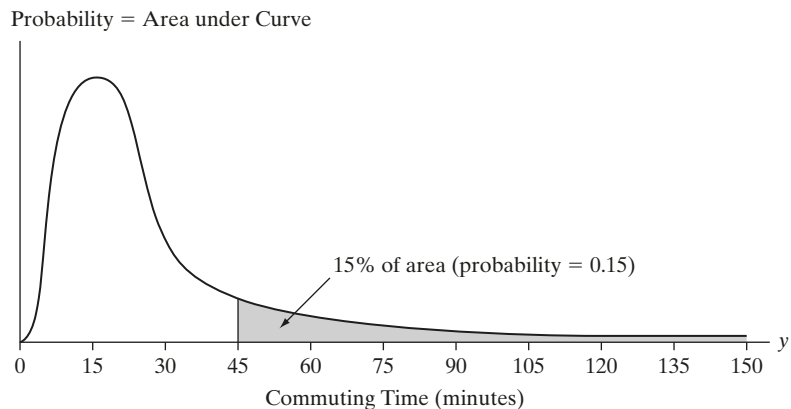
PARAMETERS DESCRIBE PROBABILITY DISTRIBUTIONS

Some probability distributions have formulas for calculating probabilities. For others, tables or software provide the probabilities. Section 4.3 shows how to find probabilities for the most important probability distribution.

Section 3.1 introduced the *population distribution* of a variable. This is, equivalently, the probability distribution of the variable for a subject selected randomly from the population. For example, if 0.12 is the population proportion of adults who believe the ideal number of children is 4, then the probability that an adult selected randomly from that population believes this is also 0.12.

Like a population distribution, a probability distribution has *parameters* describing center and variability. The *mean* describes center and the *standard deviation* describes variability. The parameter values are the values these measures would assume, *in the long run*, if the randomized experiment or random sample repeatedly took observations on the variable y having that probability distribution.

For example, suppose we take observations from the distribution in Table 4.1. Over the long run, we expect $y = 0$ to occur 1% of the time, $y = 1$ to occur 3% of the time, and so forth. In 100 observations, for instance, we expect about

one 0, 3 1's, 60 2's, 23 3's, 12 4's, and one 5.

_____

[1] Mathematically, this calculation uses integral calculus. The probability that y falls in the interval between points a and b is the integral over that interval of the function for the curve.
[2] See www.census.gov/hhes/commuting.

In that case, since the mean equals the total of the observations divided by the sample size, the mean equals

$$\frac{0(1) + 1(3) + 2(60) + 3(23) + 4(12) + 5(1)}{100} = \frac{245}{100} = 2.45.$$

This calculation has the form

$$0(0.01) + 1(0.03) + 2(0.60) + 3(0.23) + 4(0.12) + 5(0.01),$$

the sum of the possible outcomes times their probabilities. In fact, for any discrete variable y, the mean of its probability distribution has this form.

Mean of a Probability Distribution (Expected Value)

> The *mean of the probability distribution* for a discrete variable y is
>
> $$\mu = \sum y P(y).$$
>
> The sum is taken over all possible values of the variable. This parameter is also called the *expected value of y* and denoted by $E(y)$.

For Table 4.1, for example,

$$\mu = \sum y P(y) = 0P(0) + 1P(1) + 2P(2) + 3P(3) + 4P(4) + 5P(5)$$

$$= 0(0.01) + 1(0.03) + 2(0.60) + 3(0.23) + 4(0.12) + 5(0.01)$$

$$= 2.45.$$

This is also the *expected value* of y, $E(y) = 2.45$. The terminology reflects that $E(y)$ represents what we expect for the average value of y in a long series of observations.

The *standard deviation* of a probability distribution, denoted by σ, measures its variability. The more spread out the distribution, the larger the value of σ. The Empirical Rule (Section 3.3) helps us to interpret σ. If a probability distribution is bell shaped, about 68% of the probability falls between $\mu - \sigma$ and $\mu + \sigma$, about 95% falls between $\mu - 2\sigma$ and $\mu + 2\sigma$, and all or nearly all falls between $\mu - 3\sigma$ and $\mu + 3\sigma$.

The standard deviation is the square root of the *variance* of the probability distribution. The variance measures the average squared deviation of an observation from the mean. That is, it is the expected value of $(y - \mu)^2$. In the discrete case, the formula is

$$\sigma^2 = E(y - \mu)^2 = \sum (y - \mu)^2 P(y).$$

We shall not need to calculate σ^2, so we shall not further consider this formula here.

4.3 The Normal Probability Distribution

Some probability distributions are important because they approximate well sample data in the real world. Some are important because of their uses in statistical inference. This section introduces the *normal probability distribution*, which is important for both reasons.

Normal Distribution

> The *normal distribution* is symmetric, bell shaped, and characterized by its mean μ and standard deviation σ. The probability within any particular number of standard deviations of μ is the same for all normal distributions. This probability (rounded off) equals 0.68 within 1 standard deviation, 0.95 within 2 standard deviations, and 0.997 within 3 standard deviations.

Each normal distribution[3] is specified by its mean μ and standard deviation σ. For any real number for μ and any nonnegative number for σ, there is a normal distribution having that mean and standard deviation. Figure 4.3 illustrates this. Essentially the entire distribution falls between $\mu - 3\sigma$ and $\mu + 3\sigma$.

FIGURE 4.3: For Every Normal Distribution, the Probability (Rounded) Equals 0.68 within σ of μ, 0.95 within 2σ of μ, and 0.997 within 3σ of μ
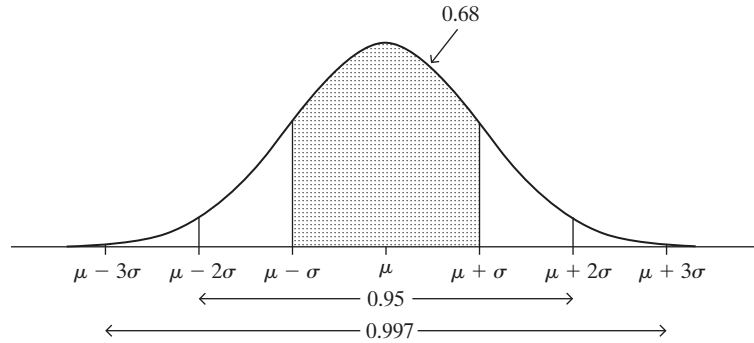
For example, heights of adult females in North America have approximately a normal distribution with $\mu = 65.0$ inches and $\sigma = 3.5$. The probability is nearly 1.0 that a randomly selected female has height between $\mu - 3\sigma = 65.0 - 3(3.5) = 54.5$ inches and $\mu + 3\sigma = 65.0 + 3(3.5) = 75.5$ inches. Adult male height has a normal distribution with $\mu = 70.0$ and $\sigma = 4.0$ inches. So, the probability is nearly 1.0 that a randomly selected male has height between $\mu - 3\sigma = 70.0 - 3(4.0) = 58$ inches and $\mu + 3\sigma = 70.0 + 3(4.0) = 82$ inches. See Figure 4.4.

FIGURE 4.4: Normal Distributions for Women's Height ($\mu = 65, \sigma = 3.5$) and for Men's Height ($\mu = 70, \sigma = 4.0$)
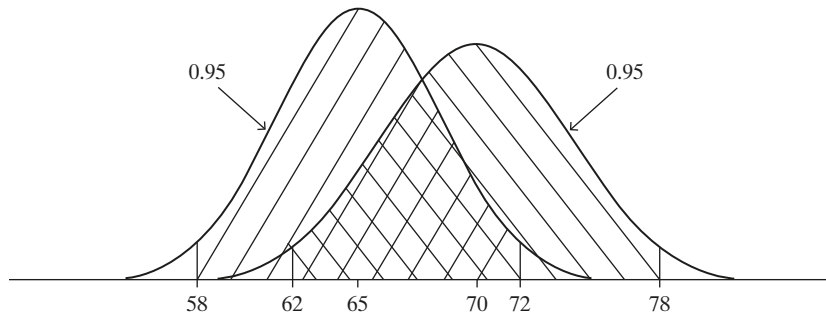
FINDING NORMAL PROBABILITIES: TABLES, SOFTWARE, AND APPLETS

For the normal distribution, for each fixed number z, the probability that is within z standard deviations of the mean depends only on the value of z. This is the area under the normal curve between $\mu - z\sigma$ and $\mu + z\sigma$. For every normal distribution, this probability is 0.68 for $z = 1$, 0.95 for $z = 2$, and nearly 1.0 for $z = 3$.

For a normal distribution, the probability concentrated within $z\sigma$ of μ is the same for all normal curves even if z is not a whole number—for instance, $z = 1.43$ instead of 1, 2, or 3. Table A, also shown next to the inside back cover, determines probabilities

---

[3] More technically, the normal distribution with mean μ and standard deviation σ is represented by a bell-shaped curve that has the formula

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-[(y-\mu)^2/2\sigma^2]}.$$

for any region of values. It tabulates the probability for the values falling in the right tail, at least z standard deviations above the mean. The left margin column of the table lists the values for z to one decimal point, with the second decimal place listed above the columns.

Table 4.2 displays a small excerpt from Table A. The probability for $z = 1.43$ falls in the row labeled 1.4 and in the column labeled .03. It equals 0.0764. This means that for every normal distribution, the right-tail probability above $\mu + 1.43\sigma$ (i.e., more than 1.43 standard deviations above the mean) equals 0.0764.

TABLE 4.2: Part of Table A Displaying Normal Right-Tail Probabilities

					Second Decimal Place of z					
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
					....					
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559

Since the entries in Table A are probabilities for the right half of the normal distribution above $\mu + z\sigma$, they fall between 0 and 0.50. By the symmetry of the normal curve, these right-tail probabilities also apply to the left tail below $\mu - z\sigma$. For example, the probability below $\mu - 1.43\sigma$ also equals 0.0764. The left-tail probabilities are called ***cumulative probabilities***.

We can also use statistical software to find normal probabilities. The free software R has a function `pnorm` that gives the cumulative probability falling below $\mu + z\sigma$. For example, *pnorm(2.0)* provides the cumulative probability falling below $\mu + 2.0\sigma$:

```
> pnorm(2.0)    # cumulative probability below mu + 2.0(sigma)
[1] 0.97724987 # right-tail probability = 1 - 0.977 = 0.023
```

In the Stata software, we can use the `display normal` command:

```
. display normal(2.0) # cumulative probability below mu + 2.0(sigma)
.97724987              # right-tail probability = 1 - 0.977 = 0.023
```

We subtract the cumulative probability from 1 to find the right-tail probability above $\mu + 2.0\sigma$. That is, the probability $1 - 0.97725 = 0.02275$ falls more than two standard deviations above the mean. By the symmetry of the normal distribution, this is also the probability falling more than two standard deviations below the mean. The probability falling *within* two standard deviations of the mean is $1 - 2(0.02275) = 0.954$. (Here, we've used rule (1) of the probability rules at the end of Section 4.1, that $P(\text{not } A) = 1 - P(A)$.) You can also find normal probabilities with SPSS and SAS software.

Normal probabilities are also available on the Internet, such as with the easy-to-use *Normal Distribution* applet[4] for which there is a link at `www.pearsonglobal editions.com/Agresti`. See Figure 4.5.

---

[4] This is one of several applets we shall use that were developed by Prof. Bernhard Klingenberg for the text *Statistics: The Art and Science of Learning from Data*, 4th ed., by A. Agresti, C. Franklin, and B. Klingenberg (Pearson, 2017).

NORMAL PROBABILITIES AND THE EMPIRICAL RULE

Probabilities for the normal distribution apply *approximately* to other bell-shaped distributions. They yield the probabilities for the Empirical Rule. Recall (page 56) that that rule states that for bell-shaped histograms, about 68% of the data fall within one standard deviation of the mean, 95% within two standard deviations, and all or nearly all within three standard deviations. For example, we've just used software to find that for normal distributions the probability falling within two standard deviations of the mean is 0.954. For one and for three standard deviations, we find central probabilities of 0.683 and 0.997, respectively.

The approximate percentages in the Empirical Rule are the actual percentages for the normal distribution, rounded to two decimal places. The Empirical Rule stated the percentages as being *approximate* rather than *exact*. Why? Because that rule referred to *all approximately bell-shaped distributions*, not only the normal distribution. Not all bell-shaped distributions are normal, only those described by the formula shown in the footnote on page 85. We won't need that formula, but we will use probabilities for it throughout the text.

FINDING z-VALUES FOR CERTAIN TAIL PROBABILITIES

Many inferential methods use z-values corresponding to certain normal curve probabilities. This entails the reverse use of Table A or software or applets. Starting with a tail probability, we find the z-value that provides the number of standard deviations that that number falls from the mean.

To illustrate, let's first use Table A to find the z-value having a right-tail probability of 0.025. We look up 0.025 in the body of Table A, which contains tail probabilities. It corresponds to $z = 1.96$ (i.e., we find .025 in the row of Table A labeled 1.9 and in the column labeled .06). This means that a probability of 0.025 falls above $\mu + 1.96\sigma$. Similarly, a probability of 0.025 falls below $\mu - 1.96\sigma$. So, a total probability of $0.025 + 0.025 = 0.050$ falls more than 1.96σ from μ. We saw in the previous subsection that 95% of a normal distribution falls within two standard deviations of the mean. More precisely, 0.954 falls within 2.00 standard deviations, and here we've seen that 0.950 falls within 1.96 standard deviations.

R software has a function `qnorm` that gives the z-value for a particular cumulative probability. The right-tail probability of 0.025 corresponds to a cumulative probability of $1 - 0.025 = 0.975$, for which the z-value is

```
> qnorm(0.975) # q denotes "quantile"; .975 quantile = 97.5 percentile
[1] 1.959964    # The z-value is 1.96, rounded to two decimals
```

Here is how you can find the z-value for a cumulative probability using the Stata software:

```
. display invnormal(0.975)    /* invnormal = "inverse normal" */
1.959964
```

The `qnorm` function in R is equivalent to the `invnormal` (inverse normal) function in Stata. You can also find this z-value using an Internet applet, such as Figure 4.6 shows with the *Normal Distribution* applet at `www.pearsonglobal editions.com/Agresti`. It is also possible to find z-values with SPSS and SAS software.

FIGURE 4.6: Using the *Normal Distribution* Applet at `www.pearsonglobal editions.com/Agresti` to Find the z-value for a Normal Tail Probability of 0.025 (i.e., 2.5 percent)
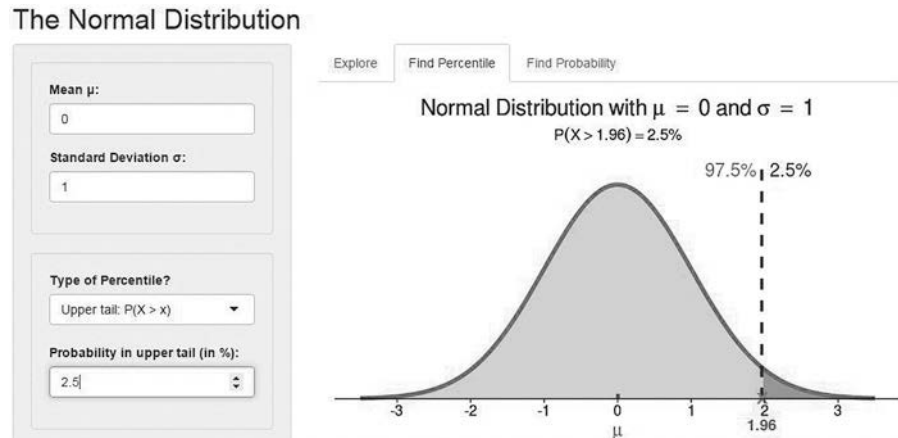
To check that you understand this reasoning, use Table A, software, or the applet to verify that the z-value for a right-tail probability of (1) 0.05 is $z = 1.64$, (2) 0.01 is $z = 2.33$, and (3) 0.005 is $z = 2.58$. Show that 90% of a normal distribution falls between $\mu - 1.64\sigma$ and $\mu + 1.64\sigma$.

Example 4.3

Finding the 99th Percentile of IQ Scores Stanford-Binet IQ scores have approximately a normal distribution with mean $= 100$ and standard deviation $= 16$. What is the 99th percentile of IQ scores? In other words, what is the IQ score that falls above 99% of the scores?

To answer this, we need to find the value of z such that $\mu + z\sigma$ falls above 99% of a normal distribution. Now, for $\mu + z\sigma$ to represent the 99th percentile, the probability below $\mu + z\sigma$ must equal 0.99, by the definition of a percentile. So, 1% of the distribution is above the 99th percentile. The right-tail probability equals 0.01, as Figure 4.7 shows.

FIGURE 4.7: The 99th Percentile for a Normal Distribution Has 99% of the Distribution below that Point and 1% above It
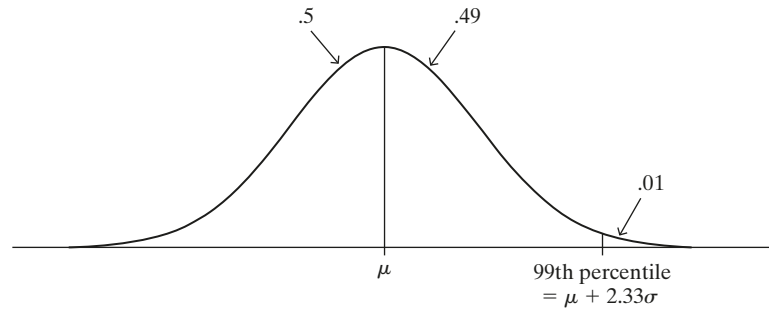
With Table A, software, or the Internet, you can find that the z-value for a cumulative probability of 0.99 or right-tail probability of 0.01 is $z = 2.33$. Thus, the 99th percentile is 2.33 standard deviations above the mean. In summary, 99% of any normal distribution is located below $\mu + 2.33\sigma$.

For IQ scores with mean $= 100$ and standard deviation $= 16$, the 99th percentile equals

$$\mu + 2.33\sigma = 100 + 2.33(16) = 137.$$

That is, about 99% of IQ scores fall below 137. ∎

To check that you understand the reasoning above, show that the 95th percentile of a normal distribution is $\mu + 1.64\sigma$, and show that the 95th percentile for the IQ distribution equals 126.

z-SCORE REPRESENTS THE NUMBER OF STANDARD DEVIATIONS FROM THE MEAN

The z symbol in a normal table refers to the distance between a possible value y of a variable and the mean μ of its probability distribution, in terms of the *number of standard deviations* that y falls from μ.

For example, scores on each portion of the Scholastic Aptitude Test (SAT) have traditionally been approximately normal with mean $\mu = 500$ and standard deviation $\sigma = 100$. The test score of $y = 650$ has a z-score of $z = 1.50$, because 650 is 1.50 standard deviations above the mean. In other words, $y = 650 = \mu + z\sigma = 500 + z(100)$, where $z = 1.50$.

For sample data, Section 3.4 introduced the z-score as a measure of position. Let's review how to find it. The distance between y and the mean μ equals $y - \mu$. The z-score expresses this difference in units of standard deviations.

z-Score

> The z-score for a value y of a variable is the *number of standard deviations* that y falls from the mean. For a probability distribution with mean μ and standard deviation σ, it equals
>
> $$z = \frac{\text{Variable value} - \text{Mean}}{\text{Standard deviation}} = \frac{y - \mu}{\sigma}.$$

To illustrate, when $\mu = 500$ and $\sigma = 100$, a value of $y = 650$ has the z-score of

$$z = \frac{y - \mu}{\sigma} = \frac{650 - 500}{100} = 1.50.$$

Positive z-scores occur when the value for y falls above the mean μ. Negative z-scores occur when the value for y falls below the mean. For example, for SAT scores with $\mu = 500$ and $\sigma = 100$, a value of $y = 350$ has a z-score of

$$z = \frac{y - \mu}{\sigma} = \frac{350 - 500}{100} = -1.50.$$

The test score of 350 is 1.50 standard deviations below the mean. The value $y = 350$ falls below the mean, so the z-score is negative.

The next example shows that z-scores provide a useful way to compare positions for different normal distributions.

Example 4.4

Comparing SAT and ACT Test Scores Suppose that when you applied to college, you took a SAT exam, scoring 550. Your friend took the ACT exam, scoring 30. If the SAT has $\mu = 500$ and $\sigma = 100$ and the ACT has $\mu = 18$ and $\sigma = 6$, then which score is relatively better?

We cannot compare the test scores of 550 and 30 directly, because they have different scales. We convert them to z-scores, analyzing how many standard deviations each falls from the mean. The SAT score of $y = 550$ converts to a z-score of

$$z = \frac{y - \mu}{\sigma} = \frac{550 - 500}{100} = 0.50.$$

The ACT score of $y = 30$ converts to a z-score of $(30 - 18)/6 = 2.0$.

The ACT score of 30 is relatively higher than the SAT score of 550, because 30 is 2.0 standard deviations above its mean whereas 550 is only 0.5 standard deviations above its mean. The SAT and ACT scores both have approximate normal distributions. From Table A, $z = 2.0$ has a right-tail probability of 0.0228 and $z = 0.5$ has a right-tail probability of 0.3085. Of all students taking the ACT, only about 2% scored higher than 30, whereas of all students taking the SAT, about 31% scored higher than 550. In this relative sense, the ACT score is higher. ∎

USING z-SCORES TO FIND PROBABILITIES OR y-VALUES

Here's a summary of how we use z-scores:

- If we have a value y and need to find a probability, convert y to a z-score using $z = (y - \mu)/\sigma$, and then convert z to the probability of interest using a table of normal probabilities, software, or the Internet.

- If we have a probability and need to find a value of y, convert the probability to a tail probability (or cumulative probability) and find the z-score (using a normal table, software, or the Internet), and then evaluate $y = \mu + z\sigma$.

For example, we used the equation $z = (y - \mu)/\sigma$ to determine how many standard deviations a SAT test score of 650 fell from the mean of 500, when $\sigma = 100$ (namely, 1.50). Example 4.3 used the equation $y = \mu + z\sigma$ to find a percentile score for a normal distribution of IQ scores.

THE STANDARD NORMAL DISTRIBUTION

Many inferential statistical methods use a particular normal distribution, called the *standard normal distribution*.

Standard Normal Distribution

> The *standard normal distribution* is the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

For the standard normal distribution, the number falling z standard deviations above the mean is $\mu + z\sigma = 0 + z(1) = z$. It is simply the z-score itself. For instance, the value of 2 is two standard deviations above the mean, and the value of -1.3 is 1.3 standard deviations below the mean. The original values are the same as the z-scores. See Figure 4.8.

FIGURE 4.8: The Standard Normal Distribution Has Mean 0 and Standard Deviation 1. Its ordinary scores are the same as its z-scores.
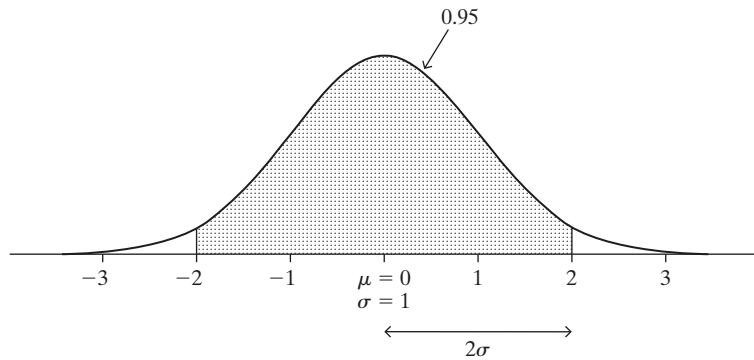
When the values for an arbitrary normal distribution are converted to z-scores, those z-scores are centered around 0 and have a standard deviation of 1. The z-scores have the standard normal distribution.

z-Scores and the Standard Normal Distribution

> If a variable has a normal distribution, and if its values are converted to z-scores by subtracting the mean and dividing by the standard deviation, then the z-scores have the standard normal distribution.

Suppose we convert each SAT score y to a z-score by using $z = (y - 500)/100$. For instance, $y = 650$ converts to $z = 1.50$, and $y = 350$ converts to $z = -1.50$. Then, the entire set of z-scores has a normal distribution with a mean of 0 and a standard deviation of 1. This is the standard normal distribution.

Many inferential methods convert values of statistics to z-scores and then to normal curve probabilities. We use z-scores and normal probabilities often throughout the rest of the book.

BIVARIATE PROBABILITY DISTRIBUTIONS: COVARIANCE AND CORRELATION*

Section 3.5 introduced *bivariate* descriptive statistics that apply to a pair of variables. An example is the sample correlation. Likewise, *bivariate probability distributions* determine joint probabilities for pairs of random variables. For example, the *bivariate normal distribution* generalizes the bell curve over the real line for a single variable y to a bell-shaped surface in three dimensions over the plane for possible values of two variables (x, y).

Each variable in a bivariate distribution has a mean and a standard deviation. Denote them by (μ_x, σ_x) for x and by (μ_y, σ_y) for y. The way that x and y vary together is described by their ***covariance***, which is defined to be

$$\text{Covariance}(x, y) = E[(x - \mu_x)(y - \mu_y)],$$

which represents the average of the cross products about the population means (weighted by their probabilities). If y tends to fall *above* its mean when x falls *above* its mean, the covariance is *positive*. If y tends to fall *below* its mean when x falls *above* its mean, the covariance is *negative*.

The covariance can be any real number. For interpretation, it is simpler to use

$$\text{Correlation}(x, y) = \frac{\text{Covariance}(x, y)}{(\text{Standard deviation of } x)(\text{Standard deviation of } y)}.$$

But this equals

$$\frac{E[(x - \mu_x)(y - \mu_y)]}{\sigma_x \sigma_y} = E\left[\left(\frac{x - \mu_x}{\sigma_x}\right)\left(\frac{y - \mu_y}{\sigma_y}\right)\right] = E(z_x z_y),$$

where $z_x = (x - \mu_x)/\sigma_x$ denotes the z-score for the variable x and $z_y = (y - \mu_y)/\sigma_y$ denotes the z-score for the variable y. That is, the population correlation equals the average cross product of the z-score for x times the z-score for y. It falls between -1 and $+1$. It is positive when positive z-scores for x tend to occur with positive z-scores for y and when negative z-scores for x tend to occur with negative z-scores for y.

We shall not need to calculate these expectations. We can use software to find sample values, as we showed in Table 3.10 for the correlation.

4.4 Sampling Distributions Describe How Statistics Vary

We've seen that probability distributions summarize probabilities of possible outcomes for a variable. Let's now look at an example that illustrates the connection between statistical inference and probability calculations.

Example 4.5

Predicting an Election from an Exit Poll Television networks sample voters on election day to help them predict the winners early. For the fall 2014 election for Governor of California, CBS News[5] reported results of an exit poll of 1824 voters. They stated that 60.5% of their *sample* reported voting for the Democratic party candidate, Jerry Brown. In this example, the probability distribution for a person's vote would state the probability that a randomly selected voter voted for Brown. This equals the proportion of the *population* of voters who voted for him. When the exit poll was taken, this was an unknown population parameter.

To judge whether this is sufficient information to predict the outcome of the election, the network can ask, "Suppose only half the population voted for Brown. Would it then be surprising that 60.5% of the sampled individuals voted for him?" If this would be very unlikely, the network infers that Brown received more than half the population votes and won the election. The inference about the election outcome is based on finding the probability of the sample result under the supposition that the population parameter, the percentage of voters preferring Brown, equals 50%. ∎

---

[5] See www.cbsnews.com/elections/2014/governor/california/exit/.

About 7.3 million people voted in this race. The exit poll sampled only 1824 voters, yet TV networks used it to predict that Brown would win. How could there possibly have been enough information from this poll to make a prediction? We next see justification for making a prediction.

SIMULATING THE SAMPLING PROCESS

A **simulation** can show us how close an exit poll result tends to be to the population proportion voting for a candidate. One way to simulate the vote of a voter randomly chosen from the population is to select a random number using software. Suppose exactly 50% of the population voted for Brown and 50% voted for the Republican candidate, Neel Kashkari. Identify all 50 two-digit numbers between 00 and 49 as Democratic votes and all 50 two-digit numbers between 50 and 99 as Republican votes. Then, each candidate has a 50% chance of selection on each choice of two-digit random number. For instance, the first two digits of the first column of the random numbers table on page 27 provide the random numbers 10, 53, 24, and 42. So, of the first four voters selected, three voted Democratic (i.e., have numbers between 00 and 49) and one voted Republican. Selecting 1824 two-digit random numbers simulates the process of observing the votes of a random sample of 1824 voters of the much larger population (which is actually treated as infinite in size).

To do this, we can use software that generates random numbers or that uses such numbers to simulate flipping a coin repeatedly, where we regard one outcome (say, head) as representing a person who votes for the Democrat and the other outcome (say, tail) as representing a person who votes for the Republican. Here is how we simulated, using an applet on the Internet. We suggest you try this also, to see how it works.

- Go to `www.pearsonglobaleditions.com/Agresti` and click on *Random Numbers*.

- Click on *Coin Flips*.

- The box for *Probability of a Head (in %)* should say 50. Then, random numbers between 00 and 49 correspond to head and random numbers between 50 and 99 correspond to tail. In the box for *How many flips do you want to generate in one simulation?* enter 1824. See Figure 4.9.

- Click *Simulate*.

FIGURE 4.9: The *Random Numbers* Applet for Simulating at `www.pearsonglobal editions.com/Agresti`. When we click on *Simulate*, we see the results of flipping a coin 1824 times when the probability on each flip of getting a head equals 0.50.
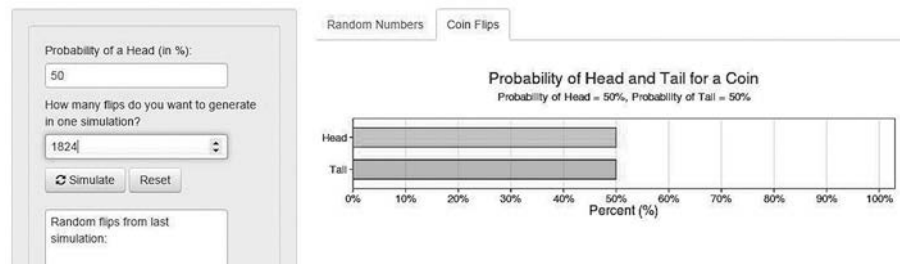
When we performed this simulation, we got 901 heads (Democratic votes) and 923 tails (Republican votes). The sample proportion of Democratic votes was $901/1824 = 0.494$, quite close to the population proportion of 0.50. This particular estimate was good. Were we merely lucky? We repeated the process and simulated

1824 more flips. (In this app, click again on *Simulate*.) This time the sample proportion of Democratic votes was 0.498, also quite good.

Using software,[6] we next performed this process of picking 1824 people 10,000 times so that we could search for a pattern in the results. Figure 4.10 shows a histogram of the 10,000 values of the sample proportion. Nearly all the simulated proportions fell between 0.46 and 0.54, that is, within 0.04 of the population proportion of 0.50. Apparently a sample of size 1824 provides quite a good estimate of a population proportion.

FIGURE 4.10: Results of 10,000 Simulations of the Sample Proportion Favoring the Democratic Candidate, for Random Samples of 1824 Subjects from a Population in which Half Voted for the Democrat and Half Voted for the Republican. In nearly all cases, the sample proportion fell within 0.04 of the population proportion of 0.50 (i.e., between 0.46 and 0.54).
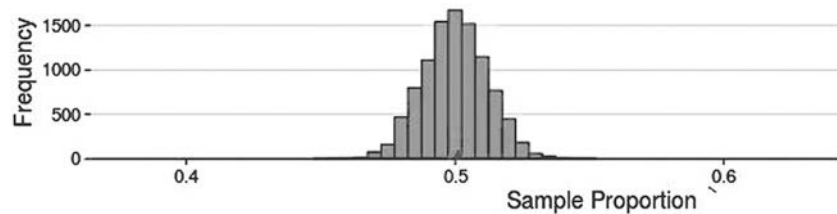
In summary, if half the population of voters had voted for Brown, we would have expected between 46% and 54% of voters in an exit poll of size 1824 to have voted for him. It would have been very unusual to observe 60.5% voting for him, as happened in the actual exit poll. If *less than half* the population had voted for Brown, it would have been even more unusual to have this outcome. This is the basis of the network's exit poll prediction that Brown won the election.

You can perform this simulation using *any* population proportion value, corresponding to flipping a coin in which head and tail have different probabilities. For instance, you could simulate sampling when the population proportion voting for the Democrat is 0.45 by changing the probability of a head in the applet to 45%. Likewise, we could change the size of each random sample in the simulation to study the impact of the sample size. From results of the next section, for a random sample of size 1824 the sample proportion has probability close to 1 of falling within 0.04 of the population proportion, regardless of its value.

REPRESENTING SAMPLING VARIABILITY BY A SAMPLING DISTRIBUTION

Voter preference is a variable, varying among voters. Likewise, so is the sample proportion voting for some candidate a variable: Before the sample is obtained, its value is unknown, and that value varies from sample to sample. If we could select several random samples of size $n = 1824$ each, a certain predictable amount of variation would occur in the sample proportion values. A probability distribution with appearance similar to Figure 4.10 describes the variation that occurs from repeatedly selecting samples of a certain size n and forming a particular statistic. This distribution is called a ***sampling distribution***. It also provides probabilities of the possible values of the statistic for a *single* sample of size n.

Sampling Distribution

A ***sampling distribution*** of a statistic (such as a sample proportion or a sample mean) is the probability distribution that specifies probabilities for the possible values the statistic can take.

---

[6] The *Sampling Distribution for the Sample Proportion* applet at `www.pearsonglobaleditions.com/Agresti` does this efficiently.

Each sample statistic has a sampling distribution. There is a sampling distribution of a sample mean, a sampling distribution of a sample proportion, a sampling distribution of a sample median, and so forth. A sampling distribution is merely a type of probability distribution. Unlike the probability distributions studied so far, a sampling distribution specifies probabilities not for individual observations but for possible values of a statistic computed from the observations. A sampling distribution allows us to calculate, for example, probabilities about the sample proportions of individuals in an exit poll who voted for the different candidates. Before the voters are selected for the exit poll, this is a variable. It has a sampling distribution that describes the probabilities of the possible values.

The sampling distribution is important in inferential statistics because it helps us predict how close a statistic falls to the parameter it estimates. From Figure 4.10, for instance, with a sample of size 1824 the probability is apparently close to 1 that a sample proportion falls within 0.04 of the population proportion.

Example 4.6 | **Constructing a Sampling Distribution** It is sometimes possible to construct the sampling distribution without resorting to simulation or complex mathematical derivations. To illustrate, we construct the sampling distribution of the sample proportion for an exit poll of $n = 4$ voters from a population in which half voted for each candidate. (Such a small n would not be used in practice, but it enables us to more easily explain this process.)

We use a symbol with four entries to represent the votes for a potential sample of size 4. For instance, (R, D, D, R) represents a sample in which the first and fourth subjects voted for the Republican and the second and third subjects voted for the Democrat. The 16 possible samples are

$$
\begin{array}{cccc}
(R, R, R, R) & (R, R, R, D) & (R, R, D, R) & (R, D, R, R) \\
(D, R, R, R) & (R, R, D, D) & (R, D, R, D) & (R, D, D, R) \\
(D, R, R, D) & (D, R, D, R) & (D, D, R, R) & (R, D, D, D) \\
(D, R, D, D) & (D, D, R, D) & (D, D, D, R) & (D, D, D, D)
\end{array}
$$

When half the population voted for each candidate, the 16 samples are equally likely.

Let's construct the sampling distribution of the sample proportion that voted for the Republican candidate. For a sample of size 4, that proportion can be 0, 0.25, 0.50, 0.75, or 1.0. The proportion 0 occurs with only one of the 16 possible samples, (D, D, D, D), so its probability equals $1/16 = 0.0625$. The proportion 0.25 occurs for four samples, (R, D, D, D), (D, R, D, D), (D, D, R, D), and (D, D, D, R), so its probability equals $4/16 = 0.25$. Based on this reasoning, Table 4.3 shows the probability for each possible sample proportion value.

TABLE 4.3: Sampling Distribution of Sample Proportion Voting Republican, for Random Sample of Size $n = 4$ when Population Proportion Is 0.50. For example, a sample proportion of 1.0 occurs for only 1 of 16 possible samples, namely (R, R, R, R), so its probability is 1/16 = 0.0625.

Sample Proportion	Probability
0.00	0.0625
0.25	0.2500
0.50	0.3750
0.75	0.2500
1.00	0.0625

Figure 4.11 portrays the sampling distribution of the sample proportion for $n = 4$. It is much more spread out than the one in Figure 4.10 for samples of size $n = 1824$, which falls nearly entirely between 0.46 and 0.54. With such a small sample ($n = 4$), the sample proportion need not be near the population proportion. This is not surprising. In practice, samples are usually much larger than $n = 4$. We used a small value in this example, so it was simpler to write down all the potential samples and find probabilities for the sampling distribution.[7] ∎

FIGURE 4.11: Sampling Distribution of Sample Proportion Voting Republican, for Random Sample of Size $n = 4$ when Population Proportion Is 0.50
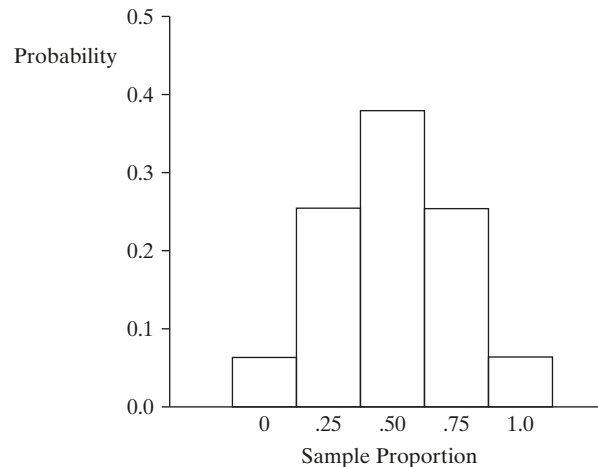
Suppose we denoted the two possible outcomes by 0 for Democrat and by 1 for Republican. From Section 3.2 (page 52), *the proportion of times that 1 occurs is the sample mean of the data*. For instance, for the sample (0, 1, 0, 0) in which only the second subject voted for the Republican, the sample mean equals $(0 + 1 + 0 + 0)/4 = 1/4 = 0.25$, the sample proportion voting for the Republican. So, Figure 4.11 is also an example of a sampling distribution of a sample mean. Section 4.5 presents properties of the sampling distribution of a sample mean.

REPEATED SAMPLING INTERPRETATION OF SAMPLING DISTRIBUTIONS

Sampling distributions portray the sampling variability that occurs in collecting data and using sample statistics to estimate parameters. If different polling organizations each take their own exit poll and estimate the population proportion voting for the Republican candidate, they will get different estimates, because the samples have different people. Likewise, Figure 4.10 describes the variability in sample proportion values that occurs in selecting a huge number of samples of size $n = 1824$ and constructing a histogram of the sample proportions. By contrast, Figure 4.11 describes the variability for a huge number of samples of size $n = 4$.

A sampling distribution of a statistic for n observations is the relative frequency distribution for that statistic resulting from repeatedly taking samples of size n, each time calculating the statistic value. It's possible to form such a distribution empirically, as in Figure 4.10, by repeated sampling or through simulation. In practice, this is not necessary. The form of sampling distributions is often known theoretically, as shown in the previous example and in the next section. We can then find probabilities about the value of the sample statistic for one random sample of the given size n.

---

[7] Section 6.7 presents a formula for probabilities in this sampling distribution, called the *binomial distribution*, but we do not need the formula here.

4.5 Sampling Distributions of Sample Means

Because the sample mean $\bar{y}$ is used so much, with the sample proportion also being a sample mean, its sampling distribution merits special attention. In practice, when we analyze data and find $\bar{y}$, we do not know how close it falls to the population mean μ, because we do not know the value of μ. Using information about the spread of the sampling distribution, though, we can predict how close it falls. For example, the sampling distribution might tell us that with high probability, $\bar{y}$ falls within 10 units of μ.

This section presents two main results about the sampling distribution of the sample mean. One provides formulas for the center and spread of the sampling distribution. The other describes its shape.

MEAN AND STANDARD ERROR OF SAMPLING DISTRIBUTION OF $\bar{y}$

The sample mean $\bar{y}$ is a variable, because its value varies from sample to sample. For random samples, it fluctuates around the population mean μ, sometimes being smaller and sometimes being larger. In fact, the mean of the sampling distribution of $\bar{y}$ equals μ. If we repeatedly took samples, then in the long run, the mean of the sample means would equal the population mean μ.

The spread of the sampling distribution of $\bar{y}$ is described by its standard deviation, which is called the **standard error** of $\bar{y}$.

Standard Error

> The standard deviation of the sampling distribution of $\bar{y}$ is called the **standard error** of $\bar{y}$ and is denoted by $\sigma_{\bar{y}}$.

The standard error describes how much $\bar{y}$ varies from sample to sample. Suppose we repeatedly selected samples of size n from the population, finding $\bar{y}$ for each set of n observations. Then, in the long run, the standard deviation of the $\bar{y}$-values would equal the standard error. The symbol $\sigma_{\bar{y}}$ (instead of σ) and the terminology *standard error* (instead of *standard deviation*) distinguish this measure from the standard deviation σ of the population distribution.

In practice, we do not need to take samples repeatedly to find the standard error of $\bar{y}$, because a formula is available. For a random sample of size n, the standard error of $\bar{y}$ depends on n and the population standard deviation σ by[8]

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}.$$

Figure 4.12 displays a population distribution having $\sigma = 10$ and shows the sampling distribution of $\bar{y}$ for $n = 100$. When $n = 100$, the standard error is $\sigma_{\bar{y}} = \sigma/\sqrt{n} = 10/\sqrt{100} = 1.0$. The sampling distribution has only a tenth of the spread of the population distribution. This means that individual observations tend to vary much more than sample means vary from sample to sample.

In summary, the following result describes the center and spread of the sampling distribution of $\bar{y}$:

Mean and Standard Error of $\bar{y}$

> For sampling a population, the sampling distribution of $\bar{y}$ states the probabilities for the possible values of $\bar{y}$. For a random sample of size n from a population having mean μ and standard deviation σ, the sampling distribution of $\bar{y}$ has mean μ and standard error $\sigma_{\bar{y}} = \sigma/\sqrt{n}$.

---

[8] Exercise 4.58 shows the basis of this formula.

FIGURE 4.12: A
Population Distribution
and the Sampling
Distribution of $\bar{y}$ for
$n = 100$

Sampling distribution for $n = 100$

$$\left(\text{Standard error } \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1 \right)$$

Population distribution
$(\mu = 100, \sigma = 10)$

$\mu = 100$
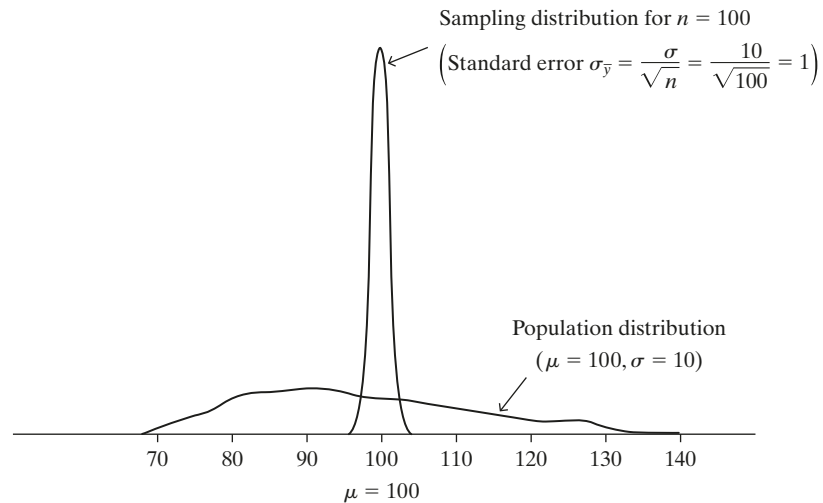
**Example
4.7**

Standard Error of Sample Proportion in Exit Poll   Following Example 4.5 (page 92),
we conducted a simulation to investigate how much variability to expect from sample
to sample in an exit poll of 1824 voters. Instead of conducting a simulation, we can
get similar information directly by finding a standard error. Knowing the standard
error helps us answer the following question: If half the population voted for each
candidate, how much would a sample proportion for an exit poll of 1824 voters tend
to vary from sample to sample?

Let the variable y denote a vote outcome. As at the end of Example 4.6, we let
$y = 0$ represent a vote for the Democrat and $y = 1$ represent a vote for the Repub-
lican. Figure 4.13 shows the population distribution for which half the population
voted for each, so that $P(0) = 0.50$ and $P(1) = 0.50$. The mean of the distribution
equals 0.50, which is the population proportion voting for each. (Or, from the for-
mula near the end of Section 4.2, $\mu = \sum y P(y) = 0(0.50) + 1(0.50) = 0.50$.) The
squared deviation of y from the mean, $(y - \mu)^2$, equals $(0 - 0.50)^2 = 0.25$ when
$y = 0$, and it equals $(1 - 0.50)^2 = 0.25$ when $y = 1$. The variance is the expected
value of this squared deviation. Thus, it equals $\sigma^2 = 0.25$. So, the standard deviation
of the population distribution of y is $\sigma = \sqrt{0.25} = 0.50$.

FIGURE 4.13: The
Population Distribution
when $y = 0$ or 1, with
Probability 0.50 Each. This
is the probability
distribution for a vote, with
$0 =$ vote for Democratic
candidate and $1 =$ vote for
Republican candidate.
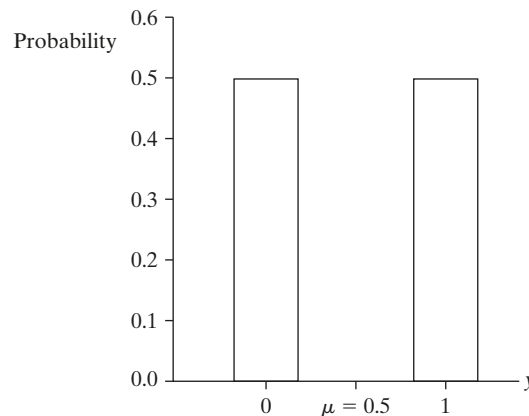
For a sample, the mean of the 0 and 1 values is the sample proportion of votes for
the Republican. Its sampling distribution has mean that is the mean of the population
distribution of y, namely, $\mu = 0.50$. For repeated samples of a fixed size n, the sample

proportions fluctuate around 0.50, being larger about half the time and smaller half the time. The standard deviation of the sampling distribution is the standard error. For a sample of size 1824, this is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.50}{\sqrt{1824}} = 0.0117.$$

A result from later in this section says that this sampling distribution is bell shaped. Thus, with probability close to 1.0 the sample proportion falls within three standard errors of μ, that is, within $3(0.0117) = 0.035$ of 0.50, or between about 0.46 and 0.54. For a random sample of size 1824 from a population in which 50% voted for each candidate, it would be surprising if fewer than 46% or more than 54% voted for one of them. We've now seen how to get this result either using simulation, as shown in Figure 4.10, or using the information about the mean and standard error of the sampling distribution. ∎

EFFECT OF SAMPLE SIZE ON SAMPLING DISTRIBUTION AND PRECISION OF ESTIMATES

The standard error gets smaller as the sample size n gets larger. The reason for this is that the denominator ($\sqrt{n}$) of the standard error formula $\sigma_{\bar{y}} = \sigma/\sqrt{n}$ increases as n increases. For instance, when the population standard deviation $\sigma = 0.50$, we've just seen that the standard error is 0.0117 when $n = 1824$. When $n = 100$, a less typical size for a poll, the standard error is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.50}{\sqrt{n}} = \frac{0.50}{\sqrt{100}} = 0.050.$$

With $n = 100$, since three standard errors equal $3(0.050) = 0.15$, the probability is very high that the sample proportion falls within 0.15 of 0.50, or between 0.35 and 0.65.

Figure 4.14 shows the sampling distributions of the sample proportion when $n = 100$ and when $n = 1824$. As n increases, the standard error decreases and the sampling distribution gets narrower. This means that the sample proportion tends to fall closer to the population proportion. It's more likely that the sample proportion closely approximates a population proportion when $n = 1824$ than when $n = 100$. This agrees with our intuition that larger samples provide more precise estimates of population characteristics.

FIGURE 4.14: The Sampling Distributions of the Sample Proportion, when $n = 100$ and when $n = 1824$. These refer to sampling from the population distribution in Figure 4.13.
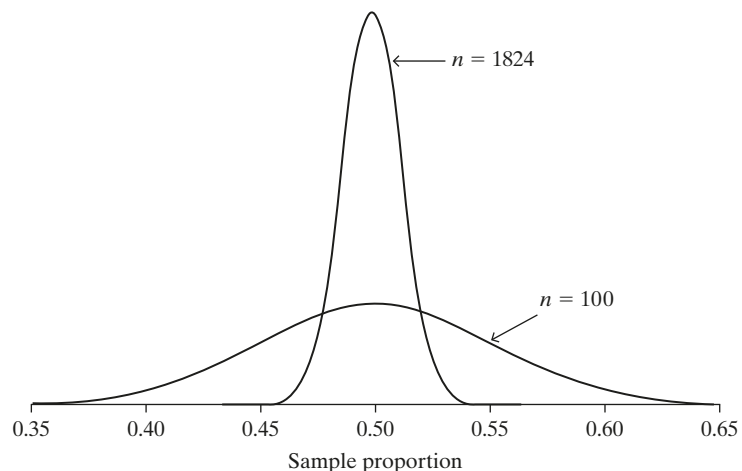
In summary, error occurs when we estimate μ by $\bar{y}$, because we sampled only part of the population. This error, which is the ***sampling error***, tends to decrease as the sample size n increases. The standard error is fundamental to inferential procedures that predict the sampling error in using $\bar{y}$ to estimate μ.

SAMPLING DISTRIBUTION OF SAMPLE MEAN IS APPROXIMATELY NORMAL

For the population distribution for the vote in an election, shown in Figure 4.13, the outcome has only two possible values. It is highly discrete. Nevertheless, the two sampling distributions shown in Figure 4.14 have bell shapes. This is a consequence of the second main result of this section, which describes the *shape* of the sampling distribution of $\bar{y}$. This result can be proven mathematically, and it is often called the *Central Limit Theorem*.

Central Limit Theorem	For random sampling with a large sample size n, the sampling distribution of the sample mean $\bar{y}$ is approximately a normal distribution.

Here are some implications and interpretations of this result:

- The bell shape of the sampling distribution applies *no matter what the shape* of the population distribution. This is remarkable. For large random samples, the sampling distribution of $\bar{y}$ has a normal bell shape even if the population distribution is very skewed or highly discrete such as the binary distribution in Figure 4.13. We'll learn how this enables us to make inferences even when the population distribution is highly irregular. This is helpful, because many social science variables are very skewed or highly discrete.

  Figure 4.15 displays sampling distributions of $\bar{y}$ for four different shapes for the population distribution, shown at the top of the figure. Below them are

FIGURE 4.15: Four Different Population Distributions and the Corresponding Sampling Distributions of $\bar{y}$. As n increases, the sampling distributions get narrower and have more of a bell shape.
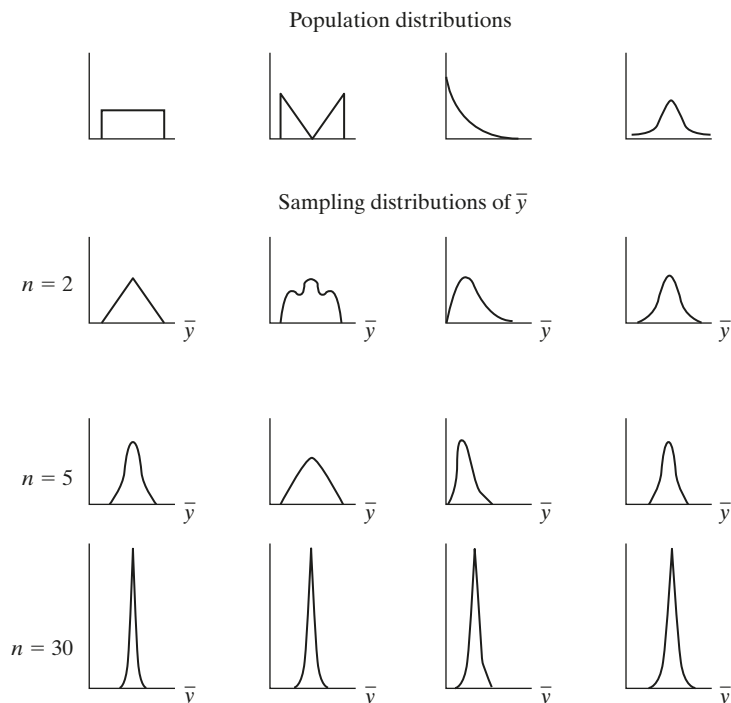
portrayed the sampling distributions for random samples of sizes $n = 2, 5$, and 30. As n increases, the sampling distribution has more of a bell shape.

- How large n must be before the sampling distribution is bell shaped largely depends on the skewness of the population distribution. If the *population* distribution is bell shaped, then the sampling distribution is bell shaped for *all* sample sizes. The rightmost panel of Figure 4.15 illustrates this. More skewed distributions require larger sample sizes. For most cases, n of about 30 is sufficient (although it may not be large enough for precise inference). So, in practice, with random sampling the sampling distribution of $\bar{y}$ is nearly always approximately bell shaped.

- Knowing that the sampling distribution of $\bar{y}$ can be approximated by a normal distribution helps us to find probabilities for possible values of $\bar{y}$. For instance, $\bar{y}$ almost certainly falls within $3\sigma_{\bar{y}} = 3\sigma/\sqrt{n}$ of μ. Reasoning of this nature is vital to inferential statistical methods.

Example 4.8

Simulating a Sampling Distribution You can verify the Central Limit Theorem empirically by repeatedly selecting random samples, calculating $\bar{y}$ for each sample of n observations. Then, the histogram of the $\bar{y}$-values is approximately a normal curve.

- Go to `www.pearsonglobaleditions.com/Agresti` and click on the *Sampling Distribution for the Sample Mean* applet for continuous variables.

- Select a skewed population distribution. You can specify how skewed the distribution is. Here, we'll use the value 2 for skewness.

- We'll consider what happens for sample sizes of $n = 200$, relatively modest for a social science study. Enter 200 in the *Select sample size* box. When you click on *Draw sample*, the applet will randomly sample 200 observations, find the sample mean and standard deviation, and plot a histogram.

- Next, change the number of samples of size $n = 200$ that you draw from 1 to 10,000. When you again click on *Draw Sample*, the applet will select 10,000 samples, each of size 200. It will find the sample mean for each sample of 200 observations, overall then finding 10,000 sample means and plotting their histogram. See Figure 4.16 (page 102) for a result. It shows the skewed population distribution on top, the sample data distribution for one of the samples of size 200 below that, and the empirical sampling distribution for the 10,000 sample means at the bottom.

Even though the population distribution in Figure 4.16 is skewed, the sampling distribution is bell shaped. It is also much less spread out, because its spread is described by the standard error, which is the population standard deviation divided by $\sqrt{200} = 14.1$. ∎

Example 4.9

Is Sample Mean Income of Migrant Workers Close to Population Mean? For the population of migrant workers doing agricultural labor in Florida, suppose that weekly income has a distribution that is skewed to the right with a mean of $\mu = \$380$ and a standard deviation of $\sigma = \$80$. A researcher, unaware of these values, plans to randomly sample 100 migrant workers and use the sample mean income $\bar{y}$ to estimate μ. What is the sampling distribution of the sample mean? Where is $\bar{y}$ likely to fall, relative to μ? What is the probability that $\bar{y}$ overestimates μ by more than $\$20$, falling above $\$400$?

By the Central Limit Theorem, the sampling distribution of the sample mean $\bar{y}$ is approximately normal, even though the population distribution is skewed. The

FIGURE 4.16: An Applet for Simulating a Sampling Distribution. Here, in clicking on *Draw Sample*, we take 10,000 samples of size 200 each. The graphic shows the population distribution, the sample data distribution for one sample of size 200, and the empirical sampling distribution that shows the 10,000 values of $\bar{y}$ for the 10,000 samples of size $n = 200$ each.
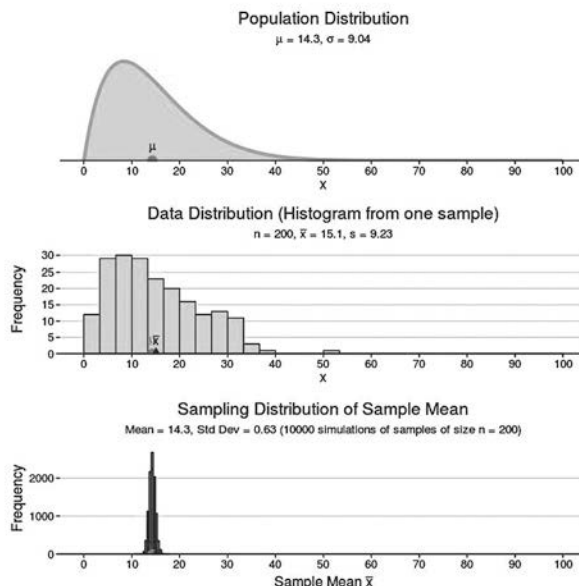
sampling distribution has the same mean as the population distribution, namely, $\mu = \$380$. Its standard error is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{100}} = 8.0 \text{ dollars.}$$

Thus, it is highly likely that $\bar{y}$ falls within about $24 (three standard errors) of μ, that is, between about $356 and $404.

For the normal sampling distribution with mean 380 and standard error 8, the possible $\bar{y}$ value of 400 has a *z*-score of

$$z = (400 - 380)/8 = 2.5.$$

From a table of normal probabilities (such as Table A) or software, the corresponding right-tail probability above 400 is 0.0062. It is very unlikely that the sample mean would fall above $400. ∎

This example is unrealistic, because we assumed knowledge of the population mean μ. In practice, μ would be unknown. However, the sampling distribution of $\bar{y}$ provides the probability that the sample mean falls within a certain distance of the population mean μ, even when μ is unknown. We illustrate by finding the probability that the sample mean weekly income $\bar{y}$ falls within $10 of the unknown true mean income μ for all such workers.

Now, the sampling distribution of $\bar{y}$ is approximately normal in shape and is centered about μ. We have just seen that when $n = 100$, the standard error is $\sigma_{\bar{y}} = \$8.0$. Hence, the probability that $\bar{y}$ falls within $10 of μ is the probability that a normally distributed variable falls within $10/8 = 1.25$ standard deviations of its mean. That is, the number of standard errors that $\mu + 10$ (or $\mu - 10$) falls from μ is

$$z = \frac{(\mu + 10) - \mu}{8} = \frac{10}{8} = 1.25.$$

See Figure 4.17. From a normal table, the probability that $\bar{y}$ falls *more than* 1.25 standard errors from μ (in either direction) is $2(0.1056) = 0.21$. Thus, the probability that $\bar{y}$ falls no more than \$10 from μ equals $1 - 0.21 = 0.79$.

FIGURE 4.17: Sampling Distribution of $\bar{y}$ for Unknown μ and Standard Error $\sigma_{\bar{y}} = 8$
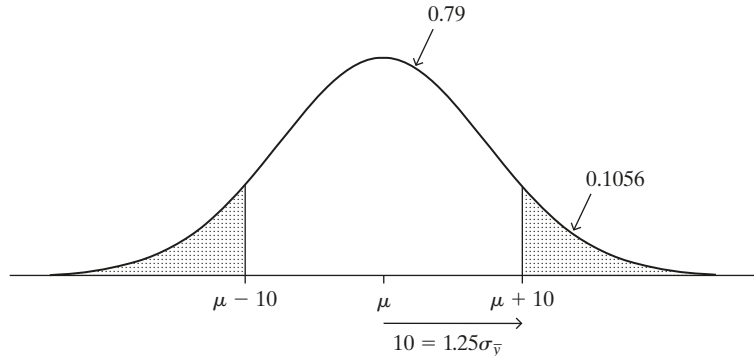
This example is still unrealistic, because we assumed knowledge of the population standard deviation σ. In practice, we'd need to estimate σ. The next chapter shows that to conduct inference, we estimate σ by the sample standard deviation s.

To get a feel for the Central Limit Theorem and how sampling distributions become more bell shaped as n increases, we suggest that you try out an applet on the Internet, as in Exercises 4.38 and 4.39.

4.6 Review: Population, Sample Data, and Sampling Distributions

Sampling distributions are fundamental to statistical inference and to methodology presented in the rest of this text. Because of this, we now review the distinction between sampling distributions and the two types of distributions presented in Section 3.1—the **population** distribution and the **sample data** distribution.

Here is a capsule description of the three types of distribution:

- **Population distribution**: This is the distribution from which we select the sample. It is usually unknown. We make inferences about its characteristics, such as the parameters μ and σ that describe its center and spread.

- **Sample data distribution**: This is the distribution of data that we actually observe, that is, the sample observations $y_1, y_2, \ldots, y_n$. We describe it by statistics such as the sample mean $\bar{y}$ and sample standard deviation s. The larger the sample size n, the closer the sample data distribution resembles the population distribution, and the closer the sample statistics such as $\bar{y}$ fall to the population parameters such as μ.

- **Sampling distribution** of a statistic: This is the probability distribution for the possible values of a sample statistic, such as $\bar{y}$. A sampling distribution describes the variability that occurs in the statistic's value among samples of a certain size. This distribution determines the probability that the statistic falls within a certain distance of the population parameter it estimates.

In Figure 4.16 on page 102, the *population distribution* is the skewed distribution shown at the top. The distribution in the middle of the figure is a *sample data distribution* based on one particular sample of $n = 200$ observations. It has similar

appearance to the population distribution, also being somewhat skewed to the right. It has $\bar{y} = 13.4$ and $s = 8.9$, similar to $\mu = 14.3$ and $\sigma = 9.0$ for the population. The distribution at the bottom of the figure describes the *sampling distribution* of the sample mean for random samples of size 200. It is an empirical sampling distribution, showing a histogram of 10,000 values of the sample mean for 10,000 random samples of size $n = 200$ each. It is bell shaped, as a consequence of the Central Limit Theorem, and very narrow, as a consequence of the standard error formula $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. Following is an example in which the three distributions would have shape like those in Figure 4.16.

Example 4.10	**Three Distributions for a General Social Survey Item** In 2014, the GSS asked about the number of hours a week spent on the Internet, excluding e-mail. The *sample data distribution* for the $n = 1399$ subjects in the sample was very highly skewed to the right. It is described by the sample mean $\bar{y} = 11.6$ and sample standard deviation $s = 15.0$.

Because the GSS cannot sample the entire population of adult Americans (about 250 million people), we don't know the *population distribution*. Because the sample data distribution had a large sample size, probably the population distribution looks like it. Most likely the population distribution would also be highly skewed to the right. Its mean and standard deviation would be similar to the sample values. Values such as $\mu = 12.0$ and $\sigma = 14.0$ would be plausible.

If the GSS repeatedly took random samples[9] of 1399 adult Americans, the sample mean time $\bar{y}$ spent on the Internet would vary from survey to survey. The *sampling distribution* describes how $\bar{y}$ would vary. For example, if the population has mean $\mu = 12.0$ and standard deviation $\sigma = 14.0$, then the sampling distribution of $\bar{y}$ also has mean 12.0, and it has a standard error of

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{14.0}{\sqrt{1399}} = 0.37.$$

Unlike the population and sample data distributions, the sampling distribution would be bell shaped and narrow. Nearly all of that distribution would fall within $3(0.37) = 1.12$ of the mean of 12.0. So, it's very likely that any sample of size 1399 would have a sample mean within 1.12 of 12.0. In summary, the sample data and population distributions are highly skewed and spread out, whereas the sampling distribution of $\bar{y}$ is bell shaped and has nearly all its probability in a narrow range. ∎

Example 4.11	**Three Distributions for Exit Poll Example** We consider, once again, the variable $y =$ vote in the 2014 California gubernatorial election for a randomly selected voter. Let $y = 1$ for Jerry Brown (the Democrat) and $y = 0$ for Neel Kashkari (the Republican). In fact, of the 7,317,581 adult residents of California who voted, 60.0% voted for Brown. So, the probability distribution for y has probability 0.600 at $y = 1$ and probability 0.400 at $y = 0$. The mean of this distribution is $\mu = 0.600$, which is the population proportion of votes for Brown. From a formula we'll study in the next chapter, the standard deviation of this two-point distribution is $\sigma = 0.490$.

The population distribution of candidate preference consists of the 7,317,581 values for y, of which 40.0% are 0 and 60.0% are 1. This distribution is described by the parameters $\mu = 0.600$ and $\sigma = 0.490$. Figure 4.18 portrays this distribution, which is highly discrete (binary). It is not at all bell shaped.

---

[9] In reality, the GSS uses a multistage cluster sample, so the true standard error is a bit larger than $\sigma/\sqrt{n}$. For purposes of illustration, we'll treat GSS data as if they come from a simple random sample, keeping in mind that in practice some adjustment is necessary as explained at the GSS website, `sda.berkeley.edu/GSS`.

FIGURE 4.18: The Population Distribution (7,317,581 votes) and the Sample Data Distribution ($n = 1824$ votes) in the 2014 California Gubernatorial Election, where $1 = $ Vote for Brown and $0 = $ Vote for Kashkari
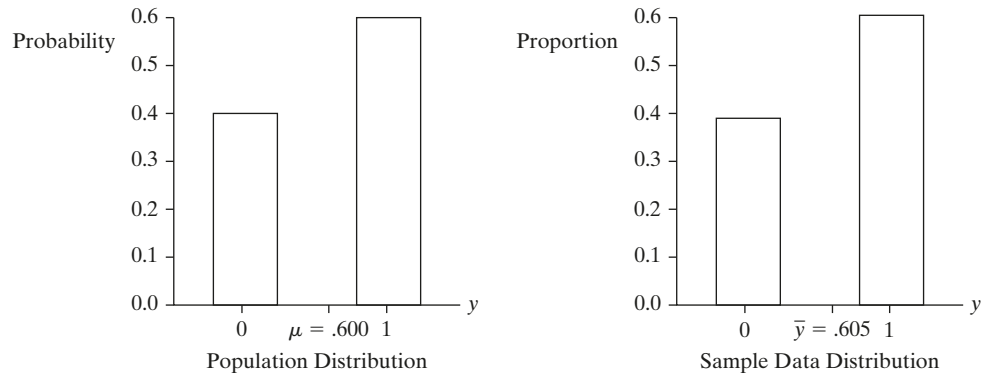
Before all the votes were counted, the population distribution was unknown. When polls closed, CBS News reported results of an exit poll of size $n = 1824$ to predict the outcome. A histogram of the 1824 votes in the sample describes the sample data distribution. Of the 1824 voters, 60.5% said they voted for Brown (i.e., have $y = 1$) and 39.5% said they voted for Kashkari ($y = 0$). Figure 4.18 also displays the histogram of these sample data values. Like the population distribution, the sample data distribution concentrates at $y = 0$ and $y = 1$. It is described by sample statistics such as $\bar{y} = 0.605$, which is the sample proportion voting for Brown. The larger the sample size, the more this sample data distribution tends to resemble the population distribution, since the sample observations are a random subset of the population values. If the entire population is sampled, as when all the votes are counted, then the two distributions are identical.

For a random sample of size $n = 1824$, the sampling distribution of $\bar{y}$ is approximately a normal distribution. Its mean is $\mu = 0.600$, and its standard error is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{0.490}{\sqrt{1824}} = 0.011.$$

Figure 4.19 portrays this sampling distribution, relative to the population distribution of votes.

FIGURE 4.19: The
Population Distribution
(where $y = 1$ is Vote for
Brown and $y = 0$ is Vote
for Kashkari) and the
Sampling Distribution of $\bar{y}$
for $n = 1824$
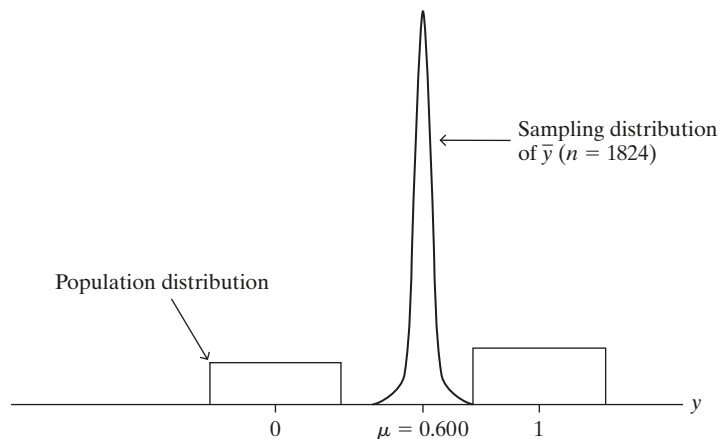
The population distribution and sample data distribution of votes are concentrated at the values 0 and 1. The sampling distribution looks completely different, being much less spread out and bell shaped. The population and sample data distributions of the vote are not bell shaped. They are highly discrete, concentrated at 0 and 1. For $n = 1824$, the sample proportion can take a large number of values

between 0 and 1, and its sampling distribution is essentially continuous, being bell shaped by the Central Limit Theorem. Although the individual values of y are 0 and 1, according to the sampling distribution it is practically impossible that a random sample of size 1824 has a sample mean anywhere near 0 or 1; nearly all the probability falls between 0.57 and 0.63, that is, within three standard errors of the mean $\mu = 0.600$. ∎

THE KEY ROLE OF SAMPLING DISTRIBUTIONS IN STATISTICAL INFERENCE

By the Central Limit Theorem, we can often use the normal distribution to find probabilities about $\bar{y}$. The next two chapters show how statistical inferences rely on this theorem.

The result about sample means having approximately normal sampling distributions is important also because similar results hold for many other statistics. For instance, most sample statistics used to estimate population parameters have approximately normal sampling distributions, for large random samples. This is the primary reason for the key role of the normal distribution in statistical science.

4.7 Chapter Summary

For an observation in a random sample or a randomized experiment, the ***probability*** of a particular outcome is the proportion of times that the outcome would occur in a very long sequence of observations.

- A ***probability distribution*** specifies probabilities for the possible values of a variable. We let $P(y)$ denote the probability of the value y. The probabilities are nonnegative and sum to 1.0.

- Probability distributions have summary parameters, such as the mean μ and standard deviation σ. The mean for a probability distribution of a discrete variable is

$$\mu = \sum yP(y).$$

  This is also called the ***expected value*** of y.

- The ***normal distribution*** has a graph that is a symmetric bell-shaped curve specified by the mean μ and standard deviation σ. For any z, the probability falling within z standard deviations of the mean is the same for every normal distribution.

- In a probability distribution, the ***z-score*** for a value y is

$$z = (y - \mu)/\sigma.$$

  It measures the number of standard deviations that y falls from the mean μ. For a normal distribution, the z-scores have the ***standard normal distribution***, which has mean $= 0$ and standard deviation $= 1$.

- A ***sampling distribution*** is a probability distribution of a sample statistic, such as the sample mean or sample proportion. It specifies probabilities for the possible values of the statistic for samples of the particular size n.

- The sampling distribution of the sample mean $\bar{y}$ centers at the population mean μ. Its standard deviation, called the ***standard error***, relates to the standard deviation σ of the population by $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. As the sample size n increases,

the standard error decreases, so the sample mean tends to be closer to the population mean.

- The ***Central Limit Theorem*** states that for large random samples on a variable, the sampling distribution of the sample mean is approximately a normal distribution. This holds no matter what the shape of the population distribution, both for continuous variables and for discrete variables. The result applies also to proportions, since the sample proportion is a special case of the sample mean for observations coded as 0 and 1 (such as for two candidates in an election).

The bell shape for the sampling distribution of many statistics is the main reason for the importance of the normal distribution. The next two chapters show how the Central Limit Theorem is the basis of methods of statistical inference.

Exercises

Practicing the Basics

4.1. In a Child Health Survey conducted in a school, 1567 children answered "yes" and 433 children answered "no" in response to the question: "Do you drink milk?"

(a) Estimate the probability that a randomly selected child drinks milk.

(b) Estimate the probability that a child does not drink milk.

(c) Of those who drink milk, 40% like it with chocolate. Estimate the probability that a randomly chosen child likes to drink chocolate milk.

4.2. Suppose a software for statistical inference methods often sets the default probability of a correct inference to be 0.95. Suppose we make an inference about the population proportion of people who like to go trekking, and we consider this separately for men and for women. Let A denote the outcome that the inference about men is correct, and let B denote the outcome that the inference about women is correct. Treating these as independent samples and inferences, find the probability that *both* inferences are correct.

4.3. Save Your Earth, an NGO, conducted a survey asking subjects whether they are a member of any NGO working for environmental protection and whether they would be willing to purchase a sapling. Table 4.4 shows the results.

(a) Estimate the probability that a randomly selected adult is a member of an NGO.

(b) Show that the estimated probability of being willing to plant a tree is (i) 0.238, given that the person is a member of an NGO; (ii) 0.1407, given that the person is not a member of an NGO.

(c) Show that the estimated probability that a person is both a member of an NGO *and* willing to purchase a sapling is 0.0564 (i) directly using the counts in the table, (ii) using the probability estimates from (a) and (b).

(d) Show that the estimated probability that a person answers yes to both questions or no to both questions is 0.712.

TABLE 4.4

| | | Willing to Plant a Tree | | |
		Yes	No	Total
Member of	Yes	50	160	210
the NGO	No	95	580	675
Total		145	740	885

4.4. Let y = number of sports a person can play. For residents of Country A, y has probability distribution $P(0) = 0.05$, $P(1) = 0.85$, and $P(2) = 0.10$, with negligible probability for higher values of y.

(a) Is y a discrete or a continuous variable? Why?

(b) Construct a table showing the probability distribution of y.

(c) Find the probability that an Indian does *not play* more than one sport.

(d) Find the mean of this probability distribution.

4.5. Let y denote the number of times a person was admitted to a hospital within the past 12 months. According to results from recent Region Health Surveys, for a randomly chosen person in Country A, the probability distribution of y is approximately $P(0) = 0.85$, $P(1) = 0.13$, $P(2) = 0.01$, and $P(3) = 0.01$.

(a) Explain why it is not valid to find the mean of this probability distribution as $(0 + 1 + 2 + 3)/4 = 1.5$.

(b) Find the correct mean of the probability distribution.

4.6. A ticket for a lottery costs $1. With probability 0.0000001, you win 2 million dollars ($2,000,000), and with probability 0.9999999 you win nothing. Let y denote

the winnings from buying one ticket. Construct the probability distribution for y. Show that the mean of the distribution equals 0.20, corresponding to an expected return of 20 cents for the dollar paid.

4.7. Let y be the outcome of selecting a single digit using a random number generator.

(a) Construct the probability distribution for y. (This type of distribution is called a *uniform* distribution, because of the uniform spread of probabilities across the possible outcomes.)

(b) Find the mean of this probability distribution.

4.8. For a normal distribution, find the probability that an observation falls **(a)** at least one standard deviation above the mean; **(b)** at least one standard deviation below the mean.

4.9. For a normal distribution, verify that the probability between

(a) $\mu - \sigma$ and $\mu + \sigma$ equals 0.68.

(b) $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$ equals 0.95.

(c) $\mu - 3\sigma$ and $\mu + 3\sigma$ equals 0.997.

(d) $\mu - 0.67\sigma$ and $\mu + 0.67\sigma$ equals 0.50.

4.10. Find the z-value for which the probability that a normal variable exceeds $\mu + z\sigma$ equals **(a)** 0.01, **(b)** 0.025, **(c)** 0.05, **(d)** 0.10, **(e)** 0.25, **(f)** 0.50.

4.11. Find the z-value such that for a normal distribution the interval from $\mu - z\sigma$ to $\mu + z\sigma$ contains **(a)** 50%, **(b)** 90%, **(c)** 95%, **(d)** 99% of the probability.

4.12. Find the z-values corresponding to the **(a)** 90th, **(b)** 95th, **(c)** 99th percentiles of a normal distribution.

4.13. If z is the number such that the interval from $\mu - z\sigma$ to $\mu + z\sigma$ contains 90% of a normal distribution, then explain why $\mu + z\sigma$ is the 95th percentile.

4.14. If z is the positive number such that the interval from $\mu - z\sigma$ to $\mu + z\sigma$ contains 50% of a normal distribution, then

(a) Which percentile is (i) $\mu + z\sigma$? (ii) $\mu - z\sigma$?

(b) Find this value of z. Using this result, explain why the upper quartile and lower quartile of a normal distribution are $\mu + 0.67\sigma$ and $\mu - 0.67\sigma$, respectively.

4.15. What proportion of a normal distribution falls

(a) above a z-score of 2.10?

(b) below a z-score of -2.10?

(c) between z-scores of -2.10 and 2.10?

4.16. Find the z-score for the number that is less than only 1% of the values of a normal distribution.

4.17. A teacher wants to know how many students have a test score at the 96th percentile or higher.

(a) How many standard deviations above the mean is the 96th percentile?

(b) For the normal distribution of scores with mean 75 and standard deviation 8, find the score for the 96th percentile.

4.18. According to a recent Population Survey, self-employed individuals in a country work an average of 55 hours per week, with a standard deviation of 18. If this variable is approximately normally distributed, what proportion averaged more than 40 hours per week?

4.19. The Mental Development Index (MDI) of the Bayley Scales of Infant Development is a standardized measure used in studies with high-risk infants. It has approximately a normal distribution with a mean of 100 and a standard deviation of 16.

(a) What proportion of children have an MDI of at least 140?

(b) Find the MDI score that is the 90th percentile.

(c) Find and interpret the lower quartile, median, and upper quartile.

4.20. For a study in a State University Hospital, 3856 pregnant women who reported information on length of gestation until birth had mean = 284 days and standard deviation = 12 days. A baby is classified as premature if the gestation time is 258 days or less.

(a) If mothers' gestation times are normally distributed, what proportion of babies would be born prematurely?

(b) The actual proportion born prematurely during this period was 0.045. Based on this information, how would you expect the distribution of gestation time to differ from the normal distribution?

4.21. Suppose that the weekly use of ground water in a village is approximately normally distributed, with a mean of 20 gallons and a standard deviation of 5 gallons.

(a) What proportion of villagers use more than 25 gallons per week?

(b) Assuming that the standard deviation and the normal form would remain constant, to what level must the mean reduce so that only 5% use more than 25 gallons per week?

(c) If the distribution of ground water use is not actually normal, how would you expect it to deviate from normal?

4.22. On the midterm exam in mathematics, an instructor always gives a grade of A to students who score between 90 and 95. One year, the scores have approximately a normal distribution with mean 83 and standard deviation 7. About what proportion of the students get an A?

4.23. Student A's scores in Country X have a distribution with $\mu = 400$ and $\sigma = 75$, and Student B's scores in Country Y have a distribution with $\mu = 25$ and $\sigma = 4.0$. Which score is relatively higher, Student A $= 500$ or Student B $= 36$? Explain.

4.24. Suppose that consumption expenditure by households in a city has an approximately normal distribution with a mean of $5400 and a standard deviation of $1800. The consumption expenditure by one particular household is $8000.

(a) Find the z-score corresponding to that value.

(b) What proportion of the consumption expenditure exceeds $8000?

4.25. An energy study found that household use of electricity had a mean of 583 and a standard deviation of 450 kWh (kilowatt-hours).

(a) If the distribution was normal, what percentage of the households used above 800 kWh?

(b) Do you think the distribution is truly normal? Why or why not?

4.26. Five students—the females Angeline and Belle and the males Charlie, David, and Edmund—are rated equally qualified for admission to medical school, ahead of other applicants. However, all but two positions have been filled for entering the class. The admissions committee can admit only two more students, so it decides to randomly select two of these five candidates. For this strategy, let $y =$ number of males admitted. Using the first letter of the name to denote a student, the different combinations that could be admitted are (A, B), (A, C), (A, D), (A, E), (B, C), (B, D), (B, E), (C, D), (C, E), and (D, E).

(a) Construct the probability distribution for y.

(b) Construct the sampling distribution of the sample *proportion* of the students selected who are male.

4.27. Construct the sampling distribution of the sample proportion of heads, for flipping a balanced coin

(a) Once.

(b) Twice. (*Hint*: The possible samples are (H, H), (H, T), (T, H), (T, T).)

(c) Three times. (*Hint*: There are 8 possible samples.)

(d) Four times. (*Hint*: There are 16 possible samples.)

(e) Describe how the shape of the sampling distribution seems to be changing as the number of flips increases.

4.28. The probability distribution associated with the outcome of rolling a balanced die has probability 1/6 attached to each integer, {1, 2, 3, 4, 5, 6}. Let (y_1, y_2) denote the outcomes for rolling the die twice.

(a) Enumerate the 36 possible (y_1, y_2) pairs (e.g., (2, 1) represents a 2 followed by a 1).

(b) Treating the 36 pairs as equally likely, construct the sampling distribution for the sample mean $\bar{y}$ of the two numbers rolled.

(c) Construct a histogram of the (i) probability distribution for each roll, (ii) sampling distribution of $\bar{y}$ in (b). Describe their shapes.

(d) What are the means of the two distributions in (c)? Why are they the same?

(e) Explain why the sampling distribution of $\bar{y}$ has relatively more probability near the middle than at the minimum and maximum values. (*Hint*: Note there are many more (y_1, y_2) pairs that have a sample mean near the middle than near the minimum or maximum.)

4.29. An exit poll of 547 voters in the state election indicated that 55% voted for Candidate X, with most of the rest voting for Candidate Y.

(a) If actually 50% of the population voted for Candidate X, find the standard error of the sample proportion voting for this candidate, for this exit poll. (Recall from Example 4.5 on page 92 that the population standard deviation is 0.50.)

(b) If actually 50% of the population voted for Candidate Y, would it have been surprising to obtain the results in this exit poll? Why?

(c) Based on your answer in (b), would you be willing to predict the outcome of this election? Explain.

4.30. According to *Population Reports*, population distribution of number of years of education for women in Country C has a mean of 10.6 and a standard deviation of 4.0. Find the mean and standard error of the sampling distribution of $\bar{y}$ for a random sample of **(a)** 6 women, **(b)** 49 women, **(c)** 100 women. Describe the pattern as n increases.

4.31. Refer to Exercise 4.6. The mean and standard deviation of the probability distribution for the lottery winnings y are $\mu = 0.20$ and $\sigma = 632.45$. Suppose you play the lottery 1 million times. Let $\bar{y}$ denote your average winnings.

(a) Find the mean and standard error of the sampling distribution of $\bar{y}$.

(b) About how likely is it that you would "come out ahead," with your average winnings exceeding $1, the amount you paid to play each time?

4.32. According to a General Regional Survey, in South Asia, the distribution of $y =$ pair of shoes has a mean of 6.5 and a standard deviation of 4.9. Suppose these are the population mean and standard deviation.

(a) Does y have a normal distribution? Explain.

(b) For a random sample of 700 adults (the size of the General Research Survey for this variable), describe the sampling distribution of $\bar{y}$ by giving its shape, mean, and standard error.

(c) Refer to (b). Report an interval within which the sample mean would almost surely fall.

4.33. The scores on the Psychomotor Development Index (PDI), a scale of infant development, are approximately normal with mean 100 and standard deviation 15.

(a) An infant is selected at random. Find the probability that PDI is below 95.

(b) A study uses a random sample of 36 infants. Specify the sampling distribution of the sample mean PDI, and find the probability that the sample mean is below 95.

(c) Would you be surprised to observe a PDI score of 95? Would you be surprised to observe a sample mean PDI of 95? Why?

(d) Sketch the population distribution for the PDI. Superimpose a sketch of the sampling distribution for $n = 25$.

4.34. A study plans to sample randomly 100 government records of farms to estimate the mean acreage of farms in Place A. Results from an earlier study suggest that 400 acres is a reasonable guess for the population standard deviation of farm size.

(a) Approximate the probability that the sample mean acreage falls within 20 acres of the population mean acreage.

(b) If in reality the population standard deviation is larger than 400, would the probability be larger, or smaller, than you found in (a)?

4.35. According to the Census Report in Country A, the number of children in a household has a mean of 3.5 and a standard deviation of 2.0. Suppose the Census Report instead had estimated this mean using a random sample of 300 homes, and that sample had a mean of 4.2 and standard deviation of 1.5.

(a) Identify the variable y.

(b) Describe the center and spread of the population distribution.

(c) Describe the center and spread of the sample data distribution.

(d) Describe the center and spread of the sampling distribution of the sample mean for 300 homes. What does that distribution describe?

4.36. At a university, 55% of the 6500 students are female. The student newspaper reports results of a survey of a random sample of 60 students about the activity clubs they wish to join. They report that their sample contained 32 females.

(a) Explain how you can set up a variable y to represent gender.

(b) Identify the population distribution of gender at this university.

(c) Identify the sample data distribution of gender for this sample.

(d) The sampling distribution of the sample proportion of females in the sample is approximately a normal distribution with mean 0.55 and standard error 0.07. Explain what this means.

4.37. The distribution of family size in a particular society is skewed to the right, with $\mu = 10.4$ and $\sigma = 6.0$. These values are unknown to an anthropologist, who samples families to estimate mean family size. For a random sample of 40 families, the anthropologist gets a mean of 8.6 and a standard deviation of 5.2.

(a) Identify the population distribution and its mean and standard deviation.

(b) Identify the sample data distribution and its mean and standard deviation.

(c) Identify the sampling distribution of $\bar{y}$ and its mean and standard error, and explain what it describes.

(d) Find the probability that the anthropologist's sample mean falls within 0.5 of the population mean.

(e) If the sample were truly random, would you be surprised if the anthropologist obtained $\bar{y} = 3.0$? Why? (This could well happen if the sample were not random.)

Concepts and Applications

4.38. Use the applet for the *Sampling Distribution for the Sample Proportion* at www.pearsonglobaleditions.com/Agresti to illustrate this concept. Set the population proportion as 0.50 and sample size $n = 100$.

(a) Simulate once (setting the number of samples to 1 and clicking on *Draw Sample*) and report the counts and the proportions for the two categories. Did you get a sample proportion close to 0.50? Perform this simulation of a random sample of size 100 ten times, each time observing from the graphs the counts and the corresponding sample proportion of yes votes. Summarize.

(b) Now plot the results of simulating a random sample of size 100 and finding the sample proportion 10,000 times, by setting 10,000 for the number of samples of size n. How does this plot reflect the Central Limit Theorem?

4.39. Use the applet for the *Sampling Distribution for the Sample Mean* for continuous variables at www.pearsonglobaleditions.com/Agresti to investigate the sampling distribution of $\bar{y}$.

(a) Select the skewed population distribution and set the skewness = 2. Take 10,000 samples of size 50 each. How does the empirical sampling distribution of sample means compare to the population distribution? What does this reflect?

(b) Repeat, this time choosing a sample size of only 2 for each sample. Why is the sampling distribution not symmetric and bell shaped?

4.40. Go to the applet for the *Sampling Distribution for the Sample Mean* for discrete variables at www.pearsonglobaleditions.com/Agresti.

(a) Construct a population distribution that you think is plausible for $y =$ number of cars in a house.

(b) Draw a single sample of size $n = 1000$ to reflect results of a typical sample survey. Summarize how the sample data mean and standard deviation resemble those for the population.

(c) Now draw 10000 samples of size 1000 each to approximate the sampling distribution of $\bar{y}$. Report the mean and standard deviation of this empirical sampling

distribution and compare to the theoretical values of μ and $\sigma - y = \sigma/\sqrt{n}$. Explain what this sampling distribution represents.

4.41. For a single roll of a die, let $y = 1$ for an even number and $y = 0$ for an odd number, to simulate the vote in an election with two equally preferred candidates.

(a) Construct the probability distribution for y, and find its mean.

(b) The die is rolled 10 times, yielding an even number six times and an odd number four times. Construct the sample data distribution.

(c) Use the applet for the *Sampling Distribution for the Sample Proportion* at www.pearsonglobaleditions.com/Agresti to simulate what would happen if everyone in a university with 10,000 students rolled a die 10 times and observed the proportion of times an even number comes up in the sample. Describe the shape and spread of the empirical sampling distribution compared to the distributions in (a) and (b).

(d) What does the applet report for the mean and the standard deviation of the empirical sampling distribution in (c)? What are the theoretical values for the true sampling distribution? (In finding this, you can use 0.50 as the population standard deviation of the distribution in (a).)

4.42. (*Class Exercise*) Refer to Exercises 1.11 and 1.12 (pages 21 and 22). Using the population defined by your class or using the Students data file, the instructor will select a variable, such as weekly time watching television.

(a) Construct a histogram or stem-and-leaf plot of the population distribution of the variable for the class.

(b) By generating random numbers, each student should select nine students at random and compute the sample mean response for those students. (Each student should use different random numbers.) Plot a histogram of the sample means obtained by all the students. How do the spread and shape compare to the histogram in (a)? What does this illustrate?

4.43. Rainbow City was designed to attract orphan children. Its current population of 30,000 children has a mean age of 6 years and a standard deviation of 10 years. The distribution of ages is skewed to the left, reflecting the predominance of older children. A random sample of 100 children of Rainbow City has $\bar{y} = 8.3$ and $s = 7.0$.

(a) Describe the center and spread of the population distribution.

(b) Describe the center and spread of the sample data distribution. What shape does it probably have?

(c) Find the center and spread of the sampling distribution of $\bar{y}$ for $n = 100$. What shape does it have and what does it describe?

(d) Explain why it would not be unusual to observe a child of age 5 years in Rainbow City, but it would be highly unusual to observe a sample mean of 5, for a random sample size of 100.

4.44.* Refer to the previous exercise. Describe the sampling distribution of $\bar{y}$ **(a)** for a random sample of size $n = 1$; **(b)** if you sample all 30,000 children.

4.45. (*Class Exercise*) Table 4.5 provides the ages of all 50 heads of households in a small Nova Scotian fishing village. The data are in the data file Ages at the text website. The distribution of these ages is characterized by $\mu = 47.18$ and $\sigma = 14.74$.

(a) Construct a stem-and-leaf plot or histogram of the population distribution.

(b) Each student should generate nine random numbers between 01 and 50 and use them to sample nine heads of households. Compute their sample mean age. Plot the empirical sampling distribution of the $\bar{y}$-values. Compare it to the distribution in (a).

(c) What do you expect for the mean of the $\bar{y}$-values in a long run of repeated samples of size 9?

(d) What do you expect for the standard deviation of the $\bar{y}$-values in a long run of repeated samples of size 9?

TABLE 4.5

Name	Age	Name	Age	Name	Age	Name	Age
Alexander	50	Griffith	66	McTell	49	Staines	33
Bell	45	Grosvenor	51	MacLeod	30	Stewart	36
Bell	23	Ian	57	McNeil	28	Stewart	25
Bok	28	Jansch	40	McNeil	31	Thames	29
Clancy	67	Keelaghan	36	McNeil	45	Thomas	57
Cochran	62	Lavin	38	McNeil	43	Todd	39
Fairchild	41	Lunny	81	Mitchell	43	Trickett	50
Finney	68	MacColl	27	Muir	54	Trickett	64
Fisher	37	McCusker	37	Oban	62	Tyson	76
Francey	60	McCusker	56	Reid	67	Watson	63
Fricker	41	McDonald	71	Renbourn	48	Young	29
Gaughan	70	McDonald	39	Rogers	32		
Graham	47	McDonald	46	Rush	42		

4.46. (a) Which distribution does the sample data distribution tend to resemble more closely—the sampling distribution or the population distribution? Explain.

(b) Explain carefully the difference between a *sample data distribution* and the *sampling distribution* of $\bar{y}$. Illustrate your answer for a variable y that can take only values of 0 and 1.

4.47. The Central Bureau of Statistics in Sunshine Village asked mothers aged between 22 and 26 about the ideal number of children. The probability distribution is approximately $P(1) = 0.02$, $P(2) = 0.14$, $P(3) = 0.07$, $P(4) = 0.33$, $P(5) = 0.15$, and $P(6 \text{ or more}) = 0.28$.

(a) Because the last category is open ended, it is not possible to calculate the mean exactly. Find a lower bound for the mean.

(b) Explain why you can find the median of the distribution, and find it.

4.48. For a normal distribution, show that

(a) The upper quartile equals $\mu + 0.67\sigma$.

(b) According to the 1.5(IQR) criterion, an outlier is an observation falling more than 2.7 standard deviations below or above the mean, and this happens for only 0.7% of the data.

4.49. In an exit poll of 3646 voters in the 2017 election in a town, 50.8% said they voted for Candidate R and 49.2% said they voted for Candidate S. Based on this information, would you be willing to predict the winner of the election? Explain your reasoning.

4.50. For an election exit poll that uses random sampling, find the standard error of the sample proportion voting for a candidate for whom the population proportion is 0.50, when $n = 100$, when $n = 1000$, and when $n = 10000$. In each case, predict an interval within which the sample proportion is almost certain to fall. Notice that the interval shrinks in width as the sample size increases. This is a consequence of the **law of large numbers**, which states that with random sampling, the sample proportion tends to get closer and closer to the population proportion as n increases indefinitely.

Select the correct response(s) in multiple-choice questions 4.51−4.52. There may be more than one correct answer.

4.51. The standard error of a statistic describes

(a) The standard deviation of the sampling distribution of that statistic.

(b) The standard deviation of the sample data.

(c) How close that statistic is likely to fall to the parameter that it estimates.

(d) The variability in the values of the statistic for repeated random samples of size n.

(e) The error that occurs due to nonresponse and measurement errors.

4.52. The Central Limit Theorem implies that

(a) All variables have bell-shaped sample data distributions if a random sample contains at least about 30 observations.

(b) Population distributions are normal whenever the population size is large.

(c) For large random samples, the sampling distribution of $\bar{y}$ is approximately normal, regardless of the shape of the population distribution.

(d) The sampling distribution looks more like the population distribution as the sample size increases.

4.53. True or False: As the sample size increases, the standard error of the sampling distribution of $\bar{y}$ increases. Explain your answer.

4.54.* A junior college admits students only if they score above 500 on a standardized achievement test. Applicants from Region A have a mean of 550 and a standard deviation of 120 on this test, and applicants from Region B have a mean of 600 and a standard deviation of 120. Both distributions are approximately normal, and both regions have the same size.

(a) Find the proportion not admitted for students from each region.

(b) Of the students who are not admitted, what proportion are from Region A?

(c) A state legislator proposes that the college lower the cutoff point for admission to 300, thinking that the proportion of the students who are not admitted who are from Region A would decrease. If this policy is implemented, determine the effect on the answer to (b), and comment.

4.55.* From the formula on page 84, the standard deviation of a discrete probability distribution is

$$\sigma = \sqrt{\sum (y - \mu)^2 P(y)}.$$

(a) When y can equal only 0 and 1, letting $\pi = P(y = 1)$ and $1 - \pi = P(y = 0)$, show that $\mu = \pi$ and that $\sigma = \sqrt{\pi(1 - \pi)}$.

(b) Show that the standard error of a sample proportion for a random sample of size n equals $\sqrt{\pi(1 - \pi)/n}$.

4.56.* Refer to the formula for the normal distribution curve shown in the footnote on page 85. Show that this curve is symmetric, by showing that for any constant c, the curve has the same value at $y = \mu + c$ as at $y = \mu - c$. (The integral of $f(y)$ for y between $\mu + z\sigma$ and ∞ equals the tail probability tabulated in Table A.)

4.57.* The standard error formula $\sigma_{\bar{y}} = \sigma / \sqrt{n}$ treats the population size as *infinitely* large relative to the sample size n. The formula for $\sigma_{\bar{y}}$ for a *finite* population size denoted by N is

$$\sigma_{\bar{y}} = \sqrt{\frac{N - n}{N - 1}} \left(\frac{\sigma}{\sqrt{n}} \right).$$

The term $\sqrt{(N - n)/(N - 1)}$ is called the **finite population correction**.

(a) When $n = 300$ students are selected from a college student body of size $N = 30,000$, show that $\sigma_{\bar{y}} = 0.995\sigma / \sqrt{n}$. (In practice, n is usually small relative to N, so the correction has little influence.)

(b) If $n = N$ (i.e., we sample the entire population), show that $\sigma_{\bar{y}} = 0$. In other words, no sampling error occurs, because $\bar{y} = \mu$.

(c) For $n = 1$, explain why the sampling distribution of $\bar{y}$ and its standard error are identical to the population distribution and its standard deviation.

4.58.* A general rule states that for independent observations, the variance of $\sum y_i$ is the sum of the variances, which is $n\sigma^2$ for n observations.

(a) Explain intuitively why $\sum y_i$ would have a larger variance than a single observation y.

(b) Since the variance of a probability distribution is $\sigma^2 = E(y - \mu)^2$, explain why the variance of the sampling distribution of $\bar{y}$ is

$$E\left[\frac{\sum y_i}{n} - \mu\right]^2 = E\left[\frac{\sum y_i - n\mu}{n}\right]^2 =$$

$$\frac{1}{n^2} E\left(\sum y_i - n\mu\right)^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}.$$

(*Hint*: The second expression represents a sum with n^2 in the denominator, which is a constant that can be put in front of the summation.)

(c) From (b), explain why the standard error equals $\sigma_{\bar{y}} = \sigma/\sqrt{n}$.

4.59.* Ellenberg (2014) noted that when you use sample data to rank states by brain cancer rates, the highest ranking state (South Dakota) and the nearly lowest ranking state (North Dakota) had relatively small sample sizes. Also, when schools in North Carolina were ranked by their average improvement in test scores, the best and the worst schools were schools with very small sample sizes. Explain how these results could merely reflect sample means and proportions having larger variability when sample sizes are smaller. (*Hint*: What would you expect with the sample if all the population means were identical?)

This page intentionally left blank

STATISTICAL INFERENCE: ESTIMATION

This chapter shows how to use sample data to estimate population parameters. With categorical variables, we estimate population proportions for the categories. For example, a study dealing with binge drinking by college students might estimate the proportion of college students who participate in binge drinking. With quantitative variables, we estimate the population mean. For example, the study might estimate the mean number of alcoholic drinks taken in a typical binge-drinking experience for the population of college students who do this.

We first learn about two types of estimates: One is a single point and the other is an interval of points, called a ***confidence interval***. We construct confidence intervals for population proportions and means by taking a point estimate and adding and subtracting a margin of error that depends on the sample size. We also learn how to find the sample size needed to achieve the desired precision of estimation. The final section presents two general-purpose methods for estimation—***maximum likelihood*** and the ***bootstrap***—that apply to nearly all other parameters, such as the population median.

5.1 Point and Interval Estimation

We use sample data to estimate a parameter in two ways:

- A ***point estimate*** is a *single number* that is the best guess for the parameter value.

- An ***interval estimate*** is an *interval of numbers* around the point estimate that we believe contains the parameter value. This interval is also called a ***confidence interval***.

For example, a General Social Survey asked, "Do you believe there is a life after death?" For 1958 subjects sampled, the point estimate for the proportion of all Americans who would respond *yes* equals 0.73. An interval estimate predicts that the population proportion responding *yes* falls between 0.71 and 0.75. That is, this confidence interval tells us that the point estimate of 0.73 has a *margin of error* of 0.02. Thus, an interval estimate helps us gauge the precision of a point estimate.

The term *estimate* alone is often used as short for *point estimate*. The term *estimator* then refers to a particular type of statistic for estimating a parameter and *estimate* refers to its value for a particular sample. For example, the sample proportion is an estimator of a population proportion. The value 0.73 is the estimate for the population proportion believing in life after death.

POINT ESTIMATION OF PARAMETERS

Estimates are the most common statistical inference reported by the mass media. For example, a Gallup Poll in May 2016 reported that 53% of the American public

approved of President Barack Obama's performance in office. This is an estimate rather than a parameter, because it was based on interviewing a sample of about 1500 people rather than the entire population.

Any particular parameter has many possible estimators. For a normal population distribution, for example, the center is the mean and the median, since that distribution is symmetric. So, with sample data, two possible estimators of that center are the sample mean and the sample median.

UNBIASED AND EFFICIENT POINT ESTIMATORS

A good estimator has a sampling distribution that (1) is centered around the parameter and (2) has as small a standard error as possible.

An estimator is ***unbiased*** if its sampling distribution centers around the parameter. Specifically, the parameter is the mean of the sampling distribution. From page 97, for random sampling the mean of the sampling distribution of the sample mean $\bar{y}$ equals the population mean μ. Thus, $\bar{y}$ is an unbiased estimator of the population mean μ. Figure 5.1 illustrates this. For any particular sample, the sample mean may underestimate μ or may overestimate it. If the sample mean were found repeatedly with different samples, however, in the long run the overestimates would counterbalance the underestimates.

FIGURE 5.1: Sampling Distributions of Two Estimators of the Population Mean, for a Skewed Population Distribution
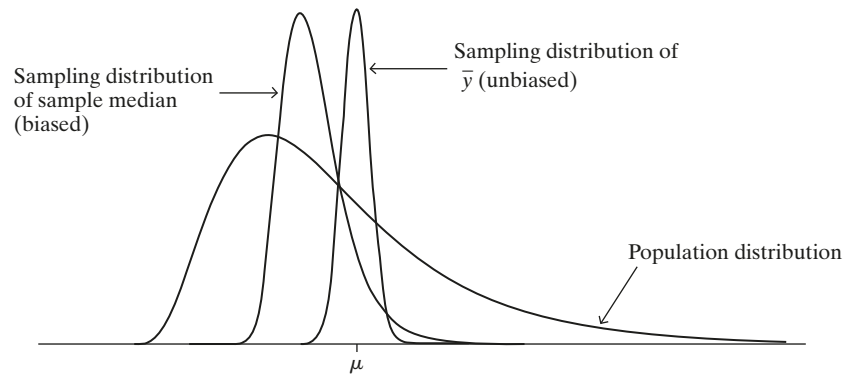
By contrast, a ***biased*** estimator tends to underestimate the parameter, on the average, or it tends to overestimate the parameter. For example, the sample range cannot be larger than the population range, because the sample minimum and maximum cannot be more extreme than the population minimum and maximum. Thus, the sample range tends to underestimate the population range. It is a biased estimator of the population range.

A second desirable property for an estimator is a relatively small standard error. An estimator having standard error that is smaller than those of other estimators is said to be ***efficient***. An efficient estimator tends to fall closer than other estimators to the parameter. For example, when a population distribution is normal, the standard error of the sample median is 25% larger than the standard error of the sample mean. The sample mean tends to be closer than the sample median to the population center. The sample mean is an efficient estimator. The sample median is inefficient.

In summary, a good estimator of a parameter is *unbiased*, or nearly so, and *efficient*. Statistical methods use estimators that possess these properties. The final

section of this chapter introduces a general method, called *maximum likelihood*, for constructing estimators that have these properties.

ESTIMATORS OF MEAN, STANDARD DEVIATION, AND PROPORTION

It is common, but not necessary, to use the sample analog of a population parameter as its estimator. For instance, to estimate a population proportion, the sample proportion is an estimator that is unbiased and efficient. For estimating a population mean μ, the sample mean $\bar{y}$ is unbiased. It is efficient for the most common population distributions. Likewise, we use the sample standard deviation s as the estimator of the population standard deviation σ.

The symbol " ˆ " over a parameter symbol is often used to represent an estimate of that parameter. The symbol " ˆ " is called a *caret*, and is usually read as *hat*. For example, $\hat{\mu}$ is read as *mu-hat*. Thus, $\hat{\mu}$ denotes an estimate of the population mean μ.

CONFIDENCE INTERVAL FORMED BY POINT ESTIMATE $\pm$ MARGIN OF ERROR

To be truly informative, an inference about a parameter should provide not only a point estimate but should also indicate how close the estimate is likely to fall to the parameter value. For example, since 1996 each year the Gallup Poll has asked, "Do you think marriages between same-sex couples should or should not be recognized by the law as valid, with the same rights as traditional marriages?" The percentage saying they should be valid has increased from 27% in 1996 to 60% in 2015. How accurate are these estimates? Within 2%? Within 5%? Within 10%?

The information about the precision of a point estimate determines the width of an *interval estimate* of the parameter. This consists of an interval of numbers around the point estimate. It is designed to contain the parameter with some chosen probability close to 1. Because interval estimates contain the parameter with a certain degree of confidence, they are referred to as *confidence intervals*.

Confidence Interval

> A *confidence interval* for a parameter is an interval of numbers within which the parameter is believed to fall. The probability that this method produces an interval that contains the parameter is called the *confidence level*. This is a number chosen to be close to 1, such as 0.95 or 0.99.

The key to constructing a confidence interval is the sampling distribution of the point estimator. Often, the sampling distribution is approximately normal. The normal distribution then determines the probability that the estimator falls within a certain distance of the parameter. With probability about 0.95, the estimator falls within two standard errors. To construct a confidence interval, we add and subtract from the point estimate a z-score multiple of its standard error. This is the *margin of error*. That is,

Form of confidence interval: **Point estimate $\pm$ Margin of error.**

To construct a confidence interval having "95% confidence," we take the point estimate and add and subtract a margin of error that equals about two standard errors. We'll see the details in the next two sections.

5.2 Confidence Interval for a Proportion

For categorical data, an observation occurs in one of a set of categories. This type of measurement occurs when the variable is nominal, such as preferred candidate (Democrat, Republican, Independent), or ordinal, such as opinion about how much the government should address global warming (less, the same, more). It also occurs when inherently continuous variables are measured with categorical scales, such as when annual income has categories ($0–$24,999, $25,000–$49,999, $50,000–$74,999, at least $75,000).

To summarize categorical data, we record the *proportions* (or *percentages*) of observations in the categories. For example, a study might provide a point or interval estimate of

- The proportion of Americans who lack health insurance.

- The proportion of Canadians who favor independent status for Quebec.

- The proportion of Australian young adults who have taken a "gap year," that is, a break of a year between high school and college or between college and regular employment.

THE SAMPLE PROPORTION AND ITS STANDARD ERROR

Let π denote a population proportion.[1] Then, π falls between 0 and 1. Its point estimator is the *sample proportion*. We denote the sample proportion by $\hat{\pi}$, since it estimates π.

Recall that the sample proportion is a mean when we let $y = 1$ for an observation in the category of interest and $y = 0$ otherwise. (See the discussion about Table 3.6 on page 52 and following Example 4.4 on page 90.) Similarly, the population proportion π is the mean μ of the probability distribution having probabilities

$$P(1) = \pi \quad \text{and} \quad P(0) = 1 - \pi.$$

The standard deviation of this probability distribution is[2] $\sigma = \sqrt{\pi(1-\pi)}$. Since the formula for the standard error of a sample mean is $\sigma_{\bar{y}} = \sigma/\sqrt{n}$, the standard error $\sigma_{\hat{\pi}}$ of the sample proportion $\hat{\pi}$ is

$$\sigma_{\hat{\pi}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\pi(1-\pi)}{n}}.$$

As the sample size increases, the standard error gets smaller. The sample proportion then tends to fall closer to the population proportion.

CONFIDENCE INTERVAL FOR A PROPORTION

Since the sample proportion $\hat{\pi}$ is a sample mean, the Central Limit Theorem applies: For large random samples, the sampling distribution of $\hat{\pi}$ is approximately normal about the parameter π it estimates. Figure 5.2 illustrates this. Recall that 95% of a normal distribution falls within two standard deviations of the mean, or, more precisely, 1.96 standard deviations. So, with probability 0.95, $\hat{\pi}$ falls within $1.96\sigma_{\hat{\pi}}$ units of the parameter π, that is, between $\pi - 1.96\sigma_{\hat{\pi}}$ and $\pi + 1.96\sigma_{\hat{\pi}}$, as Figure 5.2 shows.

---

[1] Here, π is the Greek analog of p for proportion, *not* the mathematical constant, 3.1415….
[2] From page 84, the variance is $\sigma^2 = \sum(y-\mu)^2 P(y) = (0-\pi)^2 P(0) + (1-\pi)^2 P(1) = (0-\pi)^2(1-\pi) + (1-\pi)^2\pi$, which simplifies to $\pi(1-\pi)$. Thus, $\sigma = \sqrt{\pi(1-\pi)}$.

FIGURE 5.2: Sampling
Distribution of $\hat{\pi}$ and
Possible 95% Confidence
Intervals for π
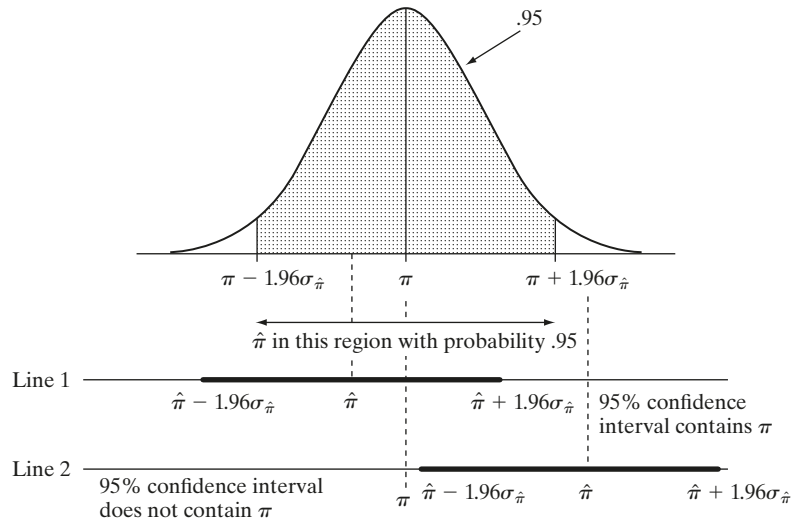
Once the sample is selected, if $\hat{\pi}$ does fall within $1.96\sigma_{\hat{\pi}}$ units of π, then the interval from $\hat{\pi} - 1.96\sigma_{\hat{\pi}}$ to $\hat{\pi} + 1.96\sigma_{\hat{\pi}}$ contains π. See line 1 of Figure 5.2. In other words, with probability 0.95, a $\hat{\pi}$ value occurs such that the interval $\hat{\pi} \pm 1.96\sigma_{\hat{\pi}}$ contains the population proportion π. On the other hand, the probability is 0.05 that $\hat{\pi}$ does *not* fall within $1.96\sigma_{\hat{\pi}}$ of π. If that happens, then the interval from $\hat{\pi} - 1.96\sigma_{\hat{\pi}}$ to $\hat{\pi} + 1.96\sigma_{\hat{\pi}}$ does *not* contain π (see Figure 5.2, line 2). Thus, the probability is 0.05 that $\hat{\pi}$ is such that $\hat{\pi} \pm 1.96\sigma_{\hat{\pi}}$ does *not* contain π.

The interval $\hat{\pi} \pm 1.96\sigma_{\hat{\pi}}$ is an interval estimate for π with confidence level 0.95. It is called a ***95% confidence interval***. In practice, the value of the standard error $\sigma_{\hat{\pi}} = \sqrt{\pi(1 - \pi)/n}$ for this formula is unknown, because it depends on the unknown parameter π. So, we estimate this standard error by substituting the sample proportion, using

$$se = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}.$$

We've used the symbol *s* to denote a sample standard deviation, which estimates the population standard deviation σ. *In the remainder of this text, we use the symbol se to denote a sample estimate of a standard error.*

The confidence interval formula uses this estimated standard error. In summary, the 95% confidence interval for π is

$$\hat{\pi} \pm 1.96(se), \quad \text{where} \quad se = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}.$$

**Example
5.1**

Estimating the Proportion Who Favor Restricting Legalized Abortion For many years, the Florida Poll[3] conducted by Florida International University asked, "In general, do you think it is appropriate for state government to make laws restricting access to abortion?" In the most recent poll, of 1200 randomly chosen adult Floridians, 396 said *yes* and 804 said *no*. We shall estimate the population proportion who would respond *yes* to this question.

Let π represent the population proportion of adult Floridians who would respond *yes*. Of the $n = 1200$ respondents in the poll, 396 said *yes*, so $\hat{\pi} = 396/1200 = 0.330$. Then, $1 - \hat{\pi} = 0.670$. That is, 33% of the sample said *yes* and 67% said *no*.

---

[3] See www2.fiu.edu/~ipor/ffp/abort1.htm.

The estimated standard error of the sample proportion $\hat{\pi}$ is

$$se = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} = \sqrt{\frac{(0.33)(0.67)}{1200}} = \sqrt{0.000184} = 0.0136.$$

A 95% confidence interval for π is

$$\hat{\pi} \pm 1.96(se) = 0.330 \pm 1.96(0.0136) = 0.330 \pm 0.027, \quad \text{or} \quad (0.30, 0.36).$$

We conclude that the population percentage supporting restricting access to abortion appears to be at least 30% but no more than 36%. All numbers in the confidence interval (0.30, 0.36) fall below 0.50. Thus, at the time of this poll, apparently fewer than half the Florida adult population supported restricting access to abortion. ∎

You can obtain this confidence interval using software with your data file. In Stata, you can also find it directly from the summary results, by applying the `cii` command[4] to n and the count in the category of interest:

```
. cii proportions 1200 396, wald

                                         -- Binomial Wald ---
    Variable |     Obs    Mean   Std. Err.    [95% Conf. Interval]
             |   1,200     .33   .0135739     .3033957    .3566043
```

The software R uses a confidence interval for the proportion that has a more complex formula[5] than the one we gave, so it gives slightly different results:

```
> prop.test(396, 1200)$conf.int
[1] 0.3035683 0.3575336
```

Calculators for such confidence intervals are also available with Internet applets. See Figure 5.3 for an example.

FIGURE 5.3: Applets at www.pearsonglobal editions.com/Agresti Perform Inference Procedures Presented in Chapters 5–9. The *Inference for a Proportion* applet can construct confidence intervals for proportions.
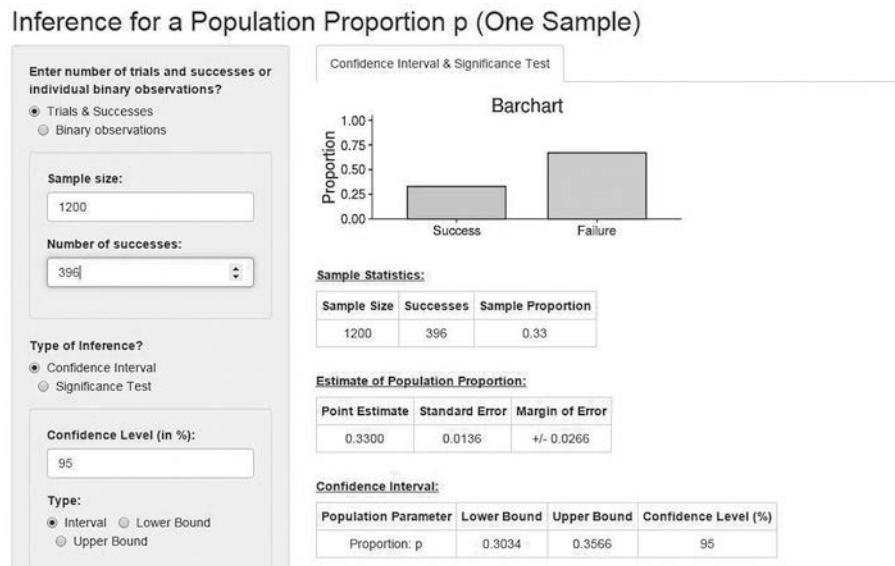
Inference for a Population Proportion p (One Sample)

Results in such surveys vary greatly depending on the question wording and where the poll is conducted. For instance, when the 2014 General Social Survey asked whether a pregnant woman should be able to obtain a legal abortion if the woman wants it *for any reason*, 907 said *no* and 746 said *yes*. The 95% confidence interval for the population proportion saying *no* equals (0.53, 0.57).

If you construct a confidence interval using a hand calculator, don't round off while doing the calculation or your answer may be affected, but do round off when you report the final answer. Likewise, in reporting results from software output, you should use only the first two or three significant digits. Report the confidence interval as (0.30, 0.36) rather than (0.303395, 0.356605). Software's extra precision provides accurate calculations in finding *se* and the confidence interval. However, the extra digits are distracting in reports and not useful. They do not tell us anything extra in a practical sense about the population proportion, and their validity is shaky because the sampling distribution is only *approximately* normal.

<table>
<tr><td>

**Example
5.2**

</td><td>

Estimating Proportion Who "Oppose" from Proportion Who "Favor"   In the Florida Poll, for estimating the population proportion who supported restricting access to abortion, we obtained $se = 0.0136$ for the point estimate $\hat{\pi} = 0.33$. Similarly, the estimated standard error for $1 - \hat{\pi} = 0.67$, the proportion of voters who say *no* to restricting access to abortion, is

</td></tr>
</table>

$$se = \sqrt{(1 - \hat{\pi})\hat{\pi}/n} = \sqrt{(0.67)(0.33)/1200} = 0.0136.$$

Both proportions have the same *se*.

A 95% confidence interval for the population proportion of *no* responses to restricting access to abortion is

$$0.67 \pm 1.96(0.0136) = 0.67 \pm 0.03, \quad \text{or} \quad (0.64, 0.70).$$

Now, $0.64 = 1 - 0.36$ and $0.70 = 1 - 0.30$, where $(0.30, 0.36)$ is the 95% confidence interval for π. Thus, inferences for the proportion $1 - \pi$ follow directly from those for the proportion π by subtracting each endpoint of the confidence interval from 1.0. ∎

CONTROLLING THE CONFIDENCE LEVEL BY CHOICE OF z-SCORE

With a confidence level of 0.95, that is, "95% confidence," there is a 0.05 probability that the method produces a confidence interval that does *not* contain the parameter value. In some applications, a 5% chance of an incorrect inference is unacceptable. To increase the chance of a correct inference, we use a larger confidence level, such as 0.99.

The general form for the confidence interval for a population proportion π is

$$\hat{\pi} \pm z(se), \quad \text{with} \quad se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n},$$

where z depends on the confidence level. The higher the confidence level, the greater the chance that the confidence interval contains the parameter. High confidence levels are used in practice so that the chance of error is small. The most common confidence level is 0.95, with 0.99 used when it is more crucial not to make an error.

**Example
5.3**

Finding a 99% Confidence Interval   For the data in Example 5.1 (page 119), let's find a 99% confidence interval for the population proportion who favor laws restricting access to abortion. Now, 99% of a normal distribution occurs within 2.58 standard deviations of the mean. So, the probability is 0.99 that the sample proportion $\hat{\pi}$ falls within 2.58 standard errors of the population proportion π. A 99% confidence interval for π is $\hat{\pi} \pm 2.58(se)$.

In Example 5.1, the sample proportion was 0.33, with $se = 0.0136$. So, the 99% confidence interval is

$$\hat{\pi} \pm 2.58(se) = 0.33 \pm 2.58(0.0136) = 0.33 \pm 0.04, \quad \text{or} \quad (0.29, 0.37).$$

Compared to the 95% confidence interval of (0.30, 0.36), this interval estimate is less precise, being a bit wider. To be more sure of enclosing the parameter, we must sacrifice precision of estimation by using a wider interval.   ∎

The z-value multiplied by se is the *margin of error*. With greater confidence, the confidence interval is wider because the z-score in the margin of error is larger—for instance, $z = 1.96$ for 95% confidence and $z = 2.58$ for 99% confidence.

Why do we settle for anything less than 100% confidence? To be absolutely 100% certain of a correct inference, the interval must contain all possible values for π. A 100% confidence interval for the population proportion in favor of limiting access to abortion goes from 0.0 to 1.0. This is not helpful. In practice, we settle for less than perfection in order to estimate much more precisely the parameter value. In forming a confidence interval, we compromise between the desired confidence that the inference is correct and the desired precision of estimation. As one gets better, the other gets worse. This is why you would not typically see a 99.9999% confidence interval. It would usually be too wide to say much about where the population parameter falls (its z-value is 4.9).

LARGER SAMPLE SIZES GIVE NARROWER INTERVALS

We can estimate a population proportion π more precisely with a larger sample size. The margin of error is $z(se)$, where $se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$. The larger the value of n, the smaller the margin of error and the narrower the interval.

To illustrate, suppose that $\hat{\pi} = 0.33$ in Example 5.1 on estimating the proportion who favor restricting legalized abortion was based on $n = 300$, only a fourth as large as the actual sample size of $n = 1200$. Then, the estimated standard error of $\hat{\pi}$ is

$$se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n} = \sqrt{(0.33)(0.67)/300} = 0.027,$$

twice as large as the se in Example 5.1. The resulting 95% confidence interval is

$$\hat{\pi} \pm 1.96(se) = 0.33 \pm 1.96(0.027) = 0.33 \pm 0.053.$$

This is twice as wide as the confidence interval formed from the sample of size $n = 1200$.

Since the margin of error is inversely proportional to the square root of n, and since $\sqrt{4n} = 2\sqrt{n}$, the sample size must *quadruple* in order to *double* the precision (i.e., halve the width). Section 5.4 shows how to find the sample size needed to achieve a certain precision.

In summary, two factors affect the width of a confidence interval:

- Increases as the confidence level increases.
- Decreases as the sample size increases.

These properties apply to all confidence intervals, not only the one for a proportion.

ERROR PROBABILITY = 1 − CONFIDENCE LEVEL

The probability that an interval estimation method yields a confidence interval that does *not* contain the parameter is called the ***error probability***. This equals 1 minus the confidence level. For confidence level 0.95, the error probability equals $1 - 0.95 = 0.05$. In statistical inference, the Greek letter α (alpha) denotes the error probability, and $1 - \alpha$ is the confidence level. For an error probability of $\alpha = 0.05$, the confidence level equals $1 - \alpha = 0.95$.

The z-value for the confidence interval is such that the probability is α that $\hat{\pi}$ falls *more than* z standard errors from π. The z-value corresponds to a total probability of α in the two tails of a normal distribution, or $\alpha/2$ (half the error probability) in each tail. For example, for a 95% confidence interval, $\alpha = 0.05$, and the z-score is the one with probability $\alpha/2 = 0.05/2 = 0.025$ in each tail. This is $z = 1.96$.

CONFIDENCE LEVEL IS LONG-RUN PROPORTION CORRECT

The confidence level for a confidence interval describes how the method performs when used over and over with many different random samples. The unknown population proportion π is a fixed number. A confidence interval constructed from any particular sample either does or does not contain π. If we repeatedly selected random samples of that size and each time constructed a 95% confidence interval, then in the long run about 95% of the intervals would contain π. This happens because about 95% of the sample proportions would fall within 1.96(se) of π, as does the $\hat{\pi}$ in line 1 of Figure 5.2 (page 119). Saying that a particular interval contains π with "95% confidence" signifies that *in the long run* 95% of such intervals would contain π. That is, 95% of the time the inference is correct.

Figure 5.4 shows the results of selecting 10 separate samples and calculating the sample proportion for each and a 95% confidence interval for the population proportion. The confidence intervals jump around because $\hat{\pi}$ varies from sample to sample. However, 9 of the 10 intervals contain the population proportion π. On the average, only about 1 out of 20 times does a 95% confidence interval fail to contain the population parameter.

FIGURE 5.4: Ten 95% Confidence Intervals for a Population Proportion π. In the long run, only 5% of the intervals fail to contain π.
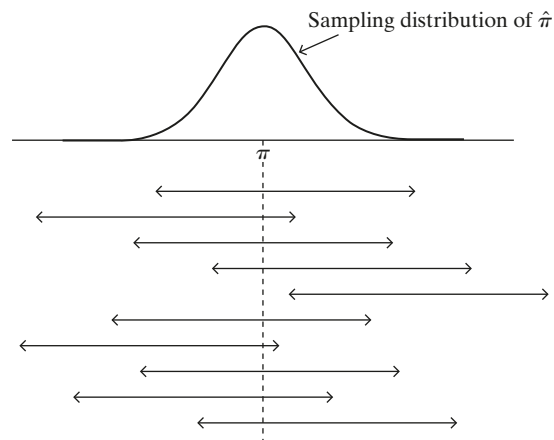
You can get a feel for this using an applet designed to illustrate the performance of confidence intervals for proportions:

- Go to `www.pearsonglobaleditions.com/Agresti` and click on *Explore Coverage*. Use the *Confidence Interval for a Proportion* option.

- The default is forming a 95% confidence interval when $n = 50$ and the true parameter value is $\pi = 0.30$. To better reflect Example 5.1, set the sample size to 1200. Choose 10 samples of size 1200 each. Click on *Draw Sample*. You will see a plot of the 10 confidence intervals, with ones drawn in red that do not contain the parameter value of 0.30. The output also summarizes the number and percentage of the confidence intervals that contain $\pi = 0.30$. What is this?

- Now select 1000 for the number of samples to draw, each of size 1200. Now the proportion of the intervals that actually contain the parameter value is probably closer to 0.95.

In practice, we select only *one* sample of some fixed size n and construct *one* confidence interval using the observations in that sample. We do not know whether that confidence interval truly contains π. Our confidence in that interval is based on long-term properties of the procedure. We can control, by our choice of the confidence level, the chance that the interval contains π. If an error probability of 0.05 makes us nervous, we can instead form a 99% confidence interval, for which the method makes an error only 1% of the time.

LARGE SAMPLE SIZE NEEDED FOR VALIDITY OF METHOD

In practice, the probability that the confidence interval contains π is *approximately* equal to the chosen confidence level. The approximation is better for larger samples. As n increases, the sampling distribution of $\hat{\pi}$ is more closely normal in form, by the Central Limit Theorem. This is what allows us to use z-scores from the normal distribution in finding the margin of error. Also as n increases, the *estimated* standard error $se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$ gets closer to the *true* standard error $\sigma_{\hat{\pi}} = \sqrt{\pi(1 - \pi)/n}$.

For this reason, the confidence interval formula applies with *large* random samples. How large is "large"? A general guideline states you should have at least 15 observations both in the category of interest and not in it.[6] This is true in most social science studies. In Example 5.1, the counts in the two categories were 396 and 804, so the sample size requirement was easily satisfied. Section 5.4 and Exercise 5.77 show methods that work well when the guideline is not satisfied.

Here is a summary of the confidence interval for a proportion:

Confidence Interval for Population Proportion π

For a random sample with sample proportion $\hat{\pi}$, a confidence interval for a population proportion π is

$$\hat{\pi} \pm z(se), \quad \text{which is} \quad \hat{\pi} \pm z\sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}.$$

The z-value is such that the probability under a normal curve within z standard errors of the mean equals the confidence level. For 95% and 99% confidence intervals, z equals 1.96 and 2.58. The sample size n should be sufficiently large that at least 15 observations are in the category and at least 15 are not in it.

---

[6] For justification of this guideline, download the article at `www.stat.ufl.edu/~aa/ci_proportion.pdf`

5.3 Confidence Interval for a Mean

We've learned how to construct a confidence interval for a population proportion for categorical data. We now learn how to construct one for the population mean for quantitative data.

ESTIMATED STANDARD ERROR FOR THE MARGIN OF ERROR

Like the confidence interval for a proportion, the confidence interval for a mean has the form

$$\text{Point estimate} \pm \text{Margin of error,}$$

where the margin of error is a multiple of the standard error. The point estimate of the population mean μ is the sample mean, $\bar{y}$. For large random samples, by the Central Limit Theorem, the sampling distribution of $\bar{y}$ is approximately normal. So, for large samples, we can again find a margin of error by multiplying a z-score from the normal distribution times the standard error.

From Section 4.5, the standard error of the sample mean is

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}},$$

where σ is the population standard deviation. Like the standard error of a sample proportion, this depends on an unknown parameter, in this case σ. In practice, we estimate σ by the sample standard deviation s. So, confidence intervals use the *estimated* standard error

$$se = s/\sqrt{n}.$$

Example 5.4

Estimating Mean Number of Sex Partners   When the 2014 General Social Survey asked respondents how many male sex partners they have had since their 18th birthday, the 129 females in the sample between the ages of 23 and 29 reported a median of 3 and mean of 6.6. Software output summarizes the results:

Variable	n	Mean	StDev	SE Mean	95.0% CI
NUMMEN	129	6.6	13.3	1.17	(4.4, 8.8)

How did software get the standard error reported of 1.17? How do we interpret it and the confidence interval shown?

The sample standard deviation is $s = 13.3$. The sample size is $n = 129$. So, the estimated standard error of the sample mean is

$$se = s/\sqrt{n} = 13.3/\sqrt{129} = 1.17.$$

In several random samples of 129 women in this age grouping, the sample mean number of male sex partners would vary from sample to sample with a standard deviation of about 1.17.

The 95% confidence interval reported of (4.4, 8.8) is an interval estimate of μ, the mean number of male sex partners since the 18th birthday for the corresponding population. We can be 95% confident that this interval contains μ. The point estimate of μ is 6.6, and the interval estimate predicts that μ is likely to be greater than 4.4 but smaller than 8.8.

This example highlights a couple of cautions: First, the sample mean of 6.6 and standard deviation of 13.3 suggest that the sample data distribution is very highly skewed to the right. The mean may be misleading as a measure of center. The median response of 3 is perhaps a more useful summary. It's also worth noting that the mode was 1, with 20.2% of the sample. Second, the margin of error refers only to sampling error. Other potential errors include those due to nonresponse or measurement error (lying or giving an inaccurate response). If such errors are not negligible, the estimate and margin of error may be invalid. ∎

How did software find the margin of error for this confidence interval? As with the proportion, for a 95% confidence interval this is approximately two times the estimated standard error. We'll next find the precise margin of error by multiplying *se* by a score that is very similar to a *z*-score.

THE *t* DISTRIBUTION

We'll now learn about a confidence interval that applies for *any* random sample size. To achieve this generality, it has the disadvantage of assuming that the population distribution is normal. In that case, the sampling distribution of $\bar{y}$ is normal even for small sample sizes.[7]

Suppose we knew the exact standard error of the sample mean, $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. Then, with the additional assumption that the population is normal, for any *n* the appropriate confidence interval formula is

$$\bar{y} \pm z\sigma_{\bar{y}}, \quad \text{which is} \quad \bar{y} \pm z\sigma/\sqrt{n},$$

for instance, with $z = 1.96$ for 95% confidence. In practice, we don't know the *population* standard deviation σ, so we don't know the *exact* standard error. Substituting the *sample* standard deviation *s* for σ to get the *estimated* standard error, $se = s/\sqrt{n}$, then introduces extra error. This error can be sizeable when *n* is small. To account for this increased error, we must replace the *z*-score by a slightly larger score, called a *t*-score. The confidence interval is then a bit wider. The *t*-score is like a *z*-score, but it comes from a bell-shaped distribution that is slightly more spread out than the standard normal distribution. This distribution is called the *t* ***distribution***.

PROPERTIES OF THE *t* DISTRIBUTION

Here are the main properties of the *t* distribution:

- The *t* distribution is bell shaped and symmetric about a mean of 0.

- The standard deviation is a bit larger than 1. The precise value depends on what is called the ***degrees of freedom***, denoted by *df*. The *t* distribution has a slightly different spread for each distinct value of *df*, and different *t*-scores apply for each *df* value.

- For inference about a population mean, the degrees of freedom equal $df = n - 1$, one less than the sample size.

---

[7] The right panel of Figure 4.15 on page 100, which showed sampling distributions for various population distributions, illustrated this.

- The t distribution has thicker tails and is more spread out than the standard normal distribution. The larger the df value, however, the more closely it resembles the standard normal. Figure 5.5 illustrates this. When df is about 30 or more, the two distributions are nearly identical.

FIGURE 5.5: t Distribution Relative to Standard Normal Distribution. The t gets closer to the normal as the degrees of freedom (df) increase, and the two distributions are practically identical when $df > 30$.
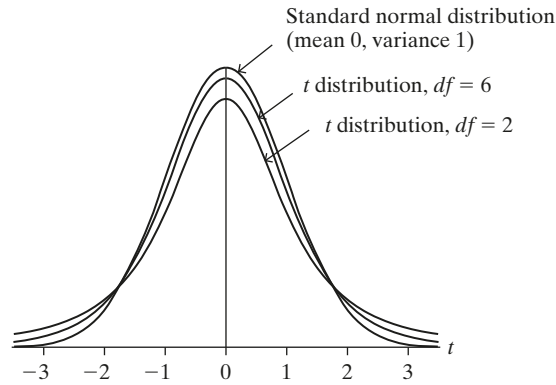
- A t-score multiplied by the estimated standard error gives the margin of error for a confidence interval for the mean.

Table B at the end of the text lists t-scores from the t distribution for various right-tail probabilities. Table 5.1 is an excerpt. The column labeled $t_{.025}$, which has probability 0.025 in the right tail and a two-tail probability of 0.05, is the t-score used in 95% confidence intervals.

TABLE 5.1: t-Scores for Various Confidence Levels and Degrees of Freedom (df). The scores, obtained with the qt function in R software, have right-tail probabilities of 0.100, 0.050, 0.025, 0.010, 0.005, and 0.001.

	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.657	318.3
10	1.372	1.812	2.228	2.764	3.169	4.144
28	1.313	1.701	2.048	2.467	2.763	3.408
30	1.310	1.697	2.042	2.457	2.750	3.385
100	1.290	1.660	1.984	2.364	2.626	3.174
Infinity	1.282	1.645	1.960	2.326	2.576	3.090

To illustrate, when the sample size is 29, the degrees of freedom are $df = n - 1 = 28$. With $df = 28$, we see that $t_{.025} = 2.048$. This means that 2.5% of the t distribution falls in the right tail above 2.048. By symmetry, 2.5% also falls in the left tail below $-t_{.025} = -2.048$. See Figure 5.6. When $df = 28$, the probability equals 0.95 between -2.048 and 2.048. These are the t-scores for a 95% confidence interval when $n = 29$. The confidence interval is $\bar{y} \pm 2.048(se)$.

The t-scores are also supplied by software. For example, the free software R has a function qt that gives the t-score for a particular cumulative probability. For

FIGURE 5.6: The t Distribution with $df = 28$
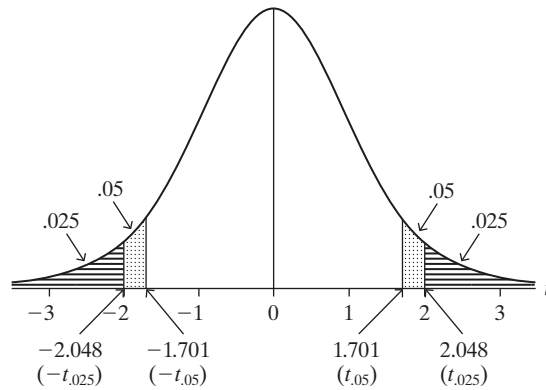
example, the right-tail probability of 0.025 corresponds to a cumulative probability of 0.975, for which the t-score when $df = 28$ is

```
> qt(0.975, 28) # q = "quantile" (percentile) for t distribution
[1] 2.048407
```

With Stata software, we can find this with the `invt` (inverse t) command:

```
. display invt(28, 0.975)
2.0484071
```

It is also possible to find t-scores with SPSS and SAS statistical software, but it is simpler to use Internet sites and statistical calculators. See Figure 5.7.

FIGURE 5.7: The t *Distribution* Applet at www.pearsonglobal editions.com/Agresti Can Supply t Cumulative and Tail Probabilities
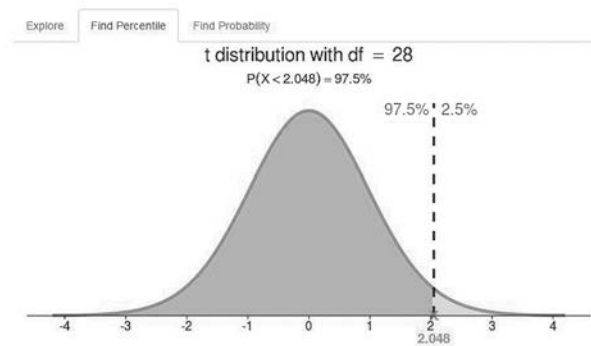
t-SCORES IN THE CONFIDENCE INTERVAL FOR A MEAN

Confidence intervals for a mean resemble those for proportions, except that they use t-scores from the t distribution instead of z-scores from the standard normal distribution.

> For a random sample from a normal population distribution, a
> 95% confidence interval for μ is
>
> $$\bar{y} \pm t_{.025}(se), \quad \text{where} \quad se = s/\sqrt{n}$$
>
> and $df = n - 1$ for the t-score.

   Like the confidence interval for a proportion, this confidence interval has margin of error that is a score multiplied by the estimated standard error. Besides substituting the t-score for the z-score, the t method also makes the assumption of a normal population distribution. In practice, the population distribution may not be close to normal. We discuss the importance of this assumption later in the section, where we'll find that this is mainly relevant for very small samples.

**Example
5.5**

Estimating Mean Weight Change for Anorexic Girls   This example comes from an experimental study that compared various treatments for young girls suffering from anorexia, an eating disorder. For each girl, weight was measured before and after a fixed period of treatment. The variable of interest was the change in weight, that is, weight at the end of the study minus weight at the beginning of the study. The change in weight was positive if the girl gained weight and negative if she lost weight. The treatments were designed to aid weight gain. The weight changes for 29 girls undergoing the cognitive behavioral treatment were[8]

$$1.7, 0.7, -0.1, -0.7, -3.5, 14.9, 3.5, 17.1, -7.6, 1.6,$$
$$11.7, 6.1, 1.1, -4.0, 20.9, -9.1, 2.1, 1.4, -0.3, -3.7,$$
$$-1.4, -0.8, 2.4, 12.6, 1.9, 3.9, 0.1, 15.4, -0.7.$$

Software used to analyze the data from a data file reports the summary results:

```
Variable |    Obs        Mean     Std. Dev.      Min        Max
  change |     29    3.006896     7.308504       -9.1       20.9
```

For the $n = 29$ girls who received this treatment, their mean weight change was $\bar{y} = 3.01$ pounds with a standard deviation of $s = 7.31$. The sample mean had an estimated standard error of $se = s/\sqrt{n} = 7.31/\sqrt{29} = 1.36$.
   Let μ denote the population mean change in weight for the cognitive behavioral treatment, for the population represented by this sample. If this treatment has a beneficial effect, then μ is positive. Since $n = 29$, $df = n - 1 = 28$. For a 95% confidence interval, we use $t_{.025} = 2.048$. The 95% confidence interval is

$$\bar{y} \pm t_{.025}(se) = 3.01 \pm 2.048(1.36) = 3.0 \pm 2.8, \text{ or } (0.2, 5.8).$$

It is simple to do this using software. For example, with R applied to a data file having a variable called *change* for the change in weight:

```
> t.test(change, conf.level=0.95)$conf.int
[1] 0.2268902 5.7869029
```

---

[8] Courtesy of Prof. Brian Everitt, King's College, London; data available in `Anorexia_CB` data file at text website.

With Stata, we apply the command `ci` to the variable name. If you have only summary statistics, Stata can construct the interval using them, with the `cii` command or using a dialog box, by entering n, $\bar{y}$, and s:

```
. cii means 29 3.007 7.309

 Variable |    Obs    Mean    Std. Err.    [95% Conf. Interval]
          |     29    3.007    1.357247    .2268051    5.787195
```

SPSS output is similar (see page 160). Calculators for t confidence intervals are also available online.[9]

With 95% confidence, we infer that this interval contains the population mean weight change. It appears that the mean weight change is positive, but it may be small in practical terms. However, this experimental study used a volunteer sample, because it is not possible to identify and randomly sample a population of anorexic girls. Because of this, inferences are tentative and "95% confidence" in the results may be overly optimistic. The results are more convincing if researchers can argue that the sample was representative of the population. The study did employ randomization in assigning girls to three therapies (only one of which is considered here), which is reassuring for analyses conducted later in the text that compare the therapies.

Another caveat about our conclusion is shown by Figure 5.8, a histogram that software shows for the data. This reveals that the sample data distribution is skewed to the right. The assumption of a normal population distribution may be violated—more about that below. The median weight change is only 1.4 pounds, somewhat less than the mean of 3.0 because of the skew to the right. The sample median is another indication that the size of the effect could be small. ∎

FIGURE 5.8: Histogram of Weight Change Values for Anorexia Study
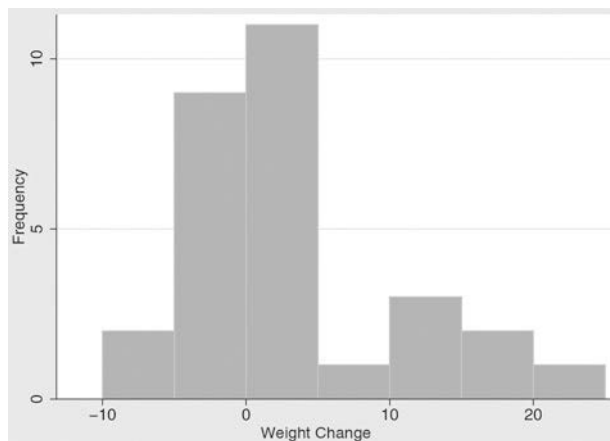
EFFECT OF CONFIDENCE LEVEL AND SAMPLE SIZE

We've used the t distribution to find a 95% confidence interval. Other confidence levels use the same formula but with a different t-score.

---

[9] Such as the *Inference for a Mean* applet at www.pearsonglobaleditions.com/Agresti.

To be safer in estimating the population mean weight change for the anorexia study in Example 5.5, we could instead use a 99% confidence interval. We then use the t-score with total probability 0.01 in the two tails, so 0.005 in each tail. Since $df = 28$ when $n = 29$, this t-score is $t_{.005} = 2.763$. The standard error does not change. The 99% confidence interval is

$$\bar{y} \pm 2.763(se) = 3.01 \pm 2.763(1.36), \quad \text{which is} \quad (-0.7, 6.8).$$

The confidence interval is wider than the 95% interval of (0.2, 5.8). This is the cost of having greater confidence. The 99% confidence interval contains 0. This tells us it is plausible, at the 99% confidence level, that the population mean change is 0, that is, that the therapy may not result in *any* change in the population mean weight.

Like the width of the confidence interval for a proportion, the width of a confidence interval for a mean also depends on the sample size n. Larger sample sizes result in narrower intervals.

ROBUSTNESS FOR VIOLATIONS OF NORMAL POPULATION ASSUMPTION

The assumptions for the confidence interval for a mean are (1) randomization for collecting the sample and (2) normal population distribution. Under the normality assumption, the sampling distribution of $\bar{y}$ is normal even for small n. Likewise, the z-score measuring the number of standard errors that $\bar{y}$ falls from μ then has the standard normal distribution. In practice, when we use the *estimated* standard error $se = s/\sqrt{n}$ (rather than the true one, $\sigma/\sqrt{n}$), the number of *se* that $\bar{y}$ falls from μ has the t distribution.

For the anorexia study, the sample data histogram in Figure 5.8 is not a precise indication of the population distribution because n is only 29, but it showed evidence of skew. Generally, the normal population assumption seems worrisome for social science application of this statistical method, because variables often have distributions that are far from normal.

A statistical method is said to be ***robust*** with respect to a particular assumption if it performs adequately even when that assumption is violated. Statisticians have shown that the confidence interval for a mean using the t distribution is robust against violations of the normal population assumption. Even if the population is not normal, confidence intervals based on the t distribution still work quite well, especially when n exceeds about 15. As the sample size gets larger, the normal population assumption becomes less important, because of the Central Limit Theorem. The sampling distribution of the sample mean is then bell shaped even when the population distribution is not. The actual probability that the 95% confidence interval method contains μ is close to 0.95 and gets closer as n increases.

An important case when the method does not work well is when the data are extremely skewed or contain extreme outliers. Partly this is because of the effect on the method, but also because the mean itself may not then be a representative summary of the center.

In practice, assumptions are rarely perfectly satisfied. Thus, it is important to know whether a statistical method is robust when a particular assumption is violated. The t confidence interval method is *not* robust to violations of the randomization assumption. Like all inferential statistical methods, the method has questionable validity if the method for producing the data did not use randomization.

STANDARD NORMAL IS THE *t* DISTRIBUTION WITH *df* = INFINITY

Look at a table of *t*-scores, such as Table 5.1 or Table B. As *df* increases, you move down the table. The *t*-score decreases and gets closer and closer to the *z*-score for a standard normal distribution. This reflects the *t* distribution becoming less spread out and more similar in appearance to the standard normal distribution as *df* increases. You can think of the standard normal distribution as a *t* distribution with $df = \infty$ (infinity).

For instance, when *df* increases from 1 to 100 in Table 5.1, the *t*-score $t_{.025}$ with right-tail probability equal to 0.025 decreases from 12.706 to 1.984. The *z*-score with right-tail probability of 0.025 for the standard normal distribution is $z = 1.96$. The *t*-scores are not printed for $df > 100$, but they are close to the *z*-scores. The last row of Table 5.1 and Table B lists the *z*-scores for various confidence levels, opposite $df = \infty$. As we showed, you can get *t*-scores for *any df* value using software, so you are not restricted to those in Table B.

Why does the *t* distribution look more like the standard normal distribution as *n* (and hence *df*) increases? Because *s* is increasingly precise as a point estimate of σ in approximating the true standard error $\sigma/\sqrt{n}$ by $se = s/\sqrt{n}$. The additional sampling error for small samples results in the *t* sampling distribution being more spread out than the standard normal.

The *t* distribution was discovered in 1908 by the statistician and chemist W. S. Gosset. At the time, Gosset was employed by Guinness Breweries in Dublin, Ireland, designing experiments pertaining to the selection, cultivation, and treatment of barley and hops for the brewing process. Due to company policy forbidding the publishing of trade secrets, Gosset used the pseudonym *Student* in articles he wrote about his discovery. The *t* distribution became known as *Student's t*, a name still sometimes used today. The method for constructing *t* confidence intervals for a mean was introduced 20 years after Gosset's discovery.

USING SOFTWARE FOR STATISTICAL METHODS

The examples in this section used output from statistical software to help us analyze the data. We'll show software output increasingly in future chapters as we cover methods that require substantial computation. You should use software yourself for some exercises and to get a feel for how researchers analyze data in practice.

When you start to use software for a given method, we suggest that you first use it for the example of that method in this book. Note whether you get the same results, as a way to check whether you are using the software correctly.

5.4 Choice of Sample Size

Polling organizations such as the Gallup Poll take samples that typically contain about a thousand subjects. This is large enough for a sample proportion estimate to have a margin of error of about 0.03. At first glance, it seems astonishing that a sample of this size from a population of perhaps many millions is adequate for predicting outcomes of elections, summarizing opinions on controversial issues, showing relative sizes of television audiences, and so forth.

Recall that the margin of error for a confidence interval depends on the *standard error* of the point estimate. Thus, the basis for this inferential power lies in the formulas for the standard errors. As long as a random sampling scheme is properly

executed, good estimates result from relatively small samples, no matter how large the population size.[10] Polling organizations use sampling methods that are more complex than simple random samples, often involving some clustering and/or stratification. However, the standard errors under their sampling plans are approximated reasonably well either by the formulas for simple random samples or by inflating those formulas by a certain factor (such as by 25%) to reflect the sample design effect.

Before data collection begins, most studies attempt to determine the sample size that will provide a certain degree of precision in estimation. A relevant measure is the value of *n* for which a confidence interval for the parameter has margin of error equal to some specified value. The key results for finding the sample size are as follows:

- The *margin of error* depends directly on the *standard error* of the sampling distribution of the point estimator.

- The *standard error* itself depends on the *sample size*.

DETERMINING SAMPLE SIZE FOR ESTIMATING PROPORTIONS

To determine the sample size, we must decide on the margin of error. Highly precise estimation is more important in some studies than in others. An exit poll in a close election requires a precise estimate to predict the winner. If, on the other hand, the goal is to estimate the proportion of U.S. citizens who do not have health insurance, a larger margin of error might be acceptable. So, we must first decide whether the margin of error should be about 0.03 (three percentage points), 0.05, or whatever.

We must also specify the *probability* with which the margin of error is achieved. For example, we might decide that the error in estimating a population proportion should not exceed 0.04, with 0.95 probability. This probability is the confidence level for the confidence interval.

Example 5.6

Sample Size for a Survey on Single-Parent Children   A social scientist wanted to estimate the proportion of school children in Boston who live in a single-parent family. Since her report was to be published, she wanted a reasonably precise estimate. However, her funding was limited, so she did not want to collect a larger sample than necessary. She decided to use a sample size such that, with probability 0.95, the error would not exceed 0.04. So, she needed to determine *n* such that a 95% confidence interval for π equals $\hat{\pi} \pm 0.04$.

Since the sampling distribution of the sample proportion $\hat{\pi}$ is approximately normal, $\hat{\pi}$ falls within 1.96 standard errors of π with probability 0.95. Thus, if the sample size is such that 1.96 standard errors equal 0.04, then with probability 0.95, $\hat{\pi}$ falls within 0.04 units of π. See Figure 5.9.

Recall that the true standard error is $\sigma_{\hat{\pi}} = \sqrt{\pi(1 - \pi)/n}$. How do we find the value of *n* that provides a value of $\sigma_{\hat{\pi}}$ for which $0.04 = 1.96\sigma_{\hat{\pi}}$? We must solve for *n* in the expression

$$0.04 = 1.96\sqrt{\frac{\pi(1 - \pi)}{n}}.$$

---

[10] In fact, the methods actually treat the population size as infinite; see Exercise 4.57 in Chapter 4.

FIGURE 5.9: Sampling Distribution of $\hat{\pi}$ with the Error of Estimation No Greater than 0.04, with Probability 0.95
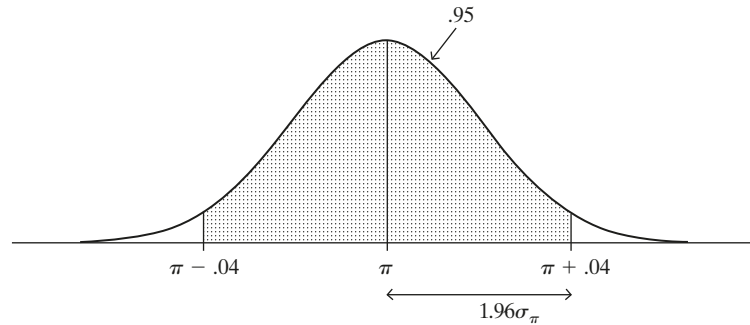
Multiplying both sides of the expression by $\sqrt{n}$ and dividing both sides by 0.04, we get

$$\sqrt{n} = \frac{1.96\sqrt{\pi(1-\pi)}}{0.04}.$$

Squaring both sides, we obtain the result

$$n = \frac{(1.96)^2\pi(1-\pi)}{(0.04)^2}.$$

Now, we face a problem. We want to select n for the purpose of estimating the population proportion π, but this formula requires the value of π. This is because the spread of the sampling distribution depends on π. The distribution is less spread out, and it is easier to estimate π, if π is close to 0 or 1 than if it is near 0.50. Since π is unknown, we must substitute an educated guess for it in this equation to solve for n.

The largest possible value for $\pi(1-\pi)$ occurs when $\pi = 0.50$. Then, $\pi(1-\pi) = 0.25$. In fact, $\pi(1-\pi)$ is fairly close to 0.25 unless π is quite far from 0.50. For example, $\pi(1-\pi) = 0.24$ when $\pi = 0.40$ or $\pi = 0.60$, and $\pi(1-\pi) = 0.21$ when $\pi = 0.70$ or $\pi = 0.30$. Thus, one approach merely substitutes 0.50 for π in the above equation for n. This yields

$$n = \frac{(1.96)^2\pi(1-\pi)}{(0.04)^2} = \frac{(1.96)^2(0.50)(0.50)}{(0.04)^2} = 600.$$

This approach ensures that with confidence level 0.95, the margin of error will not exceed 0.04, no matter what the value of π. ∎

Obtaining n by setting $\pi = 0.50$ is the "safe" approach. But this n value is excessively large if π is not near 0.50. Suppose that based on other studies the social scientist believed that π was no higher than 0.25. Then, an adequate sample size is

$$n = \frac{(1.96)^2\pi(1-\pi)}{(0.04)^2} = \frac{(1.96)^2(0.25)(0.75)}{(0.04)^2} = 450.$$

A sample size of 600 is larger than needed.

SAMPLE SIZE FORMULA FOR ESTIMATING PROPORTIONS

We next provide a general formula for determining the sample size. Let M denote the desired margin of error. The formula also uses a general z-score (in place of 1.96) determined by the probability with which the error is no greater than M.

Sample Size for Estimating a Population Proportion π

The random sample size n having margin of error M in estimating π by the sample proportion $\hat{\pi}$ is

$$n = \pi(1 - \pi)\left(\frac{z}{M}\right)^2.$$

The z-score is the one for the chosen confidence level, such as $z = 1.96$ for level 0.95. You need to guess π or take the safe approach of setting $\pi = 0.50$.

To illustrate, suppose the study about single-parent children wanted to estimate the population proportion to within 0.08 with confidence level 0.95. Then the margin of error is $M = 0.08$, and $z = 1.96$. The required sample size using the safe approach is

$$n = \pi(1 - \pi)\left(\frac{z}{M}\right)^2 = (0.50)(0.50)\left(\frac{1.96}{0.08}\right)^2 = 150.$$

This sample size of 150 is one-fourth the sample size of 600 necessary to guarantee a margin of error no greater than $M = 0.04$. Reducing the margin of error by a factor of one-half requires quadrupling the sample size. Calculators for this are also available online.[11]

DETERMINING SAMPLE SIZE FOR ESTIMATING MEANS

Next we find n for estimating a population mean μ. We determine how large n needs to be so that the sampling distribution of $\bar{y}$ has margin of error M. Figure 5.10 illustrates this.

FIGURE 5.10:
Determining n So that $\bar{y}$ Has Probability 0.95 of Falling within a Margin of Error of M Units of the Population Mean μ
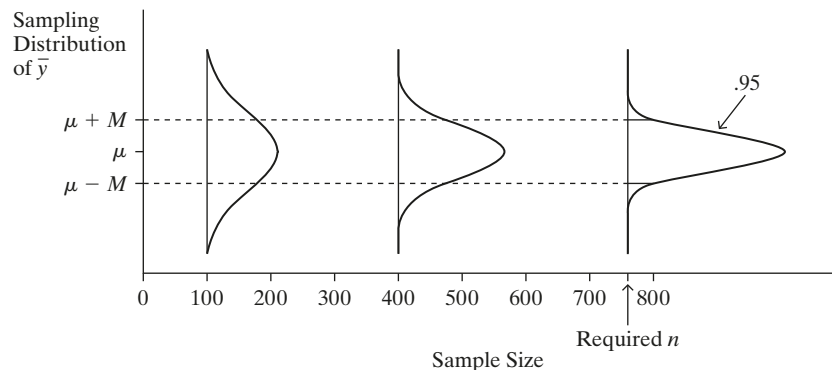
A derivation using the large-sample normal sampling distribution of $\bar{y}$ yields the following result:

Sample Size for Estimating a Population Mean μ

The random sample size n having margin of error M in estimating μ by the sample mean $\bar{y}$ is

$$n = \sigma^2\left(\frac{z}{M}\right)^2.$$

The z-score is the one for the chosen confidence level, such as $z = 1.96$ for level 0.95. You need to guess the population standard deviation σ.

---

[11] For example, at `epitools.ausvet.com.au/content.php?page=1Proportion`.

The greater the spread of the population distribution, as measured by its standard deviation σ, the larger the sample size needed to achieve a certain margin of error. If subjects have little variation (i.e., σ is small), we need less data than if they are highly heterogeneous. In practice, σ is unknown. We need to substitute an educated guess for it, perhaps based on a previous study.

A slight complication is that since we don't know σ, for inference we actually use the t distribution rather than the standard normal. But, if we don't know n, we also don't know the degrees of freedom and the t-score. We have seen, however, that unless df is small, the t-score is close to the z-score. So, we won't worry about this complication. The approximation of replacing an unknown t-score in the sample size formula by a z-score is usually much less than that in using an educated guess for σ. Calculators for the formula for determining n are also available online.[12]

Example 5.7

Estimating Mean Education of Native Americans   A study is planned of elderly Native Americans. Variables to be studied include educational level. How large a sample size is needed to estimate the mean number of years of attained education correct to within one year with probability 0.99?

If the study has no prior information about the standard deviation σ of educational attainment for Native Americans, we need to provide a guess. Perhaps nearly all educational attainment values fall within a range of 15 years, such as between 5 and 20 years. If the population distribution is approximately normal, then since the range from $\mu - 3\sigma$ to $\mu + 3\sigma$ contains nearly all of a normal distribution, the range of 15 equals about 6σ. Then, $15/6 = 2.5$ is a guess for σ.

Now, for 99% confidence, the error probability is 0.01. The z-score is 2.58, which has probability $0.01/2 = 0.005$ in each tail. Since the desired margin of error is $M = 1$ year, the required sample size is

$$ n = \sigma^2 \left(\frac{z}{M} \right)^2 = (2.5)^2 \left(\frac{2.58}{1} \right)^2 = 42. $$

A more cautious approach would select a larger value for σ. For example, if the range from 5 to 20 years encloses only about 95% of the education values, we could treat this as the range from $\mu - 2\sigma$ to $\mu + 2\sigma$ and set $15 = 4\sigma$. Then, $\sigma = 15/4 = 3.75$ and $n = (3.75)^2(2.58/1)^2 = 94$. ∎

OTHER CONSIDERATIONS IN DETERMINING SAMPLE SIZE

In summary, the necessary sample size depends on the desired *precision* for the margin of error, the *confidence level* for a confidence interval, and the *variability* in the population. For estimating means, the required sample size increases as σ increases. In most social surveys, large samples (1000 or more) are necessary, but for homogeneous populations (e.g., residents of nursing homes) smaller samples are often adequate, due to reduced population variability.

From a practical point of view, other considerations also affect the sample size. One consideration is the *complexity of analysis* planned. The more complex the analysis, such as the more variables analyzed simultaneously, the larger the sample needed. To analyze a single variable using a mean, a relatively small sample might be adequate. Planned comparisons of several groups using complex multivariate methods, however, require a larger sample. For instance, Example 5.7 showed we may be

---

[12] For example, at `http://epitools.ausvet.com.au/content.php?page=1Mean`.

able to estimate mean educational attainment quite well with only 42 people. But if we also wanted to compare the mean for several ethnic and racial groups and study how the mean depends on other variables such as gender, parents' income and education, and size of the community, we would need a much larger sample.

Another consideration concerns time, money, and other *resources*. Larger samples are more expensive and more time consuming. They may require greater resources than are available. For example, sample size formulas might suggest that 1000 cases provide the desired precision. Perhaps you can afford to gather only 400. Should you go ahead with the smaller sample and sacrifice precision and/or confidence, or should you give up unless you find additional resources? You may need to answer questions such as "Is it really crucial to study all groups, or can I reduce the sample by focusing on a couple of groups?"

The sample size formulas of this section apply to simple random sampling. Cluster samples and complex multistage samples must usually be larger to achieve the same precision, whereas stratified samples can often be smaller. In such cases, seek guidance from a statistical consultant.

In summary, no simple formula can always give an appropriate sample size. The needed sample size depends on resources and the analyses planned. This requires careful judgment. A final caveat: If the study is carried out poorly, or if data are never obtained for a substantial percentage of the sample, or if some observations are stated wrongly or incorrectly recorded by the data collector or by the statistical analyst, then the actual probability of accuracy to within the specified margin of error may be much less than intended. When someone claims to achieve a certain precision and confidence, be skeptical unless you know that the study was substantially free of such problems.

WHAT IF YOU HAVE ONLY A SMALL SAMPLE?*

Sometimes, because of financial or ethical reasons, it's just not possible to take as large a sample as we'd like. If n must be small, how does that affect the validity of confidence interval methods? The t methods for a mean can be used with any n. When n is small, though, you need to be cautious to look for extreme outliers or great departures from the normal population assumption, such as is implied by highly skewed data. These can affect the results and the validity of using the mean as a summary of center.

Recall that the confidence interval formula for a proportion requires at least 15 observations of each type. Otherwise, the sampling distribution of the sample proportion need not be close to normal, and the estimate $se = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$ of the true standard error $\sqrt{\pi(1-\pi)/n}$ may be poor. As a result, the confidence interval formula works poorly, as the next example shows.

Example 5.8 **What Proportion of Students Are Vegetarians?** For a class project, a student randomly sampled 20 fellow students at the University of Florida to estimate the proportion of undergraduate students at that university who were vegetarians. Of the 20 students she sampled, none were vegetarians. Let π denote the population proportion of vegetarians at the university. The sample proportion was $\hat{\pi} = 0/20 = 0.0$.

When $\hat{\pi} = 0.0$, then $se = \sqrt{\hat{\pi}(1-\hat{\pi})/n} = \sqrt{(0.0)(1.0)/20} = 0.0$. The 95% confidence interval for the population proportion of vegetarians is

$$\hat{\pi} \pm 1.96(se) = 0.0 \pm 1.96(0.0), \quad \text{which is} \quad 0.0 \pm 0.0, \quad \text{or} \quad (0.0, 0.0).$$

The student concluded she could be 95% confident that π falls between 0 and 0. But this confidence interval formula is valid only if the sample has at least 15 vegetarians

and at least 15 nonvegetarians. (Recall the guidelines in the box on page 124.) The sample did not have at least 15 vegetarians, so the method is not appropriate. ■

For small samples, the confidence interval formula is still valid if we use it after adding four artificial observations, two of each type. The sample of size $n = 20$ in Example 5.8 had 0 vegetarians and 20 nonvegetarians. We can apply the confidence interval formula with $0 + 2 = 2$ vegetarians and $20 + 2 = 22$ nonvegetarians. The value of the sample size for the formula is then $n = 24$. Applying the formula, we get

$$\hat{\pi} = 2/24 = 0.083, \; se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n} = \sqrt{(0.083)(0.917)/24} = 0.056.$$

The resulting 95% confidence interval is

$$\hat{\pi} \pm 1.96(se), \quad \text{which is} \quad 0.083 \pm 1.96(0.056), \quad \text{or} \quad (-0.03, 0.19).$$

A proportion cannot be negative, so we report the interval as (0.0, 0.19).

We can also find this interval using some software, or with Internet applets.[13] You can find it using Stata,[14] by applying the `cii` command to n and the count in the category of interest or using a dialog box:

```
. cii proportions 20 0, agresti

                                     -- Agresti-Coull ---
Variable |      Obs      Mean    Std. Err.    [95% Conf. Interval]
         |       20         0           0           0    .1898096
```

We can be 95% confident that the proportion of vegetarians at the University of Florida is no greater than 0.19.

Why do we add 2 to the counts of the two types? The reason is that the confidence interval then closely approximates one based on a more complex method (described in Exercise 5.77) that does not require estimating the standard error.

5.5 Estimation Methods: Maximum Likelihood and the Bootstrap*

We've focused on estimating means and proportions, but Chapter 3 showed that other statistics are also useful for describing data. These other statistics also have sampling distributions. In this section, we introduce a standard method, called *maximum likelihood*, that statisticians use to find good estimators of parameters. We also introduce a newer method, called the *bootstrap*, that uses modern computational power to find confidence intervals in cases in which it is difficult to derive the sampling distribution.

MAXIMUM LIKELIHOOD METHOD OF ESTIMATION

The most important contributions to modern statistical science were made by a British statistician and geneticist, R. A. Fisher (1890–1962). While working at an agricultural research station north of London, he developed much of the theory of point estimation as well as methodology for the design of experiments and data analysis.

---

[13] For example, with the *Inference for a Proportion* applet at www.pearsonglobaleditions.com/Agresti.
[14] Software calls it the *Agresti–Coull* confidence interval, because it was proposed in an article by A. Agresti and B. Coull, *American Statistician*, vol. 52 (1998), pp. 119–126.

For point estimation, Fisher proposed the ***maximum likelihood estimate***. This estimate is the value of the parameter that is most consistent with the observed data, in the following sense: If the parameter equaled that number (i.e., the value of the estimate), the observed data would have had greater chance of occurring than if the parameter equaled any other number.

We illustrate this method using data from a recent survey with a random sample of 1000 adult Americans, in which a sample proportion of 0.37 said that they believed in astrology. What is the maximum likelihood estimate of the population proportion who believe in astrology? Figure 5.11 plots the probability that a random sample of size 1000 has a sample proportion of 0.37, as a function of the actual population proportion believing in astrology. The probability changes dramatically as the population proportion changes. The curve, called a *likelihood function*, suggests that such a sample would be essentially impossible if the population proportion were below about 0.32 or above about 0.42. The maximum of the curve occurs at the population proportion value of 0.37. That is, the observed sample result would have been more likely to occur if the population proportion equaled 0.37 than if it equaled any other possible value between 0 and 1. So, the maximum likelihood estimate of the population proportion who believe in astrology is 0.37. In fact, with random sampling, the maximum likelihood estimate of a population proportion is necessarily the sample proportion.

FIGURE 5.11: The Probability that Exactly 37% of a Sample of Size 1000 Believe in Astrology, Plotted as a Function of the Population Proportion Believing in Astrology. The maximum probability occurs at the population proportion value of 0.37. This is the maximum likelihood estimate.
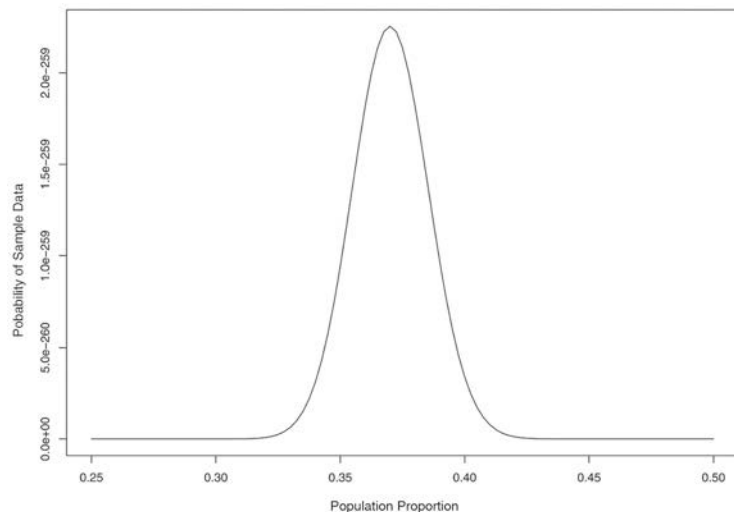
For many population distributions, such as the normal distribution, the maximum likelihood estimator of a population mean is the sample mean. The primary point estimates presented in this book are, under certain population assumptions, maximum likelihood estimates. Fisher showed that, for large samples, maximum likelihood estimators have three desirable properties:

- They are *efficient*, for relatively large samples: Other estimators do not have smaller standard errors.

- They are *consistent*, in the sense that as n increases they tend to get closer and closer to the unknown parameter value. In particular, they have little, if any, bias, with the bias diminishing to 0 as n increases.

- They have *approximately normal sampling distributions*.

Fisher also showed how to estimate standard errors for maximum likelihood estimators. Because their sampling distributions are approximately normal, confidence intervals for the parameters they estimate have the general form of taking the maximum likelihood estimate and then adding and subtracting a z-score multiplied by the estimated standard error. For instance, this is the method we used in Section 5.2 to find a confidence interval for a population proportion. To learn more about maximum likelihood, see Eliason (1993).

MAXIMUM LIKELIHOOD FOR MEAN, MEDIAN OF NORMAL DISTRIBUTION

When the population distribution is normal, the population mean and median are identical, because of the symmetry of the distribution. How should we estimate that common value, with the sample mean or the sample median? They are both point estimators of the same number. Fisher found that the maximum likelihood estimator is the sample mean, and that is preferred over the sample median.

In fact, for random samples, the standard error of the sample median equals $1.25\sigma/\sqrt{n}$. The sample median is not as efficient an estimator as the sample mean, because its standard error is 25% larger. When the population distribution is approximately normal, this is one reason the mean is more commonly used than the median in statistical inference.

When the population distribution is highly skewed, the population median is often a more useful summary than the population mean. We use the sample median to estimate the population median. However, the standard error formula $1.25\sigma/\sqrt{n}$ is valid only when the population distribution is approximately normal. We'll next learn about a general method that is useful for constructing confidence intervals even when we do not know the shape of the population distribution.

THE BOOTSTRAP

To use maximum likelihood, we need to make an assumption about the shape of the population distribution. But sometimes we do not have enough information to make a sensible assumption. In addition, some parameters do not have a confidence interval formula that works well regardless of the population distribution or sample size.

For such cases, a recent computational invention called the ***bootstrap*** is useful. This method treats the sample distribution as if it were the true population distribution and approximates by simulation the unknown sampling distribution. To do this, the method samples n observations, with replacement, from the sample distribution. That is, each of the original n data points has probability $1/n$ of selection for each "new" observation. This new sample of size n has its own point estimate of the parameter. The bootstrap method repeats this sampling process a large number of times, for instance, selecting 1000 separate samples of size n and 1000 corresponding point estimate values.

This type of empirically generated sampling distribution of the point estimate values provides information about the true parameter. For example, it generates a standard error for the point estimate we found with the actual data. This standard error is the sample standard deviation of the point estimate values from the simulations. It also generates a confidence interval for the parameter, for example, by the interval of values between the 2.5 and 97.5 percentiles of the simulated point estimate values. This is a computationally intensive process, but easily feasible with modern computing power.

**Example
5.9**

Estimating Median Shelf Time in a Library A librarian at the University of Florida wanted to estimate various characteristics of books in one of the university's special collections. Among the questions of interest were, "How old is a typical book in the collection?" and "How long has it been since a typical book has been checked out?" She suspected that the distributions of these variables were heavily skewed to the right, so she chose the median to describe the center.

Table 5.2 shows data (from the Library data file at the text website) on P = number of years since publication of book and C = number of years since book checked out, for a systematic random sample of 54 books from the collection. Figure 5.12 shows a box plot for the P values. The five starred values represent extreme outliers falling more than 3.0 (IQR) above the upper quartile. The sample median, which is 17, is more representative of the data than the sample mean of 22.6.

TABLE 5.2: Number of Years since Publication (P) and Number of Years since Checked Out (C) for 54 Books

C	P	C	P	C	P	C	P	C	P
1	3	9	9	4	4	1	18	1	5
30	30	0	17	2	7	0	12	1	13
7	19	5	5	47	47	3	15	9	17
11	140	2	19	5	8	2	10	11	18
1	5	1	22	1	11	5	19	2	3
2	97	0	10	1	21	7	7	4	19
4	4	11	11	5	20	14	14	5	43
2	19	10	10	10	10	0	18	10	17
4	13	17	71	8	19	0	17	48	48
2	19	11	11	6	6	7	20	4	4
92	92	4	44	1	5	1	54		

FIGURE 5.12: Box Plot for Number of Years since Publication for Sample of Library Books
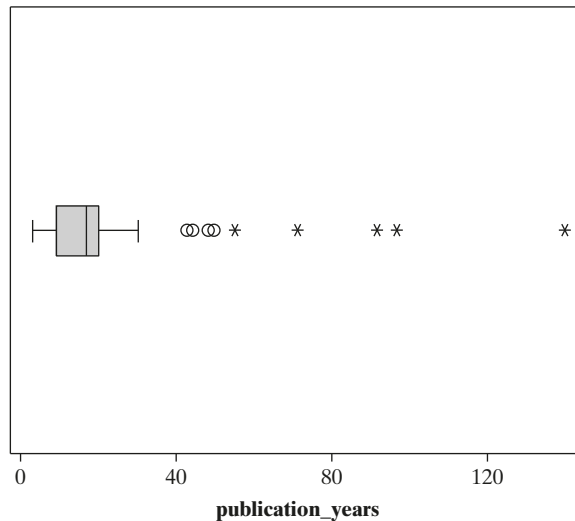
What is the standard error for this sample median estimate? There is no simple formula for this when we do not assume a shape for the population distribution. However, we can use the bootstrap to find one as well as a corresponding confidence

interval. The bootstrap is available on the Internet[15] and in software. For instance, in Stata software we find

```
. bootstrap r(p50), reps(10000): summarize P,  detail

          |   Observed   Bootstrap            Normal-based
          |      Coef.    Std. Err.       [95% Conf. Interval]
   _bs_1  |        17     2.114768        12.85513    21.14487
```

to produce 10,000 replications of a bootstrap for the median (labeled by Stata as `r(p50)` for the 50th percentile) of the variable P. The sample median of 17 has a bootstrap standard error of 2.11 and a 95% confidence interval for the population median of (12.9, 21.1). ∎

Likewise, there is not a simple formula for a confidence interval for a standard deviation unless we make rather stringent assumptions. For the library data set, in Stata we use 10,000 replications of a bootstrap for the standard deviation of the variable P:

```
. bootstrap r(sd), reps(10000): summarize P,  detail

          |   Observed   Bootstrap            Normal-based
          |      Coef.    Std. Err.       [95% Conf. Interval]
   _bs_1  |  25.91758    5.578261        14.98439    36.85077
```

The sample standard deviation of the time since publication of the book was 25.9 years, and a 95% bootstrap confidence interval for the population standard deviation is (15.0, 36.9).

5.6 Chapter Summary

This chapter presented methods of estimation, focusing on the population mean μ for quantitative variables and the population proportion π for categorical variables.

- A *point estimate* is the best single guess for the parameter value. The point estimates of the population mean μ, standard deviation σ, and proportion π are the sample values, $\bar{y}$, s, and $\hat{\pi}$.

- An *interval estimate*, called a *confidence interval*, is an interval of numbers within which the parameter is believed to fall.

Confidence Intervals

Confidence intervals for a population mean μ and for a population proportion π have the form

$$\text{Point estimate} \pm \text{Margin of error},$$
$$\text{with Margin of error} = \text{Score} \times (se),$$

where se is the estimated standard error.

---

[15] See the *Bootstrap* applet at `www.pearsonglobaleditions.com/Agresti`.

The true standard error, which is $\sigma/\sqrt{n}$, depends on the unknown population standard deviation σ. We estimate this and use it to get an *estimated* standard error, denoted by *se*. Table 5.3 shows the formula for *se* for estimating means and proportions. The score multiplied by *se* is a *z*-score from the normal distribution for confidence intervals for proportions and a *t*-score from the *t* distribution for confidence intervals for a mean. For the relatively large sample sizes of most social research, the *t*-score is essentially the same as the *z*-score.

- The probability that the method yields an interval that contains the parameter, called the **confidence level**, is controlled by the choice of the *z* or *t* score in the margin of error. Increasing the confidence level entails the use of a larger score and, hence, the sacrifice of a wider interval.

- The **t distribution** applies for statistical inference about a mean. It looks like the standard normal distribution, having a mean of 0 but being a bit more spread out. Its spread is determined by the **degrees of freedom**, which equal $n-1$ for inference about a mean.

- The width of a confidence interval also depends on the estimated standard error (*se*) of the sampling distribution of the point estimator. Larger sample sizes produce smaller *se* values and narrower confidence intervals and, hence, more precise estimates.

Confidence intervals assume random sampling. For large samples, they do not need an assumption about the population distribution, because the sampling distribution is approximately normal even if the population is highly nonnormal, by the Central Limit Theorem. Confidence intervals using the *t* distribution apply with any n but assume a normal population distribution, although the method is **robust** to violations of that assumption. Table 5.3 summarizes estimation methods.

TABLE 5.3: Summary of Estimation Methods for Means and Proportions, with Margin of Error M

Parameter	Point Estimate	Estimated Standard Error	Confidence Interval	Sample Size to Estimate to within M
Mean μ	$\bar{y}$	$se = \dfrac{s}{\sqrt{n}}$	$\bar{y} \pm t(se)$	$n = \sigma^2 \left(\dfrac{z}{M}\right)^2$
Proportion π	$\hat{\pi}$	$se = \sqrt{\dfrac{\hat{\pi}(1-\hat{\pi})}{n}}$	$\hat{\pi} \pm z(se)$	$n = \pi(1-\pi)\left(\dfrac{z}{M}\right)^2$

Note: For error probability α and confidence level $(1-\alpha)$, z-score or t-score has right-tail probability $\alpha/2$ (e.g., $\alpha/2 = 0.025$ for 95% confidence and $z = 1.96$).

Table 5.3 also shows formulas for the **sample size** needed to achieve a desired margin of error M. You must select M and the confidence level, which determines the z-score. Also, you must substitute a guess for the population standard deviation σ to determine the sample size for estimating a population mean μ. You must substitute a guess for the population proportion π to determine the sample size for estimating π. Substituting $\pi = 0.50$ guarantees that the sample size is large enough to give the desired precision and confidence.

The **maximum likelihood estimator** is an efficient estimator that has an approximately normal sampling distribution and is commonly used in statistical inference when we are willing to make an assumption about the shape of the population distribution. The **bootstrap** is a resampling method that can yield standard errors and

confidence intervals for measures, such as the median and the standard deviation, for which simple formulas are not available when we do not assume anything about the population distribution.

Exercises

Practicing the Basics

5.1. Of 288,503 people involved in motor vehicle accidents in Florida recently, 206,439 were wearing seat belts (Florida Department of Highway Safety and Motor Vehicles). Find a point estimate of the population proportion of Florida motorists wearing seat belts.

5.2. In response to a recent survey question about the number of hours daily spent playing outdoors, the responses by the seven subjects who identified themselves as Buddhists were 2, 3, 1, 1, 2, 3, 2.

(a) Find a point estimate of the population mean hours daily spent playing outdoors for Buddhists.

(b) The margin of error for this point estimate is 0.756. Explain what this represents.

5.3. In a survey by Wanderlust, a travel agency, it was estimated that of a region-wide random sample of 800 college graduates between the ages of 17 and 28, a proportion of 0.78 women like to go to a beach on their spring break. Find the standard error of this estimate, and interpret.

5.4. A survey of adults in Country A conducted in May 2017 by Alpha Research Center asked whether they were in favor of allowing laptops to school. Of the 400 conservatives sampled, 25% were in favor, whereas of the 250 liberals, 82% were in favor. Find the estimated standard errors for the sample proportions in favor. Interpret.

5.5. The World Health Survey in 2014–2016 asked if abortion was justifiable, on a scale from 1 (never) to 10 (always). The response *never* was given by 25.6% of the 3465 respondents in Country A and by 5.9% of the 1504 respondents in Country B. A report stated that the margin of error for Country A equals ±1.4%. Find the margin of error for Country B. (Although n is smaller for Country B, its margin of error is smaller, reflecting that standard errors of proportions diminish appreciably as proportions approach 0 or 1.)

5.6. One question in a recent General Survey asked, "Do you think that it should be government's responsibility to provide secondary level education to the poor?" Those answering *yes* included 85 of the 155 subjects who called themselves *strong supporters of Party A* in political party identification and 24 of the 110 who called themselves *strong supporters of Party B*.

(a) Find the point estimate of the population proportion who would answer *yes* for each group.

(b) The 95% confidence interval for the population proportion of *yes* responses is (0.47, 0.63) for strong supporters of Party A and (0.14, 0.29) for strong supporters of Party B. Explain how to interpret the intervals.

5.7. The General Survey asks whether you agree or disagree with the following statement: "It is much better for everyone involved if the girls in the family were not educated as much as the boys." The sample proportion agreeing was 0.78 in 1977 and 0.17 in 2016 ($n = 1845$).

(a) Show that the estimated standard error in 2016 was 0.008.

(b) Construct the 95% confidence interval for 2016, and interpret it.

5.8. A recent survey asked, "If the residents of an area don't vote in favor of an eatery, is it right that it should not be opened in that area?" Of 458 respondents, 355 said *yes* and 103 said *no*. Show that a 99% confidence interval for the population proportion who would say *yes* is (0.725, 0.825). Interpret.

5.9. When a 2016 research study asked whether there is solid evidence of global warming, 87% of liberals said *yes* whereas 25% of conservatives said *yes*. For conservatives, if $n = 400$, find and interpret a **(a)** 95% and **(b)** 99% confidence interval for a relevant parameter.

5.10. When a survey asked whether the government should impose strict laws to discourage people from smoking, a 95% confidence interval for the population proportion responding *yes* was (0.85, 0.89). Would a 99% confidence interval be wider, or shorter? Why?

5.11. State the z-score used in constructing a confidence interval for a proportion with confidence level **(a)** 0.98, **(b)** 0.90, **(c)** 0.50, and **(d)** 0.9973.

5.12. A 2015 survey asked respondents whether they favored or opposed the death penalty for people convicted of murder. Software shows results:

```
x        n        Sample prop        95.0% CI
900      1600       0.5625         (0.554, 0.572)
```

Here, x refers to the number of respondents who were in favor.

(a) Show how to obtain the value reported under "Sample prop."

(b) Can you conclude that more than half of all adults are in favor? Why?

(c) Find a 95% confidence interval for the proportion of adults who *opposed* the death penalty from the confidence interval shown for the proportion in favor.

5.13. The General Social Survey has asked respondents, "Do you think the use of marijuana should be made legal or not?" View results for all years at `sda.berkeley.edu/GSS` by entering the variables GRASS and YEAR.

(a) Of the respondents in 2014, what proportion said *legal* and what proportion said *not legal*?

(b) Is there enough evidence to conclude whether a majority or a minority of the population support legalization? Explain your reasoning.

(c) Describe any trend you see since about 1986 in the proportion favoring legalization.

5.14. When a survey asked whether reading novels enhanced their child's language, 65.4% of 1200 respondents answered that this was probably or definitely not true. Find a 99% confidence interval for the corresponding population proportion, and indicate whether you can conclude that a majority of respondents felt this way.

5.15. A 2016 report by the Center for Healthy Lifestyle provided a point estimate of 20.1% for the percentage of adults who regularly exercise. The sample size was 20,345. Assuming that this sample has the characteristics of a random sample, construct and interpret a 99% confidence interval for the population proportion of adults who regularly exercise. (*Note*: When n is very large, even confidence intervals with large confidence levels are narrow.)

5.16. Of the 1824 voters sampled in the exit poll discussed in the previous chapter (page 92), 60.5% said they voted for Jerry Brown. Is there enough evidence to predict the winner of the election? Base your decision on a 95% confidence interval, stating needed assumptions for that decision.

5.17. For an exit poll of people who voted in a regional election, 45% voted for Candidate A and 55% for Candidate B. Assuming this is a random sample of all voters, construct a 99% confidence interval for the proportion of votes that Candidate A received, if the sample size was **(a)** 500 and **(b)** 50. In each case, indicate whether you would be willing to predict the winner. Explain how and why the sample size affects the inference.

5.18. In 2016, a poll reported results of a survey about religious beliefs. Of 3250 adults surveyed, 24% believed in reincarnation. Treating this as a random sample, a 95% confidence interval for the population proportion believing in reincarnation is (0.225, 0.255). Without doing any new calculation, explain how the interval would change if the sample size had been only a fourth as large, $n = 813$.

5.19. Report the t-score that multiplies by the standard error to form a

(a) 95% confidence interval for μ with 15 observations.

(b) 95% confidence interval for μ with 25 observations.

(c) 95% confidence interval for μ with $df = 25$.

(d) 99% confidence interval for μ with $df = 25$.

5.20. Find and interpret the 95% confidence interval for μ, if $\bar{y} = 70$ and $s = 10$, based on a sample size of **(a)** 5, **(b)** 20.

5.21. A survey in 2015 asked respondents how many ice creams they have had since last week. For the 200 children in the sample between the ages of 13 and 16, the median $= 4$ and mode $= 2$ (16.8% of the sample). Software summarizes other results:

```
-------------------------------------------------
Variable      n    Mean  StDev  SE Mean    95.0% CI
NUMCHILDREN  200  10.25  15.68    1.11   (8.07, 12.43)
-------------------------------------------------
```

(a) Show how software got the standard error reported, and interpret.

(b) Interpret the reported confidence interval.

(c) State a statistical factor that might make you skeptical about the usefulness of this confidence interval.

5.22. A survey asked, "What do you think is the ideal number of hours a child should study?" The 528 female students who responded had a median of 5, mean of 6.02, and standard deviation of 2.75.

(a) Find and interpret the standard error of the sample mean.

(b) The 95% confidence interval is (5.79, 6.25). Interpret.

(c) Is it plausible that the population mean $= 5$? Explain.

5.23. Refer to the previous exercise. For the 450 male students in the sample, the mean was 6.80 and the standard deviation was 1.24.

(a) Show that the standard error of the sample mean is 0.058.

(b) Show that the 95% confidence interval for the population mean is (6.69, 6.91), and explain what "95% confidence" means.

5.24. Example 5.5 (page 129) analyzed data from a study that compared therapies for anorexia. For the 17 girls who received the family therapy, the changes in weight during the study (which are in the data file `Anorexia` at the text website) were

11, 11, 6, 9, 14, −3, 0, 7, 22, −5, −4, 13, 13, 9, 4, 6, 11.

(a) Verify that $\bar{y} = 7.29$, $s = 7.18$, and $se = 1.74$.

(b) To use the t distribution, explain why $df = 16$ and a 95% confidence interval uses the t-score of 2.120.

(c) Verify that the 95% confidence interval for the population mean change in weight μ for this therapy is (3.6, 11.0). Interpret.

5.25. A 2014 survey asked, "On an average, about how many hours a day do you watch television?" Stata software reports:

```
--------------------------------------------------
Mean estimation              Number of obs = 1500
           Mean   Std. Err.  [95.0% Conf. Interval]
TVHOURS    3.56   0.085         (3.55, 3.57)
--------------------------------------------------
```

What's wrong with the interpretation "In the population, 95% of the time, subjects watched between 3.55 and 3.57 hours of TV a day?" State the correct interpretation.

5.26. In response to a survey question in 2016 about the number of hours daily spent jogging, the responses by the seven subjects were 1, 1, 1, 2, 2, 3, 4.

(a) Estimate the mean, standard deviation, and standard error.

(b) Construct a 95% confidence interval for the population mean, specifying its assumptions. Interpret.

5.27. A survey asked subjects, "How long have you lived in the city, town, or community where you live now?" The responses of the 1600 subjects had a median of 14 years, a mean of 19.8, and a standard deviation of 17.2.

(a) Do you think that the population distribution is normal? Why or why not?

(b) Based on your answer in (a), can you construct a 99% confidence interval for the population mean? If not, explain why not. If so, do so and interpret.

5.28. A recent survey asked, "How many times in the past 8 months have you fallen sick?" The 900 women who responded had a median of 2, mean of 2.71, and standard deviation of 1.43. The 683 men who responded had a median of 1, mean of 1.24, and standard deviation of 1.34.

(a) Find a 95% confidence interval for the population mean for women. Interpret.

(b) Explain why the $\bar{y}$ and s values suggest that this variable does not have a normal distribution. Does this cause a problem with the confidence interval method in (a)? Explain.

5.29. A survey asked respondents how many jobs they had changed in the last 5 years. Software reports:

```
--------------------------------------------------
Variable   N    Mean    StDev   SE Mean    95.0% CI
Jobs     3000   0.885   0.825   0.015   (0.856, 0.914)
--------------------------------------------------
```

(a) Interpret the confidence interval reported.

(b) Based on these results, explain why the distribution was probably skewed to the right. Explain why the skew need not cause a problem with the validity of the confidence interval, unless there are extreme outliers.

(c) Upon closer look at the data file, we see that of the eight available responses (0, 1, 2, · · · · · ·, 7), 5 stands for 5–10 jobs, 6 stands for 11–20 jobs, and 7 stands for 21–100 jobs. If we instead had

the actual numbers of jobs, would the mean and standard deviation be larger, or smaller? Why? What would the impact be on the confidence interval?

5.30. For the `Students` data file mentioned in Exercise 1.11, software reports the results for responses on the number of times a week the subject reads a newspaper:

```
--------------------------------------------------
Variable   N    Mean   Std Dev   SE Mean    95.0% CI
News      60    4.1     3.0       0.387   (3.32, 4.88)
--------------------------------------------------
```

(a) Interpret the confidence interval shown.

(b) Does it seem plausible that the population distribution of this variable is normal? Why? Explain the implications of the term *robust* regarding the normality assumption for this analysis.

5.31. The General Social Survey asks respondents to rate their political views on a seven-point scale, where 1 = extremely liberal, 4 = moderate, and 7 = extremely conservative. A researcher analyzing data from the 2014 GSS gets software output:

```
--------------------------------------------------
Mean estimation              Number of obs = 2500
           Mean   Std. Err.  [99.0% Conf. Interval]
PolViews   4.20   0.030         (4.123, 4.277)
--------------------------------------------------
```

(a) Show how to construct the confidence interval from the other information provided.

(b) Would the confidence interval be wider, or narrower, (i) if you constructed a 95% confidence interval? (ii) if you found the 99% confidence interval only for the 411 respondents who called themselves *strong Democrat* on political party identification (PARTYID), for whom the mean was 3.150 with standard deviation 1.496?

(c) What assumption are you making about the scale of measurement for political ideology when you use the sample mean and standard deviation?

5.32. At `sda.berkeley.edu/GSS`, consider responses to the question "On how many days in the past 7 days have you felt lonely" (coded LONELY) for the most recent survey in which this was asked.

(a) Find a point estimate of the population mean.

(b) Construct the 95% confidence interval, and interpret.

5.33. A study estimates the mean annual family income for families living in public housing in Chicago. For a random sample of 30 families, the annual incomes (in hundreds of dollars) are in the `Chicago` data file at the text website and here:

> 133 140 127 150 133 114 128 142 123 172
> 146 110 135 136 158 120 189 106 144 134
> 161 143 170 120 142 150 174 109 162 129

(a) Based on a descriptive graphic, what do you predict about the shape of the population distribution?

(b) Find and interpret point estimates of μ and σ, the population mean and standard deviation.

(c) Construct and interpret a 95% confidence interval for μ.

5.34. A hospital administrator wants to estimate the mean number of admissions of patients in that hospital. Based on a systematic random sample of 100 records of patients for the previous year, she reports that "the sample mean was 7.4. In repeated random samples of this size, the sample mean could be expected to fall within 1.2 of the true mean about 95% of the time."

(a) Construct and interpret a 95% confidence interval for the mean.

(b) The administrator decides that this interval is too wide, and she prefers one of only half this width for a new study this year. How large a sample size does she need?

5.35. To estimate the proportion of traffic deaths in a region last year that were due to overspeeding, determine the necessary sample size for the estimate to be accurate to within 0.05 with probability 0.90. Based on results of studies reported by the region's Traffic Safety Administration, we expect the proportion to be about 0.25.

5.36. A television network plans to predict the outcome of an election between Candidate A and Candidate B. They will do this with an exit poll on election day. They decide to use a random sample size for which the margin of error is 0.05 for 95% confidence intervals for population proportions.

(a) What sample size should they use?

(b) If the pollsters think that the election will be close, they might use a margin of error of 0.025. How large should the sample size be?

5.37. A public health unit wants to sample death records for the past year in South Asia to estimate the proportion of deaths that were due to accidents. They want the estimate to be accurate to within 0.03 with probability 0.95.

(a) Find the necessary sample size if, based on the data published, they believe that this proportion does not exceed 0.10.

(b) Suppose that in determining n, they use the safe approach that sets $\pi = 0.50$ in the appropriate formula. Then how many records need to be sampled? Compare the result to the answer in part (a), and note the reduction in sample size that occurs by making an educated guess for π.

5.38. A 2015 poll in Country X indicated that 48% of its residents favored that all citizens who do not vote should pay a fine. A report by related polls did not report the sample size but stated, "Polls of this size are considered to be accurate within 2.5 percentage points 95% of the time." If this is the case, about how large was the sample size?

5.39. In 2006, a survey reported that the percentage of people in a university who reported *a lot of*, or *some* confidence (instead of *not too much*, or *no* confidence) in the elected Student Body President was 15% in House Red,

29% in House Blue, 26% in House Green, and 5% in House Yellow. In 2014, the corresponding percentages reported for the new president were 85%, 71%, 73%, and 56%. The reported margin of error in House Yellow was 1.5% for the 2006 result and 3.0% for the 2014 result. Find the approximate sample size in House Yellow for the studies.

5.40. An estimate is needed of the mean acreage of farms in Region X. The estimate should be correct to within 100 acres with probability 0.95. A preliminary study suggests that 400 acres is a reasonable guess for the standard deviation of farm size.

(a) How large a sample of farms is required?

(b) A random sample is selected of the size found in (a). The sample has a standard deviation of 800 acres, rather than 400. What is the margin of error for a 95% confidence interval for the mean farm size?

5.41. A social scientist plans a study of adult South Asians to investigate educational attainment in women. In some parts, the women are not required to attend school, so some of them had very little education. How large a sample size is needed so that a 95% confidence interval for the mean number of years of education completed has margin of error equal to 1 year? There is no information about the standard deviation of educational attainment, but researchers expect that nearly all values fall between 0 and 18 years.

5.42. How large a sample size is needed to estimate the mean annual income of natives of Country X correct to within $1,000 with probability 0.99? There is no prior information about the standard deviation of annual income of natives, but we guess that about 95% of their incomes are between $6,000 and $66,000 and that this distribution of incomes is skewed but approximately bell shaped.

5.43. An anthropologist wants to estimate the proportion of children in a tribe who die before reaching adulthood. For families she knew who had children born between 1990 and 1995, 4 of 32 children died before reaching adulthood. Can you use the ordinary large sample formula to construct a 95% confidence interval for the population proportion? Why or why not? Construct an appropriate confidence interval, and interpret.

5.44. You randomly sample five students at your school to estimate the proportion of students who like broccoli. None of the five students say they like it.

(a) Find the sample proportion who like it and its standard error. Does the usual interpretation of *se* make sense?

(b) Why is it not appropriate to use the ordinary confidence interval formula (from page 124) for these data? Use a more appropriate approach, and interpret.

5.45. Refer to Exercise 5.33. Use the bootstrap to construct a 95% confidence interval for the median annual income of the public housing residents. Interpret.

5.46. Refer to Example 5.9 (page 141). Construct a 95% confidence interval for the median time since a book was last checked out. Interpret.

Concepts and Applications

5.47. Use the *Explore Coverage* applet at `www.pearsonglobaleditions.com/Agresti` to repeatedly generate random samples and construct confidence intervals for a proportion, to illustrate their behavior when used for many samples. Set the population proportion value (labeled as *p*) to 0.50, the sample size to 200, the confidence level to 90%, and the number of samples to 10.

(a) Click on *Draw Sample*. In your simulation, what percentage of the ten 90% confidence intervals generated actually contained the parameter value? How many would be expected to contain the parameter?

(b) To get a feel for what happens "in the long run," reset and draw 1000 samples. What percentage actually contained the true parameter value? Copy results, and interpret.

5.48. Refer to the previous exercise. Using this applet, let's check that the confidence interval for a proportion may work poorly with small samples. Set the population proportion $\pi = 0.10$, with $n = 10$. Draw 100 random samples, each of size 10, forming 95% confidence intervals for π for each one.

(a) How many intervals failed to contain the true value, $\pi = 0.10$? How many would you expect not to contain the true value? What does this suggest? (Notice that many of the intervals contain only the value 0.0, which happens when $\hat{\pi} = 0.0$.)

(b) To see that this is not a fluke, reset and draw 1000 confidence intervals. What percentage contain $\pi = 0.10$? (*Note*: For every interval formed, the number of "successes" is smaller than 15, so the large-sample formula is not adequate.)

(c) Using the *Sampling Distribution of the Sample Proportion* applet at `www.pearsonglobaleditions.com/Agresti`, select a population proportion of 0.10. Draw 10,000 random samples of size 10 each. Look at the empirical sampling distribution of the sample proportion values. Is it bell shaped and symmetric? Use this to help explain why the confidence interval performs poorly in this case.

5.49. Refer to the `Students` data file (Exercise 1.11 on page 21). Using software, construct and interpret a 95% confidence interval for **(a)** the mean weekly number of hours spent watching TV, **(b)** the proportion believing in life after death. Interpret.

5.50. Refer to the data file created in Exercise 1.12 (page 22). For variables chosen by your instructor, pose a research question, and conduct inferential statistical analyses using basic estimation methods. Summarize and interpret your findings, and explain how you could use them to answer the research question.

5.51. In 2016, a survey asked about the number of hours a week spent on the social networking sites, excluding e-mail (variable denoted WWWHR). State a research question you could address about this response variable and a relevant explanatory variable. Go to `sda.berkeley.edu/GSS` and analyze the data. Prepare a short report summarizing your analysis and answering the question you posed.

5.52. A recent survey asked married respondents, "Did you live with your husband/wife before you got married?" The responses were 60 *yes* and 118 *no* for those who called themselves politically liberal, and 50 *yes* and 245 *no* for those who called themselves politically conservative. Analyze these data, identifying the response variable and explanatory variable. Summarize your analysis in a report of no more than 200 words.

5.53. When parents in a recent survey were asked whether they agreed with the following statements, the (yes, no) counts under various conditions were as follows:

- Homework should be banned: (250, 1545).
- Single-sex schools are better for students: (640, 1189).
- Teachers should not be allowed to contact students through social media: (756, 1150).

Analyze these data. Prepare a one-page report stating the assumptions, showing results of description and inference, and summarizing conclusions.

5.54. The observations on daily TV watching for the 15 subjects in a 2016 survey who were obese were 0, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 8. A 95% confidence interval for the mean of the corresponding population is (2.84, 4.76). Suppose the observation of 8 was incorrectly recorded as 80. What would have been obtained for the 95% confidence interval? Compare to the interval (2.84, 4.76). How does this warn you about potential effects of outliers on confidence intervals for means?

5.55. (a) Explain what it means for an estimator to be unbiased.

(b) Explain why the sample range is a biased estimator of the population range.

5.56. What is the purpose of forming a confidence interval for a parameter? What can you learn from it that you could not learn from a point estimate?

5.57. An interval estimate for a mean is more informative than a point estimate, because with an interval estimate you can figure out the point estimate, but with the point estimate alone you have no idea about the width of the interval estimate.

(a) Explain why this statement is correct, illustrating using the reported 95% confidence interval of (4.0, 5.6) for the mean number of dates in the previous month for women at a particular college.

(b) The confidence interval in (a) used a sample size of 50. What were the sample mean and standard deviation?

5.58. Explain why confidence intervals are wider with **(a)** larger confidence levels, **(b)** smaller sample sizes.

5.59. Why would it be unusual to see a **(a)** 99.9999%, **(b)** 25% confidence interval?

5.60. Give an example of a study in which it would be important to have

(a) A high degree of confidence.

(b) A high degree of precision.

5.61. How does population heterogeneity affect the sample size required to estimate a population mean? Illustrate with an example.

5.62. Explain the reasoning behind the following statement: Studies about more diverse populations require larger sample sizes. Illustrate for the problem of estimating mean income for all medical doctors in a country compared to estimating mean income for all entry-level employees at a restaurant in the same country.

5.63. You would like to find the proportion of bills passed by a ruling party in a certain country that were vetoed by the president in the last five years. After checking the records, you see that for the population of all 44 bills passed, 2 were vetoed. Does it make sense to construct a confidence interval using these data? Explain. (*Hint*: Identify the sample and population.)

5.64. The publication *Attitudes towards European Union Enlargement* from Eurobarometer states, "The readers are reminded that survey results are *estimations*, the accuracy of which rests upon the sample size and upon the observed percentage. With samples of about 1000 interviews, the real percentages vary within the following confidence limits:"

Observed limits	10% or 90% ± 1.9	20%, 80% ± 2.5	30%, 70% ± 2.7	40%, 60% ± 3.0	50% ± 3.1

(a) Explain how they got 3.0 points for 40% or 60%.

(b) Explain why the margin of error differs for different observed percentages.

(c) Explain why the accuracy is the same for a particular percentage and for 100 minus that value (e.g., both 40% and 60%).

(d) Explain why it is more difficult to estimate a population proportion when it is near 0.50 than when it is near 0 or 1.

5.65. To use the large-sample confidence interval for a proportion, you need at least 15 outcomes of each type. Show that the smallest value of n for which the method can be used is **(a)** 30 when $\hat{\pi} = 0.50$, **(b)** 50 when $\hat{\pi} = 0.30$, **(c)** 150 when $\hat{\pi} = 0.10$. That is, the overall n must increase as $\hat{\pi}$ moves toward 0 or 1. (When the true proportion is near 0 or 1, the sampling distribution can be highly skewed unless n is quite large.)

Select the best response in Exercises 5.66–5.69.

5.66. The reason we use a z-score from a normal distribution in constructing a confidence interval for a proportion is that

(a) For large random samples, the sampling distribution of the sample proportion is approximately normal.

(b) The population distribution is normal.

(c) For large random samples, the sample data distribution is approximately normal.

(d) If in doubt about the population distribution, it's safest to assume that it is the normal distribution.

5.67. Increasing the confidence level causes the width of a confidence interval to **(a)** increase, **(b)** decrease, **(c)** stay the same.

5.68. Other things being equal, quadrupling the sample size causes the width of a confidence interval to **(a)** double, **(b)** halve, **(c)** be one quarter as wide, **(d)** stay the same.

5.69. Based on responses of 1467 subjects in General Social Surveys, a 95% confidence interval for the mean number of close friends equals (6.8, 8.0). Which of the following interpretations is (are) correct?

(a) We can be 95% confident that $\bar{y}$ is between 6.8 and 8.0.

(b) We can be 95% confident that μ is between 6.8 and 8.0.

(c) Ninety-five percent of the values of y = number of close friends (for this sample) are between 6.8 and 8.0.

(d) If random samples of size 1467 were repeatedly selected, then 95% of the time $\bar{y}$ would fall between 6.8 and 8.0.

(e) If random samples of size 1467 were repeatedly selected, then in the long run 95% of the confidence intervals formed would contain the true value of μ.

5.70. A random sample of 100 records yields a 95% confidence interval for the mean age at first marriage of women in a certain country of 20.5 to 24.0 years. Explain what is wrong with each of the following interpretations of this interval.

(a) If random samples of 100 records were repeatedly selected, then 95% of the time the sample mean age at first marriage for women would be between 20.5 and 24.0 years.

(b) Ninety-five percent of the ages at first marriage for women in the county are between 20.5 and 24.0 years.

(c) We can be 95% confident that $\bar{y}$ is between 20.5 and 24.0 years.

(d) If we repeatedly sampled the entire population, then 95% of the time the population mean would be between 20.5 and 24.0 years.

5.71. Refer to the previous exercise. Provide the proper interpretation.

5.72.* For a random sample of n subjects, explain why it is about 95% likely that the sample proportion has error no more than $1/\sqrt{n}$ in estimating the population proportion. (*Hint*: To show this "$1/\sqrt{n}$ rule," find two standard errors when $\pi = 0.50$, and explain how this compares to two standard errors at other values of π.) Using this result, show that $n = 1/M^2$ is a safe sample size for estimating a proportion to within M with 95% confidence.

5.73.* You know the sample mean $\bar{y}$ of n observations. Once you know $(n-1)$ of the observations, show that you can find the remaining one. In other words, for a given value of $\bar{y}$, the values of $(n-1)$ observations determine the remaining one. In summarizing scores on a quantitative variable, having $(n-1)$ degrees of freedom means that only that many observations are independent.

5.74.* Find the true standard error of the sample proportion when $\pi = 0$ or $\pi = 1$. What does this reflect?

5.75.* Let π be the probability that a randomly selected voter prefers the Republican candidate. You sample two people, and neither prefers the Republican. Find the point estimate of π. Does this estimate seem sensible? Why? (The *Bayesian* estimator is an alternative one that uses a *subjective* approach, combining the sample data with your prior beliefs about π before seeing the data. For example, if you believed π was equally likely to fall anywhere from 0 to 1, the Bayesian estimate adds two observations, one of each type, thus yielding the estimate 1/4.)

5.76.* To encourage subjects to make responses on sensitive questions, the method of **randomized response** is often used. The subject is asked to flip a coin, in secret. If it is a head, the subject tosses the coin once more and reports the outcome, head or tails. If, instead, the first flip is a tail, the subject reports instead the response to the sensitive question, for instance, reporting the response *head* if the true response is *yes* and reporting the response *tail* if the true response is *no*. Let π denote the true probability of the *yes* response on the sensitive question.

(a) Explain why the numbers in Table 5.4 are the probabilities of the four possible outcomes.

(b) Let p denote the sample proportion of subjects who report *head* for the second response. Explain why $\hat{\pi} = 2p - 0.5$ estimates π.

(c) Using this approach, 200 subjects are asked whether they have ever knowingly cheated on their income tax.

Report the estimate of π if the number of reported heads equals (i) 50, (ii) 70, (iii) 100, (iv) 150.

TABLE 5.4

First Coin	Second Response	
	Head	Tail
Head	0.25	0.25
Tail	$\pi/2$	$(1-\pi)/2$

5.77.* To construct a confidence interval for a proportion π, it is not necessary to substitute $\hat{\pi}$ for the unknown value of π in the formula for the true standard error of $\hat{\pi}$. A less approximate method (called the *score* confidence interval for a proportion) finds the endpoints for a 95% interval by determining the π values that are 1.96 standard errors from the sample proportion, by solving for π in the equation

$$|\hat{\pi} - \pi| = 1.96\sqrt{\frac{\pi(1-\pi)}{n}}.$$

For Example 5.8 (page 137) with no vegetarians in a sample of size 20, substitute $\hat{\pi}$ and n in this equation and show the equation is satisfied at $\pi = 0$ and at $\pi = 0.161$. So, the 95% confidence interval is (0, 0.161), compared to (0, 0) based on the ordinary formula.

5.78. Refer to the previous exercise. At the *Inference for a Proportion* applet at `www.pearsonglobal editions.com/Agresti`, enter 0 successes for $n = 20$. Click on *Confidence Interval* for type of inference. Why are the reported standard error and confidence interval not sensible? Clicking on the *Agresti–Coull adjustment* (introduced on page 138) gives a sensible alternative to the score confidence interval.

5.79.* This exercise presents a confidence interval for the population median that requires no assumption about the population distribution other than it is essentially continuous.

(a) Explain why for a random sample of size n the sample proportion $\hat{\pi}$ falling below the median has expected value 0.50 and standard error $\sigma_{\hat{\pi}} = 0.50/\sqrt{n}$, and so the probability is about 0.95 that the *number* of observations falling below (above) the median is within $n(1/\sqrt{n}) = \sqrt{n}$ of half the sample.

(b) For the ordered sample of size n, explain why a 95% confidence interval for the median goes from the observation that has the index $(n+1)/2 - \sqrt{n}$ to the observation with the index $(n+1)/2 + \sqrt{n}$. For Example 5.9 on median shelf life in a library, show that this interval is 11 to 19 years.

Chapter 6

STATISTICAL INFERENCE: SIGNIFICANCE TESTS

An aim of many studies is to check whether the data agree with certain predictions. The predictions, which often result from the theory that drives the research, are *hypotheses* about the study population.

Hypothesis

> In statistics, a **hypothesis** is a statement about a population. It takes the form of a prediction that a parameter takes a particular numerical value or falls in a certain range of values.

Examples of hypotheses are the following: "For restaurant managerial employees, the mean salary is the same for women and for men"; "There is no difference between Democrats and Republicans in the probabilities that they vote with their party leadership"; and "A majority of adult Canadians are satisfied with their national health service."

A statistical **significance test** uses data to summarize the evidence about a hypothesis. It does this by comparing point estimates of parameters to the values predicted by the hypothesis. The following example illustrates concepts behind significance tests.

Example 6.1

Testing for Gender Bias in Selecting Managers   A large supermarket chain in Florida periodically selects employees to receive management training. A group of women employees recently claimed that the company selects males at a disproportionally high rate for such training. The company denied this claim. In past years, similar claims of gender bias have been made about promotions and pay for women who work for various companies.[1] How could the women employees statistically back up their assertion?

Suppose the employee pool for potential selection for management training is half male and half female. Then, the company's claim of a lack of gender bias is a hypothesis. It states that, other things being equal, at each choice the probability of selecting a female equals 1/2 and the probability of selecting a male equals 1/2. If the employees truly are selected for management training randomly in terms of gender, about half the employees picked should be females and about half should be male. The women's claim is an alternative hypothesis that the probability of selecting a male exceeds 1/2.

Suppose that 9 of the 10 employees chosen for management training were male. We might be inclined to believe the women's claim. However, we should analyze whether these results would be unlikely if there were *no* gender bias. Would it be highly unusual that 9/10 of the employees chosen would have the same gender if they were truly selected at random from the employee pool?

---

[1] For example, Wal-Mart, see `http://now.org/blog/walmart-and-sex-discrimination`.

Due to sampling variation, not exactly 1/2 of the sample need be male. How far above 1/2 must the sample proportion of males chosen be before we believe the women's claim? ∎

This chapter introduces statistical methods for summarizing evidence and making decisions about hypotheses. We first present the parts that all significance tests have in common. The rest of the chapter presents significance tests about population means and population proportions. We'll also learn how to find and how to control the probability of an incorrect decision about a hypothesis.

6.1 The Five Parts of a Significance Test

Now let's take a closer look at the significance test method, also called a *hypothesis test*, or *test* for short. All tests have five parts:

Assumptions, Hypotheses, Test statistic, P-value, Conclusion.

ASSUMPTIONS

Each test makes certain assumptions or has certain conditions for the test to be valid. These pertain to

- *Type of data*: Like other statistical methods, each test applies for either quantitative data or categorical data.

- *Randomization*: Like other methods of statistical inference, a test assumes that the data gathering employed randomization, such as a random sample.

- *Population distribution*: Some tests assume that the variable has a particular probability distribution, such as the normal distribution.

- *Sample size*: Many tests employ an approximate normal or t sampling distribution. The approximation is adequate for any n when the population distribution is approximately normal, but it also holds for highly nonnormal populations when the sample size is relatively large, by the Central Limit Theorem.

HYPOTHESES

Each significance test has two hypotheses about the value of a population parameter.

Null Hypothesis, Alternative Hypothesis

> The **null hypothesis**, denoted by the symbol H_0, is a statement that the parameter takes a particular value. The ***alternative hypothesis***, denoted by H_a, states that the parameter falls in some alternative range of values. Usually the value in H_0 corresponds, in a certain sense, to *no effect*. The values in H_a then represent an effect of some type.

In Example 6.1 about possible gender discrimination in selecting management trainees, let π denote the probability that any particular selection is a male. The company claims that $\pi = 1/2$. This is an example of a null hypothesis, *no effect* referring to a lack of gender bias. The alternative hypothesis reflects the skeptical women employees' belief that this probability actually exceeds 1/2. So, the hypotheses are H_0: $\pi = 1/2$ and H_a: $\pi > 1/2$. Note that H_0 has a *single* value whereas H_a has a range of values.

A significance test analyzes the sample evidence about H_0, by investigating whether the data contradict H_0, hence suggesting that H_a is true. The approach taken

is the indirect one of *proof by contradiction*. The null hypothesis is presumed to be true. Under this presumption, if the data observed would be very unusual, the evidence supports the alternative hypothesis. In the study of potential gender discrimination, we presume that H_0: $\pi = 1/2$ is true. Then we determine whether the sample result of 9 men selected for management training in 10 choices would be unusual, under this presumption. If so, then we may be inclined to believe the women's claim. But, if the difference between the sample proportion of men chosen (9/10) and the H_0 value of 1/2 could easily be due to ordinary sampling variability, there's not enough evidence to accept the women's claim.

A researcher usually conducts a test to gauge the amount of support for the alternative hypothesis, as that typically reflects an effect that he or she predicts. Thus, H_a is sometimes called the **research hypothesis**. The hypotheses are formulated *before* collecting or analyzing the data.

TEST STATISTIC

The parameter to which the hypotheses refer has a point estimate. The **test statistic** summarizes how far that estimate falls from the parameter value in H_0. Often this is expressed by the *number of standard errors* between the estimate and the H_0 value.

P-VALUE

To interpret a test statistic value, we create a probability summary of the evidence against H_0. This uses the sampling distribution of the test statistic, under the presumption that H_0 is true. The purpose is to summarize how unusual the observed test statistic value is compared to what H_0 predicts.

Specifically, if the test statistic falls well out in a tail of the sampling distribution in a direction predicted by H_a, then it is far from what H_0 predicts. We can summarize how far out in the tail the test statistic falls by the tail probability of that value and of more extreme values. These are the possible test statistic values that provide *at least as much evidence against H_0 as the observed test statistic*, in the direction predicted by H_a. This probability is called the **P-value**.

P-value

> The **P-value** is the probability that the test statistic equals the observed value or a value even more extreme in the direction predicted by H_a. It is calculated by presuming that H_0 is true. The P-value is denoted by P.

A small P-value (such as $P = 0.01$) means that the data we observed would have been unusual if H_0 were true. *The smaller the P-value, the stronger the evidence is against H_0.*

For Example 6.1 on potential gender discrimination in choosing managerial trainees, π is the probability of selecting a male. We test H_0: $\pi = 1/2$ against H_a: $\pi > 1/2$. One possible test statistic is the sample proportion of males selected, which is $9/10 = 0.90$. The values for the sample proportion that provide this much or even more extreme evidence against H_0: $\pi = 1/2$ and in favor of H_a: $\pi > 1/2$ are the right-tail sample proportion values of 0.90 and higher. See Figure 6.1. A formula from Section 6.7 calculates this probability as 0.01, so the P-value equals $P = 0.01$. If the selections truly were random with respect to gender, the probability is only 0.01 of such an extreme sample result, namely, that 9 or all 10 selections would be males. Other things being equal, this small P-value provides considerable evidence against H_0: $\pi = 1/2$ and supporting the alternative H_a: $\pi > 1/2$ of discrimination against females.
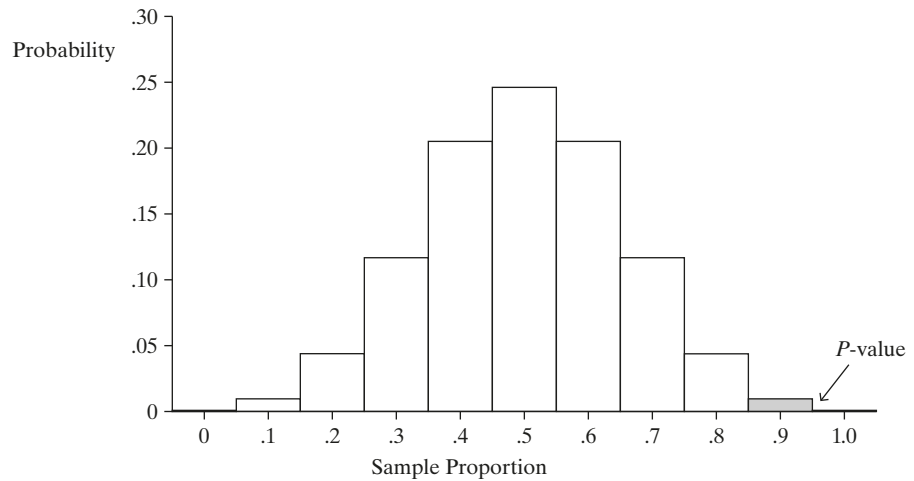
By contrast, a moderate to large *P*-value means the data are consistent with H_0. A *P*-value such as 0.26 or 0.83 indicates that, if H_0 were true, the observed data would not be unusual.

CONCLUSION

The *P*-value summarizes the evidence against H_0. Our conclusion should also *interpret* what the *P*-value tells us about the question motivating the test. Sometimes it is necessary to make a decision about the validity of H_0. If the *P*-value is sufficiently small, we reject H_0 and accept H_a.

Most studies require very small *P*-values, such as $P \leq 0.05$, in order to reject H_0. In such cases, results are said to be *significant at the 0.05 level*. This means that if H_0 were true, the chance of getting such extreme results as in the sample data would be no greater than 0.05.

Making a decision by rejecting or not rejecting a null hypothesis is an optional part of the significance test. We defer discussion of it until Section 6.4. Table 6.1 summarizes the parts of a significance test.

TABLE 6.1: The Five Parts of a Statistical Significance Test
1. **Assumptions** Type of data, randomization, population distribution, sample size condition
2. **Hypotheses** Null hypothesis, H_0 (parameter value for "no effect") Alternative hypothesis, H_a (alternative parameter values)
3. **Test statistic** Compares point estimate to H_0 parameter value
4. **P-value** Weight of evidence against H_0; smaller *P* is stronger evidence
5. **Conclusion** Report and interpret *P*-value Formal decision (optional; see Section 6.4)

6.2 Significance Test for a Mean

For quantitative variables, significance tests usually refer to population means. The five parts of the significance test for a single mean follow:

THE FIVE PARTS OF A SIGNIFICANCE TEST FOR A MEAN

1. Assumptions
The test assumes the data are obtained using randomization, such as a random sample. The quantitative variable is assumed to have a normal population distribution. We'll see that this is mainly relevant for small sample sizes and certain types of H_a.

2. Hypotheses
The null hypothesis about a population mean μ has the form

$$H_0: \mu = \mu_0,$$

where μ_0 is a particular value for the population mean. In other words, the hypothesized value of μ in H_0 is a single value. This hypothesis usually refers to *no effect* or *no change* compared to some standard. For example, Example 5.5 in the previous chapter (page 129) estimated the population mean weight change μ for teenage girls after receiving a treatment for anorexia. The hypothesis that the treatment has *no effect* is a null hypothesis, $H_0: \mu = 0$. Here, the H_0 value μ_0 for the parameter μ is 0.

The alternative hypothesis contains alternative parameter values from the value in H_0. The most common alternative hypothesis is

$$H_a: \mu \neq \mu_0, \quad \text{such as} \quad H_a: \mu \neq 0.$$

This alternative hypothesis is called **two-sided**, because it contains values both below and above the value listed in H_0. For the anorexia study, $H_a: \mu \neq 0$ states that the treatment has *some effect*, the population mean equaling some value other than 0.

3. Test Statistic
The sample mean $\bar{y}$ estimates the population mean μ. When the population distribution is normal, the sampling distribution of $\bar{y}$ is normal about μ. This is also approximately true when the population distribution is *not* normal but the random sample size is relatively large, by the Central Limit Theorem.

Under the presumption that $H_0: \mu = \mu_0$ is true, the center of the sampling distribution of $\bar{y}$ is the value μ_0, as Figure 6.2 shows. A value of $\bar{y}$ that falls far out in the tail provides strong evidence against H_0, because it would be unusual if truly $\mu = \mu_0$.

FIGURE 6.2: Sampling Distribution of $\bar{y}$ if H_0: $\mu = \mu_0$ Is True. For large random samples, it is approximately normal, centered at the null hypothesis value, μ_0.
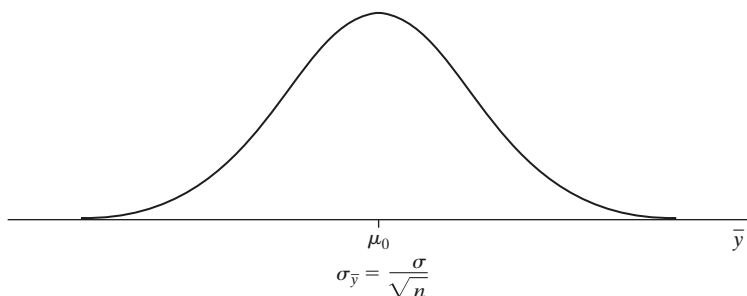
The evidence about H_0 is summarized by the number of standard errors that $\bar{y}$ falls from the null hypothesis value μ_0.

Recall that the *true* standard error is $\sigma_{\bar{y}} = \sigma/\sqrt{n}$. As in Chapter 5, we substitute the sample standard deviation s for the unknown population standard deviation σ to get the *estimated* standard error, $se = s/\sqrt{n}$. The test statistic is the t-score

$$t = \frac{\bar{y} - \mu_0,}{se} \quad \text{where} \quad se = \frac{s}{\sqrt{n}}.$$

The farther $\bar{y}$ falls from μ_0, the larger the absolute value of the t test statistic. Hence, the larger the value of $|t|$, the stronger the evidence against H_0.

We use the symbol t rather than z because, as in forming a confidence interval, using s to estimate σ in the standard error introduces additional error. The null sampling distribution of the t test statistic is the t *distribution* (see Section 5.3). It looks like the standard normal distribution, having mean equal to 0 but being more spread out, more so for smaller n. It is specified by its degrees of freedom, $df = n - 1$.

4. P-Value

The test statistic summarizes how far the data fall from H_0. Different tests use different test statistics, though, and simpler interpretations result from transforming it to the probability scale of 0 to 1. The P-value does this.

We calculate the P-value under the presumption that H_0 is true. That is, we give the benefit of the doubt to H_0, analyzing how unusual the observed data would be if H_0 were true. The P-value is the probability that the test statistic equals the observed value or a value in the set of more extreme values that provide even stronger evidence against H_0. For H_a: $\mu \neq \mu_0$, the more extreme t-values are the ones even farther out in the tails of the t distribution. So, the P-value is the two-tail probability that the t test statistic is at least as large in absolute value as the observed test statistic. This is also the probability that $\bar{y}$ falls at least as far from μ_0 *in either direction* as the observed value of $\bar{y}$.

Figure 6.3 shows the sampling distribution of the t test statistic when H_0 is true. A test statistic value of $t = (\bar{y} - \mu_0)/se = 0$ results when $\bar{y} = \mu_0$. This is the t-value most consistent with H_0. The P-value is the probability of a t test statistic value at least as far from this consistent value as the one observed. To illustrate its calculation, suppose $t = 1.283$ for a sample size of 369. (This is the result in the example below.) This t-score means that the sample mean $\bar{y}$ falls 1.283 estimated standard errors above μ_0. The P-value is the probability that $t \geq 1.283$ or $t \leq -1.283$ (i.e., $|t| \geq 1.283$). Since $n = 369$, $df = n - 1 = 368$ is large, and the t distribution is nearly identical to the standard normal. The probability in one tail above 1.28 is 0.10, so the two-tail probability is $P = 2(0.10) = 0.20$.

Software can supply tail probabilities for the t distribution. For example, the free software R has a function `pt` that gives the cumulative probability for a particular t-score. When $df = 368$, the right-tail probability above $t = 1.283$ is 1 − the cumulative probability:

FIGURE 6.3: Calculation of P-Value when $t = 1.283$, for Testing H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$. The P-value is the two-tail probability of a more extreme result than the observed one.

P = Sum of tail probabilities
= 2(0.10) = 0.20

Sampling distribution of $t = \dfrac{\bar{y} - \mu_0}{se}$ when H_0 is true (t distribution)

0.10   0.10

−1.283   0   1.283 t
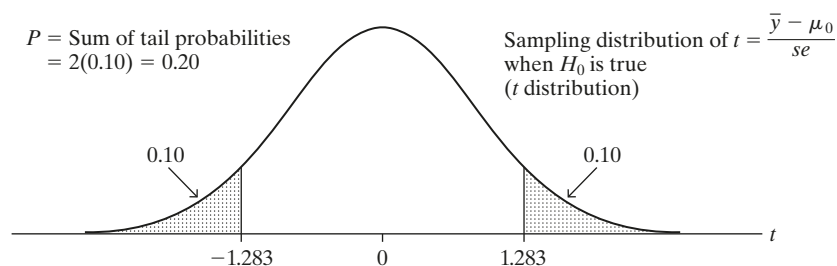
```
> 1 - pt(1.283, 368)
[1] 0.1001498   # right-tail probability above t=1.283, when df=368
```

With Stata software, we can find the right-tail probability with the `ttail` function:

```
. display ttail(368, 1.283)
.10014975
```

We double the single-tail probability to get the *P*-value, $P = 2(0.10014975) = 0.2002995$. Round such a value, say to 0.20, before reporting it. Reporting the *P*-value with many decimal places makes it seem as if more accuracy exists than actually does. In practice, the sampling distribution is only *approximately* the *t* distribution, because the population distribution is not exactly normal as is assumed with the *t* test.

Tail probabilities for the *t* distribution are also available using SPSS and SAS and Internet applets, such as Figure 5.7 showed with the *t Distribution* applet at `www.pearsonglobaleditions.com/Agresti`.

5. Conclusion

Finally, the study should interpret the *P*-value in context. The smaller *P* is, the stronger the evidence against H_0 and in favor of H_a.

Example 6.2

Significance Test about Political Ideology   Some political commentators have remarked that citizens of the United States are increasingly conservative, so much so that many treat "liberal" as a dirty word. We can study political ideology by analyzing responses to certain items on the General Social Survey. For instance, that survey asks where you would place yourself on a seven-point scale of political views ranging from extremely liberal, point 1, to extremely conservative, point 7. Table 6.2 shows the scale and the distribution of responses among the levels for the 2014 survey. Results are shown separately according to subjects classified as white, black, or Hispanic.

Political ideology is an ordinal scale. Often, we treat such scales in a quantitative manner by assigning scores to the categories. Then we can use quantitative summaries such as means, allowing us to detect the extent to which observations gravitate toward the conservative or the liberal end of the scale. If we assign the category

TABLE 6.2: Responses of Subjects on a Scale of Political Ideology

Response	Race		
	Black	White	Hispanic
1. Extremely liberal	16	73	5
2. Liberal	52	209	49
3. Slightly liberal	42	190	46
4. Moderate, middle of road	182	705	155
5. Slightly conservative	43	260	50
6. Conservative	25	314	50
7. Extremely conservative	11	84	14
	$n = 371$	$n = 1835$	$n = 369$

scores shown in Table 6.2, then a mean below 4 shows a propensity toward liberalism and a mean above 4 shows a propensity toward conservatism. We can test whether these data show much evidence of either of these by conducting a significance test about how the population mean compares to the moderate value of 4. We'll do this here for the Hispanic sample and in Section 6.5 for the entire sample.

1. *Assumptions*: The sample is randomly selected. We are treating political ideology as quantitative with equally spaced scores. The t test assumes a normal population distribution for political ideology, which seems inappropriate because the measurement of political ideology is discrete. We'll discuss this assumption further at the end of this section.

2. *Hypotheses*: Let μ denote the population mean ideology for Hispanic Americans, for this seven-point scale. The null hypothesis contains one specified value for μ. Since we conduct the analysis to check how, if at all, the population mean departs from the moderate response of 4, the null hypothesis is

$$H_0: \mu = 4.0.$$

The alternative hypothesis is then

$$H_a: \mu \neq 4.0.$$

The null hypothesis states that, on the average, the population response is politically "moderate, middle of road." The alternative states that the mean falls in the liberal direction ($\mu < 4.0$) or in the conservative direction ($\mu > 4.0$).

3. *Test statistic*: The 369 observations in Table 6.2 for Hispanics are summarized by $\bar{y} = 4.089$ and $s = 1.339$. The estimated standard error of the sampling distribution of $\bar{y}$ is

$$se = \frac{s}{\sqrt{n}} = \frac{1.339}{\sqrt{369}} = 0.0697.$$

The value of the test statistic is

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{4.089 - 4.0}{0.0697} = 1.283.$$

The sample mean falls 1.283 estimated standard errors above the null hypothesis value of the mean. The $df = 369 - 1 = 368$.

4. *P-value*: The *P*-value is the two-tail probability, presuming H_0 is true, that t would exceed 1.283 in absolute value. From the t distribution with $df = 368$, this two-tail probability is $P = 0.20$. If the population mean ideology were 4.0, then the probability equals 0.20 that a sample mean for $n = 368$ subjects would fall at least as far from 4.0 as the observed $\bar{y}$ of 4.089.

5. *Conclusion*: The *P*-value of $P = 0.20$ is not very small, so it does not contradict H_0. If H_0 were true, the data we observed would not be unusual. It is plausible that the population mean response for Hispanic Americans in 2014 was 4.0, not leaning in the conservative or liberal direction. ∎

CORRESPONDENCE BETWEEN TWO-SIDED TESTS AND CONFIDENCE INTERVALS

Conclusions using two-sided significance tests are consistent with conclusions using confidence intervals. If a test says that a particular value is believable for the parameter, then so does a confidence interval.

Example
6.3

Confidence Interval for Mean Political Ideology   For the data in Example 6.2, let's construct a 95% confidence interval for the Hispanic population mean political ideology. With $df = 368$, the multiple of the standard error ($se = 0.0697$) is $t_{.025} = 1.966$. Since $\bar{y} = 4.089$, the confidence interval is

$$\bar{y} \pm 1.966(se) = 4.089 \pm 1.966(0.0697) = 4.089 \pm 0.137, \quad \text{or} \quad (3.95, 4.23).$$

At the 95% confidence level, these are the plausible values for μ.   ■

This confidence interval indicates that μ may equal 4.0, since 4.0 falls inside the confidence interval. Thus, it is not surprising that the P-value ($P = 0.20$) in testing H_0: $\mu = 4.0$ against H_a: $\mu \neq 4.0$ in Example 6.2 was not small. In fact,

> Whenever the $P > 0.05$ in a two-sided test about a mean μ, a 95% confidence interval for μ *necessarily contains* the H_0 value for μ.

By contrast, suppose the P-value = 0.02 in testing H_0: $\mu = 4.0$. Then, a 95% confidence interval would tell us that 4.0 is implausible for μ, with 4.0 falling *outside* the confidence interval.

> Whenever $P \leq 0.05$ in a two-sided test about a mean μ, a 95% confidence interval for μ *does not contain* the H_0 value for μ.

ONE-SIDED SIGNIFICANCE TESTS

We can use a different alternative hypothesis when a researcher predicts a deviation from H_0 in a particular direction. It has one of the forms

$$H_a: \mu > \mu_0 \quad \text{or} \quad H_a: \mu < \mu_0.$$

We use the alternative H_a: $\mu > \mu_0$ to detect whether μ is *larger* than the particular value μ_0, whereas we use H_a: $\mu < \mu_0$ to detect whether μ is *smaller* than that value. These hypotheses are called *one-sided*. By contrast, we use the *two-sided* H_a to detect *any* type of deviation from H_0. This choice is made before analyzing the data.

For H_a: $\mu > \mu_0$, the P-value is the probability (presuming H_0 is true) of a t-score *above* the observed t-score, that is, to the right of it on the real number line. These t-scores provide more extreme evidence than the observed value in favor of H_a: $\mu > \mu_0$. So, P equals the right-tail probability under the t curve. See Figure 6.4. A t-score of 1.283 with $df = 368$ results in $P = 0.10$ for this alternative.

FIGURE 6.4: Calculation of P-Value in Testing H_0: $\mu = \mu_0$ against H_a: $\mu > \mu_0$. The P-value is the probability of values to the right of the observed test statistic.
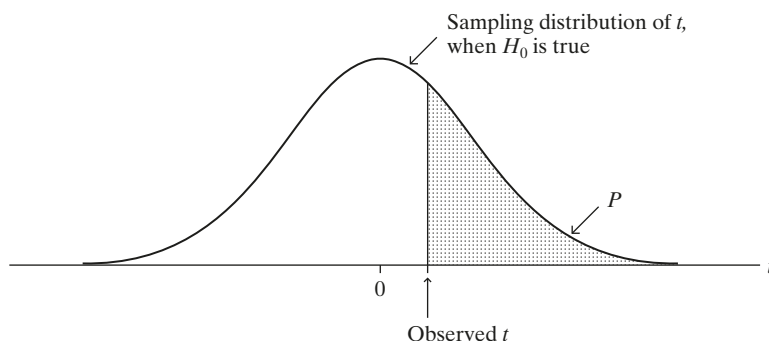
For H_a: $\mu < \mu_0$, the P-value is the left-tail probability, *below* the observed t-score. A t-score of $t = -1.283$ with $df = 368$ results in $P = 0.10$ for this alternative. A t-score of 1.283 results in $P = 1 - 0.10 = 0.90$.

Example **6.4**	**Test about Mean Weight Change in Anorexic Girls** Example 5.5 in Chapter 5 (page 129) analyzed data (available in the Anorexia_CB data file at the text website) from a study comparing treatments for teenage girls suffering from anorexia. For each girl, the study observed her change in weight while receiving the therapy. Let μ denote the population mean change in weight for the cognitive behavioral treatment. If this treatment has beneficial effect, as expected, then μ is positive. To test for no treatment effect versus a positive mean weight change, we test H_0: $\mu = 0$ against H_a: $\mu > 0$.

In the Chapter 5 analysis, we found that the $n = 29$ girls had a sample mean weight change of 3.007 pounds, a standard deviation of 7.309 pounds, and an estimated standard error of $se = 1.357$. The test statistic is

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{3.007 - 0}{1.357} = 2.22.$$

For this one-sided H_a, the P-value is the right-tail probability above 2.22. Why do we use the right tail? Because H_a: $\mu > 0$ has values *above* (i.e., to the right of) the null hypothesis value of 0. It's the positive values of t that support this alternative hypothesis.

Now, for $n = 29$, $df = n - 1 = 28$. The P-value equals 0.02. Software can find the P-value for you. For instance, for the one-sided and two-sided alternatives with a data file with variable *change* for weight change, R reports

```
> t.test(change, mu = 0, alternative = "greater")$p.value
[1] 0.0175113
> t.test(change, mu = 0, alternative = "two.sided")$p.value
[1] 0.0350226
```

Using its `ttest` command with the data file, Stata also reports $P = 0.0175$ for the one-sided H_a: $\mu > 0$. See Table 6.3. If you have only summary statistics rather than a data file, Stata can conduct the test using them, with the `ttesti` command (or a dialog box), by entering n, $\bar{y}$, s, and μ_0 as shown in Table 6.3. Internet applets can also do this.[2]

Some software reports the P-value for a two-sided alternative as the default, unless you request otherwise. SPSS reports results for the two-sided test and confidence interval as

	t	df	Test Value = 0 Sig.(2-tailed)	Mean Difference	95% Confidence Interval of the Difference Lower	Upper
change	2.216	28	.035	3.00690	.2269	5.7869

The one-sided P-value is $0.035/2 = 0.018$. The evidence against H_0 is relatively strong. It seems that the treatment has an effect.

The significance test concludes that the mean weight gain was not equal to 0. But the 95% confidence interval of (0.2, 5.8) is more informative. It shows just how

---

[2] Such as the *Inference for a Mean* applet at www.pearsonglobaleditions.com/Agresti.

different from 0 the population mean change is likely to be. The effect could be very small. Also, keep in mind that this experimental study (like many medically oriented studies) had to use a volunteer sample. So, these results are highly tentative, another reason that it is silly for studies like this to report *P*-values to several decimal places. ∎

TABLE 6.3: Stata Software Output (Edited) for Performing a Significance Test about a Mean

```
. ttest change == 0

One-sample t test
Variable | Obs      Mean    Std. Err.  Std. Dev.  [95% Conf. Interval]
  change |  29   3.006896   1.357155   7.308504   .2268896    5.786902

    mean = mean(change)                                    t =   2.2156
Ho: mean = 0                                     degrees of freedom = 28
    Ha: mean < 0              Ha: mean != 0                Ha: mean > 0
 Pr(T < t) = 0.9825      Pr(|T| > |t|) = 0.0350      Pr(T > t) = 0.0175

/*  Can also perform test with n, mean, std. dev., null value  */
. ttesti 29 3.007 7.309 0

    Ha: mean < 0              Ha: mean != 0                Ha: mean > 0
 Pr(T < t) = 0.9825      Pr(|T| > |t|) = 0.0350      Pr(T > t) = 0.0175
```

IMPLICIT ONE-SIDED H_0 FOR ONE-SIDED H_a

From Example 6.4, the one-sided *P*-value = 0.018. So, if $\mu = 0$, the probability equals 0.018 of observing a sample mean weight gain of 3.01 or greater. Now, suppose $\mu < 0$; that is, the population mean weight change is negative. Then, the probability of observing $\bar{y} \geq 3.01$ would be even smaller than 0.018. For example, a sample value of $\bar{y} = 3.01$ is even less likely when $\mu = -5$ than when $\mu = 0$, since 3.01 is farther out in the tail of the sampling distribution of $\bar{y}$ when $\mu = -5$ than when $\mu = 0$. Thus, rejection of H_0: $\mu = 0$ in favor of H_a: $\mu > 0$ also inherently rejects the broader null hypothesis of H_0: $\mu \leq 0$. In other words, one concludes that $\mu = 0$ is false *and* that $\mu < 0$ is false.

THE CHOICE OF ONE-SIDED VERSUS TWO-SIDED TESTS

In practice, two-sided tests are more common than one-sided tests. Even if a researcher predicts the direction of an effect, two-sided tests can also detect an effect that falls in the opposite direction. In most research articles, significance tests use two-sided *P*-values. Partly this reflects an objective approach to research that recognizes that an effect could go in either direction. In using two-sided *P*-values, researchers avoid the suspicion that they chose H_a when they saw the direction in which the data occurred. That is not ethical.

Two-sided tests coincide with the usual approach in estimation. Confidence intervals are two sided, obtained by adding and subtracting some quantity from the point estimate. One can form one-sided confidence intervals, for instance, having 95% confidence that a population mean weight change is *at least* equal to 0.8 pounds (i.e., between 0.8 and ∞), but in practice one-sided intervals are rarely used.

In deciding whether to use a one-sided or a two-sided H_a in a particular exercise or in practice, consider the context. An exercise that says "Test whether the mean has *changed*" suggests a two-sided alternative, to allow for increase or decrease. "Test whether the mean has *increased*" suggests the one-sided H_a: $\mu > \mu_0$.

In either the one-sided or two-sided case, hypotheses always refer to population parameters, not sample statistics. So, *never* express a hypothesis using sample statistic notation, such as H_0: $\bar{y} = 0$. There is no uncertainty or need to conduct statistical inference about sample statistics such as $\bar{y}$, because we can calculate their values exactly from the data.

THE α-LEVEL: USING THE *P*-VALUE TO MAKE A DECISION

A significance test analyzes the strength of the evidence against the null hypothesis, H_0. We start by presuming that H_0 is true. We analyze whether the data would be unusual if H_0 were true by finding the *P*-value. If the *P*-value is small, the data contradict H_0 and support H_a. Generally, researchers do not regard the evidence against H_0 as strong unless P is very small, say, $P \leq 0.05$ or $P \leq 0.01$.

Why do smaller *P*-values indicate stronger evidence against H_0? Because the data would then be more unusual if H_0 were true. When H_0 is true, the *P*-value is roughly equally likely to fall anywhere between 0 and 1. By contrast, when H_0 is false, the *P*-value is more likely to be near 0 than near 1.

Sometimes we need to decide whether the evidence against H_0 is strong enough to reject it. We base the decision on whether the *P*-value falls below a prespecified cutoff point. For example, we could reject H_0 if $P \leq 0.05$ and conclude that the evidence is not strong enough to reject H_0 if $P > 0.05$. The boundary value 0.05 is called the α-*level* of the test.

α-Level	The **α-level** is a number such that we reject H_0 if the *P*-value is less than or equal to it. The α-level is also called the **significance level**. In practice, the most common α-levels are 0.05 and 0.01.

Like the choice of a confidence level for a confidence interval, the choice of α reflects how cautious you want to be. The smaller the α-level, the stronger the evidence must be to reject H_0. To avoid bias in the decision-making process, you select α *before* analyzing the data.

**Example
6.5**

Examples of Decisions about H_0   Let's use $\alpha = 0.05$ to guide us in making a decision about H_0 for the examples of this section. Example 6.2 (page 157) tested H_0: $\mu = 4.0$ about mean political ideology. With sample mean $\bar{y} = 4.089$, the *P*-value was 0.20. The *P*-value is not small, so if truly $\mu = 4.0$, it would not be unusual to observe $\bar{y} = 4.089$. Since $P = 0.20 > 0.05$, there is insufficient evidence to reject H_0. It is believable that the population mean ideology was 4.0.

Example 6.4 tested H_0: $\mu = 0$ about mean weight gain for teenage girls suffering from anorexia. The *P*-value was 0.018. Since $P = 0.018 < 0.05$, there is sufficient evidence to reject H_0 in favor of H_a: $\mu > 0$. We conclude that the treatment results in an increase in mean weight. Such a conclusion is sometimes phrased as "The increase in mean weight is *statistically significant* at the 0.05 level." Since $P = 0.018$ is *not* less than 0.010, the result is *not* statistically significant at the 0.010 level. In fact, *the P-value is the smallest level for α at which the results are statistically significant*. So, with *P*-value $= 0.018$, we reject H_0 if $\alpha = 0.02$ or 0.05 or 0.10, but not if $\alpha = 0.010$ or 0.001. ■

Table 6.4 summarizes significance tests for population means.

TABLE 6.4: The Five Parts of Significance Tests for Population Means
1. **Assumptions**   Quantitative variable   Randomization   Normal population (robust, especially for two-sided H_a, large n)   2. **Hypotheses** $H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$ (or $H_a: \mu > \mu_0$ or $H_a: \mu < \mu_0$)   3. **Test statistic** $t = \dfrac{\bar{y} - \mu_0}{se}$, where $se = \dfrac{s}{\sqrt{n}}$ 4. ***P*-value**   With the t distribution, use P = Two-tail probability for $H_a: \mu \neq \mu_0$ P = Probability to right of observed t-value for $H_a: \mu > \mu_0$ P = Probability to left of observed t-value for $H_a: \mu < \mu_0$ 5. **Conclusion**   Report P-value. Smaller P provides stronger evidence against H_0 and supporting H_a. Can reject H_0 if $P \leq \alpha$-level

ROBUSTNESS FOR VIOLATIONS OF NORMALITY ASSUMPTION

The t test for a mean assumes that the population distribution is normal. This ensures that the sampling distribution of the sample mean $\bar{y}$ is normal (even for small n) and, after using s to estimate σ in finding the se, the t test statistic has the t distribution. As n increases, this assumption of a normal population becomes less important. We've seen that when n is roughly about 30 or higher, an approximate normal sampling distribution occurs for $\bar{y}$ regardless of the population distribution, by the Central Limit Theorem.

From Section 5.3 (page 125), a statistical method is ***robust*** if it performs adequately even when an assumption is violated. *Two-sided* inferences for a mean using the t distribution are robust against violations of the normal population assumption. Even if the population is not normal, two-sided t tests and confidence intervals still work quite well. The test does not work so well for a one-sided test with small n when the population distribution is highly skewed.

Figure 6.5 shows a histogram and a box plot of the data from the anorexia study of Example 6.4 (page 160). They suggest skew to the right. The box plot highlights (as outliers) six girls who had considerable weight gains. As just mentioned, a two-sided

FIGURE 6.5: Histogram and Box Plot of Weight Change for Anorexia Sufferers
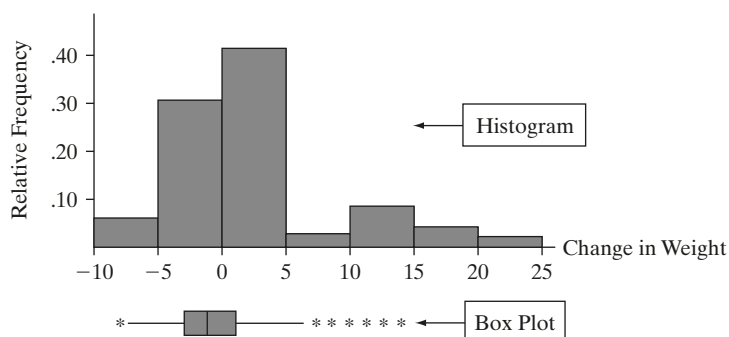
t test works quite well even if the population distribution is skewed. However, this plot makes us wary about using a one-sided test, since the sample size is not large ($n = 29$). Given this and the discussion in the previous subsection about one-sided versus two-sided tests, we're safest with that study to report a two-sided *P*-value of 0.035. Also, the median may be a more relevant summary for these data.

6.3 Significance Test for a Proportion

For a categorical variable, the parameter is the population proportion for a category. For example, a significance test could analyze whether a majority of the population support legalizing same-sex marriage by testing $H_0: \pi = 0.50$ against $H_a: \pi > 0.50$, where π is the population proportion π supporting it. The test for a proportion, like the test for a mean, finds a *P*-value for a test statistic that measures the number of standard errors a point estimate falls from a H_0 value.

THE FIVE PARTS OF A SIGNIFICANCE TEST FOR A PROPORTION

1. Assumptions
Like other tests, this test assumes that the data are obtained using randomization. The sample size must be sufficiently large that the sampling distribution of $\hat{\pi}$ is approximately normal. For the most common case, in which the H_0 value of π is 0.50, a sample size of at least 20 is sufficient.[3]

2. Hypotheses
The null hypothesis of a test about a population proportion has the form

$$H_0: \pi = \pi_0, \quad \text{such as} \quad H_0: \pi = 0.50.$$

Here, π_0 denotes a particular proportion value between 0 and 1, such as 0.50. The most common alternative hypothesis is

$$H_a: \pi \neq \pi_0, \quad \text{such as} \quad H_a: \pi \neq 0.50.$$

This *two-sided* alternative states that the population proportion *differs* from the value in H_0. The *one-sided* alternatives

$$H_a: \pi > \pi_0 \quad \text{and} \quad H_a: \pi < \pi_0$$

apply when the researcher predicts a deviation in a certain direction from the H_0 value.

3. Test Statistic
From Section 5.2, the sampling distribution of the sample proportion $\hat{\pi}$ has mean π and standard error $\sqrt{\pi(1-\pi)/n}$. When H_0 is true, $\pi = \pi_0$, so the standard error is $se_0 = \sqrt{\pi_0(1-\pi_0)/n}$. We use the notation se_0 to indicate that this is the standard error under the presumption that H_0 is true.

The test statistic is

$$z = \frac{\hat{\pi} - \pi_0}{se_0}, \quad \text{where} \quad se_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}.$$

---

[3] Section 6.7, which presents a small-sample test, gives a precise guideline.

This measures the number of standard errors that the sample proportion $\hat{\pi}$ falls from π_0. When H_0 is true, the sampling distribution of the z test statistic is approximately the standard normal distribution.

The test statistic has a similar form as in tests for a mean.

Form of Test Statistic in Test for a Proportion

$$z = \frac{\text{Estimate of parameter} - \text{Null hypothesis value of parameter}}{\text{Standard error of estimate}}$$

Here, the estimate $\hat{\pi}$ of the proportion replaces the estimate $\bar{y}$ of the mean, and the null hypothesis proportion π_0 replaces the null hypothesis mean μ_0.

Note that in the standard error formula, $\sqrt{\pi(1-\pi)/n}$, we substitute the null hypothesis value π_0 for the population proportion π. The parameter values in sampling distributions for tests presume that H_0 is true, since the P-value is based on that presumption. This is why, for tests, we use $se_0 = \sqrt{\pi_0(1-\pi_0)/n}$ rather than the estimated standard error, $se = \sqrt{\hat{\pi}(1-\hat{\pi})/n}$. If we instead used the estimated se, the normal approximation for the sampling distribution of z would be poorer. This is especially true for proportions close to 0 or 1. By contrast, the confidence interval method does not have a hypothesized value for π, so that method uses the estimated se rather than a H_0 value.

4. *P*-Value

The P-value is a one- or two-tail probability, as in tests for a mean, except using the standard normal distribution rather than the t distribution. For $H_a: \pi \neq \pi_0$, P is the two-tail probability. See Figure 6.6. This probability is double the single-tail probability beyond the observed z-value.

For one-sided alternatives, the P-value is a one-tail probability. Since $H_a: \pi > \pi_0$ predicts that the population proportion is *larger* than the H_0 value, its P-value is the probability *above* (i.e., to the right) of the observed z-value. For $H_a: \pi < \pi_0$, the P-value is the probability *below* (i.e., to the left) of the observed z-value.

5. Conclusion

As usual, the smaller the P-value, the more strongly the data contradict H_0 and support H_a. When we need to make a decision, we reject H_0 if $P \leq \alpha$ for a prespecified α-level such as 0.05.

Example 6.6

Reduce Services, or Raise Taxes? These days, whether at the local, state, or national level, government often faces the problem of not having enough money to pay for the various services that it provides. One way to deal with this problem is to raise taxes. Another way is to reduce services. Which would you prefer? When the Florida Poll recently asked a random sample of 1200 Floridians, 52% (624 of the 1200) said raise taxes and 48% said reduce services.

FIGURE 6.6: Calculation of P-Value in Testing H_0: $\pi = \pi_0$ against $H_a: \pi \neq \pi_0$. The two-sided alternative hypothesis uses a two-tail probability.

Sampling distribution of $z = \dfrac{\hat{\pi} - \pi_0}{se_0}$ when H_0 is true (standard normal distribution)

0   Observed z z
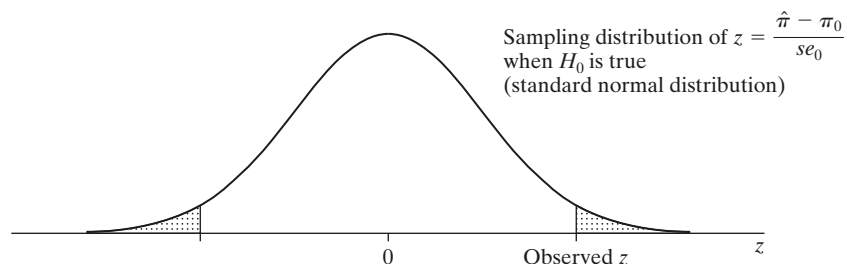
Let π denote the population proportion in Florida who would choose raising taxes rather than reducing services. If $\pi < 0.50$, this is a minority of the population, whereas if $\pi > 0.50$, it is a majority. To analyze whether π is in either of these ranges, we test $H_0: \pi = 0.50$ against $H_a: \pi \neq 0.50$.

The estimate of π is $\hat{\pi} = 0.52$. Presuming $H_0: \pi = 0.50$ is true, the standard error of $\hat{\pi}$ is

$$se_0 = \sqrt{\frac{\pi_0(1 - \pi_0)}{n}} = \sqrt{\frac{(0.50)(0.50)}{1200}} = 0.0144.$$

The value of the test statistic is

$$z = \frac{\hat{\pi} - \pi_0}{se_0} = \frac{0.52 - 0.50}{0.0144} = 1.386.$$

The two-tail P-value is about $P = 0.17$. If H_0 is true (i.e., if $\pi = 0.50$), the probability equals 0.17 that sample results would be as extreme in one direction or the other as in this sample.

This P-value is not small, so there is not much evidence against H_0. It seems believable that $\pi = 0.50$. With an α-level such as 0.05, since $P = 0.17 > 0.05$, we would not reject H_0. We cannot determine whether those favoring raising taxes rather than reducing services are a majority or minority of the population. ∎

We can conduct the test using software. Table 6.5 shows some output (edited) using the free software R applied to the number in the category, n, and the null value π_0. With Stata, you can do this for a variable in a data file, or also directly using the summary statistics as shown in Table 6.6 with the command `prtesti`. The test is also easy to conduct with an Internet applet.[4]

TABLE 6.5: R Software for Performing a Significance Test about a Proportion

```
> prop.test(624, 1200, p=0.50, alt="two.sided", correct=FALSE)

data:  624 out of 1200, null probability 0.5
p-value = 0.1659
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval: 0.4917142 0.5481581
sample estimates:  p  0.52
```

TABLE 6.6: Stata Software for Performing a Significance Test about a Proportion

```
. prtesti 1200 0.52 0.50 // provide n, sample prop., H0 prop.

One-sample test of proportion          x: Number of obs =     1200
  Variable |    Mean   Std. Err.        [95% Conf. Interval]
        x |     .52    .0144222          .491733      .548267
    p = proportion(x)                        z =    1.3856
Ho: p = 0.5
    Ha: p < 0.5               Ha: p != 0.5              Ha: p > 0.5
 Pr(Z < z) = 0.9171      Pr(|Z| > |z|) = 0.1659     Pr(Z > z) = 0.0829
```

---

[4] For example, with the *Inference for a Proportion* applet at www.pearsonglobaleditions.com/Agresti.

NEVER "ACCEPT H_0"

In Example 6.6 about raising taxes or reducing services, the P-value of 0.17 was not small. So, $H_0: \pi = 0.50$ is plausible. In this case, the conclusion is sometimes reported as "Do not reject H_0," since the data do not contradict H_0.

It is better to say "Do not reject H_0" than "Accept H_0." The population proportion has many plausible values besides the number in H_0. For instance, the software output above reports a 95% confidence interval for the population proportion π as (0.49, 0.55). This interval shows a range of plausible values for π. Even though insufficient evidence exists to conclude that $\pi \neq 0.50$, it is improper to conclude that $\pi = 0.50$.

In summary, H_0 contains a single value for the parameter. When the P-value is larger than the α-level, saying "Do not reject H_0" instead of "Accept H_0" emphasizes that that value is merely one of *many* believable values. Because of sampling variability, there is a range of believable values, so we can never accept H_0. The reason "accept H_a" terminology is permissible for H_a is that when the P-value is sufficiently small, the entire range of believable values for the parameter falls within the range of values that H_a specifies.

EFFECT OF SAMPLE SIZE ON P-VALUES

In Example 6.6 on raising taxes or cutting services, suppose $\hat{\pi} = 0.52$ had been based on $n = 4800$ instead of $n = 1200$. The standard error then decreases to 0.0072 (half as large), and you can verify that the test statistic $z = 2.77$. This has two-sided P-value $= 0.006$. That P-value provides strong evidence against $H_0: \pi = 0.50$ and suggests that a majority support raising taxes rather than cutting services. In that case, though, the 95% confidence interval for π equals (0.506, 0.534). This indicates that π is quite close to 0.50 in practical terms.

A given difference between an estimate and the H_0 value has a smaller P-value as the sample size increases. The larger the sample size, the more certain we can be that sample deviations from H_0 are indicative of true population deviations. In particular, notice that even a small departure between $\hat{\pi}$ and π_0 (or between $\bar{y}$ and μ_0) can yield a small P-value if the sample size is very large.

6.4 Decisions and Types of Errors in Tests

When we need to decide whether the evidence against H_0 is strong enough to reject it, we reject H_0 if $P \leq \alpha$, for a prespecified α-level. Table 6.7 summarizes the two possible conclusions for α-level = 0.05. The null hypothesis is either "rejected" or "not rejected." If H_0 is rejected, then H_a is accepted. If H_0 is not rejected, then H_0 is plausible, but other parameter values are also plausible. Thus, H_0 is never "accepted." In this case, results are inconclusive, and the test does not identify either hypothesis as more valid.

TABLE 6.7: Possible Decisions in a Significance Test with α-Level = 0.05

P-Value	Conclusion H_0	H_a
$P \leq 0.05$	Reject	Accept
$P > 0.05$	Do not reject	Do not accept

It is better to report the P-value than to indicate merely whether the result is "statistically significant." Reporting the P-value has the advantage that the reader can tell whether the result is significant at any level. The P-values of 0.049 and 0.001 are both "significant at the 0.05 level," but the second case provides much stronger evidence than the first case. Likewise, P-values of 0.049 and 0.051 provide, in practical terms, the same amount of evidence about H_0. It is a bit artificial to call one result "significant" and the other "nonsignificant." Some software places the symbol * next to a test statistic that is significant at the 0.05 level, ** next to a test statistic that is significant at the 0.01 level, and *** next to a test statistic that is significant at the 0.001 level.

TYPE I AND TYPE II ERRORS FOR DECISIONS

Because of sampling variability, decisions in tests always have some uncertainty. The decision could be erroneous. The two types of potential errors are conventionally called *Type I* and *Type II* errors.

Type I and Type II Errors	When H_0 is true, a ***Type I error*** occurs if H_0 is rejected. When H_0 is false, a ***Type II error*** occurs if H_0 is not rejected.

The two possible decisions cross-classified with the two possibilities for whether H_0 is true generate four possible results. See Table 6.8.

TABLE 6.8: The Four Possible Results of Making a Decision in a Significance Test. Type I and Type II errors are the incorrect decisions.

		Decision	
		Reject H_0	Do Not Reject H_0
Condition of H_0	H_0 true	Type I error	Correct decision
	H_0 false	Correct decision	Type II error

REJECTION REGIONS: STATISTICALLY SIGNIFICANT TEST STATISTIC VALUES

The collection of test statistic values for which the test rejects H_0 is called the ***rejection region***. For example, the rejection region for a test of level $\alpha = 0.05$ is the set of test statistic values for which $P \leq 0.05$.

For two-sided tests about a proportion, the two-tail P-value is ≤ 0.05 whenever the test statistic $|z| \geq 1.96$. In other words, the rejection region consists of values of z resulting from the estimate falling at least 1.96 standard errors from the H_0 value.

THE α-LEVEL IS THE PROBABILITY OF TYPE I ERROR

When H_0 is true, let's find the probability of Type I error. Suppose $\alpha = 0.05$. We've just seen that for the two-sided test about a proportion, the rejection region is $|z| \geq 1.96$. So, the probability of rejecting H_0 is exactly 0.05, because the probability of the values in this rejection region under the standard normal curve is 0.05. But this is precisely the α-level.

> The probability of a Type I error is the α-level for the test.

With $\alpha = 0.05$, if H_0 is true, the probability equals 0.05 of making a Type I error and rejecting H_0. We control P(Type I error) by the choice of α. The more serious the consequences of a Type I error, the smaller α should be. In practice, $\alpha = 0.05$ is most common, just as an error probability of 0.05 is most common with confidence intervals (i.e, 95% confidence). However, this may be too high when a decision has serious implications.

For example, consider a criminal legal trial of a defendant. Let H_0 represent innocence and H_a represent guilt. The jury rejects H_0 and judges the defendant to be guilty if it decides the evidence is sufficient to convict. A Type I error, rejecting a true H_0, occurs in convicting a defendant who is actually innocent. In a murder trial, suppose a convicted defendant may receive the death penalty. Then, if a defendant is actually innocent, we would hope that the probability of conviction is much smaller than 0.05.

When we make a decision, we do not know whether we have made a Type I or Type II error, just as we do not know whether a particular confidence interval truly contains the parameter value. However, we can control the probability of an incorrect decision for either type of inference.

AS P(TYPE I ERROR) GOES DOWN, P(TYPE II ERROR) GOES UP

In an ideal world, Type I or Type II errors would not occur. However, errors do happen. We've all read about defendants who were convicted but later determined to be innocent. When we make a decision, why don't we use an extremely small P(Type I error), such as $\alpha = 0.000001$? For instance, why don't we make it almost impossible to convict someone who is really innocent?

When we make α smaller in a significance test, we need a smaller P-value to reject H_0. It then becomes harder to reject H_0. But this means that it will also be harder even if H_0 is false. The stronger the evidence required to convict someone, the more likely we will fail to convict defendants who are actually guilty. In other words, the smaller we make P(Type I error), the larger P(Type II error) becomes, that is, failing to reject H_0 even though it is false.

If we tolerate only an extremely small P(Type I error), such as $\alpha = 0.000001$, the test may be unlikely to reject H_0 even if it is false—for instance, unlikely to convict someone even if they are guilty. This reasoning reflects the fundamental relation:

- The smaller P(Type I error) is, the larger P(Type II error) is.

For instance, in an example in Section 6.6, when P(Type I error) = 0.05 we'll find that P(Type II error) = 0.02, but when P(Type I error) decreases to 0.01, P(Type II error) increases to 0.08. Except in Section 6.6, we shall not find P(Type II error), as it is beyond our scope. In practice, making a decision requires setting only α, the P(Type I error).

Section 6.6 shows that P(Type II error) depends on just how far the true parameter value falls from H_0. If the parameter is nearly equal to the value in H_0, P(Type II error) is relatively high. If it falls far from H_0, P(Type II error) is relatively low. The farther the parameter falls from the H_0 value, the less likely the sample is to result in a Type II error.

For a fixed P(Type I error), P(Type II error) depends also on the sample size n. The larger the sample size, the more likely we are to reject a false H_0. To keep both P(Type I error) and P(Type II error) at low levels, it may be necessary to use a very

large sample size. The P(Type II error) may be quite large when the sample size is small, unless the parameter falls quite far from the H_0 value.

EQUIVALENCE BETWEEN CONFIDENCE INTERVALS AND TEST DECISIONS

We now elaborate on the equivalence for means[5] between decisions from two-sided tests and conclusions from confidence intervals, first alluded to in Example 6.3 (page 159). Consider the significance test of

$$H_0: \mu = \mu_0 \qquad \text{versus} \qquad H_a: \mu \neq \mu_0.$$

When $P < 0.05$, H_0 is rejected at the $\alpha = 0.05$ level. When n is large (so the t distribution is essentially the same as the standard normal), this happens when the test statistic $t = (\bar{y} - \mu_0)/se$ is greater in absolute value than 1.96, that is, when $\bar{y}$ falls more than 1.96(se) from μ_0. But if this happens, then the 95% confidence interval for μ, namely, $\bar{y} \pm 1.96(se)$, does not contain the null hypothesis value μ_0. See Figure 6.7. These two inference procedures are consistent.

FIGURE 6.7: Relationship between Confidence Interval and Significance Test. For large n, the 95% confidence interval does not contain the H_0 value μ_0 when the sample mean falls more than 1.96 standard errors from μ_0, in which case the test statistic $|t| > 1.96$ and the P-value < 0.05.
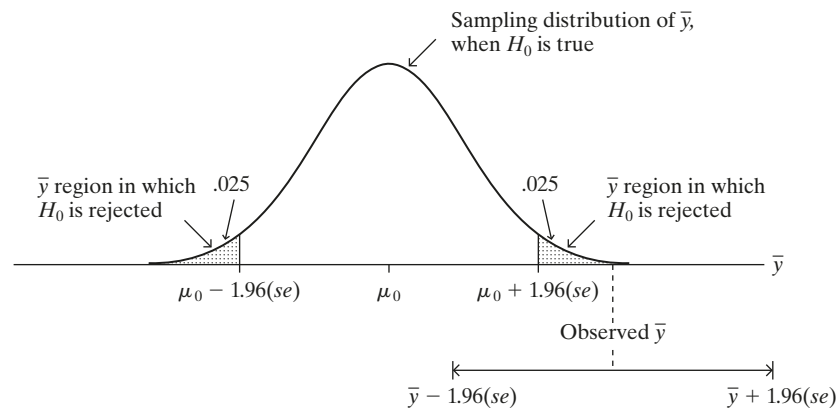
Significance Test Decisions and Confidence Intervals

In testing $H_0: \mu = \mu_0$ against $H_a: \mu \neq \mu_0$, when we reject H_0 at the 0.05 α-level, the 95% confidence interval for μ does not contain μ_0. The 95% confidence interval consists of those μ_0 values for which we do not reject H_0 at the 0.05 α-level.

In Example 6.2 about mean political ideology (page 157), the P-value for testing $H_0: \mu = 4.0$ against $H_a: \mu \neq 4.0$ was $P = 0.20$. At the $\alpha = 0.05$ level, we do not reject $H_0: \mu = 4.0$. It is believable that $\mu = 4.0$. Example 6.3 (page 159) showed that a 95% confidence interval for μ is (3.95, 4.23), which contains $\mu_0 = 4.0$.

Rejecting H_0 at a particular α-level is equivalent to the confidence interval for μ with the same error probability not containing μ_0. For example, if a 99% confidence interval does not contain 0, then we would reject $H_0: \mu = 0$ in favor of $H_a: \mu \neq 0$ at the $\alpha = 0.01$ level with the test. The α-level is P(Type I error) for the test and the probability that the confidence interval method does not contain the parameter.

---

[5] This equivalence also holds for proportions when we use the two-sided test of Section 6.3 and the confidence interval method presented in Exercise 5.77.

MAKING DECISIONS VERSUS REPORTING THE *P*-VALUE

The approach to hypothesis testing that incorporates a formal decision with a fixed P(Type I error) was developed by the statisticians Jerzy Neyman and Egon Pearson in the late 1920s and early 1930s. In summary, this approach formulates null and alternative hypotheses, selects an α-level for the P(Type I error), determines the rejection region of test statistic values that provide enough evidence to reject H_0, and then makes a decision about whether to reject H_0 according to what is actually observed for the test statistic value. With this approach, it's not even necessary to find a P-value. The choice of α-level determines the rejection region, which together with the test statistic determines the decision.

The alternative approach of finding a P-value and using it to summarize evidence against a hypothesis is due to the great British statistician R. A. Fisher. He advocated merely reporting the P-value rather than using it to make a formal decision about H_0. Over time, this approach has gained favor, especially since software can now report precise P-values for a wide variety of significance tests.

This chapter has presented an amalgamation of the two approaches (the decision-based approach using an α-level and the P-value approach), so you can interpret a P-value yet also know how to use it to make a decision when that is needed. These days, most research articles merely report the P-value rather than a decision about whether to reject H_0. From the P-value, readers can view the strength of evidence against H_0 and make their own decision, if they want to.

6.5 Limitations of Significance Tests

A significance test makes an inference about whether a parameter differs from the H_0 value and about its direction from that value. In practice, we also want to know whether the parameter is sufficiently different from the H_0 value to be practically important. In this section, we'll learn that a test does not tell us as much as a confidence interval about practical importance.

STATISTICAL SIGNIFICANCE VERSUS PRACTICAL SIGNIFICANCE

It is important to distinguish between *statistical significance* and *practical significance*. A small P-value, such as $P = 0.001$, is highly statistically significant. It provides strong evidence against H_0. It does not, however, imply an *important* finding in any practical sense. The small P-value merely means that if H_0 were true, the observed data would be very unusual. It does not mean that the true parameter value is far from H_0 in practical terms.

Example 6.7

Mean Political Ideology for All Americans The political ideology $\bar{y} = 4.089$ reported in Example 6.2 (page 157) refers to a sample of Hispanic Americans. We now consider the entire 2014 GSS sample who responded to the political ideology question. For a scoring of 1.0 through 7.0 for the ideology categories with 4.0 = moderate, the $n = 2575$ observations have $\bar{y} = 4.108$ and standard deviation $s = 4.125$. On the average, political ideology was the same for the entire sample as it was for Hispanics alone.[6]

As in Example 6.2, we test H_0: $\mu = 4.0$ against H_a: $\mu \neq 4.0$ to analyze whether the population mean differs from the moderate ideology score of 4.0. Now,

---

[6] And it seems stable over time, equaling 4.13 in 1980, 4.16 in 1990, and 4.10 in 2000.

$$se = s/\sqrt{n} = 1.425/\sqrt{2575} = 0.028, \text{ and}$$

$$t = \frac{\bar{y} - \mu_0}{se} = \frac{4.108 - 4.0}{0.028} = 3.85.$$

The two-sided *P*-value is $P = 0.0001$. There is *very* strong evidence that the true mean exceeds 4.0, that is, that the true mean falls on the conservative side of moderate. But, on a scale of 1.0 to 7.0, 4.108 is close to the moderate score of 4.0. Although the difference of 0.108 between the sample mean of 4.108 and the H_0 mean of 4.0 is highly significant statistically, the magnitude of this difference is very small in practical terms. The mean response on political ideology for all Americans is essentially a moderate one. ∎

In Example 6.2, the sample mean of $\bar{y} = 4.1$ for $n = 369$ Hispanic Americans had a *P*-value of $P = 0.20$, not much evidence against H_0. But now with $\bar{y} = 4.1$ based on $n = 2575$, we have instead found $P = 0.0001$. This is highly *statistically significant*, but not *practically significant*. For practical purposes, a mean of 4.1 on a scale of 1.0 to 7.0 for political ideology does not differ from 4.00.

A way of summarizing practical significance is to measure the ***effect size*** by the number of standard deviations (*not* standard errors) that $\bar{y}$ falls from μ_0. In this example, the estimated effect size is $(4.108 - 4.0)/1.425 = 0.08$. This is a tiny effect. Whether a particular effect size is small, medium, or large depends on the substantive context, but an effect size of about 0.2 or less in absolute value is usually not practically important.

SIGNIFICANCE TESTS ARE LESS USEFUL THAN CONFIDENCE INTERVALS

We've seen that, with large sample sizes, *P*-values can be small even when the point estimate falls near the H_0 value. The size of *P* merely summarizes the extent of evidence about H_0, not how far the parameter falls from H_0. Always inspect the difference between the estimate and the H_0 value to gauge the practical implications of a test result.

Null hypotheses containing single values are rarely true. That is, rarely is the parameter *exactly* equal to the value listed in H_0. With sufficiently large samples, so that a Type II error is unlikely, these hypotheses will normally be rejected. What is more relevant is whether the parameter is sufficiently different from the H_0 value to be of practical importance.

Although significance tests can be useful, most statisticians believe they are overemphasized in social science research. It is preferable to construct confidence intervals for parameters instead of performing only significance tests. A test merely indicates whether the particular value in H_0 is plausible. It does not tell us which other potential values are plausible. The confidence interval, by contrast, displays the entire set of believable values. It shows the extent to which reality may differ from the parameter value in H_0 by showing whether the values in the interval are far from the H_0 value. Thus, it helps us to determine whether rejection of H_0 has practical importance.

To illustrate, for the complete political ideology data in Example 6.7, a 95% confidence interval for μ is

$$\bar{y} \pm 1.96(se) = 4.108 \pm 1.96(0.028), \text{ or } (4.05, 4.16).$$

This indicates that the difference between the population mean and the moderate score of 4.0 is very small. Although the *P*-value of $P = 0.0001$ provides very strong evidence against H_0: $\mu = 4.0$, in practical terms the confidence interval shows that

H_0 is not wrong by much. By contrast, if $\bar{y}$ had been 6.108 (instead of 4.108), the 95% confidence interval would equal (6.05, 6.16). This indicates a substantial practical difference from 4.0, the mean response being near the conservative score rather than the moderate score.

When a *P*-value is not small but the confidence interval is quite wide, this forces us to realize that the parameter might well fall far from H_0 even though we cannot reject it. This also supports why it does not make sense to "accept H_0," as we discussed on page 167.

The remainder of the text presents significance tests for a variety of situations. It is important to become familiar with these tests, if for no other reason than their frequent use in social science research. However, we'll also introduce confidence intervals that describe how far reality is from the H_0 value.

SIGNIFICANCE TESTS AND *P*-VALUES CAN BE MISLEADING

We've seen it is improper to "accept H_0." We've also seen that statistical significance does not imply practical significance. Here are other ways that results of significance tests can be misleading:

- **It is misleading to report results only if they are statistically significant.** Some research journals have the policy of publishing results of a study only if the *P*-value ≤ 0.05. Here's a danger of this policy: Suppose there truly is no effect, but 20 researchers independently conduct studies. We would expect about $20(0.05) = 1$ of them to obtain significance at the 0.05 level merely by chance. (When H_0 is true, about 5% of the time we get a *P*-value below 0.05 anyway.) If that researcher then submits results to a journal but the other 19 researchers do not, the article published will be a Type I error. It will report an effect when there really is not one.

- **Some tests may be statistically significant just by chance.** You should never scan software output for results that are statistically significant and report only those. If you run 100 tests, even if all the null hypotheses are correct, you would expect to get *P*-values ≤ 0.05 about $100(0.05) = 5$ times. Be skeptical of reports of significance that might merely reflect ordinary random variability.

- **It is incorrect to interpret the *P*-value as the probability that H_0 is true.** The *P*-value is *P*(test statistic takes value like observed or even more extreme), presuming that H_0 is true. It is not $P(H_0$ true). Classical statistical methods calculate probabilities about variables and statistics (such as test statistics) that vary randomly from sample to sample, not about parameters. Statistics have sampling distributions, parameters do not. In reality, H_0 is not a matter of probability. It is either true or not true. We just don't know which is the case.

- **True effects are often smaller than reported estimates.** Even if a statistically significant result is a real effect, the true effect may be smaller than reported. For example, often several researchers perform similar studies, but the results that receive attention are the most extreme ones. The researcher who decides to publicize the result may be the one who got the most impressive sample result, perhaps way out in the tail of the sampling distribution of all the possible results. See Figure 6.8.

Example 6.8

Are Many Medical "Discoveries" Actually Type I Errors? In medical research studies, suppose that an actual population effect exists only 10% of the time. Suppose also that when an effect truly exists, there is a 50% chance of making a Type II error

FIGURE 6.8: When Many Researchers Conduct Studies about a Hypothesis, the Statistically Significant Result Published in a Journal and Reported by Popular Media Often Overestimates the True Effect
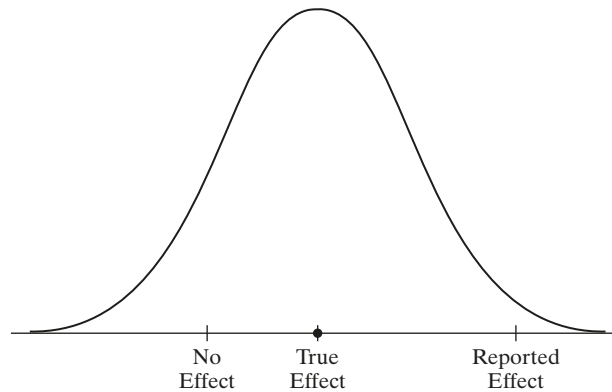
and failing to detect it. These were the hypothetical percentages used in an article in a medical journal.[7] The authors noted that many medical studies have a high Type II error rate because they are not able to use a large sample size. Assuming these rates, could a substantial percentage of medical "discoveries" actually be Type I errors?

Figure 6.9 is a ***tree diagram*** showing what we would expect with 1000 medical studies that test various hypotheses. If a population effect truly exists only 10% of the time, this would be the case for 100 of the 1000 studies. We do not obtain a small enough P-value to detect this true effect 50% of the time, that is, in 50 of these 100 studies. An effect will be reported for the other 50 of the 100 that do truly have an effect. For the 900 cases in which there truly is no effect, with the usual significance level of 0.05 we expect 5% of the 900 studies to incorrectly reject H_0. This happens for $(0.05)900 = 45$ studies. In summary, of the 1000 studies, we expect 50 to report an effect that is truly there, but we also expect 45 to report an effect that does not actually exist. So, a proportion of $45/(45 + 50) = 0.47$ of medical studies that report effects are actually reporting Type I errors.

FIGURE 6.9: Tree Diagram of 1000 Hypothetical Medical Studies. This assumes a population effect truly exists 10% of the time and a 50% chance of a Type II error when an effect truly exists.
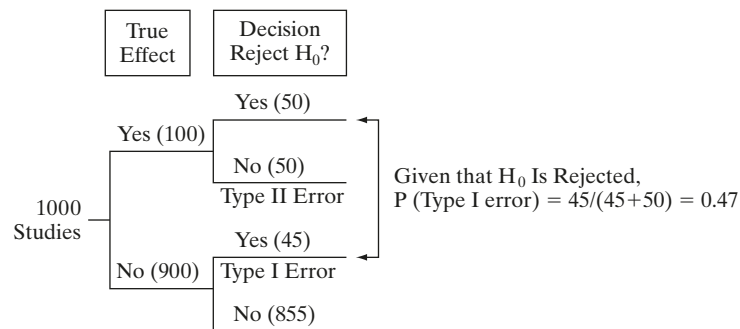
The moral is to be skeptical when you hear reports of new medical advances. The true effect may be weaker than reported, or there may actually be no effect at all. ∎

Related to this is the *publication bias* that occurs when results of some studies never appear in print because they did not obtain a small enough P-value to seem important. One investigation[8] of this reported that 94% of medical studies that had positive results found their way into print whereas only 14% of those with disappointing or uncertain results did.

---

[7] By J. Sterne, G. Smith, and D. R. Cox, *BMJ*, vol. 322 (2001), pp. 226–231.
[8] Reported in *The New York Times*, January 17, 2008.

6.6 Finding P(Type II Error)*

We've seen that decisions in significance tests have two potential types of error. A Type I error results from rejecting H_0 when it is actually true. Given that H_0 is true, the probability of a Type I error is the α-level of the test; when $\alpha = 0.05$, the probability of rejecting H_0 equals 0.05.

When H_0 is false, a Type II error results from *not* rejecting it. This probability has more than one value, because H_a contains a range of possible values. Each value in H_a has its own P(Type II error). This section shows how to calculate P(Type II error) at a particular value.

Example 6.9

Testing whether Astrology Really Works   One scientific test of the pseudoscience astrology used the following experiment[9]: For each of 116 adult subjects, an astrologer prepared a horoscope based on the positions of the planets and the moon at the moment of the person's birth. Each subject also filled out a California Personality Index survey. For each adult, his or her birth data and horoscope were shown to an astrologer with the results of the personality survey for that adult and for two other adults randomly selected from the experimental group. The astrologer was asked which personality chart of the three subjects was the correct one for that adult, based on their horoscope.

Let π denote the probability of a correct prediction by an astrologer. If the astrologers' predictions are like random guessing, then $\pi = 1/3$. To test this against the alternative that the guesses are better than random guessing, we can test H_0: $\pi = 1/3$ against H_a: $\pi > 1/3$. The alternative hypothesis reflects the astrologers' belief that they can predict better than random guessing. In fact, the National Council for Geocosmic Research, which supplied the astrologers for the experiment, claimed π would be 0.50 or higher. So, let's find P(Type II error) if actually $\pi = 0.50$, for an $\alpha = 0.05$-level test. That is, if actually $\pi = 0.50$, we'll find the probability that we'd fail to reject H_0: $\pi = 1/3$.

To determine this, we first find the sample proportion values for which we would not reject H_0. For the test of H_0: $\pi = 1/3$, the sampling distribution of $\hat{\pi}$ is the curve shown on the left in Figure 6.10. With $n = 116$, this curve has standard error

$$se_0 = \sqrt{\frac{\pi_0(1 - \pi_0)}{n}} = \sqrt{\frac{(1/3)(2/3)}{116}} = 0.0438.$$

FIGURE 6.10: Calculation of P(Type II Error) for Testing H_0: $\pi = 1/3$ against H_a: $\pi > 1/3$ at $\alpha = 0.05$ Level, when True Proportion Is $\pi = 0.50$ and $n = 116$. A Type II error occurs if $\hat{\pi} < 0.405$, since then the P-value > 0.05 even though H_0 is false.

Sampling distribution of $\hat{\pi}$ when $\pi = \frac{1}{3}$ (H_0 true)

.02 = Probability that H_0 is not rejected when $\pi = 0.50$ (H_0 false)

Sampling distribution of $\hat{\pi}$ when $\pi = 0.50$ (H_0 false)

$\alpha = .05$

$\frac{1}{3}$ 0.405   0.50 $\hat{\pi}$
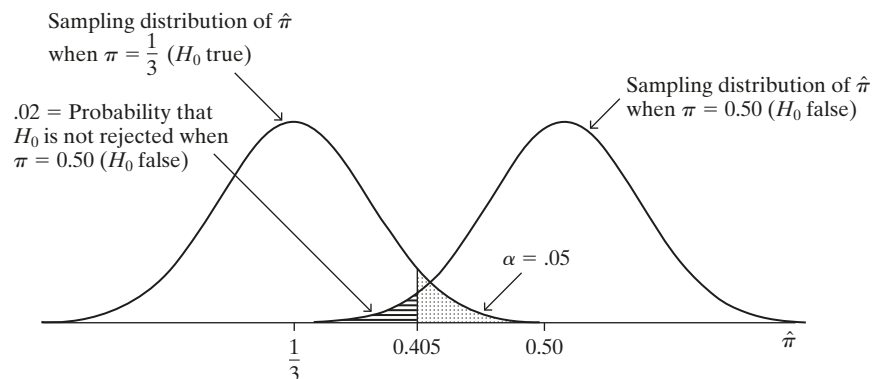
[9] S. Carlson, *Nature*, vol. 318 (1985), pp. 419–425.

For H_a: $\pi > 1/3$, the P-value equals 0.05 if the test statistic $z = 1.645$. That is, 1.645 is the z-score that has a right-tail probability of 0.05. So, we *fail to reject* H_0, getting a P-value *above* 0.05, if $z < 1.645$. In other words, we fail to reject H_0: $\pi = 1/3$ if the sample proportion $\hat{\pi}$ falls less than 1.645 standard errors above 1/3, that is, if

$$\hat{\pi} < 1/3 + 1.645(se_0) = 1/3 + 1.645(0.0438) = 0.405.$$

So, the right-tail probability above 0.405 is $\alpha = 0.05$ for the curve on the left in Figure 6.10.

To find P(Type II error) if π actually equals 0.50, we must find $P(\hat{\pi} < 0.405)$ when $\pi = 0.50$. This is the left-tail probability *below* 0.405 for the curve on the right in Figure 6.10, which is the curve that applies when $\pi = 0.50$. When $\pi = 0.50$, the standard error for a sample size of 116 is $\sqrt{[(0.50)(0.50)]/116} = 0.0464$. (This differs a bit from se_0 for the test statistic, which uses 1/3 instead of 0.50 for π.) For the normal distribution with a mean of 0.50 and standard error of 0.0464, the $\hat{\pi}$ value of 0.405 has a z-score of

$$z = \frac{0.405 - 0.50}{0.0464} = -2.04.$$

The probability that $\hat{\pi} < 0.405$ is the probability that a standard normal variable falls below -2.04, which equals 0.02. So, for a sample of size 116, the probability of not rejecting H_0: $\pi = 1/3$ is 0.02, if in fact $\pi = 0.50$. In other words, if astrologers truly had the predictive power they claimed, the chance of failing to detect this with this experiment would have only been about 0.02. To see what actually happened in the experiment, see Exercise 6.17. ∎

This probability calculation of P(Type II error) was rather involved. Such calculations can be performed easily with an Internet applet.[10]

The probability of Type II error increases when the parameter value moves closer to H_0. To verify this, you can check that P(Type II error) $= 0.55$ at $\pi = 0.40$. So, if the parameter falls near the H_0 value, there may be a substantial chance of failing to reject H_0. Likewise, the farther the parameter falls from H_0, the less likely a Type II error. Figure 6.11 plots P(Type II error) for the various π values in H_a.

FIGURE 6.11: Probability of Type II Error for Testing H_0: $\pi = 1/3$ against H_a: $\pi > 1/3$ at $\alpha = 0.05$ Level, Plotted for the Potential π Values in H_a
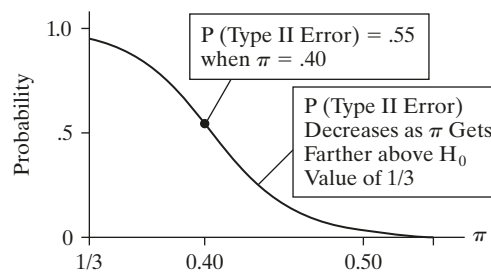
For a fixed α-level and alternative parameter value, P(Type II error) decreases when the sample size increases. If you can obtain more data, you will be less likely to make this sort of error.

TESTS WITH SMALLER α HAVE GREATER P(TYPE II ERROR)

As explained on page 169, the smaller $\alpha = P$(Type I error) is in a test, the larger P(Type II error) is. To illustrate, suppose the astrology study in Example 6.9 used

---

[10] See, for example, the *Errors and Power* applet at www.pearsonglobaleditions.com/Agresti.

$\alpha = 0.01$. Then, when $\pi = 0.50$ you can verify that $P(\text{Type II error}) = 0.08$, compared to $P(\text{Type II error}) = 0.02$ when $\alpha = 0.05$.

The reason that extremely small values are not normally used for α, such as $\alpha = 0.0001$, is that $P(\text{Type II error})$ is too high. We may be unlikely to reject H_0 even if the parameter falls far from the null hypothesis. In summary, for fixed values of other factors,

- $P(\text{Type II error})$ decreases as
  - the parameter value is farther from H_0.
  - the sample size increases.
  - $P(\text{Type I error})$ increases.

THE POWER OF A TEST

When H_0 is false, you want the probability of rejecting H_0 to be high. The probability of rejecting H_0 is called the ***power*** of the test. For a particular value of the parameter from within the H_a range,

$$\text{Power} = 1 - P(\text{Type II error}).$$

In Example 6.9, for instance, the test of H_0: $\pi = 1/3$ has $P(\text{Type II error}) = 0.02$ at $\pi = 0.50$. Therefore, the power of the test at $\pi = 0.50$ is $1 - 0.02 = 0.98$.

The power increases for values of the parameter falling farther from the H_0 value. Just as the curve for $P(\text{Type II error})$ in Figure 6.11 decreases as π gets farther above $\pi_0 = 1/3$, the curve for the power increases.

In practice, studies should ideally have high power. Before granting financial support for a planned study, research agencies often expect principal investigators to show that reasonable power (usually, at least 0.80) exists at values of the parameter that are practically significant.

When you read that results of a study are not statistically significant, be skeptical if no information is given about the power. The power may be low, especially if n is small or the effect is not large.

6.7 Small-Sample Test for a Proportion— The Binomial Distribution*

For a population proportion π, Section 6.3 presented a significance test that is valid for relatively large samples. The sampling distribution of the sample proportion $\hat{\pi}$ is then approximately normal, which justifies using a z test statistic.

For small n, the sampling distribution of $\hat{\pi}$ occurs at only a few points. If $n = 5$, for example, the only possible values for the sample proportion $\hat{\pi}$ are $0, 1/5, 2/5, 3/5, 4/5,$ and 1. A continuous approximation such as the normal distribution is inappropriate. In addition, the closer π is to 0 or 1 for a given sample size, the more skewed the actual sampling distribution becomes.

This section introduces a small-sample test for proportions. It uses the most important probability distribution for discrete variables, the *binomial distribution*.

THE BINOMIAL DISTRIBUTION

For categorical data, often the following three conditions hold:

- Each observation falls into one of two categories.

- The probabilities for the two categories are the same for each observation. We denote the probabilities by π for category 1 and $(1 - \pi)$ for category 2.

- The outcomes of successive observations are independent. That is, the outcome for one observation does not depend on the outcomes of other observations.

Flipping a coin repeatedly is a prototype for these conditions. For each flip, we observe whether the outcome is head (category 1) or tail (category 2). The probabilities of the outcomes are the same for each flip (0.50 for each if the coin is balanced). The outcome of a particular flip does not depend on the outcome of other flips.

Now, for n observations, let x denote the number of them that occur in category 1. For example, for $n = 5$ coin flips, $x =$ number of heads could equal 0, 1, 2, 3, 4, or 5. When the observations satisfy the above three conditions, the probability distribution of x is the ***binomial distribution***.

The binomial variable x is discrete, taking one of the integer values 0, 1, 2, ..., n. The formula for the binomial probabilities follows:

Probabilities for a Binomial Distribution

Denote the probability of category 1, for each observation, by π. For n independent observations, the probability that x of the n observations occur in category 1 is

$$P(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

The symbol $n!$ is called ***n factorial***. It represents $n! = 1 \times 2 \times 3 \times \cdots \times n$. For example, $1! = 1$, $2! = 1 \times 2 = 2$, $3! = 1 \times 2 \times 3 = 6$, and so forth. Also, $0!$ is defined to be 1.

For particular values for π and n, substituting the possible values for x into the formula for $P(x)$ provides the probabilities of the possible outcomes. The sum of the probabilities equals 1.0.

Example 6.10

Gender and Selection of Managerial Trainees Example 6.1 (page 151) discussed a case involving potential bias against females in selection of management trainees for a large supermarket chain. The pool of employees is half female and half male. The company claims to have selected 10 trainees at random from this pool. If they are truly selected at random, how many females would we expect to be chosen?

The probability that any one person selected is a female is $\pi = 0.50$, the proportion of available trainees who are female. Similarly, the probability that any one person selected is male is $(1 - \pi) = 0.50$. Let $x =$ number of females selected. This has the binomial distribution with $n = 10$ and $\pi = 0.50$. For each x between 0 and 10, the probability that x of the 10 people selected are female equals

$$P(x) = \frac{10!}{x!(10-x)!}(0.50)^x(0.50)^{10-x}, \quad x = 0, 1, 2, \ldots, 10.$$

For example, the probability that no females are chosen ($x = 0$) is

$$P(0) = \frac{10!}{0!10!}(0.50)^0(0.50)^{10} = (0.50)^{10} = 0.001.$$

(Recall that any number raised to the power of 0 equals 1.) The probability that exactly one female is chosen is

$$P(1) = \frac{10!}{1!9!}(0.50)^1(0.50)^9 = 10(0.50)(0.50)^9 = 0.010.$$

Table 6.9 lists the entire binomial distribution for $n = 10$, $\pi = 0.50$. Binomial probabilities for any n, π, and x value can be found with Internet applets.[11]

TABLE 6.9: The Binomial Distribution for $n = 10$, $\pi = 0.50$. The binomial variable x can take any value between 0 and 10.

x	$P(x)$	x	$P(x)$
0	0.001	6	0.205
1	0.010	7	0.117
2	0.044	8	0.044
3	0.117	9	0.010
4	0.205	10	0.001
5	0.246		

In Table 6.9, the probability is about 0.98 that x falls between 2 and 8, inclusive. The least likely values for x are 0, 1, 9, and 10, which have a combined probability of only 0.022. If the sample were randomly selected, somewhere between about two and eight females would probably be selected. It is especially unlikely that none or 10 would be selected.

The probabilities for females determine those for males. For instance, the probability that 9 of the 10 people selected are male equals the probability that 1 of the 10 selected is female. ∎

PROPERTIES OF THE BINOMIAL DISTRIBUTION

The binomial distribution is perfectly symmetric only when $\pi = 0.50$. In Example 6.10, for instance, since the population proportion of females equals 0.50, $x = 10$ has the same probability as $x = 0$.

The sample proportion $\hat{\pi}$ relates to the binomial variable x by

$$\hat{\pi} = x/n.$$

For example, for $x = 1$ female chosen out of $n = 10$, $\hat{\pi} = 1/10 = 0.10$. The sampling distribution of $\hat{\pi}$ is also symmetric when $\pi = 0.50$. When $\pi \neq 0.50$, the distribution is skewed, the degree of skew increasing as π gets closer to 0 or 1. Figure 6.12 illustrates this. When $\pi = 0.10$, for instance, the sample proportion $\hat{\pi}$ can't fall much below 0.10 since it must be positive, but it could fall considerably above 0.10.

Like the normal distribution, the binomial can be characterized by its mean and standard deviation.

Binomial Mean and Standard Deviation

> The binomial distribution for $x =$ how many of n observations fall in a category having probability π has mean and standard deviation
> $$\mu = n\pi \quad \text{and} \quad \sigma = \sqrt{n\pi(1 - \pi)}.$$

For example, suppose the probability of a female in any one selection for management training is 0.50, as the supermarket chain claims. Then, out of 10 trainees, we expect $\mu = n\pi = 10(0.50) = 5.0$ females.

We've seen (in Sections 5.2 and 6.3) that the sampling distribution of the sample proportion $\hat{\pi}$ has mean π and standard error $\sqrt{\pi(1 - \pi)/n}$. To obtain these formulas,

---

[11] For example, with the *Binomial Distribution* applet at www.pearsonglobaleditions.com/Agresti.
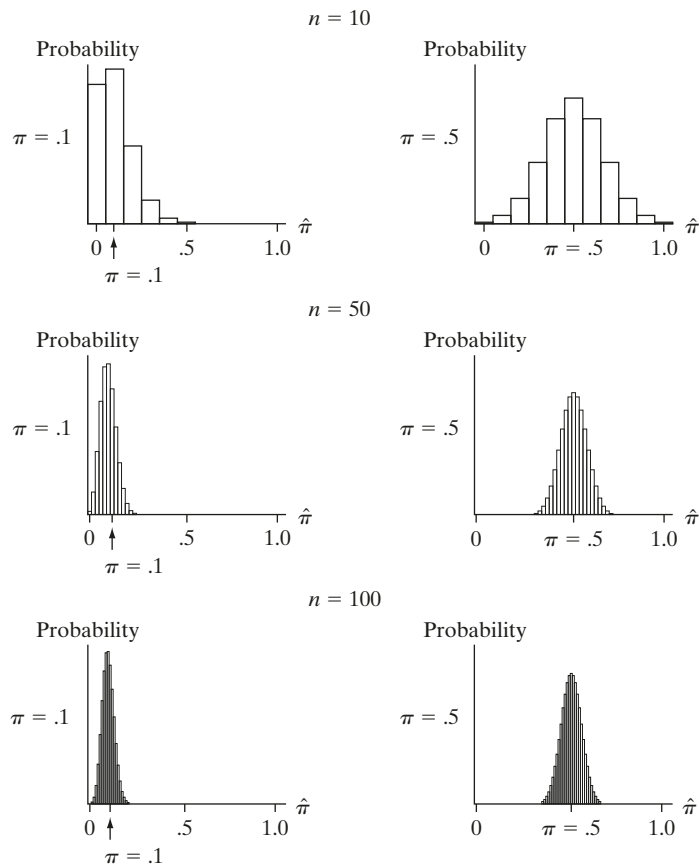
we divide the binomial mean $\mu = n\pi$ and standard deviation $\sigma = \sqrt{n\pi(1-\pi)}$ by n, since $\hat{\pi}$ divides x by n.

**Example
6.11**

How Much Variability Can an Exit Poll Show? Example 4.6 (page 95) discussed an exit poll of 1824 voters for the 2014 California gubernatorial election. Let x denote the number in the exit poll who voted for Jerry Brown. In the population of more than 7 million voters, 60.0% voted for him. If the exit poll was randomly selected, then the binomial distribution for x has $n = 1824$ and $\pi = 0.600$. The distribution is described by

$$\mu = 1824(0.600) = 1094, \quad \sigma = \sqrt{1824(0.600)(0.400)} = 21.$$

Almost certainly, x would fall within three standard deviations of the mean. This is the interval from 1031 to 1157. In fact, in that exit poll, 1104 people of the 1824 sampled reported voting for Brown. ■

THE BINOMIAL TEST

The binomial distribution and the sampling distribution of $\hat{\pi}$ are approximately normal for large n. This approximation is the basis of the large-sample test of Section 6.3. How large is "large"? A guideline is that the expected number of observations should be at least 10 for both categories. For example, if $\pi = 0.50$, we need at least about $n = 20$, because then we expect $20(0.50) = 10$ observations in

one category and $20(1 - 0.50) = 10$ in the other category. For testing $H_0: \pi = 0.90$ or $H_0: \pi = 0.10$, we need $n \geq 100$. The sample size requirement reflects the fact that a symmetric bell shape for the sampling distribution of $\hat{\pi}$ requires larger sample sizes when π is near 0 or 1 than when π is near 0.50.

If the sample size is not large enough to use the normal test, we can use the binomial distribution directly. Refer to Example 6.10 (page 178) about potential gender discrimination. For random sampling, the probability π that a person selected for management training is female equals 0.50. If there is bias against females, then $\pi < 0.50$. Thus, we can test the company's claim of random sampling by testing

$$H_0: \pi = 0.50 \qquad \text{versus} \qquad H_a: \pi < 0.50.$$

Of the 10 employees chosen for management training, let x denote the number of women. Under H_0, the sampling distribution of x is the binomial distribution with $n = 10$ and $\pi = 0.50$. Table 6.9 tabulated it. As in Example 6.1 (page 151), suppose $x = 1$. The P-value is then the left-tail probability of an outcome at least this extreme; that is, $x = 1$ or 0. From Table 6.9, the P-value is

$$P = P(0) + P(1) = 0.001 + 0.010 = 0.011.$$

If the company selected trainees randomly, the probability of choosing one or fewer females is only 0.011. This result provides evidence against the null hypothesis of a random selection process. We can reject H_0 for $\alpha = 0.05$, though not for $\alpha = 0.010$.

Even if we suspect bias in a particular direction, the most even-handed way to perform a test uses a two-sided alternative. For $H_a: \pi \neq 0.50$, the P-value is $2(0.011) = 0.022$. This is a two-tail probability of the outcome that one or fewer *of either sex* is selected. Figure 6.13 shows the formation of this P-value.

FIGURE 6.13: Calculation of P-Value in Testing H_0: $\pi = 0.50$ against H_a: $\pi \neq 0.50$, when $n = 10$ and $x = 1$
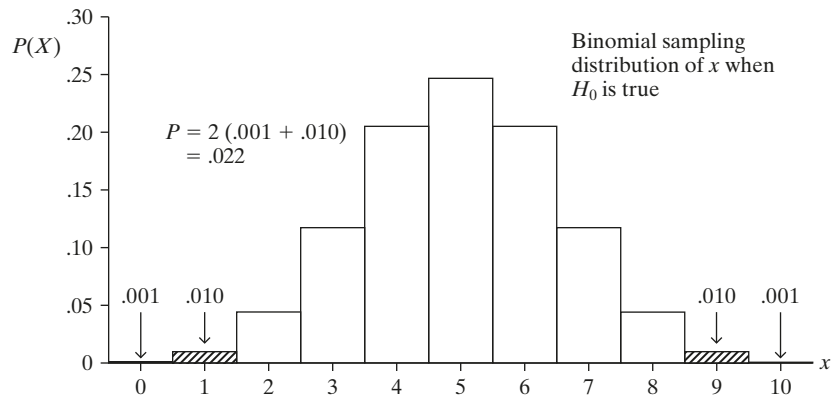
The assumptions for the binomial test are the three conditions for the binomial distribution. Here, the conditions are satisfied. Each observation has only two possible outcomes, female or male. The probability of each outcome is the same for each selection, 0.50 for selecting a female and 0.50 for selecting a male (under H_0). For random sampling, the outcome of any one selection does not depend on any other one.

6.8 Chapter Summary

Chapter 5 and this chapter have introduced two methods for using sample data to make inferences about populations—*confidence intervals* and *significance tests*.

A confidence interval provides a range of plausible values for a parameter. A significance test judges whether a particular value for the parameter is plausible. Both methods utilize the sampling distribution of the estimator of the parameter.

Significance tests have five parts:

1. ***Assumptions:***
   - Tests for *means* apply with quantitative variables whereas tests for *proportions* apply with categorical variables.
   - Tests assume *randomization*, such as a random sample.
   - Large-sample tests about proportions require no assumption about the population distribution, because the Central Limit Theorem implies approximate normality of the sampling distribution of the sample proportion.
   - Tests for means use the t distribution, which assumes the population distribution is normal. In practice, two-sided tests (like confidence intervals) are *robust* to violations of the normality assumption.

2. ***Null and alternative hypotheses*** about the parameter: The null hypothesis has the form $H_0: \mu = \mu_0$ for a mean and $H_0: \pi = \pi_0$ for a proportion. Here, μ_0 and π_0 denote values hypothesized for the parameters, such as 0.50 in $H_0: \pi = 0.50$. The most common alternative hypothesis is *two sided*, such as $H_a: \pi \neq 0.50$. Hypotheses such as $H_a: \pi > 0.50$ and $H_a: \pi < 0.50$ are *one sided*, designed to detect departures from H_0 in a particular direction.

3. A ***test statistic*** describes how far the point estimate falls from the H_0 value. The z statistic for proportions and t statistic for means measure the number of standard errors that the point estimate ($\hat{\pi}$ or $\bar{y}$) falls from the H_0 value.

4. The ***P-value*** describes the evidence about H_0 in probability form.
   - We calculate the *P*-value by presuming that H_0 is true. It equals the probability that the test statistic equals the observed value or a value even more extreme.
   - The "more extreme" results are determined by the alternative hypothesis. For two-sided H_a, the *P*-value is a two-tail probability.
   - Small *P*-values result when the point estimate falls far from the H_0 value, so that the test statistic is large. When the *P*-value is small, it would be unusual to observe such data if H_0 were true. The smaller the *P*-value, the stronger the evidence against H_0.

5. A ***conclusion*** based on the sample evidence about H_0: We report and interpret the *P*-value. When we need to make a decision, we reject H_0 when the *P*-value is less than or equal to a fixed α-level (such as $\alpha = 0.05$). Otherwise, we cannot reject H_0.

When we make a decision, two types of errors can occur.

- When H_0 is true, a Type I error results if we reject it.

- When H_0 is false, a Type II error results if we fail to reject it.

The choice of α, the cutoff point for the *P*-value in making a decision, equals P(Type I error). Normally, we choose small values such as $\alpha = 0.05$ or 0.01. For fixed α, P(Type II error) decreases as the distance increases between the parameter and the H_0 value or as the sample size increases.

Table 6.10 summarizes the five parts of the tests this chapter presented.

Sample size is a critical factor in both estimation and significance tests. With small sample sizes, confidence intervals are wide, making estimation imprecise.

TABLE 6.10: Summary of Significance Tests for Means and Proportions

Parameter	Mean	Proportion
1. Assumptions	Random sample, quantitative variable, normal population	Random sample, categorical variable, null expected counts at least 10
2. Hypotheses	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$ $H_a: \mu > \mu_0$ $H_a: \mu < \mu_0$	$H_0: \pi = \pi_0$ $H_a: \pi \neq \pi_0$ $H_a: \pi > \pi_0$ $H_a: \pi < \pi_0$
3. Test statistic	$t = \dfrac{\bar{y} - \mu_0}{se}$ with $se = \frac{s}{\sqrt{n}}$, $df = n - 1$	$z = \dfrac{\hat{\pi} - \pi_0}{se_0}$ with $se_0 = \sqrt{\pi_0(1 - \pi_0)/n}$
4. P-value	Two-tail probability in sampling distribution for two-sided test ($H_0: \mu \neq \mu_0$ or $H_a: \pi \neq \pi_0$); one-tail probability for one-sided test	
5. Conclusion	Reject H_0 if P-value $\leq \alpha$-level such as 0.05	

Small sample sizes also make it difficult to reject false null hypotheses unless the true parameter value is far from the null hypothesis value. P(Type II error) may be high for parameter values of interest.

This chapter presented significance tests about a single parameter for a single variable. In practice, it is usually artificial to have a particular fixed number for the H_0 value of a parameter. One of the few times this happens is when the response score results from taking a difference of two values, such as the change in weight in Example 6.4 (page 160). In that case, $\mu_0 = 0$ is a natural baseline. Significance tests much more commonly refer to comparisons of means for two samples than to a fixed value of a parameter for a single sample. The next chapter shows how to compare means or proportions for two groups.

Exercises

Practicing the Basics

6.1. For (a)–(c), is it a null hypothesis, or an alternative hypothesis?

(a) In India, the proportion of teenagers who watch horror movies equals 0.45.

(b) The proportion of all Indian school students with access to the Internet is now more than 0.30 (the value it was 5 years ago).

(c) The mean IQ of newly recruited workers at ABC Inc. is less than 100.

(d) Introducing notation for a parameter, state the hypotheses in (a)–(c) in terms of the parameter values.

6.2. You want to know whether employees in a company consider the ideal average working hours to be 10, or higher or lower than that.

(a) Define notation and state the null and alternative hypotheses for studying this.

(b) For responses in a recent survey to the question "What do you think are the ideal average working hours?" software shows results:

```
-------------------------------------------------
Test of mu = 10 vs mu not = 10

Variable   n     Mean   StDev  SE Mean   T    P-value
Hours     1500   10.5   0.750  0.036    25.82  0.0000
-------------------------------------------------
```

Report the test statistic value, and show how it was obtained from other values reported in the table.

(c) Explain what the P-value represents, and interpret its value.

6.3. For a test of $H_0: \mu = 0$ against $H_a: \mu \neq 0$ with $n = 1500$, the t test statistic equals 0.92.

(a) Find the P-value, and interpret it.

(b) Suppose $t = -3.60$ rather than 0.92. Find the P-value. Does this provide stronger, or weaker, evidence against the null hypothesis? Explain.

(c) When $t = 0.92$, find the P-value for (i) $H_a: \mu > 0$, (ii) $H_a: \mu < 0$.

6.4. The P-value for a test about a mean with $n = 31$ is $P = 0.01$.

(a) Find the t test statistic value that has this P-value for (i) H_a: $\mu \neq 0$, (ii) H_a: $\mu > 0$, (iii) H_a: $\mu < 0$.

(b) Does this P-value provide stronger, or weaker, evidence against the null hypothesis than $P = 0.001$? Explain.

6.5. Find and interpret the P-value for testing H_0: $\mu = 50$ against H_0: $\mu \neq 50$ if a sample has

(a) $n = 200$, $\bar{y} = 53$, and $s = 20$.

(b) $n = 800$, $\bar{y} = 53$, and $s = 20$. Comment on the effect of n on the results of a significance test.

6.6. Example 6.4 (page 160) described a study about therapies for teenage girls suffering from anorexia. For the 17 girls who received the family therapy, the changes in weight were
11, 11, 6, 9, 14, −3, 0, 7, 22, −5, −4, 13, 13, 9, 4, 6, 11.
Some Stata output (edited) for the data shows

```
-----------------------------------------------
Variable |  Obs      Mean   Std. Err.  Std. Dev.
  change |   17   7.294118      ??     7.183007

    mean = mean(change)                t = ??
Ho: mean = 0                 degrees of freedom = ??
   Ha: mean != 0                 Ha: mean > 0
 Pr(|T| > |t|) = 0.0007    Pr(T > t) = ??
-----------------------------------------------
```

Fill in the four missing results.

6.7. According to a union agreement, the mean income for all managers in a large manufacturing company equals \$2000 per month. A representative of a graduate group decides to analyze whether the mean income μ for graduate managers matches this norm. For a random sample of 10 graduate employees, $\bar{y} = \$1500$ and $s = 100$.

(a) Test whether the mean income of graduate employees differs from \$2000 per month. Include assumptions, hypotheses, test statistic, and P-value. Interpret the result.

(b) Report the P-value for H_a: $\mu < 2000$. Interpret.

(c) Report and interpret the P-value for H_a: $\mu > 2000$. (Hint: The P-values for the two possible one-sided tests must sum to 1.)

6.8. By law, a chemical plant can discharge no more than 400 units of chemical waste per hour, on an average, into a neighboring river. Based on other infractions they have noticed, an environmental action group believes this limit is being exceeded. Monitoring the plant is expensive, and a random sample of five hours is taken over a period of a week. Software reports

```
-----------------------------------------------
Variable  No. Cases   Mean   StDev   SE of Mean
   WASTE       5      960.0   500.0     150.0
-----------------------------------------------
```

(a) Test whether the mean discharge equals 400 units per hour against the alternative that the limit is being exceeded. Find the P-value, and interpret.

(b) Explain why the test may be highly approximate or even invalid if the population distribution of discharge is far from normal.

(c) Explain how your one-sided analysis implicitly tests the broader null hypothesis that $\mu \leq 400$.

6.9. In response to the statement "An effective leader must be able to feel the emotions of his group-members," the response categories (Strongly agree, Agree, Disagree, Strongly disagree) had counts (103, 375, 425, 97) for responses in an Emotional Quotient Survey. To treat this ordinal variable as quantitative, we assign scores to the categories. For the scores (2, 1, −1, −2), which treat the distance between Agree and Disagree as twice the distance between Strongly agree and Agree or between Disagree and Strongly disagree, software reports

```
-----------------------------------------------
   n      Mean    Std Dev    Std Err
 1000    -.048    1.543      0.0413
-----------------------------------------------
```

(a) Set up null and alternative hypotheses to test whether the population mean response differs from the neutral value, 0.

(b) Find the test statistic and P-value. Interpret, and make a decision about H_0, using $\alpha = 0.05$.

(c) Based on (b), can you "accept" H_0: $\mu = 0$? Why or why not?

(d) Construct a 95% confidence interval for μ. Show the correspondence between whether 0 falls in the interval and the decision about H_0.

6.10. In Example 6.2 on political ideology (page 157), suppose we use the scores (−3, −2, −1, 0, 1, 2, 3) instead of (1, 2, 3, 4, 5, 6, 7). We then test H_0: $\mu = 0$. Explain the effect of the change in scores on **(a)** the sample mean and standard deviation, **(b)** the test statistic, **(c)** the P-value and interpretation.

6.11. Results of 95% confidence intervals for means are consistent with results of two-sided tests with which α-level? Explain the connection.

6.12. For a test of H_0: $\pi = 0.45$, the z test statistic equals 0.80.

(a) Find the P-value for H_a: $\pi > 0.45$.

(b) Find the P-value for H_a: $\pi \neq 0.45$.

(c) Find the P-value for H_a: $\pi < 0.45$.

(d) Do any of the P-values in (a), (b), or (c) give strong evidence against H_0? Explain.

6.13. For a test of H_0: $\pi = 0.60$, the sample proportion is 0.45 with $n = 125$.

(a) Show that the test statistic is $z = −3.423$.

(b) Find and interpret the P-value for H_a: $\pi < 0.60$.

(c) For a significance level of $\alpha = 0.01$, what decision do you make?

(d) If the decision in (c) was an error, what type of error was it? What could you do to reduce the chance of that type of error?

6.14. Illegal trafficking of some items in a country was banned by its government in 2007. Is this supported by a majority, or a minority, of the population of the country? In a global poll conducted by an international agency for a reputed newspaper in January 2010 involving 900 residents of the country that asked whether the ban should stand or be repealed, 55% supported the banning. Let π denote the population proportion of the country that supports banning. For testing $H_0: \pi = 0.5$ against $H_a: \pi \neq 0.5$,

(a) Find the standard error, and interpret.

(b) Find the test statistic, and interpret.

(c) Find the *P*-value, and interpret in context.

6.15. When the 2015 Heritage Conservation Survey asked, "Would you be willing to pay higher taxes in order to protect the heritage buildings?" 429 people answered *yes* and 726 answered *no*. Software shows the following results to analyze whether a majority or a minority of people would answer *yes*:

```
---------------------------------------------------
Test of proportion = 0.5 vs not = 0.5

   n    Sample prop     95% CI      z-Value  P-Value
 1155     0.453     (0.424, 0.481)   -3.21    0.0007
---------------------------------------------------
```

(a) Specify the hypotheses that are tested.

(b) Report and interpret the test statistic value.

(c) Report and interpret the *P*-value as a probability.

(d) Explain an advantage of the confidence interval shown over the significance test.

6.16. The International EQ Survey asked, "Indicate the importance of honesty in your life." Of the people sampled in a country, Stata reports the following output for the proportion who answered "very important" or "rather important" (instead of "not very important" or "not at all important").

```
-------------------------------
One-sample test of proportion:
Number of obs = 2000
Variable  |   Mean   Std. Err.
Honesty   |   .35     .0099288
p = proportion(honesty)
z = -13.4164
Ho: p = 0.5
Ha: p < 0.5
Pr(Z < z) = 0.0000
Ha: p ! =  0.5
Pr(|Z| > |z|) = 0.0000
Ha: p > 0.5
Pr(Z > z) = 1.0000
-------------------------------
```

(a) Explain how to interpret all results shown.

(b) By contrast, the results for Country B had a sample proportion of 0.88 who answered "very important" or "rather important." Here are the test results:

```
---------------------
Ha: p < 0.5
Pr(Z < z) = 1.0000
Ha: p != 0.5
Pr(|Z| > |z|) = 0.0000
Ha: p > 0.5
Pr(Z > z) = 0.0000
---------------------
```

6.17. In the scientific test of astrology discussed in Example 6.9 (page 175), the astrologers were correct with 40 of their 116 predictions. Test $H_0: \pi = 1/3$ against $H_a: \pi > 1/3$. Find the *P*-value, make a decision using $\alpha = 0.05$, and interpret.

6.18. The previous exercise analyzed whether astrologers could predict the correct personality chart for a given horoscope better than by random guessing. In the words of that study, what would be a **(a)** Type I error? **(b)** Type II error?

6.19. A secretary election in a labor union has two candidates. Exactly half the members currently prefer each candidate.

(a) For a random sample of 500 voters, 290 voted for a particular candidate. Are you willing to predict the winner? Why?

(b) For a random sample of 50 voters, 29 voted for a particular candidate. Would you be willing to predict the winner? Why? (The sample proportion is the same in (a) and (b), but the sample sizes differ.)

6.20. The authorship of an old document is in doubt. An academician hypothesizes that the author was a sociologist named Olivia Munn. Upon a thorough investigation of Munn's known works, it is observed that one unusual feature of her writing was that she consistently began 5% of her sentences with the word *therefore*. To test the academician's hypothesis, it is decided to count the number of sentences in the disputed document that begin with *therefore*. Out of the 400 sentences, none do. Let π denote the probability that any one sentence written by the unknown author of the document begins with *therefore*. Test H_0: $\pi = 0.05$ against H_a: $\pi \neq 0.05$. What assumptions are needed for your conclusion to be valid?[12]

6.21. A multiple-choice test question has four possible responses. The question is difficult, with none of the four responses being obviously wrong, yet with only one correct answer. It first occurs on an exam taken by 500 students. Test whether more people answer the question correctly

---

[12] F. Mosteller and D. L. Wallace conducted this type of investigation to determine whether Alexander Hamilton or James Madison authored 12 of 12. See www.worldvaluessurvey.org/WVSOnline.jsp. MAIN, Vol. 19 (17), p. 174, Chapter 6, "Statistical Inference: Significance Tests the Federalist Papers." See Frederick Mosteller and David L. Wallace, *Inference and Disputed Authorship: The Federalist* (Reading Addison-Wesley, 1964).

than would be expected just due to chance (that is, if everyone randomly guessed the correct answer).

(a) Set up the hypotheses for the test.

(b) Of the 500 students, 175 correctly answer the question. Find the P-value, and interpret.

6.22. Example 6.4 (page 160) tested a therapy for anorexia, using H_0: $\mu = 0$ and H_a: $\mu > 0$ about the population mean weight change.

(a) In the words of that example, what is a (i) Type I error? (ii) Type II error?

(b) The P-value was 0.018. If the decision for $\alpha = 0.05$ were in error, what type of error is it?

(c) Suppose instead $\alpha = 0.01$. What decision would you make? If it is in error, what type of error is it?

6.23. Nora and Ayush separately conduct studies to test H_0: $\mu = 400$ against H_A: $\mu \neq 400$, each with $n = 1200$. Nora gets $\bar{y} = 423.35$ with $se = 12.0$. Ayush gets $\bar{y} = 434.2$, with $se = 17.0$

(a) Show that $t = 1.95$ and P-value $= 0.051$ for Nora. Show that $t = 2.01$ and P-value $= 0.044$ for Ayush.

(b) Using $\alpha = 0.050$, for each study indicate whether the result is "statistically significant."

(c) Using this example, explain the misleading aspects of reporting the result of a test as "$P \leq 0.05$" versus "$P > 0.05$," or as "Reject H_0" versus "Do not reject H_0," without reporting the actual P-value.

6.24. Nora and Oliva separately conduct studies to test H_0: $\pi = 0.45$ against H_0: $\pi \neq 0.45$, each with $n = 485$. Nora gets $\hat{\pi} = 240/485 = 0.494845$ Oliva gets $\hat{\pi} = 239/485 = 0.492784$.

(a) Show that $z = 1.99$ and P-value $= 0.046$ for Nora. Show that $z = 1.89$ and P-value $= 0.058$ for Oliva.

(b) Using $\alpha = 0.05$, indicate in each case whether the result is "statistically significant." Interpret.

(c) Use this example to explain why important information is lost by reporting the result of a test as "P-value ≤ 0.05" versus "P-value > 0.05," or as "Reject H_0" versus "Do not reject H_0," without reporting the P-value.

(d) The 95% confidence interval for π is (0.451, 0.539) for Nora and (0.449, 0.537) for Oliva. Explain how this method shows that, in practical terms, the two studies had very similar results.

6.25. A study considers whether the mean score μ on an MBA entrance exam for students in 2017 was any different from the mean of 400 for students in 1966. Test H_0: $\mu = 400$ against H_a: $\mu \neq 400$ if for a nationwide random sample of 9,000 students who took the exam in 2017, $\bar{y} = 397$ and $s = 90$. Show that the result is highly significant statistically, but not practically significant.

6.26. An academic survey on employees at their job market entry level reported the presence of a notable gap between graduate and postgraduate employees in their

confidence that tenure rules are clear, with postgraduates feeling more confident. The 4500 staff members in the survey were asked to evaluate policies on a scale of 1 to 5 (very unclear to very clear). The mean response about the criteria for tenure was 4.91 for graduates and 4.95 for postgraduates, which was indicated to meet the test for statistical significance. Use this study to explain the distinction between statistical significance and practical significance.

6.27. Refer to Example 6.8 on "medical discoveries" (page 173). Using a tree diagram, approximate P(Type I error) under the assumption that a true effect exists 20% of the time and that P(Type II error) $= 0.30$.

6.28. A decision is planned in a test of H_0: $\mu = 0$ against H_a: $\mu > 0$, using $\alpha = 0.05$. If $\mu = 5$, P(Type II error) $= 0.17$.

(a) Explain the meaning of this last sentence.

(b) If the test used $\alpha = 0.01$, would P(Type II error) be less than, equal to, or greater than 0.17? Explain.

(c) If $\mu = 10$, would P(Type II error) be less than, equal to, or greater than 0.17? Explain.

6.29. Let π denote the proportion of people under stress who respond positively to treatment. A test is conducted of H_0: $\pi = 0.60$ against H_a: $\pi > 0.60$, for $n = 30$, using $\alpha = 0.05$.

(a) Find the region of sample proportion values for which H_0 is rejected.

(b) Suppose that $\pi = 0.70$. Find P (Type II error).

6.30. A follow-up to the anorexia study of Example 6.4 plans a test of H_0: $\mu = 0$ against H_a: $\mu > 0$, for the mean weight change μ for a new therapy, predicting that $\mu = 10$. The test uses $\alpha = 0.05$. For $n = 30$, suppose the standard deviation is 18. Find P(Type II error) if $\mu = 10$ by showing **(a)** a test statistic of $t = 1.699$ has a P-value of 0.05, **(b)** we fail to reject H_0 if $\bar{y} < 5.6$, **(c)** this happens if $\bar{y}$ falls more than 1.33 standard errors below 10, **(d)** this happens with probability about 0.10.

6.31. Refer to the previous exercise.

(a) Find P(Type II error) if $\mu = 5$. How does P(Type II error) depend on the value of μ?

(b) Find P(Type II error) if $\mu = 10$ and $\alpha = 0.01$. How does P(Type II error) depend on α?

(c) How does P(Type II error) depend on n?

6.32. An employee list contains the names of all employees who may be deputed with additional charges. The proportion of women on the list is 0.56. A group of size 15 is selected at random from the list. None selected are women.

(a) Find the probability of selecting 0 women.

(b) Test the hypothesis that the selections are random against the alternative of bias against women. Report the P-value, and interpret.

6.33. A person claiming to possess extrasensory perception (ESP) says she can guess more often than not the outcome of a flip of a balanced coin in another room, not visible to her.

(a) Introduce appropriate notation, and state hypotheses for testing her claim.

(b) Of five coin flips, she guesses the correct result four times. Find the *P*-value and interpret.

6.34. In a newspaper exit poll of 1500 voters in the 2015 general election in an emerging nation, let $x =$ number of voters in the exit poll who voted for a candidate, Mr. X.

(a) Explain why this scenario would seem to satisfy the three conditions needed to use the binomial distribution.

(b) If the population proportion voting for him had been 0.30, find the mean and standard deviation of the probability distribution of x.

(c) For (b), using the normal distribution approximation, give an interval in which x would almost certainly fall.

(d) Now, actually the exit poll had $x = 550$. Explain how you could make an inference about whether π is above or below 0.30.

6.35. In a given year in a developing country, the probability of death in any earthquake equals 0.0004 for highlanders and 0.0003 for plain land dwellers (*Statistical Abstract of the Developing Countries*).

(a) In a town having 2 million highland dwellers, find the mean and standard deviation of $x =$ number of deaths from any severe earthquake. State the assumptions for these to be valid.

(b) Would it be surprising if $x = 10$? Explain. (*Hint*: How many standard deviations is 10 from the expected value?)

(c) Based on the normal approximation to the binomial, find an interval within which x has probability 0.95 of occurring.

Concepts and Applications

6.36. You can use the *Errors and Power* applet at www.pearsonglobaleditions.com/Agresti to investigate the performance of significance tests, to illustrate their long-run behavior when used for many samples. For significance tests, set the null hypothesis as $H_0: \pi = 0.50$ for a one-sided test with $H_a: \pi > 0.50$ and sample size 50, and set *P*(Type I error) $= \alpha = 0.05$. The applet shows the null sampling distribution of $\hat{\pi}$ and the actual sampling distribution of $\hat{\pi}$ for various true values of π. Click on *Show Type II error*, and it also displays *P*(Type II error), which is the probability of failing to reject H_0 even though it is false.

(a) Set the true value of the proportion to be (i) 0.60, (ii) 0.70, (iii) 0.80. What happens to *P*(Type II error) as π gets farther from the H_0 value?

(b) Set the true value of the proportion to be (i) 0.53, (ii) 0.52, (iii) 0.51. What value does *P*(Type II error) approach as π gets closer to the H_0 value?

(c) Fix the true value of the proportion to be 0.60. Show how *P*(Type II error) changes for n equal to (i) 50, (ii) 100, (iii) 200.

(d) Summarize how *P*(Type II error) depends on π (for fixed n) and on n (for fixed π).

6.37. Refer to the `Students` data file (Exercise 1.11 on page 21).

(a) Test whether the population mean political ideology differs from 4.0. Report the *P*-value, and interpret.

(b) Test whether the proportion favoring legalized abortion equals, or differs, from 0.50. Report the *P*-value, and interpret.

6.38. Refer to the data file your class created in Exercise 1.12 (page 22). For variables chosen by your instructor, state a research question and conduct inferential statistical analyses. Also, use graphical and numerical methods presented earlier in this text to describe the data and, if necessary, to check assumptions for your analyses. Prepare a report summarizing and interpreting your findings.

6.39. A study considered the effects of a special class designed to improve children's verbal skills. Each child took a verbal skills test before and after attending the class for three weeks. Let $y =$ second exam score $-$ first exam score. The scores on y for a random sample of four children having learning problems were 3, 7, 3, 3. Conduct inferential statistical methods to determine whether the class has a positive effect. Summarize your analyses and interpretations in a short report. (*Note*: The scores could improve merely from the students feeling more comfortable with the testing process. A more appropriate design would also administer the exam twice to a control group that does not take the special class, comparing the changes for the experimental and control groups using methods of Chapter 7.)

6.40. Sixty students in a class at the University of Florida made blinded evaluations of pairs of dark chocolates. For 60 comparisons of two chocolates, the first one was preferred 41 times. In the population that this sample represents, is there a strong evidence that a majority prefers one of the chocolates? Refer to the following software output:

```
-------------------------------------------------
Test of parameter = 0.50 vs not = 0.50

  n   Sample prop    95.0% CI    z-Value  P-Value
  60    0.6833    (0.557, 0.810)   2.84    0.002
-------------------------------------------------
```

Explain how each result on this output was obtained. Summarize results in a way that would be clear to someone who is not familiar with statistical inference.

6.41. Researchers have studied possible abuse by security officers in treating owner-drivers of small cars at crossroads. An NGO studied whether owner-drivers of small cars were more likely than others in the population to be

targeted by officers for traffic stops. They considered 360 car stops due to security reasons in a chosen city during one week. Of those, 234 of the drivers were owners of small cars. At that time, 45% of the city's population was owner of small cars. Does the number of owner-drivers of small cars stopped give strong evidence of possible bias, being higher than you'd expect if we take into account ordinary random variation? Explain your reasoning in a report of at most 250 words.

6.42. An experiment with 26 students in an Israeli classroom consisted of giving everyone a lottery ticket and then later asking if they would be willing to exchange their ticket for another one, plus a small monetary incentive. Only 7 students agreed to the exchange. In a separate experiment, 31 students were given a new pen and then later asked to exchange it for another pen and a small monetary incentive. All 31 agreed.[13] Conduct inferential statistical methods to analyze the data. Summarize your analyses and interpretations in a short report.

6.43. Ideally, results of a statistical analysis should not depend greatly on a single observation. To check this, it's a good idea to conduct a *sensitivity study*: Redo the analysis after deleting an outlier from the data set or changing its value to a more typical value, and check whether results change much. For the anorexia data of Example 6.4 (available in the `Anorexia_CB` data file at the text website), the weight change of 20.9 pounds was a severe outlier. Suppose this observation was actually 2.9 pounds but was incorrectly recorded. Redo the one-sided test of Example 6.4 (page 160), and summarize the influence of that observation.

6.44. In making a decision in a test, a researcher worries about the possibility of rejecting H_0 when it is actually true. Explain how to control the probability of this type of error.

6.45. Consider the analogy between making a decision in a test and making a decision about the innocence or guilt of a defendant in a criminal trial.

(a) Explain what Type I and Type II errors are in the trial.

(b) Explain intuitively why decreasing P(Type I error) increases P(Type II error).

(c) Defendants are convicted if the jury finds them to be guilty "beyond a reasonable doubt." A jury interprets this to mean that if the defendant is innocent, the probability of being found guilty should be only 1 in a billion. Describe any disadvantages this strategy has.

6.46. Medical tests for diagnosing conditions such as breast cancer are fallible, just like decisions in significance tests. Identify (H_0 true, H_0 false) with disease (absent, present), and (Reject H_0, Do not reject H_0) with diagnostic test (positive, negative), where a positive diagno-

sis means that the test predicts that the disease is present. Explain the difference between Type I and Type II errors in this context. Explain why decreasing P(Type I error) increases P(Type II error), in this context.

6.47. An article in a management journal that deals with changes in social skill over time states, "For these subjects, the difference in their mean performances on a scale of skill between age 25 and the current survey was significant ($P < 0.01$)."

(a) Explain what it means for the result to be "significant."

(b) Explain why it would have been more informative if the authors provided the actual P-value rather than merely indicating that it is below 0.01. What other information might they have provided?

6.48. A popular psychology analysis journal reports that "no significant difference was found between Emotional Intelligence Quotient for the academicians and service holders ($P = 0.85$)." Hence, the population emotional intelligence quotients would be identical for academicians and service holders—true or false?

6.49. Harry conducted a significance test using Stata. The output reports a P-value of 0.3598621. In summarizing analyses in an article, explain why it would be sensible for Harry to report $P = 0.36$ rather than $P = 0.3598621$.

6.50. A social study conducts 1000 significance tests. Of these, 50 are significant at the 0.05 level. The researchers write a report stressing only the 50 "significant" results, not mentioning the other 950 tests that were "not significant." Explain what is misleading about their report.

6.51. Some journals have a policy of publishing research results only if they achieve statistical significance at the 0.05 α-level.

(a) Explain the dangers of this.

(b) When medical stories in the mass media report supposed large dangers or benefits of certain agents (e.g., coffee drinking, fiber in cereal), later research often suggests that the effects are smaller than first believed, or may not even exist. Explain why.

Select the correct response(s) in Exercises 6.52–6.56. (More than one may be correct.)

6.52. We analyze whether the true mean daily discharge of waste by a chemical plant exceeds its claim of 1500 units. For the decision in the one-sided test using $\alpha = 0.01$,

(a) If the plant is not exceeding the limit, but actually $\mu = 1500$, there is only a 1% chance that we will conclude that they are not exceeding the limit.

(b) If the plant is exceeding the limit, there is only a 1% chance that we will conclude that they are exceeding the limit.

---

[13] M. Bar-Hillel and E. Neter, *Journal of Personality and Social Psychology*, Vol. 70 (1996), pp. 17–27.

(c) The probability that the sample mean equals exactly the observed value would equal 0.01 if H_0 were true.

(d) If we accept H_a, the probability that H_0 is actually true is 0.01.

(e) All of the above.

6.53. The P-value for testing H_0: $\mu = 50$ against H_a: $\mu \neq 50$ is $P = 0.003$. This implies that

(a) There is strong evidence that $\mu \neq 50$.

(b) There is strong evidence that $\mu = 50$.

(c) There is strong evidence that $\mu > 50$.

(d) There is strong evidence that $\mu < 50$.

(e) If μ were equal to 50, it would be unusual to obtain data such as those observed.

6.54. In the previous exercise, suppose the test statistic $z = -5.48$.

(a) There is strong evidence that $\mu > 50$.

(b) There is strong evidence that $\mu < 50$.

(c) There is strong evidence that $\mu = 50$.

6.55. A 99% confidence interval for μ is (99, 116). Which two statements about significance tests for the same data are correct?

(a) In testing H_0: $\mu = 110$ against H_a: $\mu \neq 110$, $P < 0.01$.

(b) In testing H_0: $\mu = 110$ against H_a: $\mu \neq 110$, $P > 0.01$.

(c) In testing H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$, $P > 0.01$ if μ_0 is any of the numbers outside the confidence interval.

(d) In testing H_0: $\mu = \mu_0$ against H_a: $\mu \neq \mu_0$, $P > 0.01$ if μ_0 is any of the numbers inside the confidence interval.

6.56. Let β denote P (Type II error). For an $\alpha = 0.01$-level test of H_0: $\mu = 0$ against H_a: $\mu > 0$ with $n = 50$ observations, $\beta = 0.57$ at $\mu = 6$. Then,

(a) At $\mu = 6$, $\beta > 0.57$.

(b) If $n = 80$, then at $\mu = 6$, $\beta > 0.57$.

(c) If $\alpha = 0.01$, then at $\mu = 6$, $\beta > 0.57$

(d) The power of the test is 0.43 at $\mu = 6$.

(e) (d) must be false, because necessarily $\alpha + \beta = 1$.

6.57. Answer true or false for each of the following, and explain your answer:

(a) P (Type II error) always increases with P (Type I error).

(b) If we accept H_A using $\alpha = 0.01$, then we also accept it using $\alpha = 0.05$.

(c) The P-value is the probability that H_A is false. (*Hint*: Do we find probabilities about variables and their statistics, or about parameters?)

(d) A survey conducted by an NGO reports $P = 0.082$ for testing H_0: $\mu = 0$ against H_a: $\mu \neq 0$. If the surveyors had instead reported a 99% confidence interval for μ, then the interval would have contained 0, and readers could have better judged just which values are plausible for μ.

6.58. Explain the difference between one-sided and two-sided alternative hypotheses, and explain how this affects calculation of the P-value.

6.59. Explain why the terminology "do not reject H_0" is preferable to "accept H_0."

6.60. A researcher plans to survey people in your locality to study whether a majority feel that the cap on foreign direct investment should be reduced. The researcher has never studied statistics. How would you explain to him the concepts of **(a)** null and alternative hypotheses, **(b)** P-value, **(c)** α-level, **(d)** Type II error?

6.61. A random sample of size 50 has $\bar{y} = 130$. The P-value for testing H_0: $\mu = 120$ against H_a: $\mu \neq 120$ is $P = 0.069$. Explain what is incorrect about each of the following interpretations of this P-value, and provide a proper interpretation.

(a) The probability that the null hypothesis is correct equals 0.069.

(b) The probability that $\bar{y} = 130$ if H_0 is true equals 0.069.

(c) If in fact $\mu \neq 120$, the probability that the data would at least be as contradictory to H_0 as the observed data equals 0.069.

(d) The probability of Type I error equals 0.069.

(e) We can accept H_0 at the $\alpha = 0.06$ level.

(f) We can reject H_0 at the $\alpha = 0.06$ level.

6.62.*Refer to the previous exercise and the P-value of 0.069.

(a) Explain why the P-value is the smallest α-level at which H_0 can be rejected; that is, P equals the smallest level at which the data are significant.

(b) Refer to the correspondence between results of confidence intervals and two-sided tests. When the P-value is 0.069, explain why the 93.1% confidence interval is the narrowest confidence interval for μ that contains $\mu_0 = 120$.

6.63.*A surveyor conducts a significance test every time he analyzes a new survey data. Over time, he conducts 200 tests.

(a) Suppose H_0 is true in every case. What is the distribution of the number of times she rejects H_0 at the 0.01 level?

(b) Suppose she rejects H_0 in seven of the tests. Is it plausible that H_0 is correct in every case? Explain.

6.64.*Each year in an archive at New Delhi, the director estimates the mean number of times the archived documents have been checked out in the previous year. To do this, the director randomly samples computer records for 500 books and forms a 95% confidence interval for the mean. This has been done for 10 years. Find the probability that **(a)** all the confidence intervals contain the true

means; **(b)** at least one confidence interval does not contain the true mean.

6.65.*Refer to Example 5.8 on 0 vegetarians in a sample of $n = 20$ and Exercises 5.77 and 5.77.

(a) At the *Inference for a Proportion* applet at www.pearsonglobaleditions.com/Agresti, enter 0 successes for $n = 20$. Explain why the reported standard error values differ when you click on *Confidence Interval* and when you click on *Significance Test* and enter a null hypothesis value such as 0.50 for π.

(b) To test H_0: $\pi = 0.50$, show what happens if you find the z test statistic using the $se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$ formula for confidence intervals. Explain why $se_0 = \sqrt{\pi_0(1 - \pi_0)/n}$ is more appropriate for tests.

6.66.* You test H_0: $\pi = 0.50$ against H_a: $\pi > 0.50$, using $\alpha = 0.05$. In fact, H_a is true. Explain why P(Type II error) increases toward 0.95 as π moves down toward 0.50. (Assume n and α stay fixed. You may want to look at the applet introduced in Exercise 6.36.)

6.67.* Refer to the ESP experiment in Exercise 6.33, with $n = 5$.

(a) For what value(s) of $x =$ number of correct guesses can you reject H_0: $\pi = 0.50$ in favor of H_a: $\pi > 0.50$, using $\alpha = 0.05$?

(b) For what value(s) of x can you reject H_0 using $\alpha = 0.01$? (*Note*: For small samples, it may not be possible to achieve very small P-values.)

(c) Suppose you test H_0 using $\alpha = 0.05$. If $\pi = 0.50$, what is the actual P(Type I error)? (*Note*: For discrete distributions, P(Type I error) may be less than intended. It is better to report the P-value.)

6.68.* You evaluate 16 schools in your city over the past four years according to the change from the previous year in the mean student score on a standardized achievement test. Of the schools, school number 3 performed above the median in all four years. Explain what is misleading if you use this record to decide that that school is performing better than the others.

COMPARISON OF TWO GROUPS

Comparing two groups is a very common analysis in the social and behavioral sciences. A study might compare mean amount of time spent on housework for men and women. Another study might compare the proportions of Americans and Europeans who support strict gun control. For quantitative variables, we compare means. For categorical variables, we compare proportions.

7.1 Preliminaries for Comparing Groups

Do women tend to spend more time on housework than men? If so, how much more? The 2012 General Social Survey asked, "On average, how many hours a week do you personally spend on household work, not including childcare and leisure time activities?" Table 7.1 reports the mean and standard deviation, classified by sex of respondent. We use Table 7.1 in explaining basic concepts for comparing groups.

TABLE 7.1: Hours per Week Spent on Household Work			
		Household Work Time	
Sex	Sample Size	Mean	Standard Deviation
Men	583	8.3	9.4
Women	693	11.9	12.7

BIVARIATE ANALYSES WITH RESPONSE AND EXPLANATORY VARIABLES

Two groups being compared constitute a ***binary*** variable—a variable having only two categories, sometimes also called ***dichotomous***. In a comparison of mean housework time for men and women, men and women are the two categories of the binary variable, sex. Methods for comparing two groups are special cases of ***bivariate*** statistical methods—an outcome variable of some type is analyzed for each category of a second variable.

From Section 3.5 (page 63), an outcome variable about which comparisons are made is called a ***response variable***. The variable that defines the groups is called the ***explanatory variable***. In Table 7.1, weekly time spent on household work is the response variable. The sex of the respondent is the explanatory variable.

DEPENDENT AND INDEPENDENT SAMPLES

Some studies compare groups at two or more points in time. For example, a ***longitudinal study*** observes subjects at several times. An example is the Framingham Heart Study, which every two years since 1948 has observed many health characteristics of more than 5000 adults from Framingham, Massachusetts. Samples that have the same subjects in each sample are called ***dependent samples***.

More generally, two samples are *dependent* when a natural matching occurs between each subject in one sample and a subject in the other sample. Usually this happens when each sample has the same subjects. But matching can also occur when the two samples have different subjects. An example is a comparison of housework time of husbands and wives, the husbands forming one sample and their wives the other.

More commonly, comparisons use ***independent samples***. This means that the observations in one sample are *independent* of those in the other sample. The subjects in the two samples are different, with no matching between one sample and the other sample. In *observational* studies (page 29), comparisons of groups often result from dividing a sample into subsamples according to classification on a variable such as sex or race or political party. An example is Table 7.1. Subjects were randomly selected and then classified on their sex and measured on various response variables. The samples of men and women were independent. Such a ***cross-sectional*** study uses a single survey to compare groups. If the overall sample was randomly selected, then the subsamples are independent random samples from the corresponding subpopulations.

Suppose you plan to use an *experimental* study to analyze whether a tutoring program improves mathematical understanding. One study design administers an exam on math concepts to a sample of students both before and after they go through the program. The sample of exam scores before the program and the sample of exam scores after the program are then *dependent*, because each sample has the same subjects. Another study design randomly splits a class of students into two groups, one of which takes the tutoring program (the *experimental* group) and one of which does not (the *control* group). After the course, both groups take the math concepts exam, and mean scores are compared. The two samples are then *independent*, because they contain different subjects without a matching between samples.

Why do we distinguish between *independent* and *dependent* samples? Because the standard error formulas for statistics that compare means or compare proportions are different for the two types of sample. With dependent samples, matched responses are likely to be correlated. In the study about a tutoring program, the students who perform relatively well on one exam probably tend to perform well on the second exam also. This affects the standard error of statistics comparing the groups.

STANDARD ERROR OF ESTIMATED DIFFERENCE BETWEEN GROUPS

To compare two populations, we estimate the difference between their parameters. To compare population means μ_1 and μ_2, we treat $\mu_2 - \mu_1$ as a parameter and estimate it by the difference of sample means, $\bar{y}_2 - \bar{y}_1$. For Table 7.1, the estimated difference between women and men in the population mean weekly time spent on household work is $\bar{y}_2 - \bar{y}_1 = 11.9 - 8.3 = 3.6$ hours.

The sampling distribution of the estimator $\bar{y}_2 - \bar{y}_1$ has expected value $\mu_2 - \mu_1$. For large random samples, by the Central Limit Theorem this sampling distribution has a normal shape, as Figure 7.1 portrays.

An estimate has a standard error that describes how precisely it estimates a parameter. Likewise, the difference between estimates from two samples has a standard
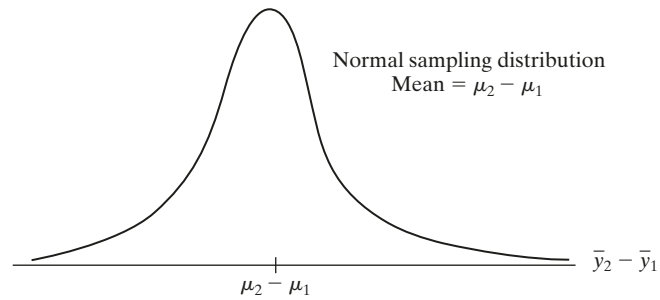
FIGURE 7.1: For Random Samples, the Sampling Distribution of the Difference between the Sample Means $\bar{y}_2 - \bar{y}_1$ Is Approximately Normal about $\mu_2 - \mu_1$

error. For Table 7.1, the standard error of the sampling distribution of $\bar{y}_2 - \bar{y}_1$ describes how precisely $\bar{y}_2 - \bar{y}_1 = 3.6$ estimates $\mu_2 - \mu_1$. If many studies had been conducted in the United States comparing weekly household work time for women and men, the estimate $\bar{y}_2 - \bar{y}_1$ would have varied from study to study. The standard error describes the variability of the estimates from different potential studies of the same size.

The following general rule enables us to find standard errors when we compare estimates from independent samples:

Standard Error of Difference between Two Estimates

> For two estimates from independent samples that have estimated standard errors se_1 and se_2, the sampling distribution of their difference has
>
> Estimated standard error $= \sqrt{(se_1)^2 + (se_2)^2}$.

Each estimate has sampling error, and the variabilities add together to determine the standard error of the difference of the estimates. The standard error formula for dependent samples differs, and Section 7.4 presents it.

Recall that the estimated standard error of a sample mean is

$$se = \frac{s}{\sqrt{n}},$$

where s is the sample standard deviation. Let n_1 denote the sample size for the first sample and n_2 the sample size for the second sample. Let s_1 and s_2 denote the sample standard deviations, which estimate the corresponding population standard deviations σ_1 and σ_2. The difference $\bar{y}_2 - \bar{y}_1$ between two sample means with independent samples has estimated standard error

$$se = \sqrt{(se_1)^2 + (se_2)^2} = \sqrt{\left(\frac{s_1}{\sqrt{n_1}}\right)^2 + \left(\frac{s_2}{\sqrt{n_2}}\right)^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

For example, from Table 7.1, the estimated standard error of the difference of 3.6 hours between the sample mean weekly household work time for women and men is

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(9.4)^2}{583} + \frac{(12.7)^2}{693}} = 0.62.$$

For such large sample sizes, the estimate $\bar{y}_2 - \bar{y}_1$ would not vary much from study to study.

From the formula, the standard error of the difference is larger than the standard error for either sample estimate alone. Why is this? In practical terms, $(\bar{y}_2 - \bar{y}_1)$ is

often farther from $(\mu_2 - \mu_1)$ than $\bar{y}_1$ is from μ_1 or $\bar{y}_2$ is from μ_2. For instance, suppose $\mu_1 = \mu_2 = 10$ (unknown to us), but the sample means are $\bar{y}_1 = 8$ and $\bar{y}_2 = 12$. Then the errors of estimation were

$$\bar{y}_1 - \mu_1 = 8 - 10 = -2 \quad \text{and} \quad \bar{y}_2 - \mu_2 = 12 - 10 = 2,$$

each estimate being off by a distance of 2. But the estimate $(\bar{y}_2 - \bar{y}_1) = 12 - 8 = 4$ falls 4 from $(\mu_2 - \mu_1) = 0$. The error of size 4 for the difference is larger than the error of size 2 for either mean individually. Suppose a sample mean that falls 2 away from a population mean is well out in the tail of a sampling distribution for a single sample mean. Then, a difference between sample means that falls 4 away from the difference between population means is well out in the tail of the sampling distribution for $\bar{y}_2 - \bar{y}_1$.

RATIOS OF MEANS AND PROPORTIONS

This chapter focuses on comparing parameters by their difference, but we can also compare them by their *ratio*. The ratio equals 1.0 when the parameters are equal. Ratios farther from 1.0 represent larger effects.

In Table 7.1, the ratio of sample mean household work time for women and for men is $11.9/8.3 = 1.43$. The sample mean for women was 1.43 times the sample mean for men. We can also express this by saying that the mean for women was 43% higher than the mean for men.

When proportions for two groups are close to 0, the ratio[1] is often more informative than the difference. For example, according to recent data from the United Nations, the annual gun homicide rate is 62.4 per one million residents in the United States and 1.3 per one million residents in Britain. In proportion form, the results are 0.0000624 in the United States and 0.0000013 in Britain. The difference between the proportions is $0.0000624 - 0.0000013 = 0.0000611$, extremely small. By contrast, the ratio is $0.000624/0.0000013 = 624/13 = 48$. The proportion of people killed by guns in the United States is 48 times the proportion in Britain. In this sense, the effect is very large.

7.2 Categorical Data: Comparing Two Proportions

Let's now learn how to compare proportions inferentially. Let π_1 denote the proportion for the first population and π_2 the proportion for the second population. Let $\hat{\pi}_1$ and $\hat{\pi}_2$ denote the sample proportions. You may wish to review Sections 5.2 and 6.3 on inferences for proportions in the one-sample case.

Example 7.1

Does Prayer Help Coronary Surgery Patients? A study used patients at six U.S. hospitals who were to receive coronary artery bypass graft surgery.[2] The patients were randomly assigned to two groups. For one group, Christian volunteers were instructed to pray for a successful surgery with a quick, healthy recovery and no complications. The praying started the night before surgery and continued for two weeks. The other group did not have volunteers praying for them. The response was whether medical complications occurred within 30 days of the surgery. Table 7.2 summarizes results.

[1] The ratio is often called the *relative risk*, because it is used in public health to compare rates for an undesirable outcome.
[2] H. Benson et al., *American Heart Journal*, vol. 151 (2006), pp. 934–952.

TABLE 7.2: Whether Complications Occurred for Heart Surgery Patients Who Did or Did Not Have Group Prayer			
	Complications		
Prayer	Yes	No	Total
Yes	315	289	604
No	304	293	597

Is there a difference in complication rates for the two groups? Let π_1 denote the probability of complications for those patients who had a prayer group. Let π_2 denote the probability of complications for the subjects not having a prayer group. These are population proportions for the conceptual population this sample represents. From Table 7.2, the sample proportions who had complications are

$$\hat{\pi}_1 = \frac{315}{604} = 0.522 \quad \text{and} \quad \hat{\pi}_2 = \frac{304}{597} = 0.509.$$ ∎

We compare the probabilities using their difference, $\pi_2 - \pi_1$. The difference of sample proportions, $\hat{\pi}_2 - \hat{\pi}_1$, estimates $\pi_2 - \pi_1$. If n_1 and n_2 are relatively large, $\hat{\pi}_2 - \hat{\pi}_1$ has a sampling distribution that is approximately normal, by the Central Limit Theorem. See Figure 7.2. The mean of the sampling distribution is the parameter $\pi_2 - \pi_1$ to be estimated.
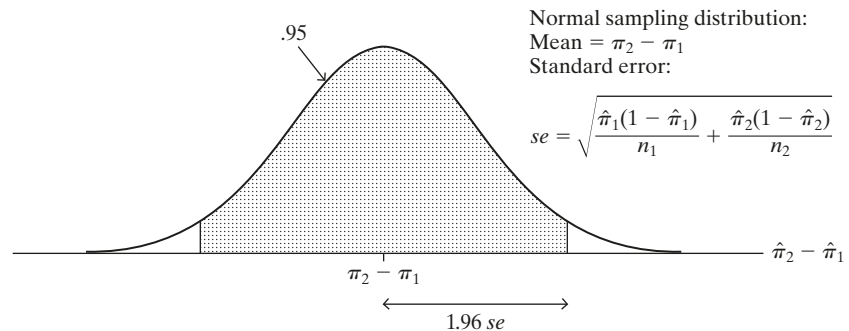
FIGURE 7.2: For Large Random Samples, the Sampling Distribution of the Estimator $\hat{\pi}_2 - \hat{\pi}_1$ of the Difference $\pi_2 - \pi_1$ of Population Proportions Is Approximately Normal, by the Central Limit Theorem

From the rule on page 193, the standard error of the difference of sample proportions equals the square root of the sum of squared standard errors of the separate sample proportions. Recall (page 119) that the estimated standard error of a single sample proportion is

$$se = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}.$$

Therefore, the difference between two proportions has estimated standard error

$$se = \sqrt{(se_1)^2 + (se_2)^2} = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}.$$

For Table 7.2, $\hat{\pi}_1 = 0.522$, $\hat{\pi}_2 = 0.509$, and $\hat{\pi}_2 - \hat{\pi}_1$ has estimated standard error

$$se = \sqrt{\frac{(0.522)(0.478)}{604} + \frac{(0.509)(0.491)}{597}} = 0.0288.$$

For samples of these sizes, the difference in sample proportions would not vary much from study to study.

CONFIDENCE INTERVAL FOR DIFFERENCE OF PROPORTIONS

As with a single proportion, the confidence interval takes the point estimate and adds and subtracts a margin of error that is a z-score times the estimated standard error, such as

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm 1.96(se)$$

for 95% confidence.

Confidence Interval for
$\pi_2 - \pi_1$

> For large, independent random samples, a confidence interval for the difference $\pi_2 - \pi_1$ between two population proportions is
>
> $$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se), \quad \text{where } se = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}.$$
>
> The z-score depends on the confidence level, such as 1.96 for 95% confidence.

The sample is large enough to use this formula if, for each sample, at least 10 observations fall in the category for which the proportion is estimated, and at least 10 observations do not fall in that category. When this is not the case, the formula is still valid when you make a simple adjustment[3]: Add four observations, by adding one observation of each type to each sample.

Example 7.2

Comparing Prayer and Non-Prayer Surgery Patients For Table 7.2, we estimate the difference $\pi_2 - \pi_1$ between the probability of complications for the non-prayer and prayer surgery patients. Since $\hat{\pi}_1 = 0.522$ and $\hat{\pi}_2 = 0.509$, the estimated difference is $\hat{\pi}_2 - \hat{\pi}_1 = -0.013$. There was a drop of 0.013 in the proportion who had complications among those not receiving prayer.

To determine the precision of this estimate, we form a confidence interval. We have found that $se = 0.0288$. A 95% confidence interval for $\pi_2 - \pi_1$ is

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm 1.96(se), \quad \text{or } (0.509 - 0.522) \pm 1.96(0.0288)$$

$$= -0.013 \pm 0.057, \quad \text{or} \quad (-0.07, 0.04).$$

It seems that the difference is close to 0, so the probability of complications is similar for the two groups. ∎

INTERPRETING A CONFIDENCE INTERVAL COMPARING PROPORTIONS

When the confidence interval for $\pi_2 - \pi_1$ contains 0, as in the previous example, it is plausible that $\pi_2 - \pi_1 = 0$. That is, it is believable that $\pi_1 = \pi_2$. Insufficient evidence exists to conclude which of π_1 or π_2 is larger. For the confidence interval for $\pi_2 - \pi_1$ of $(-0.07, 0.04)$, we infer that π_2 may be as much as 0.07 smaller or as much as 0.04 larger than π_1.

---

[3] See A. Agresti and B. Caffo, *American Statistician*, vol. 54 (2000), pp. 280–288; this is an analog of the Agresti–Coull method (page 138) of adding four observations for a single sample.

When a confidence interval for $\pi_2 - \pi_1$ contains only *negative* values, this suggests that $\pi_2 - \pi_1$ is negative. In other words, we infer that π_2 is *smaller* than π_1. When a confidence interval for $\pi_2 - \pi_1$ contains only *positive* values, we conclude that $\pi_2 - \pi_1$ is positive; that is, π_2 is *larger* than π_1.

Which group we call Group 1 and which we call Group 2 is arbitrary. If we let Group 1 be the non-prayer group rather than the prayer group, then the estimated difference would be $+0.013$ rather than -0.013. The confidence interval would have been $(-0.04, 0.07)$, the negatives of the endpoints we obtained. Similarly, for conclusions it does not matter whether we form a confidence interval for $\pi_2 - \pi_1$ or for $\pi_1 - \pi_2$. If the confidence interval for $\pi_2 - \pi_1$ is $(-0.07, 0.04)$, then the confidence interval for $\pi_1 - \pi_2$ is $(-0.04, 0.07)$.

The magnitude of values in the confidence interval tells you how large any true difference is likely to be. If all values in the confidence interval are near 0, such as the interval $(-0.07, 0.04)$, we infer that $\pi_2 - \pi_1$ is small in practical terms even if not exactly equal to 0.

As in the one-sample case, larger sample sizes contribute to a smaller *se*, a smaller margin of error, and narrower confidence intervals. In addition, higher confidence levels yield wider confidence intervals. For the prayer study, a 99% confidence interval equals $(-0.06, 0.09)$, wider than the 95% confidence interval of $(-0.04, 0.07)$.

SIGNIFICANCE TESTS ABOUT $\pi_2 - \pi_1$

To compare population proportions π_1 and π_2, a significance test specifies $H_0: \pi_1 = \pi_2$. For the difference of proportions parameter, this hypothesis is $H_0: \pi_2 - \pi_1 = 0$, *no difference*, or *no effect*.

Under the presumption for H_0 that $\pi_1 = \pi_2$, we estimate the common value of π_1 and π_2 by the sample proportion for the entire sample. Denote this by $\hat{\pi}$. To illustrate, for the data in Table 7.2 from the prayer study, $\hat{\pi}_1 = 315/604 = 0.522$ and $\hat{\pi}_2 = 304/597 = 0.509$. For the entire sample,

$$\hat{\pi} = (315 + 304)/(604 + 597) = 619/1201 = 0.515.$$

The proportion $\hat{\pi}$ is called a *pooled estimate*, because it pools together observations from the two samples.

The test statistic measures the number of standard errors between the estimate and the H_0 value. Treating $\pi_2 - \pi_1$ as the parameter, we test that $\pi_2 - \pi_1 = 0$; that is, the null hypothesis value of $\pi_2 - \pi_1$ is 0. The estimated value of $\pi_2 - \pi_1$ is $\hat{\pi}_2 - \hat{\pi}_1$. The test statistic is

$$z = \frac{\text{Estimate} - \text{Null hypothesis value}}{\text{Standard error}} = \frac{(\hat{\pi}_2 - \hat{\pi}_1) - 0}{se_0}.$$

Rather than using the standard error from the confidence interval, you should use an alternative formula based on the presumption stated in H_0 that $\pi_1 = \pi_2$. We use the notation se_0, because it is a *se* that holds under H_0. This standard error, which uses the pooled estimate $\hat{\pi}$ of a common value, is

$$se_0 = \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n_1} + \frac{\hat{\pi}(1 - \hat{\pi})}{n_2}} = \sqrt{\hat{\pi}(1 - \hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

The *P*-value depends in the usual way on whether the test is two-sided, $H_a: \pi_1 \neq \pi_2$ (i.e., $\pi_2 - \pi_1 \neq 0$), or one-sided, $H_a: \pi_1 > \pi_2$ (i.e., $\pi_2 - \pi_1 < 0$) or $H_a: \pi_1 < \pi_2$ (i.e., $\pi_2 - \pi_1 > 0$). Most common is the two-sided alternative. Its *P*-value is the two-tail probability from the standard normal distribution that falls beyond the observed test statistic value. This z test is valid when each sample has at least 10 outcomes of each type.

Example
7.3

Test Comparing Prayer and Non-Prayer Surgery Patients   For Table 7.2 on complica-
tion rates for prayer and non-prayer surgery patients, the standard error estimate for
the significance test is

$$se_0 = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{0.515(0.485)\left(\frac{1}{604} + \frac{1}{597}\right)}$$

$$= \sqrt{0.000832} = 0.0288.$$

The test statistic for $H_0: \pi_1 = \pi_2$ is

$$z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{se_0} = \frac{0.509 - 0.522}{0.0288} = -0.43.$$

This z-score has two-sided P-value equal to 0.67. There is not much evidence
against H_0.

In summary, it is plausible that the probability of complications is the same for
the prayer and non-prayer conditions. However, this study does not disprove the
power of prayer. Apart from the fact that we cannot accept a null hypothesis, the
experiment could not control many factors, such as whether friends and family were
also praying for the patients.   ∎

You can use software to construct confidence intervals and significance tests
comparing proportions. If you already have sample proportions, Stata can conduct
the inferences with the **prtesti** command (or a dialog box), by entering n and $\hat{\pi}$ for
each group. It constructs a confidence interval for each population proportion and
for the difference, and a significance test for each possible alternative hypothesis.
Table 7.3 illustrates for this example.

TABLE 7.3: Stata Software for Performing Two-Sample Inferences for Proportions. (Stata
estimates $\pi_1 - \pi_2$ instead of $\pi_2 - \pi_1$.)

```
. prtesti 604 0.52152 597 0.50921

Two-sample test of proportions            x: Number of obs =       604
                                          y: Number of obs =       597
-------------------------------------------------------------------------
Variable |   Mean    Std. Err.   z    P>|z|    [95% Conf. Interval]
      x  |  .52152   .0203259                   .481682    .561358
      y  |  .50921   .0204602                   .469109    .549311
---------+---------------------------------------------------------------
   diff  |  .01231   .0288402                  -.044216    .068836
         | under Ho:  .0288423  0.43  0.670

   Ho: diff = 0
   Ha: diff < 0               Ha: diff != 0              Ha: diff > 0
 Pr(Z < z) = 0.6652      Pr(|Z| < |z|) = 0.6695     Pr(Z > z) = 0.3348
```

With R software, you can also conduct inference by entering the counts in the
category of interest and the sample sizes, as shown in Appendix A. Internet applets
are also available.[4]

---

[4] For example, the *Comparing Two Proportions* applet at www.pearsonglobaleditions.com/Agresti.

CONTINGENCY TABLES AND CONDITIONAL PROBABILITIES

Table 7.2 is an example of a ***contingency table***. Each row is a category of the explanatory variable (whether prayed for) which defines the two groups compared. Each column is a category of the response variable (whether complications occurred). The ***cells*** of the table contain frequencies for the four possible combinations of outcomes.

The parameters π_1 and π_2 estimated using the contingency table are *conditional probabilities*. This term refers to probabilities for a response variable evaluated under two conditions, namely, the two levels of the explanatory variable. For instance, under the condition that the subject is being prayed for, the conditional probability of developing complications is estimated to be $315/604 = 0.52$.

This section has considered binary response variables. Instead, the response could have several categories. For example, the response categories might be (No complications, Slight complications, Severe complications). Then, we could compare the two groups in terms of the conditional probabilities of observations in each of the three categories. Likewise, the number of groups compared could exceed two. Chapter 8 shows how to analyze contingency tables having more than two rows or columns.

7.3 Quantitative Data: Comparing Two Means

We compare two population means μ_1 and μ_2 by making inferences about their difference. You may wish to review Sections 5.3 and 6.2 on inferences for means in the one-sample case.

CONFIDENCE INTERVAL FOR $\mu_2 - \mu_1$

For large random samples, or for small random samples from normal population distributions, the sampling distribution of $(\bar{y}_2 - \bar{y}_1)$ has a normal shape. As usual, inference for means with *estimated* standard errors uses the t distribution for test statistics and for the margin of error in confidence intervals. A confidence interval takes the point estimate and adds and subtracts a margin of error that is a t-score times the standard error.

Confidence Interval for $\mu_2 - \mu_1$

For independent random samples from two groups that have normal population distributions, a confidence interval for $\mu_2 - \mu_1$ is

$$(\bar{y}_2 - \bar{y}_1) \pm t(se), \ \ \text{where } se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The t-score is chosen to provide the desired confidence level.

The formula for the degrees of freedom for the t-score, called the *Welch–Satterthwaite approximation*, is complex. The df depends on the sample standard deviations s_1 and s_2 as well as the sample sizes n_1 and n_2. If $s_1 = s_2$ and $n_1 = n_2$, it simplifies to $df = (n_1 + n_2 - 2)$. This is the sum of the df values for separate inference about each group; that is, $df = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$. Generally, df falls somewhere between $n_1 + n_2 - 2$ and the minimum of $(n_1 - 1)$ and $(n_2 - 1)$. Software can find this df value, t-score, and confidence interval.

In practice, the method is *robust* to violations of the normal population assumption. This is especially true when both n_1 and n_2 are at least about 30, by the Central

Limit Theorem. As usual, you should be wary of extreme outliers or of extreme skew that may make the mean unsuitable as a summary measure.

Comparing Housework Time of Men and Women For Table 7.1 (page 191) on the weekly time spent on housework, denote the population mean by μ_1 for men and μ_2 for women. That table reported sample means of 8.3 hours for 583 men and 11.9 hours for 693 women, with sample standard deviations of 9.4 and 12.7, respectively. The point estimate of $\mu_2 - \mu_1$ equals $\bar{y}_2 - \bar{y}_1 = 11.9 - 8.3 = 3.6$ hours. Section 7.1 found that the estimated standard error of this difference is

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(9.4)^2}{583} + \frac{(12.7)^2}{693}} = 0.620.$$

The sample sizes are very large, so the t-score is essentially the z-score (1.96). The 95% confidence interval for $\mu_2 - \mu_1$ is

$$(\bar{y}_2 - \bar{y}_1) \pm 1.96(se) = 3.6 \pm 1.96(0.620), \quad \text{or } 3.6 \pm 1.2, \quad \text{which is } (2.4, 4.8).$$

We can be 95% confident that the population mean amount of weekly time spent on housework is between 2.4 and 4.8 hours higher for women than men. ∎

INTERPRETING A CONFIDENCE INTERVAL COMPARING MEANS

The confidence interval (2.4, 4.8) contains only positive values. Since we took the difference between the mean for women and the mean for men, we can infer that the population mean is higher for women. A confidence interval for $\mu_2 - \mu_1$ that contains only negative values suggests that $\mu_2 - \mu_1$ is negative, meaning that we can infer that μ_2 is less than μ_1. When the confidence interval contains 0, insufficient evidence exists to conclude which of μ_1 or μ_2 is larger. It is then plausible that $\mu_1 = \mu_2$.

The identification of which is group 1 and which is group 2 is arbitrary, as is whether we estimate $\mu_2 - \mu_1$ or $\mu_1 - \mu_2$. For instance, a confidence interval of (2.4, 4.8) for $\mu_2 - \mu_1$ is equivalent to one of $(-4.8, -2.4)$ for $\mu_1 - \mu_2$.

SIGNIFICANCE TESTS ABOUT $\mu_2 - \mu_1$

To compare population means μ_1 and μ_2, we can also conduct a significance test of $H_0: \mu_1 = \mu_2$. For the difference of means parameter, this hypothesis is $H_0: \mu_2 - \mu_1 = 0$ (no effect). Alternative hypotheses can be two-sided or one-sided.

As usual, the test statistic measures the number of standard errors between the estimate and the H_0 value,

$$t = \frac{\text{Estimate of parameter} - \text{Null hypothesis value of parameter}}{\text{Standard error of estimate}}.$$

Treating $\mu_2 - \mu_1$ as the parameter, we test that $\mu_2 - \mu_1 = 0$. Its estimate is $\bar{y}_2 - \bar{y}_1$. The standard error is the same as in a confidence interval. The t test statistic is

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se}, \quad \text{where} \quad se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

with the same df as in constructing a confidence interval.

**Example
7.5**

Test Comparing Mean Housework for Men and Women   Using the data from Table 7.1 (page 191), we now test for a difference between the population mean housework time, μ_1 for men and μ_2 for women. We test $H_0: \mu_1 = \mu_2$ against $H_a: \mu_1 \neq \mu_2$. We've seen that the estimate $\bar{y}_2 - \bar{y}_1 = 11.9 - 8.3 = 3.6$ has $se = 0.620$.

The t test statistic is

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se} = \frac{(11.9 - 8.3)}{0.620} = 5.8.$$

Such an enormous t-value gives a P-value that is 0 to many decimal places. We conclude that the population means differ. The sample means show that the difference takes the direction of a higher mean for women.   ■

In practice, data analysts use significance tests much more often for two-sample comparisons than for one-sample analyses. It is usually artificial to test whether the population mean equals one particular value, such as in testing $H_0: \mu = \mu_0$. However, it is often relevant to test whether a *difference* exists between two population means, such as in testing $H_0: \mu_1 = \mu_2$. For instance, we may have no idea what to hypothesize for the mean amount of housework time for men, but we may want to know whether that mean (whatever its value) is the same as, larger than, or smaller than the mean for women.

Software can construct confidence intervals and significance tests comparing means. If you already have summary statistics, some software (such as SPSS and Stata) can conduct the inferences with them. With Stata, you apply the `ttesti` command (or use a dialog box) and enter n, $\bar{y}$, and s for each group. See Table 7.4. Internet applets are also available.[5] For comparing means, most software also presents results for another method that makes the additional assumption that $\sigma_1 = \sigma_2$. We present this method on page 205.

TABLE 7.4: Stata Software for Performing Two-Sample Inferences for Means. (Stata estimates $\mu_1 - \mu_2$ instead of $\mu_2 - \mu_1$.)

```
. ttesti 583 8.3 9.4 693 11.9 12.7, unequal

Two-sample t test with unequal variances
     |   Obs     Mean    Std. Err.    Std. Dev.    [95% Conf. Interval]
   x |   583      8.3    .3893082          9.4      7.53538     9.06462
   y |   693     11.9    .4824331         12.7     10.95279    12.84721
  ---+------------------------------------------------------------------
diff |           -3.6    .6199214                 -4.816197   -2.383803

    diff = mean(x) - mean(y)                               t =  -5.8072
Ho: diff = 0                 Satterthwaite's degrees of freedom =  1254.29
   Ha: diff < 0                  Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000    Pr(|T| > |t|) = 0.0000        Pr(T > t) = 1.0000
```

CORRESPONDENCE BETWEEN CONFIDENCE INTERVALS AND TESTS

For means, the equivalence between two-sided tests and confidence intervals mentioned in Sections 6.2 and 6.4 also applies in the two-sample case. For example, since the two-sided P-value in Example 7.5 is less than 0.05, we reject $H_0: \mu_2 - \mu_1 = 0$ at the

---

[5] For example, the *Comparing Two Means* applet at www.pearsonglobaleditions.com/Agresti.

$\alpha = 0.05$ level. Similarly, a 95% confidence interval for $\mu_2 - \mu_1$ does not contain 0, the H_0 value. That interval equals (2.4, 4.8).

As in one-sample inference, confidence intervals are more informative than tests. The confidence interval tells us not only that the population mean differs for men and women, but it shows us just how large that difference is likely to be, and in which direction.

7.4 Comparing Means with Dependent Samples

Dependent samples occur when each observation in sample 1 matches with an observation in sample 2. The data are often called *matched-pairs* data, because of this matching.

Dependent samples commonly occur when each sample has the same subjects. Examples are *longitudinal* observational studies that observe a person's response at several points in time and experimental studies that take *repeated measures* on subjects. An example of the latter is a *crossover study*, in which a subject receives one treatment for a period and then the other treatment. The next example illustrates this.

Example 7.6

Cell Phone Use and Driver Reaction Time An experiment[6] used a sample of college students to investigate whether cell phone use impairs drivers' reaction times. On a machine that simulated driving situations, at irregular periods a target flashed red or green. Participants were instructed to press a brake button as soon as possible when they detected a red light. Under the cell phone condition, the student carried out a conversation about a political issue on the cell phone with someone in a separate room. In the control condition, they listened to a radio broadcast or to books-on-tape while performing the simulated driving.

For each student and each condition, Table 7.5 records their mean response time (in milliseconds) over several trials. Figure 7.3 shows box plots of the data for the two conditions. Student 28 is an outlier under each condition. ■

With matched-pairs data, for each pair we form

$$\text{Difference} = \text{Observation in sample 2} - \text{Observation in sample 1.}$$

Table 7.5 shows the difference scores for the cell phone experiment. Let $\bar{y}_d$ denote the sample mean of the difference scores. This estimates μ_d, the population mean difference. In fact, the parameter μ_d is identical to $\mu_2 - \mu_1$, the difference between the population means for the two groups. The mean of the differences equals the difference between the means.

Difference of means = Mean of differences

> For matched-pairs data, the *difference between the means* of the two groups equals the *mean of the difference* scores.

INFERENCES COMPARING MEANS USING PAIRED DIFFERENCE SCORES

We can base analyses about $\mu_2 - \mu_1$ on inferences about μ_d, using the single sample of difference scores. This simplifies the analysis, because it reduces a two-sample problem to a one-sample problem.

---

[6] Data courtesy of David Strayer, University of Utah. See D. Strayer and W. Johnston, *Psychological Science*, vol. 21 (2001), pp. 462–466.

TABLE 7.5: Reaction Times (in Milliseconds) on Driving Skills Task and Cell Phone Use (Yes or No). The difference score is the reaction time using the cell phone minus the reaction time not using it. This is the `CellPhone` data file at the text website.

Student	Cell Phone? No	Yes	Difference	Student	Cell Phone? No	Yes	Difference
1	604	636	32	17	525	626	101
2	556	623	67	18	508	501	−7
3	540	615	75	19	529	574	45
4	522	672	150	20	470	468	−2
5	459	601	142	29	512	578	66
6	544	600	56	22	487	560	73
7	513	542	29	23	515	525	10
8	470	554	84	24	499	647	148
9	556	543	−13	25	448	456	8
10	531	520	−11	26	558	688	130
11	599	609	10	27	589	679	90
12	537	559	22	28	814	960	146
13	619	595	−24	29	519	558	39
14	536	565	29	30	462	482	20
15	554	573	19	31	521	527	6
16	467	554	87	32	543	536	−7
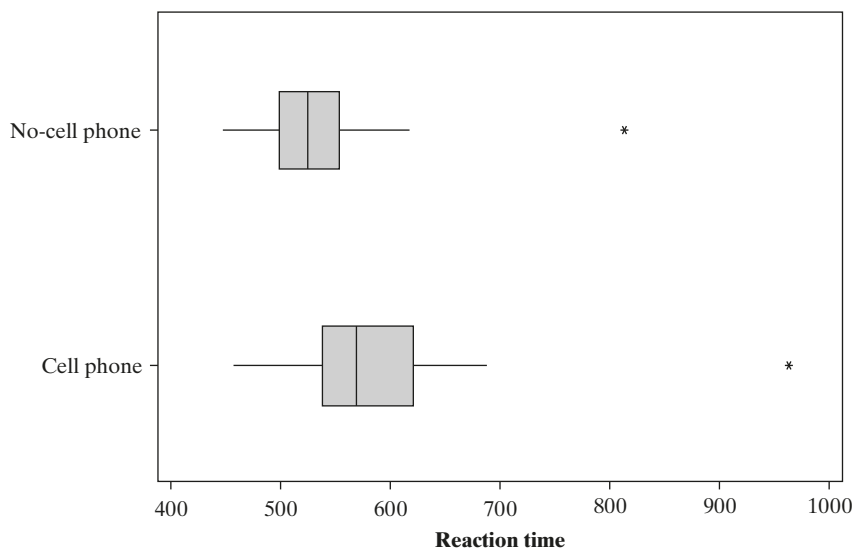
FIGURE 7.3: Box Plots of Observations for the Experiment on the Effects of Cell Phone Use on Reaction Times

Let n denote the number of observations in each sample. This equals the number of difference scores. The confidence interval for μ_d is

$$\bar{y}_d \pm t\left(\frac{s_d}{\sqrt{n}}\right).$$

Here, $\bar{y}_d$ and s_d are the sample mean and standard deviation of the difference scores, and t is the t-score for the chosen confidence level, having $df = n-1$. This confidence interval has the same form as the one Section 6.3 presented for a single mean. We

apply the formula to the single sample of n differences rather than to the original two sets of observations.

For testing H_0: $\mu_1 = \mu_2$, we express the hypothesis in terms of the difference scores as H_0: $\mu_d = 0$. The test statistic is

$$t = \frac{\bar{y}_d - 0}{se}, \quad \text{where} \quad se = \frac{s_d}{\sqrt{n}}.$$

This compares the sample mean of the differences to the null hypothesis value of 0 by the number of standard errors between them. The standard error is the same one used for a confidence interval. Since this test uses the difference scores for the pairs of observations, it is called a ***paired-difference t test***.

Example 7.7	**Comparing Driver Reaction Times by Cell Phone Use** For the matched-pairs data in Table 7.5 for the driving and cell phone experiment, the mean reaction times were 534.6 milliseconds without the cell phone and 585.2 milliseconds while using it. The 32 difference scores $(32, 67, 75, \dots)$ from the table have a sample mean of

$$\bar{y}_d = [32 + 67 + 75 + \cdots + (-7)]/32 = 50.6.$$

This equals the difference between the sample means of 585.2 and 534.6 for the two conditions. The sample standard deviation of the 32 difference scores is

$$s_d = \sqrt{\frac{(32 - 50.6)^2 + (67 - 50.6)^2 + \cdots}{32 - 1}} = 52.5.$$

The standard error of $\bar{y}_d$ is $se = s_d/\sqrt{n} = 52.5/\sqrt{32} = 9.28$.

For a 95% confidence interval for $\mu_d = \mu_2 - \mu_1$ with $df = n - 1 = 31$, we use $t_{0.025} = 2.04$. The confidence interval equals

$$\bar{y}_d \pm 2.04(se) = 50.6 \pm 2.04(9.28), \quad \text{which is} \quad (31.7, 69.5).$$

We infer that the population mean reaction time when using cell phones is between about 32 and 70 milliseconds higher than when not using cell phones. The confidence interval does not contain 0. We conclude that the population mean reaction time is greater when using a cell phone.

Next consider the significance test of H_0: $\mu_d = 0$, which corresponds to equal population means for the two conditions, against H_a: $\mu_d \neq 0$. The test statistic is

$$t = \frac{\bar{y}_d - 0}{se} = \frac{50.6}{9.28} = 5.5,$$

with $df = 31$. The P-value for the two-sided H_a: $\mu_d \neq 0$ equals 0.000005. There is extremely strong evidence that mean reaction time is greater when using a cell phone. Table 7.6 shows how SPSS software reports these results for its paired-samples t test option. ∎

TABLE 7.6: SPSS Output for Matched-Pairs Analysis of `CellPhone` Data File Comparing Driver Reaction Times for *Yes* and *No* Categories of Cell Phone Use

Paired Samples Test

	Mean	Std. Dev.	Std. Error Mean	95% Conf Int of Difference Lower	95% Conf Int of Difference Upper	t	df	Sig. (2-tailed)
Yes - No	50.625	52.486	9.278	31.702	69.548	5.456	31	.000

Paired-difference inferences make the usual assumptions for t procedures: The observations (the difference scores) are randomly obtained from a population distribution that is normal. Confidence intervals and two-sided tests are valid even if the

normality assumption is violated (their *robustness* property), unless the sample size is small and the distribution is highly skewed or has severe outliers. For the study about driving and cell phones, one subject was an outlier on both reaction times. However, the difference score for that subject, which is the observation used in the analysis, is not an outlier. The article about the study did not indicate whether the subjects were randomly selected. The subjects in the experiment were probably a volunteer sample, so inferential conclusions are highly tentative.

INDEPENDENT VERSUS DEPENDENT SAMPLES

Using dependent samples can have certain benefits. First, sources of potential bias are controlled. Using the same subjects in each sample, for instance, keeps other factors fixed that could affect the analysis. Suppose younger subjects tend to have faster reaction times. If group 1 has a lower sample mean than group 2, it is not because subjects in group 1 are younger, because both groups have the same subjects.

Second, the standard error of $\bar{y}_2 - \bar{y}_1$ may be smaller with dependent samples. In the cell phone study, the standard error was 9.3. If we had observed *independent* samples with the same scores as in Table 7.3, the standard error of $\bar{y}_2 - \bar{y}_1$ would have been 19.7. This is because the variability in the difference scores tends to be less than the variability in the original scores when the scores in the two samples are strongly positively correlated. In fact, for Table 7.5, the correlation (recall Section 3.5) between the no–cell phone reaction times and the cell phone reaction times is 0.81, strongly positive.

7.5 Other Methods for Comparing Means*

Section 7.3 presented inference comparing two means with independent samples. A slightly different inference method can be used when we expect similar variability for the two groups. For example, under a null hypothesis of "no effect," we often expect the entire distributions of the response variable to be identical for the two groups. So, we expect standard deviations as well as means to be identical.

COMPARING MEANS WHILE ASSUMING EQUAL STANDARD DEVIATIONS

In comparing the population means, this method makes the additional assumption that the population standard deviations are equal; that is, $\sigma_1 = \sigma_2$. For it, a simpler *df* expression holds for an *exact t* distribution for the test statistic. Although it seems disagreeable to make an additional assumption, confidence intervals and two-sided tests are robust against violations of this and the normality assumption, particularly when the sample sizes are similar and not extremely small. In fact, the method itself is a special case of a multiple-group method introduced in Chapter 12 (called *analysis of variance*) that, for simplicity, has a single standard deviation parameter.

We estimate the common value σ of σ_1 and σ_2 by

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{\sum(y_{i1} - \bar{y}_1)^2 + \sum(y_{i2} - \bar{y}_2)^2}{n_1 + n_2 - 2}}.$$

Here, $\sum(y_{i1} - \bar{y}_1)^2$ denotes the sum of squares about the mean for the observations in the first sample, and $\sum(y_{i2} - \bar{y}_2)^2$ denotes the sum of squares about the mean for the observations in the second sample. The estimate s pools information from the two

samples to provide a single estimate of variability. It is called the *pooled estimate*. The term inside the square root is a weighted average of the two sample variances. When $n_1 = n_2$, it is the ordinary average. The estimate s falls between s_1 and s_2. With s as the estimate of σ_1 and σ_2, the estimated standard error of $\bar{y}_2 - \bar{y}_1$ simplifies to

$$se = \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}} = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

The confidence interval for $\mu_2 - \mu_1$ has the usual form

$$(\bar{y}_2 - \bar{y}_1) \pm t(se).$$

The t-score for the desired confidence level has $df = n_1 + n_2 - 2$. The df equals the total number of observations $(n_1 + n_2)$ minus the number of parameters estimated in order to calculate s (namely, the two means, μ_1 and μ_2, estimated by $\bar{y}_1$ and $\bar{y}_2$).

To test $H_0: \mu_1 = \mu_2$, the test statistic has the usual form

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se}.$$

Now, *se* uses the pooled formula, as in the confidence interval. The test statistic has a t distribution with $df = n_1 + n_2 - 2$.

Example **7.8**	**Comparing a Therapy to a Control Group** Examples 5.5 (page 129) and 6.4 (page 160) described a study that used a cognitive behavioral therapy to treat a sample of teenage girls who suffered from anorexia. The study analyzed the mean weight change after a period of treatment. Such studies also usually have a control group that receives no treatment or a standard treatment. Then researchers can analyze how the change in weight compares for the treatment group to the control group.

In fact, the anorexia study had a control group. Teenage girls in the study were randomly assigned to the cognitive behavioral treatment (Group 1) or to the control group (Group 2). Table 7.7 summarizes the results.[7]

TABLE 7.7: Summary of Results Comparing Treatment Group to Control Group for Weight Change in Anorexia Study

Group	Sample Size	Mean	Standard Deviation
Treatment	29	3.01	7.31
Control	26	−0.45	7.99

Let μ_1 and μ_2 denote the mean weight gains for these therapies for the conceptual populations that the samples represent. We test $H_0: \mu_1 = \mu_2$ against $H_a: \mu_1 \neq \mu_2$. If treatment has no effect relative to control, then we would expect the groups to have equal means and equal standard deviations of weight change. For these data, the pooled estimate of the assumed common standard deviation is

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{28(7.31)^2 + 25(7.99)^2}{29 + 26 - 2}} = 7.64.$$

---

[7] The data for both groups are shown in Table 12.18 on page 385 and are in the `Anorexia` data file at the text website.

Now, $\bar{y}_1 - \bar{y}_2 = 3.01 - (-0.45) = 3.46$ has an estimated standard error of

$$se = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 7.64\sqrt{\frac{1}{29} + \frac{1}{26}} = 2.06.$$

The test statistic is

$$t = \frac{\bar{y}_1 - \bar{y}_2}{se} = \frac{3.01 - (-0.45)}{2.06} = 1.68.$$

This statistic has $df = n_1 + n_2 - 2 = 29 + 26 - 2 = 53$. From software, the two-sided P-value is $P = 0.10$, only weak evidence of a different mean using the cognitive behavioral therapy instead of the control condition.

When $df = 53$, the t-score for a 95% confidence interval for $(\mu_1 - \mu_2)$ is $t_{0.025} = 2.006$. The interval is

$$(\bar{y}_1 - \bar{y}_2) \pm t(se) = 3.46 \pm 2.006(2.06), \text{ which is } 3.46 \pm 4.14, \text{ or } (-0.7, 7.6).$$

We conclude that the mean weight change for the cognitive behavioral therapy could be as much as 0.7 pounds lower or as much as 7.6 pounds higher than the mean weight change for the control group. Since the interval contains 0, it is plausible that the population means are identical. This is consistent with the P-value exceeding 0.05 in the test. If the population mean weight change is less for the cognitive behavioral group than for the control group, it is just barely less (less than 1 pound), but if the population mean change is greater, it could be nearly 8 pounds greater. Since the sample sizes are not large, the confidence interval is relatively wide. ■

COMPLETELY RANDOMIZED VERSUS RANDOMIZED BLOCK DESIGN

The anorexia study used a *completely randomized* experimental design: Subjects were randomly assigned to the two therapies. With this design, there's the chance that the subjects selected for one therapy might differ in an important way from subjects selected for the other therapy. For moderate to large samples, factors that could influence results (such as initial weight) tend to balance by virtue of the randomization. For small samples, an imbalance could occur.

An alternative experimental design *matches* subjects in the two samples, such as by taking two girls of the same weight and randomly deciding which girl receives which therapy. This matched-pairs plan is a simple example of a **randomized block design**. Each pair of subjects forms a *block*, and within blocks subjects are randomly assigned to the treatments. Another example of a block design occurs when each subject is measured twice, such as before and after receiving some treatment. With such designs, we would use the methods of Section 7.4 for dependent samples.

SOFTWARE CAN PERFORM INFERENCES ASSUMING EQUAL OR UNEQUAL VARIABILITY

Software can conduct two-sample inference for means with or without the assumption of equal population standard deviations. For example, Table 7.8 illustrates the way SPSS reports results of two-sample t tests. The t test just presented assumes that $\sigma_1 = \sigma_2$. The t statistic that SPSS reports for the "equal variances not assumed" case is the t statistic of Section 7.3,

$$t = \frac{(\bar{y}_2 - \bar{y}_1)}{se}, \text{ with } se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

TABLE 7.8: SPSS Output for Performing Two-Sample t Tests with `Anorexia` Data File. The output also has confidence intervals for the two cases, not shown here.

```
                         t-test for Equality of Means
                                     Sig.      Mean      Std. Error
                    t      df   (2-tailed) Difference  Difference
Equal variances  1.676    53      0.099     3.45690     2.06259
assumed
Equal variances  1.668  50.971   0.102     3.45690     2.07279
not assumed
```

When $n_1 = n_2$, the "equal variances" and "unequal variances" test statistics are identical. They are usually similar if n_1 and n_2 are close or if s_1 and s_2 are close.

If you already have summary statistics, some software (such as SPSS and Stata) can conduct the inferences with them. In Stata, use the `ttesti` command (or a dialog box), by entering n, $\bar{y}$, and s for each group. Table 7.9 shows the analysis of this section for the anorexia data, assuming $\sigma_1 = \sigma_2$. Adding "unequal" to the command line, as shown on page 201 for the housework example, does the t inference without assuming that $\sigma_1 = \sigma_2$. Internet applets are also available.[8]

TABLE 7.9: Stata Software Output (Edited) for Performing Two-Sample t Inferences under Assumption $\sigma_1 = \sigma_2$

```
 . ttesti 29 3.01 7.31 26 -0.45 7.99

Two-sample t test with equal variances
     |  Obs    Mean    Std. Err.  Std. Dev.   [95% Conf. Interval]
   x |   29    3.01   1.357433      7.31     .2294247    5.790575
   y |   26    -.45   1.566968      7.99    -3.677231    2.777231
 ------+--------------------------------------------------------------
  diff |           3.46   2.062968             -.6777895    7.597789

     diff = mean(x) - mean(y)                          t =  1.6772
 Ho: diff = 0                              degrees of freedom =    53
    Ha: diff < 0              Ha: diff != 0               Ha: diff > 0
 Pr(T < t) = 0.9503      Pr(|T| > |t|) = 0.0994   Pr(T > t) = 0.0497
```

If the data show evidence of a potentially large difference in standard deviations (with, say, one sample standard deviation being at least double the other), it is better to use the approximate t test (Section 7.3) that does not make the $\sigma_1 = \sigma_2$ assumption. It can yield a t statistic value much different from the method that assumes $\sigma_1 = \sigma_2$ if s_1 and s_2 are quite different and the sample sizes are unequal.

Many texts and most software present a statistic denoted by F for testing that the population standard deviations are equal. It's not appropriate to conduct this test in order to determine which t method to use. In fact, we don't recommend this test even if your main purpose is to compare variability of two groups. The test assumes that the population distributions are normal, but it is *not* robust to violations of that assumption.

---

[8] For example, the *Comparing Two Means* applet at www.pearsonglobaleditions.com/Agresti.

EFFECT SIZE

In Example 7.8 on page 206 about an anorexia study, is the estimated difference of 3.46 between the mean weight gains for treatment and control groups large, or small, in practical terms? Recall that the size of an estimated difference depends on the units of measurement. These data were in pounds, but if converted to kilograms the estimated difference would be 1.57 and if converted to ounces it would be 55.4.

A standardized way to describe the difference divides it by the estimated standard deviation for each group. This is called the ***effect size*** for comparing means. With sample means of 3.01 and -0.45 pounds and an estimated common standard deviation of $s = 7.64$ pounds, the effect size is

$$\text{Effect size} = \frac{\bar{y}_1 - \bar{y}_2}{s} = \frac{3.01 - (-0.45)}{7.64} = 0.45.$$

The difference between the sample means is less than half a standard deviation. This is usually considered to be a small to moderate difference. The effect is considered to be quite large if the effect size is about 1 (or larger) in absolute value. We obtain the same value for the effect size if we measure the data in different units, such as kilograms or ounces.

A MODEL FOR MEANS

In the second half of this book, we'll learn about advanced methods for analyzing associations among variables. We'll base analyses explicitly on a *model*. For two variables, a ***model*** is a simple approximation for the true relationship between those variables in the population.

Let $N(\mu, \sigma)$ denote a normal distribution with mean μ and standard deviation σ. Let y_1 denote a randomly selected observation from group 1 and y_2 a randomly selected observation from group 2. The hypothesis tested above for comparing means under the assumption $\sigma_1 = \sigma_2$ can be expressed as the model

H_0: Both y_1 and y_2 have a $N(\mu, \sigma)$ distribution.

H_a: y_1 has a $N(\mu_1, \sigma)$ distribution and y_2 has a $N(\mu_2, \sigma)$ distribution, with $\mu_1 \neq \mu_2$.

Under H_0, the population means are equal, with some common value μ. Under H_a, the population means differ. This is a special case of a model that Chapter 12 uses for comparing *several* means.

Sampling distributions and resulting inferences are derived under the assumed model structure. But models are merely convenient simplifications of reality. We do not expect distributions to be exactly normal, for instance. One of the key parts of becoming more comfortable using statistical methods is becoming knowledgeable about which assumptions are most important in a model and how to check the assumptions. Generally, simpler models have benefits. They have fewer parameters to estimate, and inferences can be more powerful. However, when such a model is badly in error, we are better off using a more complex model.

The significance test in Section 7.3 for comparing means results from a slightly more complex model

H_0: y_1 has a $N(\mu, \sigma_1)$ distribution and y_2 has a $N(\mu, \sigma_2)$ distribution.

H_a: y_1 has a $N(\mu_1, \sigma_1)$ distribution and y_2 has a $N(\mu_2, \sigma_2)$ distribution, with $\mu_1 \neq \mu_2$.

Again, under H_0 the population means are equal. But now, no assumption is made about the population standard deviations being equal. If there is reason to expect

the population standard deviations to be very different, or if the data indicate this (with one of the sample standard deviations being at least double the other), then we're better off using analyses based on this model. If the data show that even this model is badly in error, such as when the sample data distributions are so highly skewed that the mean is an inappropriate summary, we're better off using a different model yet. The final section of this chapter presents a model that does not assume normality or use means.

7.6 Other Methods for Comparing Proportions*

Section 7.2 presented large-sample methods for comparing proportions with independent samples. This section presents methods for comparing proportions with (1) dependent samples and (2) small samples.

COMPARING PROPORTIONS FROM DEPENDENT SAMPLES

Section 7.4 presented dependent-samples methods for comparing means. We use the following example to illustrate dependent-samples methods for comparing proportions.

_____

Example 7.9

Belief in Heaven and Hell   A recent General Social Survey asked subjects whether they believed in heaven and whether they believed in hell. Table 7.10 shows results. Of 1214 subjects responding, 875 believed in both, 168 believed in neither, 162 believed in heaven but not in hell, and 9 believed in hell but not in heaven. The row marginal counts (1037, 177) are the (yes, no) totals for belief in heaven. The column marginal counts (884, 330) are the (yes, no) totals for belief in hell.

TABLE 7.10: Belief in Heaven and Belief in Hell			
	Belief in Hell		
Belief in Heaven	Yes	No	Total
Yes	875	162	1037
No	9	168	177
Total	884	330	1214

We will compare the proportions responding *yes* for heaven and for hell. The samples are dependent, because the same 1214 people responded to each question. Let π_1 denote the population proportion who believe in heaven, and let π_2 denote the population proportion who believe in hell. The sample estimates are $\hat{\pi}_1 = 1037/1214 = 0.854$ and $\hat{\pi}_2 = 884/1214 = 0.728$.

If the proportions responding *yes* were identical for heaven and hell, the number of observations in the first row of Table 7.10 would equal the number of observations in the first column. The first cell (the one containing 875 in Table 7.10) is common to both the first row and first column, so the other cell count in the first row would equal the other cell count in the first column. That is, the number of people saying *yes* to heaven but *no* to hell would equal the number of people saying *no* to heaven but *yes* to hell. We can test $H_0: \pi_1 = \pi_2$ using the counts in those two cells. If H_0 is true, then of these people, we expect half to be in each of those two cells.

As in the matched-pairs test for a mean, we reduce the inference to one about a single parameter. For the population in the two cells just mentioned, we test whether half are in each cell. In Table 7.10, of the $162 + 9 = 171$ people who believe in one but not the other, the sample proportion $162/171 = 0.947$ believed

in heaven but not in hell. Under the null hypothesis that the population proportion is 0.50, the standard error of the sample proportion for these 171 observations is $\sqrt{(0.50)(0.50)/171} = 0.038$.

From Section 6.3 (page 164), the z statistic for testing that of the population observations in those two cells half are in each cell is

$$z = \frac{\text{Sample proportion} - H_0 \text{ proportion}}{\text{Standard error}} = \frac{0.947 - 0.50}{0.038} = 11.7.$$

The two-sided P-value equals 0.000. This provides extremely strong evidence against $H_0: \pi_1 = \pi_2$. Based on the sample proportions, the evidence favors a greater population proportion of belief in heaven than in hell. ∎

McNEMAR TEST FOR COMPARING DEPENDENT PROPORTIONS

A simple formula exists for this z test statistic for comparing two dependent proportions. For a table of the form of Table 7.10, denote the cell counts in the two relevant cells by n_{12} for those in row 1 and in column 2 and by n_{21} for those in row 2 and in column 1. The test statistic is

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}.$$

When $n_{12} + n_{21}$ exceeds about 20, this statistic has approximately a standard normal distribution when H_0 is true. This test is often referred to as ***McNemar's test***. For smaller samples, use the binomial distribution (Section 6.7) to conduct the test.

For Table 7.10, the McNemar test uses $n_{12} = 162$, the number of people believing in heaven but not in hell, and $n_{21} = 9$, the number for the reverse. The test statistic is

$$z = \frac{162 - 9}{\sqrt{162 + 9}} = 11.7,$$

as we previously obtained.

CONFIDENCE INTERVAL FOR DIFFERENCE OF DEPENDENT PROPORTIONS

A confidence interval for the difference of proportions is more informative than a significance test. For large samples, this is

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se),$$

where the standard error is estimated using

$$se = \frac{1}{n}\sqrt{(n_{12} + n_{21}) - (n_{12} - n_{21})^2/n}.$$

For Table 7.10, $\hat{\pi}_1 = 1037/1214 = 0.8542$ and $\hat{\pi}_2 = 884/1214 = 0.7282$. The difference $\hat{\pi}_1 - \hat{\pi}_2 = 0.8542 - 0.7282 = 0.126$. For $n = 1214$ observations with $n_{12} = 162$ and $n_{21} = 9$,

$$se = (1/1214)\sqrt{(162 + 9) - (162 - 9)^2/1214} = 0.0101.$$

A 95% confidence interval for $\pi_1 - \pi_2$ equals $0.126 \pm 1.96(0.0101)$, or $(0.106, 0.146)$. We conclude that the population proportion who believe in heaven is between about 0.10 and 0.15 higher than the population proportion who believe in hell.

This confidence interval and McNemar's test are available in statistical software and with Internet applets. See, for instance, the *Comparing Two Proportions* applet at `www.pearsonglobaleditions.com/Agresti`, with the *Two Dependent Samples* option.

FISHER'S EXACT TEST FOR COMPARING PROPORTIONS

The inferences of Section 7.2 for comparing proportions with independent samples are valid for relatively large samples. For small sample sizes, the sampling distribution of $\hat{\pi}_2 - \hat{\pi}_1$ may not be close to normality. You can then compare two proportions π_1 and π_2 using a method called **Fisher's exact test**, due to the eminent British statistician R. A. Fisher.

The calculations for Fisher's exact test are complex and beyond the scope of this text.[9] The principle behind the test is straightforward, however, as Exercise 7.57 shows. Statistical software provides its P-value. As usual, the P-value is the probability of the sample result or a result even more extreme, under the presumption that H_0 is true.

Example 7.10

Depression and Suicide among HIV-Infected Persons A study[10] on psychological impacts of being HIV positive examined rates of major depression and suicidality for HIV-infected and uninfected persons in China. The study used a volunteer sample. In an attempt to make the sample more representative, subjects were recruited from clinics in two very different regions of China, one urban and one rural. Table 7.11 shows results based on a diagnostic interview asking whether the subject had ever attempted suicide. The table also shows output from conducting Fisher's exact test.

TABLE 7.11: Comparison of HIV-Infected and Uninfected Subjects on whether They Have Ever Attempted Suicide

```
 HIV           suicide
         |yes     |no      | Total
positive |    10 |     18 |    28
negative |     1 |     22 |    23
---------+--------+--------+
Total         11      40       51

 STATISTICS FOR TABLE OF HIV BY SUICIDE
 Statistic                         Prob
 Fisher's Exact Test (Left)       0.9995
                     (Right)      0.0068
                     (2-Tail)     0.0075
```

Denote the population proportion who had ever made a suicide attempt by π_1 for those who were HIV positive and by π_2 for those who were HIV negative. Then, $\hat{\pi}_1 = 10/28 = 0.36$ and $\hat{\pi}_2 = 1/23 = 0.04$. We test $H_0: \pi_1 = \pi_2$ against $H_a: \pi_1 > \pi_2$. One of the four counts is very small, so to be safe we use Fisher's exact test.

On the output shown, the right-sided alternative refers to $H_a: \pi_1 - \pi_2 > 0$; that is, $H_a: \pi_1 > \pi_2$. The P-value = 0.0068 gives very strong evidence that the population proportion attempting suicide is higher for those who are HIV positive. The P-value for the two-sided alternative equals 0.0075. This is not double the one-sided P-value because, except in certain special cases, the sampling distribution (called the **hypergeometric distribution**) is not symmetric. ∎

---

[9] For details about Fisher's exact test, see Agresti (2007, pp. 45–48); calculations are available at the *Fisher's Exact Test* applet at www.pearsonglobaleditions.com/Agresti.
[10] H. Jin et al., *Journal of Affective Disorders*, vol. 94 (2006), pp. 269–275.

The test is called *exact* because it uses the actual (hypergeometric) sampling distribution rather than a normal approximation. Corresponding exact confidence intervals exist. They are beyond the scope of this text, but are available in software.

7.7 Nonparametric Statistics for Comparing Groups*

Many statistics have large-sample normal sampling distributions, even when population distributions are not normal. In fact, with random sampling, nearly all parameter estimators have normal distributions, for large sample sizes. Small samples, though, often require additional assumptions. For instance, inferences for means using the *t* distribution assume normal population distributions.

A body of methods exists that makes *no* assumption about the shape of the population distribution. These methods are called **nonparametric**.[11] They contrast with the traditional (so-called *parametric*) methods that assume particular population distributions, such as normality. Nonparametric methods still apply when the normality assumption for methods using the *t* distribution is badly violated. They are primarily useful for small samples, especially for one-sided tests, as parametric methods may then be invalid when the normal population assumption is badly violated. They are also useful when the two groups have highly skewed distributions, because then the mean may not be a meaningful summary measure.

WILCOXON–MANN–WHITNEY TEST

To illustrate, Section 7.5 introduced a *t* distribution method for comparing means that assumes normal population distributions with identical standard deviations. These assumptions are mainly relevant for small samples, namely, when n_1 or n_2 is less than about 20–30. Most nonparametric comparisons of groups also assume identical shapes for the population distributions, but the shapes are not required to be normal. The model for the test is then

H_0: Both y_1 and y_2 have the same distribution.

H_a: The distributions for y_1 and y_2 have the same shape, but the one for y_1 is shifted up or shifted down compared to the one for y_2.

Here, H_a is two-sided. One-sided H_a is also possible.

The most popular test of this type is called the *Wilcoxon* test. This test is an ordinal-level method, in the sense that it uses only the rankings of the observations. The combined sample of $n_1 + n_2$ measurements is ranked from 1 to $n_1 + n_2$, and the means of the ranks are computed for observations in each sample. The test statistic compares the sample mean ranks. For large samples, a *z* test statistic has an approximate standard normal distribution. For small samples, an exact *P*-value is based on how unusual the observed difference between the mean ranks is (under the presumption that H_0 is true) compared to the differences between the mean ranks for all other possible rankings.

Another nonparametric test is the *Mann–Whitney* test. It views all the *pairs* of observations, such that one observation is from one group and the other observations is from the other group. The test statistic is based on the number of pairs for which the observation from the first group was higher. This test is equivalent to the Wilcoxon test, giving the same *P*-value.[12]

---

[11] Hollander et al. (2013) presented an overview of nonparametric statistical methods.
[12] Frank Wilcoxon developed equivalent tests as Henry Mann and D. R. Whitney at about the same time in the late 1940s.

For Example 7.8 comparing weight changes for a cognitive behavioral therapy group and a control group in the anorexia study (page 206), the parametric t test had a P-value of 0.10. The large-sample version of the Wilcoxon–Mann–Whitney test reports similar results, with a P-value of 0.11.

Some software also can report a corresponding confidence interval for the difference between the population medians. The method assumes that the two population distributions have the same shape, but not necessarily bell shaped. The median weight change was 1.4 pounds for the cognitive behavioral therapy group and -0.35 pounds for the control group. Software reports a 95% confidence interval for the difference between the medians of $(-0.6, 8.1)$ pounds.

EFFECT SIZE: PROPORTION OF BETTER RESPONSES FOR A GROUP

Section 7.5 introduced an *effect size* measure, $(\bar{y}_1 - \bar{y}_2)/s$, for summarizing the size of the difference between two groups. When the distributions are very skewed or have outliers, the means are less useful and this effect size summary may be inappropriate. A nonparametric effect size measure is the proportion of pairs of observations (one from each group) for which the observation from the first group was higher. If y_1 denotes a randomly selected observation from group 1 and y_2 a randomly selected observation from group 2, then this measure estimates $P(y_1 > y_2)$.

To illustrate, suppose the anorexia study had four girls, two using a new therapy and two in a control group. Suppose the weight changes were

Therapy group (y_1): 4, 10
Control group (y_2): 2, 6

There are four pairs of observations, with one from each group:

$y_1 = 4, y_2 = 2$ (Group 1 is higher)
$y_1 = 4, y_2 = 6$ (Group 2 is higher)
$y_1 = 10, y_2 = 2$ (Group 1 is higher)
$y_1 = 10, y_2 = 6$ (Group 1 is higher)

Group 1 is higher in three of the four pairs, so the estimate of $P(y_1 > y_2)$ is 0.75. If two observations had the same value, we would count it as y_1 being higher for half the pair (rather than 1 or 0).

Under H_0 of no effect, $P(y_1 > y_2) = 0.50$. The farther $P(y_1 > y_2)$ falls from 0.50, the stronger the effect. For the full anorexia data set analyzed on page 206, the sample estimate of $P(y_1 > y_2)$ is 0.63. The estimated probability that a girl using the cognitive behavioral therapy has a larger weight gain than a girl using the control therapy is 0.63.

When the two groups have normal distributions with the same standard deviation, a connection exists between this effect size and the parametric one, $(\mu_1 - \mu_2)/\sigma$. For example, when $(\mu_1 - \mu_2)/\sigma = 0$, then $P(y_1 > y_2) = 0.50$; when $(\mu_1 - \mu_2)/\sigma = 0.5$, then $P(y_1 > y_2) = 0.64$; when $(\mu_1 - \mu_2)/\sigma = 1$, then $P(y_1 > y_2) = 0.71$; when $(\mu_1 - \mu_2)/\sigma = 2$, then $P(y_1 > y_2) = 0.92$. The effect is relatively strong if $P(y_1 > y_2)$ is larger than about 0.70 or smaller than about 0.30.

TREATING ORDINAL VARIABLES AS QUANTITATIVE

Social scientists often use parametric statistical methods for quantitative data with variables that are only ordinal. They do this by assigning scores to the ordered categories. Example 6.2 (page 157) on political ideology showed an example of this. Sometimes the choice of scores is straightforward. For categories (liberal, moderate,

conservative) for political ideology, any set of equally spaced scores is sensible, such as (1, 2, 3) or (0, 5, 10). When the choice is unclear, such as with categories (not too happy, pretty happy, very happy) for happiness, it is a good idea to perform a sensitivity study. Choose two or three reasonable sets of potential scores, such as (0, 5, 10), (0, 6, 10), (0, 7, 10), and check whether the ultimate conclusions are similar for each. If not, any report should point out how conclusions depend on the scores chosen.

Alternatively, nonparametric methods are valid with ordinal data. The reason is that nonparametric methods do not use quantitative scores but rather rankings of the observations, and rankings are ordinal information. However, this approach works best when the response variable is continuous (or nearly so), so each observation has its own rank. When used with ordered categorical responses, such methods are often less sensible than using parametric methods that treat the response as quantitative. The next example illustrates this.

Example 7.11

Alcohol Use and Infant Malformation   Table 7.12 refers to a study of maternal drinking and congenital malformations. After the first three months of pregnancy, the women in the sample completed a questionnaire about alcohol consumption. Following childbirth, observations were recorded on presence or absence of congenital sex organ malformations. Alcohol consumption was measured as average number of drinks per day.

TABLE 7.12: Infant Malformation and Mother's Alcohol Consumption

Malformation	Alcohol Consumption				
	0	< 1	1–2	3–5	≥6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1
Total	17,114	14,502	793	127	38

Source: Graubard, B. I., and Korn, E. L., *Biometrics*, vol. 43 (1987), pp. 471–476.

Is alcohol consumption associated with malformation? One approach to investigate this is to compare the mean alcohol consumption of mothers for the cases where malformation occurred to the mean alcohol consumption of mothers for the cases where malformation did not occur. Alcohol consumption was measured by grouping values of a quantitative variable. To find means, we assign scores to alcohol consumption that are midpoints of the categories, that is, (0, 0.5, 1.5, 4.0, 7.0), the last score (for ≥6) being somewhat arbitrary. The sample means are then 0.28 for the absent group and 0.40 for the present group, and the t statistic of 2.56 has P-value of 0.01. There is strong evidence that mothers whose infants suffered malformation had a higher mean alcohol consumption.

An alternative, nonparametric, approach assigns ranks to the subjects and uses them as the category scores. For all subjects in a category, we assign the average of the ranks that would apply for a complete ranking of the sample. These are called *midranks*. For example, the 17,114 subjects at level 0 for alcohol consumption share ranks 1 through 17,114. We assign to each of them the average of these ranks, which is the midrank $(1 + 17, 114)/2 = 8557.5$. The 14,502 subjects at level <1 for alcohol consumption share ranks 17,115 through $17, 114 + 14, 502 = 31, 616$, for a midrank of $(17, 115 + 31, 616)/2 = 24, 365.5$. Similarly the midranks for the last three categories are 32,013, 32,473, and 32,555.5. Used in a large-sample Wilcoxon test, these scores yield much less evidence of an effect ($P = 0.55$).

Why does this happen? Adjacent categories having relatively few observations necessarily have similar midranks. The midranks (8557.5, 24365.5, 32013, 32473, and

32555.5) are similar for the final three categories, since those categories have considerably fewer observations than the first two categories. A consequence is that this scoring scheme treats alcohol consumption level 1–2 (category 3) as much closer to consumption level ≥ 6 (category 5) than to consumption level 0 (category 1). This seems inappropriate. It is better to use your judgment by selecting scores that reflect well the distances between categories. ∎

Although nonparametric methods have the benefit of weaker assumptions, in practice social scientists do not use them as much as parametric methods. Partly this reflects the large sample sizes for most studies, for which assumptions about population distributions are not so vital. In addition, nonparametric methods for multivariate data sets are not as thoroughly developed as parametric methods.

7.8 Chapter Summary

This chapter introduced methods for comparing two groups. For quantitative response variables, inferences apply to the difference $\mu_2 - \mu_1$ between population means. For categorical response variables, inferences apply to the difference $\pi_2 - \pi_1$ between population proportions.

In each case, the significance test analyzes whether 0 is a plausible difference. If the confidence interval contains 0, it is plausible that the parameters are equal. Table 7.13 summarizes the methods for ***independent*** random samples, for which observations in the two samples are not matched. This is the most common case in practice.

TABLE 7.13: Comparison Methods for Two Groups, for Independent Random Samples

	Type of Response Variable	
	Categorical	Quantitative
Estimation		
1. Parameter	$\pi_2 - \pi_1$	$\mu_2 - \mu_1$
2. Point estimate	$\hat{\pi}_2 - \hat{\pi}_1$	$\bar{y}_2 - \bar{y}_1$
3. Standard error	$se = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$	$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
4. Confidence interval	$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se)$	$(\bar{y}_2 - \bar{y}_1) \pm t(se)$
Significance testing		
1. Assumptions	Randomization	Randomization
	≥ 10 observations in each	Normal population dist.'s
	category, for each group	(robust, especially for large n's)
2. Hypotheses	$H_0: \pi_1 = \pi_2$	$H_0: \mu_1 = \mu_2$
	$(\pi_2 - \pi_1 = 0)$	$(\mu_2 - \mu_1 = 0)$
	$H_a: \pi_1 \neq \pi_2$	$H_a: \mu_1 \neq \mu_2$
3. Test statistic	$z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{se_0}$	$t = \frac{\bar{y}_2 - \bar{y}_1}{se}$
4. P-value	Two-tail probability from standard normal or t distribution	
	(Use one tail for one-sided alternative)	

- Both for differences of proportions and differences of means, confidence intervals have the form

$$\text{Estimated difference} \pm \text{Score}(se)$$

using a z-score for proportions and t-score for means. In each case, the test statistic equals the estimated difference divided by the standard error.

- For **dependent** samples, each observation in one sample matches with an observation in the other sample. For quantitative variables, we compare means by analyzing the mean of difference scores computed between the paired observations. The **paired-difference** confidence interval and test procedures are the one-sample methods of Chapters 5 and 6 applied to the difference scores.

- Another approach for comparing means makes the extra assumption that the normal population distributions have equal standard deviations. This approach pools the standard deviations from the two samples to find a common estimate.

- For comparing proportions, with independent samples the small-sample test is **Fisher's exact test**. For dependent samples, **McNemar's test** compares the number of subjects who are in category 1 in the first sample and category 2 in the second sample to the number of subjects who are in category 2 in the first sample and category 1 in the second.

- **Nonparametric** statistical methods make no assumption about the shape of the population distribution. Most such methods use the ranks of the observations.

At this stage, you may feel confused about which method to use for any given situation. It may help if you use the following checklist. Ask yourself, is the analysis about

- Means or proportions (quantitative or categorical response variable)?

- Independent samples or dependent samples?

- Confidence interval or significance test?

Exercises

Practicing the Basics

7.1. The biannual ABDC survey of school students indicated that 78% of tenth graders in 2016 considered having a car to be essential or very important, compared to 45% when the survey was first conducted in 1968. Are the sample percentages of 45% in 1968 and 78% in 2016 based on *independent* samples, or *dependent* samples? Explain.

7.2. An annual survey of public opinion considers the public views about adequacy of policies taken to protect global environment, with a random sample of about 5000 adults from each of 20 developing nations each year. In 2000, 15% of them expressed positive views about adequacy of the measures taken at the global level. In 2016, 75% of them expressed a positive attitude about the same.

(a) Explain what it would mean for these results to be based on (i) *independent* samples, (ii) *dependent* samples.

(b) If we compare results in 2000 and 2016, identify the response variable and the explanatory variable, and specify whether the response variable is quantitative or categorical.

7.3. An annual survey on consumer expenditure for a country estimated 56% of its citizens to own large cars in 2017 as compared to 29% in 2012.

(a) Estimate the difference between the proportions of large car owners over the periods.

(b) If standard error is 0.040 for each proportion, find the standard error of the difference. Interpret.

7.4. When a recent Eurobarometer survey asked subjects in each European Union country whether they would be willing to pay more for energy produced from renewable sources than for energy produced from other sources, the proportion answering *yes* varied from a high of 0.52 in Denmark ($n = 1008$) to a low of 0.14 in Lithuania ($n = 1002$). For this survey,

(a) Estimate the difference between Denmark and Lithuania in the population proportion of *yes* responses.

(b) From the $se = \sqrt{\hat{\pi}(1 - \hat{\pi})/n}$ formula, the proportion estimates have $se = 0.0157$ for Denmark and $se = 0.0110$ for Lithuania. Use these to find the se for the difference estimate in (a). Interpret this se.

7.5. The Regional Center for Health Survey recently estimated that the mean weight for adult Indian women was 52 kg in 1968 and 65 kg in 2012.

(a) Suppose these estimates had a standard error of 1.2 kg each year. Estimate the increase in mean weight in the population from 1968 to 2012, and find and interpret the standard error of that estimate.

(b) Show that the estimated mean in 2012 was 1.25 times the estimated mean in 1968. Express this in terms of the percentage increase.

(c) The estimated mean weights for men were 71 kg in 1968 and 81 kg in 2012. Find and interpret the difference and the ratio.

7.6. In 2015, a country's planning commission estimated its median household consumption expenditure in domestic currency to be 265,000 in urban and 51,000 in rural sectors.

(a) Identify the response variable and the explanatory variable.

(b) Compare the households using a (i) difference, (ii) ratio.

7.7. In 2016, the Educational Research Wing reported that the school dropout rate in a country was 375 per 10,000 male students (composed of 8% of urban male students and 12% of rural male students), and 425 per 10,000 female students.

(a) Find the ratio of proportions of dropouts, for females relative to males. Interpret.

(b) Find the difference of proportions dropout. Interpret.

(c) Which measure do you think better summarizes these data? Why?

7.8. According to a university database, the annual probability that a student between the ages of 20 and 24 fails to go abroad for higher studies is 0.00237 for males and 0.00069 for females.

(a) Compare these rates using the difference of proportions.

(b) Compare these rates using the ratio of proportions.

(c) Which of the two measures seems to better summarize results when both proportions are very close to 0? Explain.

7.9. The Global Charitable Trust asked, "How often do you donate?" The response *sometimes* was given by 22.1% of the 1532 respondents in Country A and by 46.7% of the 1098 respondents in Country B.

(a) Assuming random sampling, the 95% confidence interval for the difference between corresponding population proportions is (0.21, 0.28). Interpret it.

(b) The *P*-value is <0.0001 for testing the null hypothesis that the corresponding population proportions are equal. Interpret.

7.10. For a random sample of university teachers, 50% were positive about the vice chancellor's performance. A similar poll two months later indicated an approval rate of 47%. A 99% confidence interval for the change in the population proportions is (−0.09, 0.03). Explain why **(a)** there may have been no change in support, **(b)** if a decrease in support occurred, it may have been fairly important, whereas if an increase in support occurred, it was probably so small as to be substantively unimportant.

7.11. A study group at a high school has interviewed random samples of students at higher grades several times since 2010. Most recently, of the students who reported reading Encyclopedia Britannica, the percentage who reported that reading "to portray their image as studious" is an important reason for borrowing such books from the library was 35.2% of 4690 women students and 59.5% of 5120 male students. For comparing men and women,

(a) Show that the standard error for the estimated difference between the corresponding population proportions equals 0.0098.

(b) Show that the 95% confidence interval for the difference is (0.22, 0.26). Interpret.

7.12. The study mentioned in the previous exercise estimated in 2010 that 3% had done better in interschool competitions because of reading such books; this was 4.5% at the latest survey.

(a) Specify assumptions, notation, and hypotheses for a two-sided test comparing the corresponding population proportions.

(b) The test statistic $z = 5.53$ and the *P*-value close to 0.0000. Interpret the *P*-value.

(c) Some might argue that the result in (b) reflects *statistical significance* but not *practical significance*. Explain the basis of this argument, and explain why you learn more from the 95% confidence interval, which is (0.009, 0.02).

7.13. In Great Britain, the Time Use Survey (www.statistics.gov.uk) studied how a random sample of Brits spend their time on a typical day. Of those working full time, 55% of 1219 men and 74% of 733 women reported spending some time on cooking and washing up during a typical day. Find and interpret a 95% confidence interval for the difference in the population proportions.

7.14. Table 7.14 summarizes responses from General Social Surveys in 1977 and in 2014 to the statement "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family." Let π_1 denote the population proportion who agreed with this statement in 1977, and let π_2 denote the population proportion in 2014.

(a) Show that $\hat{\pi}_1 - \hat{\pi}_2 = 0.347$, with standard error 0.017.

(b) Show that the 95% confidence interval for $\pi_1 - \pi_2$ is (0.31, 0.38). Interpret.

(c) Explain how results would differ for comparing the proportions who did *not* agree in the two years.

TABLE 7.14

Year	Agree	Disagree	Total
1977	989	514	1503
2014	515	1140	1655

7.15. Refer to the previous exercise. In 2014, of 745 male respondents, 247 (33.2%) agreed. Of 910 female respondents, 268 (29.5%) agreed.

(a) Set up notation and specify hypotheses for the hypothesis of no difference between the population proportions of males and of females who would agree.

(b) Estimate the population proportion presuming H_0, find the standard error of the sample difference of proportions, and find the test statistic.

(c) Find the *P*-value for the two-sided alternative. Interpret.

(d) Of 1032 respondents having less education than a college degree, 37.1% agreed. Of 623 respondents having at least a college degree, 21.2% agreed. Which variable, gender or educational level, seems to have had the greater influence on opinion? In other words, did opinion tend to differ more between men and women, or between those with and without a degree? Explain.

7.16. In a survey conducted by the education department, college teachers were asked if they frequently used interactive teaching learning methods. Table 7.15 shows software output. Treating these observations as a random sample from the population of interest,

(a) State a research question that could be addressed with this output.

(b) Interpret the reported confidence interval.

(c) Interpret the reported *P*-value.

TABLE 7.15

Sample	Yes	No	Sample prop
1. Seniors	335	1200	0.2792
2. Juniors	425	1260	0.3373

```
estimate for p(1) - p(2): -0.058
95% CI for p(1) - p(2): (-0.021, -0.095)
Test for difference = 0 (vs not = 0): z = -3.13
P-value = 0.002
```

7.17. A study of compulsive buying behavior (uncontrolled urges to buy) conducted a national telephone survey in 2004 of adults ages 18 and over.[13] Of 800 men, 44 were judged to be compulsive buyers according to the Compulsive Buying Scale. Of 1501 women, 90 were judged to be compulsive buyers. Conduct an inference to analyze whether one sex is more likely than the other to be a compulsive buyer. Interpret, including relevant assumptions for the interpretations to be valid.

7.18. Table 7.16 shows results from the 2014 General Social Survey on belief in an afterlife, classified by sex. Conduct all steps of a significance test, using $\alpha = 0.05$, to compare the population proportions of females and males who would respond *yes* to belief in an afterlife. If you have made an error in your decision, what type of error is it, Type I or Type II?

TABLE 7.16

Sex	Belief in Afterlife		Total
	Yes	No	
Female	600	109	709
Male	424	170	594

7.19. A GSS reported that the 486 females had a mean of 8.3 close friends ($s = 15.6$) and the 354 males had a mean of 8.9 close friends ($s = 15.5$).

(a) A 95% confidence interval for the difference between the population means for males and for females is (-1.5, 2.7). Interpret.

(b) For each sex, does it seem like the distribution of number of close friends is normal? Explain why this does not invalidate the result in (a) but may affect the usefulness of the interval.

7.20. Table 7.17 summarizes the number of hours spent in laboratory per week by researchers, based on a recent study.

(a) Estimate the difference between population means for junior and senior fellows.

(b) Show that the estimated standard error of the sample difference is 0.58. Interpret.

(c) Show that a 99% confidence interval for the difference is (0.51, 3.49). Interpret.

TABLE 7.17

Research fellow	Sample Size	Laboratory Hours	
		Mean	Standard Deviation
Junior	415	5.5	8.2
Senior	560	7.5	9.8

7.21. A study evaluated the degree of addiction that children form to cartoon shows once they start watching them in their leisure hours. The study used a random sample of 445 school students in two metropolitan cities who had started watching cartoon shows by the commencement of the study. The response variable was from the Checklist on Student Behavior, a list of 15 questions such as "Have you ever tried to stop watching but couldn't?" The checklist score is the total number of questions to which a student answered *yes*. The final checklist means describing such addiction were 6.5 ($s = 2.4$) for the 350 TV show observers and 2.0 ($s = 4.3$) for the 95 ex-observers.

(a) Software reports a 95% confidence interval of (3.60, 5.40). Interpret.

(b) Was the checklist sample data distribution for ex-observers approximately normal? How does this affect inference? Why?

---

[13] Koran et al., *American Journal of Psychiatry*, vol. 163 (2006), p. 1806.

7.22. In Great Britain, the Time Use Survey[14] studied how a random sample of Brits spend their time on a typical day. For those who reported working full time, Table 7.18 reports the mean and standard deviation of the reported average number of minutes per day spent on cooking and washing up.

TABLE 7.18

| Sex | Sample Size | Cooking and Washing Up Minutes | |
		Mean	Standard Deviation
Men	1219	23	32
Women	733	37	16

(a) Estimate the difference between the means for women and men, and find its standard error.

(b) Compare the population means using a two-sided significance test. Interpret the P-value.

7.23. A current employees' behavior survey asked, "How many days in the past 15 days have you felt demotivated?" Software reported sample means of 5.9 for executives and 2.1 for nonexecutives, with a 95% confidence interval comparing them of (1.5, 3.2), a t statistic of 5.2, and a P-value of 0.000. Interpret these results.

7.24. A survey for 2016, comparing retired and in-service persons on the number of hours a day that they spent on gardening, gave

```
--------------------------------------------
 Group        n    Mean   StdDev  Std Error Mean
 Retired     875   4.65   1.80       0.073
 In-service  525   2.75   1.47       0.085
--------------------------------------------
```

(a) Conduct all parts of a significance test to analyze whether the population means differ for retired and in-service persons. Interpret the P-value, and report the conclusion for α-level = 0.05.

(b) If you were to construct a 95% confidence interval comparing the means, would it contain 0? Answer based on the result of (a), without finding the interval.

(c) Do you think that the distribution of hours on gardening is approximately normal? Why or why not? Does this affect the validity of your inferences? Why?

7.25. For a 2015 study, Table 7.19 shows software output for evaluating the number of hours of Internet surfing per day by age.

(a) Interpret the reported confidence interval. Can you conclude that one population mean is higher? If so, which one? Explain.

(b) Interpret the reported P-value.

(c) Explain the connection between the result of the significance test and the result of the confidence interval.

TABLE 7.19

```
Race          n   Mean StdDev Std Error Mean
Teenagers    355  6.13  2.57      0.1456
Retired     1365  4.78  1.80      0.0576

Difference = mu (Teenagers) - mu (Retired)
Estimate for difference : 1.19
95% CI for difference: (1.08, 1.62)
T-Test of difference = 0:
T-value = 9.64,
P-value = 0.000
```

7.26. In a study comparing various drink types (tea, energy drink, coffee with cream, lime soda, etc.) on teenagers' desire for additional drinks after various lengths of time, coffee with cream had the least effect. In one analysis for coffee with cream, the researchers reported that desire ratings were significantly higher at 20 minutes compared with the baseline ($t = 3.68$, $df = 25$, $P = 0.0011$) but not significantly different for the other time points (P-values > 0.45). Explain why these analyses used statistical methods for dependent samples, identify the sample size, and interpret the P-value for the significant result.

7.27. In a study[15] of the effect of the compound tomoxetine as a treatment for adult attention deficit hyperactivity disorder (ADHD), the 21 subjects had an ADHD rating scale mean of 30.0 ($s = 6.7$) at baseline and 21.5 ($s = 10.1$) after three weeks of treatment. The standard deviation was 9.84 for the 21 changes in rating. The authors reported a paired t statistic of 3.96 with $df = 20$. Show how the authors constructed the t statistic, and report and interpret the P-value for a two-sided test.

7.28. As part of her class project, a student at the University of Florida randomly sampled 10 fellow students to investigate their most common social activities. As part of the study, she asked the students to state how many times they had done each of the following activities during the previous year: going to a movie, going to a sporting event, or going to a party. Table 7.20 shows the data.

(a) To compare the mean movie attendance and mean sports attendance using statistical inference, should we treat the samples as independent or dependent? Why?

(b) For the analysis in (a), software shows results in Table 7.21. Interpret the 95% confidence interval shown.

(c) Show how the test statistic shown in the output was obtained from the other information given. Report the P-value, and interpret in context.

---

[14] www.statistics.gov.uk.
[15] T. Spencer et al., *American Journal of Psychiatry*, vol. 155 (1998), pp. 693–695.

TABLE 7.20

Student	Movies	Sports	Parties
	Activity		
1	10	5	25
2	4	0	10
3	12	20	6
4	2	6	52
5	12	2	12
6	7	8	30
7	45	12	52
8	1	25	2
9	25	0	25
10	12	12	4

TABLE 7.21

```
             n    Mean   Std Dev   Std Error Mean
movies      10  13.000   13.174       4.166
sports      10   9.000    8.380       2.650
Difference  10   4.000   16.166       5.112


95% CI for mean difference: (-7.56, 15.56)
T-Test of mean difference = 0 (vs not = 0):
          T-Value = 0.78    P-Value = 0.454
```

7.29. Refer to the previous exercise. For comparing parties and sports, software reports a 95% confidence interval of $(-3.33, 28.93)$ and a *P*-value of 0.106. Explain the connection between the results of the significance test and the confidence interval.

7.30. A clinical psychologist wants to choose between two therapies for treating mental depression. For six patients, she randomly selects three to receive therapy A, and the other three receive therapy B. She selects small samples for ethical reasons; if her experiment indicates that one therapy is superior, that therapy will be used on her other patients having these symptoms. After one month of treatment, the improvement is measured by the change in score on a standardized scale of mental depression severity. The improvement scores are 10, 20, 30 for the patients receiving therapy A, and 30, 45, 45 for the patients receiving therapy B.

(a) Using the method that assumes a common standard deviation for the two therapies, show that the pooled $s = 9.35$ and $se = 7.64$.

(b) When the sample sizes are very small, it may be worth sacrificing some confidence to achieve more precision. Show that the 90% confidence interval for $(\mu_2 - \mu_1)$ is (3.7, 36.3). Interpret.

(c) Estimate and interpret the effect size.

7.31. Refer to the previous exercise. To avoid bias from the samples being unbalanced with such small *n*, the psychologist redesigned the experiment. She forms three pairs of subjects, such that the patients matched in any given pair are similar in health and socioeconomic status. For each pair, she randomly selects one subject for each therapy. Table 7.22 shows the improvement scores, and Table 7.23 shows some results of using SPSS to analyze the data.

(a) Compare the means by (i) finding the difference of the sample means for the two therapies, (ii) finding the mean of the difference scores. Compare.

(b) Verify the standard deviation of the differences and standard error for the mean difference.

(c) Verify the confidence interval shown for the population mean difference. Interpret.

(d) Verify the test statistic, df, and *P*-value for comparing the means. Interpret.

TABLE 7.22

Pair	Therapy A	Therapy B
1	10	30
2	20	45
3	30	45

TABLE 7.23

```
             Paired Samples Statistics
 Variable    Mean   N  Std. Dev.   Std. Error
 Therapy A   20.00  3   10.000       5.774
 Therapy B   40.00  3    8.660       5.000


               Paired Samples Test
                        Std. Error       Sig.
         Mean  Std. Dev.  Mean    t   df  (2-tailed)
Yes - No 20.00   5.00    2.887  6.93  2    .020
```

7.32. A study[16] of treatments for obesity for rural women examined the impact of a six-month behavioral weight loss program delivered by phone either one-on-one with a counselor or to a group via conference call. A sample of rural women classified as obese by their BMI were randomly assigned to the two conditions. Table 7.24 shows software results of two-sample comparisons of mean weight loss (in kg) over the six months.

(a) State the assumptions on which each significance test is based, and explain how to interpret the results.

---

[16] C. Befort et al., *Eating Behaviors*, vol. 11 (2010), pp. 11–17.

(b) Assuming equal population standard deviations, a 95% confidence interval for the difference in mean weight changes for group versus individual treatments is (1.45, 9.35). Interpret.

TABLE 7.24

		t	df	t-test for Equality of Means Sig. (2-tailed)	Mean Difference	Std. Error Difference
wtloss	Equal variances assumed	2.82	25	0.008	5.4	0.522
	Equal variances not assumed	2.91	23	0.009	5.4	0.538

7.33. For the survey of students described in Exercise 1.11, the responses on political ideology had a mean of 3.18 and standard deviation of 1.72 for the 51 nonvegetarian students and a mean of 2.22 and standard deviation of 0.67 for the 9 vegetarian students. When we use software to compare the means with a significance test, we obtain

```
-----------------------------------------
 Variances      T       DF     P-value
 Unequal       2.915    30.9    0.0066
 Equal         1.636    58.0    0.1073
-----------------------------------------
```

Explain why the results of the two tests differ so much. What would you conclude about whether the population means are equal?

7.34. In 2014, the General Social Survey asked about the number of hours a week spent on the World Wide Web, not counting e-mail. The 778 females had a mean of 11.2 and standard deviation of 13.7. The 620 males had a mean of 11.9 and standard deviation of 16.0. Use these results to make an inference comparing males and females in the population. Since the sample standard deviations are not dramatically different, your inference can assume equality of population standard deviations.

7.35. Two new short courses have been proposed for helping students who suffer from severe math phobia, scoring at least 8 on a measure of math phobia that falls between 0 and 10 (based on responses to 10 questions). A sample of 10 such students were randomly allocated to the two courses. Following the course, the drop in math phobia score was recorded. The sample values were

Course A: 0, 2, 2, 3, 3
Course B: 3, 6, 6, 7, 8

(a) Make an inferential comparison of the means, assuming equal population standard deviations. Interpret your results.

(b) Find and interpret the effect size $(\bar{y}_B - \bar{y}_A)/s$.

(c) Using software, report and interpret the *P*-value for the two-sided Wilcoxon test.

(d) Estimate and interpret the effect size $P(y_B > y_A)$.

7.36. Recent years have seen impressive improvements in systems for automatically recognizing speech. Research in comparing the quality of different speech recognition systems often uses as a benchmark test a series of isolated words, checking how often each system makes errors recognizing the word. Table 7.25 shows an example[17] of one such test, comparing two speech recognition systems, called generalized minimal distortion segmentation (GMDS) and continuous density hidden Markov model (CDHMM).

(a) Estimate the population proportion correct for each system.

(b) Show all steps of McNemar's test to compare the population proportions. Interpret.

(c) Construct a 95% confidence interval to compare the population proportions. Interpret.

TABLE 7.25

GMDS	CDHMM Correct	Incorrect	Total
Correct	1921	58	1979
Incorrect	16	5	21
Total	1937	63	2000

7.37. A General Social Survey asked subjects their opinions about government spending on health and on law enforcement. Table 7.26 shows results.

(a) Find the sample proportion favoring increased spending, for each item.

(b) Test whether the population proportions are equal. Report the *P*-value, and interpret.

(c) Analyze the data using a 95% confidence interval. Interpret.

TABLE 7.26

Health Spending	Law Enforcement Spending	
	Increase	Decrease
Increase	292	25
Decrease	14	9

7.38. The World Values Survey has asked if homosexuality is justifiable, on a scale from 1 ("never") to 10 ("always"). In 1981–1984 surveys in the United States of 2325 subjects, 62.4% gave response 1. In 2010–2014 surveys of 2232 respondents, 24.0% gave response 1. A report claimed that this was evidence of a change in the population in tolerance toward homosexuality. Do these data support this claim? Justify your answer which an inferential analysis, in which you state relevant assumptions.

7.39. An experimental study[18] of young children's moral behavior used as subjects 32 three-year-old girls, half assigned to each of two conditions. Each girl and two actors created a picture or clay sculpture, after which one actor left the room. In the Harm condition, the remaining actor then destroyed the absent actor's picture or sculpture. In a Control condition, the actor did not harm it. Upon return of the actor, in the Harm condition 7 of 16 children acted prosocially such as tattling on the actor, whereas in the Control condition none did. The study authors concluded, "This is the first study to show that children as young as three years of age actively intervene in third-party moral transgressions." Table 7.27 shows results using software for conducting Fisher's exact test.

TABLE 7.27

```
                response
condition | yes     | no      | Total
harm      |     7 |       9 |    16
control   |     0 |      16 |    16
Total           7        25       32

        Fisher's Exact Test
Left-sided P-value        1.0000
Right-sided P-value       0.0034
Two-sided P-value         0.0068
```

(a) Why is Fisher's exact test used to compare the groups?

(b) Report and interpret the *P*-value for the alternative hypothesis that the probability of prosocial behavior is higher for the Harm condition.

7.40. In a study about lesbianism, 45 young adults were asked whether they had ever had a same-gender sexual relationship. Table 7.28 shows results. Use software to test whether the probability of this is higher for those raised by lesbian mothers. Interpret.

TABLE 7.28

Mother	Same-Gender Relationship	
	Yes	No
Lesbian	6	19
Heterosexual	0	20

Concepts and Applications

7.41. For the `Students` data file (Exercise 1.11 on page 21), use graphical and numerical summaries and inferential statistical methods for the following:

(a) Compare political ideology of students identifying as Democrats and as Republicans.

(b) Compare opinions of males and females about legalized abortion.

(c) Compare the mean weekly time spent watching TV to the mean weekly time in sports and other physical exercise.

7.42. For the data file created in Exercise 1.12, with variables chosen by your instructor, state a research question and conduct inferential statistical analyses. Prepare a report that summarizes your findings. In this report, also use graphical and numerical methods to describe the data.

7.43. Using the *Comparing Two Means* applet at `www.pearsonglobaleditions.com/Agresti`, construct two scenarios of independent samples of four men and four women with y = number of hours spent on Internet in past week having $\bar{y}_1 = 5$ and $\bar{y}_2 = 10$, such that for testing H_0: $\mu_1 = \mu_2$ against H_a: $\mu_1 \neq \mu_2$, **(a)** *P*-value > 0.10, **(b)** *P*-value < 0.01. What differs in the two cases to make the *P*-values so different?

7.44. Exercise 3.6 on page 70 showed data on carbon dioxide emissions, a major contributor to global warming, for advanced industrialized nations. State a research question for these data that involves comparing means or proportions. Conduct an investigation to answer this question.

7.45. Pose null and alternative hypotheses about the relationship between time spent on the Internet (variable WWWHR for the GSS) and a binary explanatory that you believe may be associated with Internet use. Using the most recent GSS data on these variables at `sda.berkeley.edu/GSS`, conduct the test. Prepare a short report summarizing your analysis.

7.46. Browse one or two daily online newspapers such as *The New York Times*. Find an article about a research

---

[18] A. Vaish et al., *British Journal of Developmental Psychology*, vol. 2 (2011), pp. 124–130.

study that compared two groups. Prepare a short report that answers the following questions:

(a) What was the purpose of the research study?

(b) Identify explanatory and response variables.

(c) Can you tell whether the statistical analysis used (1) independent samples or dependent samples, (2) a comparison of proportions or a comparison of means, (3) a significance test or a confidence interval? Explain.

7.47. A study[19] considered whether greater levels of TV watching by teenagers were associated with a greater likelihood of committing aggressive acts over the years. The researchers randomly sampled 707 families in two counties in northern New York State and made follow-up observations over 17 years. They observed whether a sampled teenager later conducted any aggressive act against another person, according to a self-report by that person or by their mother. Of 88 cases with less than 1 hour per day of TV watching, 5 had committed aggressive acts. Of 619 cases with at least 1 hour per day of TV, 154 had committed aggressive acts. Analyze these data, summarizing your analyses in a short report.

7.48. According to an article in *The New York Times* (December 6, 2015), the number of annual gun homicides per million people is 31.2 in the United States, 5.6 in Canada, 2.3 in the Netherlands, 0.9 in England, and 0.1 in Japan. Show an informative way to compare these rates between the United States and other countries.

7.49. The World Values Survey asks if homosexuality is justifiable, on a scale from 1 ("never") to 10 ("always"). In 1981–1984 surveys, the 2325 respondents in the United States had a mean response of 2.4 and standard deviation of 2.3. In 2010–2014 surveys, the 2232 U.S. respondents had a mean of 5.4 and standard deviation of 3.4. A report about the results stated, "The mean tolerance level for homosexuality was 3 units higher in 2010–2014 than in 1981–1984. If the true means were equal, a difference of this size could be expected less than 0.01% of the time. For samples of this size, 95% of the time one would expect this difference in means to be within 0.2 of the true value."

(a) Explain how this conclusion refers to the results of a (i) confidence interval, (ii) significance test.

(b) Describe how you would explain the results of the study to someone who has not studied inferential statistics.

7.50. The results in Table 7.29 are from a study[20] of physical attractiveness and subjective well-being. A sample of college students were rated by a panel on their physical attractiveness. The table presents the number of dates in the past three months for students rated in the top or bottom quartile of attractiveness. Analyze these data, and interpret.

TABLE 7.29

	No. Dates, Men			No. Dates, Women		
Attractiveness	Mean	Std. Dev.	n	Mean	Std. Dev.	n
More	9.7	10.0	35	17.8	14.2	33
Less	9.9	12.6	36	10.4	16.6	27

7.51. In the World Values Survey, interviews of 1902 subjects in the Netherlands found that 29.2% reported having confidence in the European Union and 33.0% reported having confidence in the government of the Netherlands. Do you have enough information to make an inferential comparison of the percentages? If so, do so. If not, what else would you need to know?

7.52. In 2011, the United States Supreme Court dealt with a sex discrimination case in which women managers at Walmart earned $14,500 a year less, on the average, than their male counterparts. If you were also given the standard errors of the annual mean salaries for male and female managers at Walmart, would you have enough information to determine whether this is a "statistically significant" difference? Explain.

7.53. The International Adult Literacy Survey[21] was a 22-country study in which nationally representative samples of adults were interviewed and tested at home, using the same literacy test having scores that could range from 0 to 500. For those of age 16–25, some of the mean prose literacy scores were United Kingdom 273.5, New Zealand 276.8, Ireland 277.7, United States 277.9, Denmark 283.4, Australia 283.6, Canada 286.9, the Netherlands 293.5, Norway 300.4, and Sweden 312.1. The website does not provide sample sizes or standard deviations. Suppose each sample size was 250 and each standard deviation was 50. How far apart do two sample means have to be before you feel confident that an actual difference exists between the population means? Explain your reasoning, giving your conclusion for Canada and the United States.

7.54. Table 7.30 compares two hospitals on the outcomes of patient admissions for severe pneumonia. Although patient status is an ordinal variable, two researchers who analyze the data treat it as an interval variable. The first researcher assigns the scores (0, 5, 10) to the three categories. The second researcher, believing that the middle category is much closer to the third category than to the first, uses the scores (0, 9, 10). Each researcher calculates the means for the two institutions and identifies the institution with the higher mean as the one having more

[19] J. G. Johnson et al., *Science*, vol. 295 (2002), pp. 2468–2471.

[20] E. Diener et al., *Journal of Personality & Social Psychology*, vol. 69 (1995), pp. 120–129.

[21] www.nifl.gov/nifl/facts/IALS.html.

success in treating its patients. Find the two means for the scoring system used by **(a)** the first researcher, **(b)** the second researcher. Interpret. (Since conclusions can depend on the scoring system, if you treat ordinal variables as quantitative, take care in selecting scores.)

TABLE 7.30

	Patient Status		
	Died in Hospital	Released after Lengthy Stay	Released after Brief Stay
Hospital A	1	29	0
Hospital B	8	8	14

7.55. From Example 6.4 (page 160), for the cognitive behavioral therapy group the sample mean change in weight of 3.0 pounds was significantly different from 0. However, Example 7.8 (page 206) showed it is not significantly different from the mean change for the control group, even though that group had a negative sample mean change. How do you explain this paradox? (*Hint*: From Sections 7.1 and 7.3, how does the *se* value for estimating a difference between two means compare to the *se* value for estimating a single mean?)

7.56. A survey by the Harris Poll of 2250 Americans in 2013 indicated that 42% believe in ghosts, 26% believe in witches, 29% believe in astrology, and 36% believe in creationism.

(a) Is it valid to compare the proportions using inferential methods for independent samples? Explain.

(b) Do you have enough information to compare them using inferential methods for dependent samples? Explain.

7.57. A pool of six candidates for three managerial positions includes three females and three males. Table 7.31 shows the results.

(a) Denote the three females by F_1, F_2, and F_3 and the 3 males by M_1, M_2, and M_3. Identify the 20 distinct samples of size three that can be chosen from these six individuals.

(b) Let $\hat{\pi}_1$ denote the sample proportion of males selected and $\hat{\pi}_2$ the sample proportion of females. For Table 7.28, $\hat{\pi}_1 - \hat{\pi}_2 = (2/3) - (1/3) = 1/3$. Of the 20 possible samples, show that 10 have $\hat{\pi}_1 - \hat{\pi}_2 \geq 1/3$. Thus, if the three managers were randomly selected, the probability would equal $10/20 = 0.50$ of obtaining $\hat{\pi}_1 - \hat{\pi}_2 \geq 1/3$. This is the

TABLE 7.31

	Chosen for Position	
Gender	Yes	No
Male	2	1
Female	1	2

reasoning that provides the one-sided *P*-value for Fisher's exact test.

(c) Find the *P*-value if all three selected are male. Interpret.

7.58. Describe a situation in which it would be more sensible to compare means using dependent samples than independent samples.

7.59. An AP story about a University of Chicago survey of 1600 people of ages 15 to 25 in several Midwest U.S. cities indicated that 58% of black youth, 45% of Hispanic youth, and 23% of white youth reported listening to rap music every day.

(a) True or false: If a 95% confidence interval comparing the population proportions for Hispanic and white youths was (0.18, 0.26), then we can infer that at least 18% but no more than 26% of the corresponding white population listens daily to rap music.

(b) The study reported that 66% of black females and 57% of black males agreed that rap music videos portray black women in bad and offensive ways. True or false: Because both these groups had the same race, inferential methods comparing them must assume dependent rather than independent samples.

7.60. True or false? If a 95% confidence interval for $(\mu_2 - \mu_1)$ contains only positive numbers, then we can conclude that both μ_1 and μ_2 are positive.

7.61. True or false? If you know the standard error of the sample mean for each of two independent samples, you can figure out the standard error of the difference between the sample means, even if you do not know the sample sizes.

In Exercises 7.62–7.64, select the correct response(s). More than one may be correct.

7.62. A 99% confidence interval for the difference $\pi_2 - \pi_1$ between the proportions of men and women in California who are alcoholics equals (0.02, 0.09).

(a) We are 99% confident that the proportion of alcoholics is between 0.02 and 0.09.

(b) We are 99% confident that the proportion of men in California who are alcoholics is between 0.02 and 0.09 larger than the proportion of women in California who are alcoholics.

(c) At this confidence level, there is insufficient evidence to infer that the population proportions are different.

(d) We are 99% confident that a minority of California residents are alcoholics.

(e) Since the confidence interval does not contain 0, it is impossible that $\pi_1 = \pi_2$.

7.63. To compare the population mean annual incomes for Hispanics (μ_1) and for whites (μ_2) having jobs in

construction, we construct a 95% confidence interval for $\mu_2 - \mu_1$.

(a) If the confidence interval is (3000, 6000), then at this confidence level we conclude that the population mean income is higher for whites than for Hispanics.

(b) If the confidence interval is $(-1000, 3000)$, then the corresponding $\alpha = 0.05$-level test of $H_0: \mu_1 = \mu_2$ against $H_a: \mu_1 \neq \mu_2$ rejects H_0.

(c) If the confidence interval is $(-1000, 3000)$, then it is plausible that $\mu_1 = \mu_2$.

(d) If the confidence interval is $(-1000, 3000)$, then we are 95% confident that the population mean annual income for whites is between $1000 less and $3000 more than the population mean annual income for Hispanics.

7.64. The Wilcoxon test differs from parametric procedures comparing means in the sense that

(a) It applies directly to ordinal as well as interval response variables.

(b) It is unnecessary to assume that the population distribution is normal.

(c) Random sampling is not assumed.

7.65.* A test consists of 100 true–false questions. Joe did not study, so on each question he randomly guesses the correct response.

(a) Find the probability that he scores at least 70, thus passing the exam. (*Hint*: Use either the binomial distribution or the sampling distribution for the proportion of correct responses.)

(b) Jane studied a little and has a 0.60 chance of a correct response for each question. Find the probability that her score is nonetheless lower than Joe's. (*Hint*: Use the sampling distribution of the difference of sample proportions.)

(c) How do the answers to (a) and (b) depend on the number of questions? Explain.

7.66.* Let y_{i1} denote the observation for subject i at time 1, y_{i2} the observation for subject i at time 2, and $y_i = y_{i2} - y_{i1}$.

(a) Letting $\bar{y}_1$, $\bar{y}_2$, and $\bar{y}_d$ denote the means of these observations, show that $\bar{y}_d = \bar{y}_2 - \bar{y}_1$.

(b) Is the median difference (i.e., the median of the y_i values) equal to the difference between the medians of the y_{i1} and y_{i2} values? Show this is true, or give a counterexample to show that it is false.

Chapter 8

ANALYZING ASSOCIATION BETWEEN CATEGORICAL VARIABLES

Recall that we say a sample shows ***association*** between two variables if certain values of one variable tend to go with certain values of the other. Many studies involve making an inference about the association between a response variable and an explanatory variable in a particular population of interest. In that population, an association exists if the probability distribution of the response variable changes in some way as the value of the explanatory variable changes.

In the previous chapter, we learned how to analyze whether population means or population proportions differ between two groups. These methods analyzed whether an association exists between a response variable and a binary explanatory variable that defines the two groups (e.g., sex). For a quantitative response variable, we compared means. For a categorical response variable, we compared proportions.

This chapter presents methods for detecting and describing associations between two categorical variables. The methods of this chapter help us answer a question such as "Is there an association between happiness and whether one is religious?" The methods of Chapter 7 for comparing two proportions are special cases of ones considered here in which both categorical variables have only two categories.

We first introduce terminology for categorical data analysis and define *statistical independence*, a type of lack of association in a population. We then present a significance test, called the *chi-squared test*, for determining whether two categorical variables are statistically independent or associated. We follow up that test by a *residual analysis* that describes the nature of that association. We then present ways of determining whether the association is strong enough to have practical importance, and we detail some specialized analyses for ordinal categorical variables.

8.1 Contingency Tables

Data for the analysis of categorical variables are displayed in ***contingency tables***. This type of table displays the number of subjects observed at all combinations of possible outcomes for the two variables.

Example 8.1

Gender Gap in Political Beliefs Does a "gender gap" exist in political beliefs? Do women and men tend to differ in their political thinking and voting behavior? To investigate this in the United States, we analyze Table 8.1, from the 2014 General Social Survey. The categorical variables are gender and political party identification (party ID, for short). Subjects indicated whether they identified more strongly with the Democratic or Republican party or as Independents. We regard party ID as the response variable and gender as the explanatory variable.

Table 8.1 contains responses for 2450 subjects, cross-classified by their gender and party ID. Table 8.1 is called a 2 × 3 (read "2-by-3") contingency table, meaning that it has two rows and three columns. The row totals and the column totals are

TABLE 8.1: Political Party Identification (ID) and Gender, for GSS Data in Data File `PartyID` at Text Website				
	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	495	590	272	1357
Males	330	498	265	1093
Total	825	1088	537	2450

called the **marginal distributions**. The sample marginal distribution for party ID, for instance, is the set of marginal frequencies (825, 1088, 537), showing that Independent was the most common response and Republican was the least common. ■

PERCENTAGE COMPARISONS: CONDITIONAL DISTRIBUTIONS

Constructing a contingency table from a data file is the first step in investigating an association between two categorical variables. As we explained on pages 64 and 199, when we distinguish between response and explanatory variables, it is natural to convert the cell frequencies to percentages for the response categories. In Table 8.1, to study how party ID differs for females and males, we find percentages within each row. For example, the percentage who identify themselves as Democrat is 36% for females (a proportion of $495/1357 = 0.36$) and 30% for males (330 out of 1093). Table 8.2 shows all the percentages. It seems that females are more likely than males to identify as Democrats.

| TABLE 8.2: Political Party Identification and Gender: Percentages Computed within Rows of Table 8.1					
	Party Identification				
Gender	Democrat	Independent	Republican	Total	n
Females	36%	43%	20%	100%	1357
Males	30%	46%	24%	100%	1093

The two sets of percentages for females and males are the sample **conditional distributions** on party ID. They describe the sample data distribution of party ID, *conditional* on gender. The females' conditional distribution on party ID is the set of percentages (36, 43, 20) for (Democrat, Independent, Republican). The percentages sum to 100 in each row, except for rounding. Figure 8.1 portrays graphically the two conditional distributions.

Another way to report percentages provides a single set for all cells in the table, using the total sample size as the base. To illustrate, in Table 8.1, of the 2450 subjects, 495 or 20% fall in the cell (Female, Democrat), 330 or 13% fall in the cell (Male, Democrat), and so forth. This percentage distribution is called the sample **joint distribution**. It is useful for comparing relative frequencies of occurrences for combinations of variable levels. When we distinguish between response and explanatory variables, however, conditional distributions are more informative than the joint distribution.

When you report a contingency table with conditional or joint distributions, include the total sample sizes on which the percentages or proportions are based. That
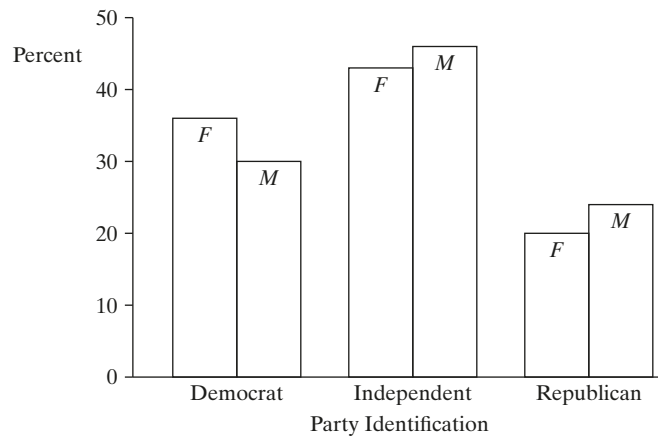
way, readers can determine standard errors to analyze the precision of sample proportion estimates.

INDEPENDENCE AND DEPENDENCE

Whether an association exists between party ID and gender is a matter of whether females and males differ in their conditional distributions on party ID. For the corresponding population, we answer the question "Is party ID associated with gender?" with reference to the concepts of statistical *independence* and *dependence*.

Statistical Independence and Statistical Dependence

> Two categorical variables are ***statistically independent*** if the population conditional distributions on one of them are identical at each category of the other. The variables are ***statistically dependent*** if the conditional distributions are not identical.

In other words, two variables are statistically independent if the percentage of the population in any particular category of one variable is the same for all categories of the other variable. In Table 8.2, the two conditional distributions are not identical. But that table describes a *sample*, and the definition of statistical independence refers to the *population*. If those observations were the entire population, then the variables would be statistically dependent.

For simplicity, we usually use the term *independent* rather than *statistically independent*. Table 8.3 is a contingency table showing independence. The table contains hypothetical population data for two variables—party ID and ethnic group. The percentage of Democrats is the same for each ethnic group, 35%. Similarly, the percentage of Independents and the percentage of Republicans are the same for each ethnic

TABLE 8.3: Population Cross-Classification Exhibiting Statistical Independence. The conditional distribution on party ID is the same in each row, (35%, 40%, 25%).

Ethnic Group	Party Identification			Total
	Democrat	Independent	Republican	
White	3500 (35%)	4000 (40%)	2500 (25%)	10,000 (100%)
Black	350 (35%)	400 (40%)	250 (25%)	1000 (100%)
Hispanic	875 (35%)	1000 (40%)	625 (25%)	2500 (100%)

group. The probability that a person has a particular party ID is the same for each ethnic group, and so party ID is independent of ethnic group.

Statistical independence is a symmetric property: If the conditional distributions within rows are identical, then so are the conditional distributions within columns. In Table 8.3, for example, you can check that the conditional distribution within each column equals (74%, 7%, 19%).

**Example
8.2** **Associations with Belief in Life after Death**   In recent General Social Surveys, the percentage of Americans who express a belief in life after death has been nearly 80%. This has been true both for females and for males and true for those who classify their race as black, white, or Hispanic. Thus, it appears that belief in life after death may be statistically independent of variables such as gender and race. On the other hand, whereas about 80% of Catholics and Protestants believe in an afterlife, only about 40% of Jews and 50% of those with no religion believe in an afterlife. We can't be sure, not having data for the entire population, but it seems that belief in life after death and religious affiliation are statistically dependent. ∎

8.2 Chi-Squared Test of Independence

The definition of statistical independence refers to the population. Two variables are independent if the *population* conditional distributions on the response variable are identical. Since Table 8.1 refers to a sample, it provides evidence but does not definitively answer whether party ID and gender are independent. Even if they are independent, we would not expect the *sample* conditional distributions to be identical. Because of sampling variability, we expect sample percentages to differ from the population percentages.

We next study whether it is plausible that party ID and gender are independent. If they are truly independent, could we expect sample differences such as Table 8.2 shows between females and males in their conditional distributions merely by sampling variation? Or, would differences of this size be unlikely? We address this with a significance test, by testing

H_0: The variables are statistically independent.
H_a: The variables are statistically dependent.

EXPECTED FREQUENCIES FOR INDEPENDENCE

The significance test compares the observed frequencies in the contingency table with values that satisfy the null hypothesis of independence. Table 8.4 shows the observed frequencies from Table 8.1, with the values (in parentheses) that satisfy H_0. These H_0 values have the same row and column marginal totals as the observed frequencies, but satisfy independence. They are called ***expected frequencies***.

TABLE 8.4: Political Party Identification by Gender, with Expected Frequencies in Parentheses

Gender	Democrat	Independent	Republican	Total
		Party Identification		
Female	495 (456.9)	590 (602.6)	272 (297.4)	1357
Male	330 (368.1)	498 (485.4)	265 (239.6)	1093
Total	825	1088	537	2450

Observed and Expected Frequencies

> Let f_o denote an ***observed*** frequency in a cell of the table. Let f_e denote an ***expected*** frequency. This is the count expected in a cell if the variables were independent. It equals the product of the row and column totals for that cell, divided by the total sample size.

For instance, for the cell in the upper left-hand corner (females who identify as Democrats), $f_o = 495$. Its expected frequency is $f_e = (1357)(825)/2450 = 456.9$, the product of the row total for Females and the column total for Democrats, divided by the overall sample size.

Let's see why this rule makes sense. In the entire sample, 825 out of 2450 people (33.7%) identify as Democrats. If the variables were independent, we would expect 33.7% of males and 33.7% of females to identify as Democrats. For instance, 33.7% of the 1357 females should be classified in the Democrat category. The expected frequency for the cell is then

$$f_e = \left(\frac{825}{2450} \right) 1357 = 0.337(1357) = 456.9.$$

THE PEARSON STATISTIC FOR TESTING INDEPENDENCE

The test statistic for H_0: independence summarizes how close the expected frequencies fall to the observed frequencies. Symbolized by X^2, it is

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e},$$

with summation over all cells in the contingency table. This is the oldest test statistic in common use today, having been introduced by the great British statistician Karl Pearson in 1900.

When H_0 is true, f_o and f_e tend to be close for each cell, and X^2 is relatively small. If H_0 is false, at least some f_o and f_e values tend not to be close, leading to large $(f_o - f_e)^2$ values and a large test statistic. The larger the X^2 value, the greater the evidence against H_0: independence.

Substituting the f_o and f_e values from Table 8.4, we find

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(495 - 456.9)^2}{456.9} + \frac{(590 - 602.6)^2}{602.6} + \frac{(272 - 297.4)^2}{297.4}$$

$$+ \frac{(330 - 368.1)^2}{368.1} + \frac{(498 - 485.4)^2}{485.4} + \frac{(265 - 239.6)^2}{239.6} = 12.57.$$

The calculation is messy, but it is simple to obtain X^2 using software. We next study how to interpret its magnitude.

THE CHI-SQUARED PROBABILITY DISTRIBUTION

As usual, for statistical inference we assume randomization, such as random sampling in a sample survey or a randomized experiment. The probability distribution of cell counts in conditional distributions or marginal distributions is then the ***multinomial distribution***, which generalizes the *binomial distribution* (Section 6.7) from two categories to several categories. The multinomial distribution for the cell counts implies a sampling distribution for the X^2 statistic. Under the presumption that

H_0: independence is true, we can use this sampling distribution to determine whether a particular sample value of X^2 is consistent with H_0 or would be unusually large.

For large sample sizes under randomization, the sampling distribution of X^2 is called the ***chi-squared probability distribution***. The symbol χ^2 for the chi-squared distribution is the Greek analog of the symbol X^2 for the test statistic. That statistic is often called the *Pearson chi-squared statistic*. Here are the main properties of the chi-squared distribution:

- It is concentrated on the positive part of the real line. The X^2 test statistic cannot be negative, since it sums squared differences divided by positive expected frequencies. The minimum possible value, $X^2 = 0$, would occur if $f_o = f_e$ in each cell.

- It is skewed to the right.

- The precise shape of the distribution depends on the ***degrees of freedom*** (*df*). The mean $\mu = df$ and the standard deviation $\sigma = \sqrt{2df}$. Thus, the distribution tends to shift to the right and become more spread out for larger *df* values. In addition, as *df* increases, the skew lessens and the chi-squared curve becomes more bell shaped. See Figure 8.2.

FIGURE 8.2: The Chi-Squared Distribution. The curve has larger mean and standard deviation as the degrees of freedom increase.
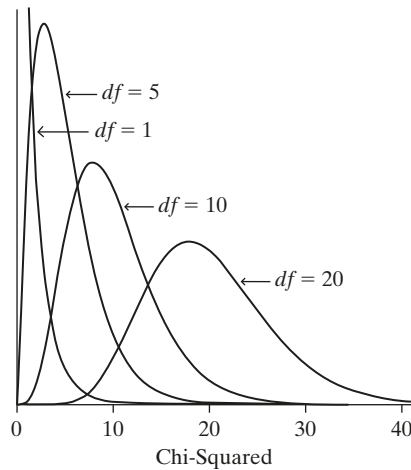
- For testing H_0: independence with a table having *r* rows and *c* columns,

$$df = (r - 1)(c - 1).$$

For a 2×3 table, $r = 2$, $c = 3$, and $df = (2 - 1)(3 - 1) = 1 \times 2 = 2$. Larger numbers of rows and columns produce larger *df* values. Since larger tables have more terms in the summation for the X^2 test statistic, the X^2 values also tend to be larger.

- The larger the X^2 value for a particular *df*, the stronger the evidence against H_0: independence. The *P*-value equals the right-tail probability above the observed X^2 value. It measures the probability, presuming H_0 is true, that X^2 is at least as large as the observed value.

Chi-squared right-tail probabilities are available with software or with Internet applets. Figure 8.3 shows an applet for finding the *P*-value when $X^2 = 12.57$ with

$df = 2$. With the software R, this *P*-value equals 1 minus the cumulative probability at 12.57 when $df = 2$:

```
> 1 - pchisq(12.57, 2)
[1] 0.001864057
```

FIGURE 8.3: The *P*-Value for the Chi-Squared Test of Independence Is the Right-Tail Probability, above the Observed Value of the Test Statistic. The *Chi-Squared Distribution* applet at `www.pearsonglobal editions.com/Agresti` can supply chi-squared tail probabilities.
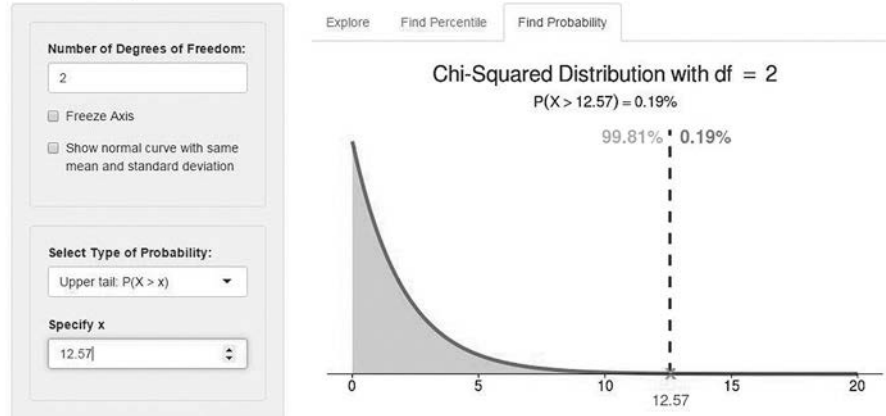
When $df = 1$ or 2, the chi-squared curve is so skewed that the mode is 0, as in Figure 8.3.

Example 8.3

Chi-Squared Statistic for Party ID and Gender To apply the chi-squared test to Table 8.1 on party ID and gender, we test the following:

H_0: Party ID and gender are statistically independent.
H_a: Party ID and gender are statistically dependent.

Using Stata with a data file of 2450 observations classified on categorical variables *partyid* and *gender*, we obtain the results shown in Table 8.5.

TABLE 8.5: Stata Software Output for Expected Frequencies and Pearson Chi-Squared Test of Independence for Data in Table 8.1 from `PartyID` Data File

```
. tab gender partyid, expected chi2

           |              partyid
    gender |        1         2         3 |     Total
         1 |      495       590       272 |     1,357
           |    456.9     602.6     297.4 |   1,357.0
         2 |      330       498       265 |     1,093
           |    368.1     485.4     239.6 |   1,093.0
-----------+--------------------------------+----------
     Total |      825     1,088       537 |     2,450
           |    825.0     1,088.0     537.0 |   2,450.0

        Pearson chi2(2) =   12.5693    Pr = 0.002
```

With some software, if we already have the cell counts we can enter them and obtain X^2 and the P-value, such as in Stata with the command

```
tabi 495 590 272 \ 330 498 265, chi2
```

or at some websites.[1]

The P-value of 0.002 provides strong evidence against H_0. It seems likely that party ID and gender are associated, in the population. If the variables were independent, it would be very unusual for a random sample to have this large a X^2 statistic. ∎

CHI-SQUARED AND DIFFERENCE OF PROPORTIONS FOR 2 × 2 TABLES

As Section 7.2 showed, we can use 2 × 2 contingency tables to compare two groups on a binary response variable. The outcomes could be, for example, (*yes, no*) on an opinion question. For convenience, we label the two possible outcomes for that binary variable by the generic labels *success* and *failure*.

Let π_1 represent the proportion of successes in population 1, and let π_2 represent the proportion of successes in population 2. Then $(1 - \pi_1)$ and $(1 - \pi_2)$ are the proportions of failures. Table 8.6 displays the notation. The rows are the groups to be compared and the columns are the response categories.

TABLE 8.6: 2 × 2 Table for Comparing Two Groups on a Binary Response Variable

Group	Population Proportion		Total
	Success	Failure	
1	π_1	$1 - \pi_1$	1.0
2	π_2	$1 - \pi_2$	1.0

If the response variable is statistically independent of the populations considered, then $\pi_1 = \pi_2$. The null hypothesis of independence corresponds to the *homogeneity* hypothesis, H_0: $\pi_1 = \pi_2$. In fact, the chi-squared test of independence is equivalent to a test for equality of two population proportions. Section 7.2 presented a z test statistic for this, based on dividing the difference of sample proportions by its standard error under H_0,

$$z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{se_0}.$$

The chi-squared statistic relates to this z statistic by $X^2 = z^2$.

The chi-squared statistic for 2 × 2 tables has $df = 1$. Its P-value from the chi-squared distribution is the same as the P-value for the two-sided test with the z test statistic. This is because of a direct connection between the standard normal distribution and the chi-squared distribution with $df = 1$: Squaring z-scores yields chi-squared scores with $df = 1$. The chi-squared right-tail probability is the same as the two-tail standard normal probability for z. For instance, $z = 1.96$ is the z-score with a two-tail probability of 0.05. The square of this, $(1.96)^2 = 3.84$, is the chi-squared score for $df = 1$ with a right-tail probability of 0.05.

---

[1] For example, the *Chi-Squared Test* applet at www.pearsonglobaleditions.com/Agresti.

Example 8.4

Women's and Men's Roles Table 8.7 summarizes responses from General Social Surveys in 1977 and in 2014 to the statement "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family." You can check that the sample proportions agreeing with the statement were $\hat{\pi}_1 = 0.658$ in 1977 and $\hat{\pi}_2 = 0.311$ in 2014, the se_0 for the test comparing them equals 0.0178, and the z test statistic for $H_0: \pi_1 = \pi_2$ is $z = (0.658 - 0.311)/0.0178 = 19.49$. You can also check that the chi-squared statistic for this table is $X^2 = 379.9$. This equals the square of the z test statistic. Both statistics show extremely strong evidence against the null hypothesis of equal population proportions. ■

TABLE 8.7: GSS Responses to the Statement "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family"			
Year	Agree	Disagree	Total
1977	989	514	1503
2014	515	1140	1655

CHI-SQUARED NEEDED FOR LARGER TABLES THAN 2 × 2

For a 2 × 2 table, why should we ever do a z test, if we can get the same result with chi-squared? An advantage of the z test is that it also applies with one-sided alternative hypotheses, such as $H_a: \pi_1 > \pi_2$. The direction of the effect is lost in squaring z and using X^2.

Why do we need the X^2 statistic? The reason is that a z statistic can only compare a single estimate to a single H_0 value. Examples are a z statistic for comparing a sample proportion to a H_0 proportion such as 0.5, or a difference of sample proportions to a H_0 value of 0 for $\pi_2 - \pi_1$. When a table is larger than 2 × 2 and thus $df > 1$, we need more than one difference parameter to describe the association. For instance, suppose Table 8.7 had three rows, for three years of data. Then H_0: independence corresponds to $\pi_1 = \pi_2 = \pi_3$, where π_i is the population proportion agreeing with the statement in year i. The comparison parameters are $(\pi_1 - \pi_2)$, $(\pi_1 - \pi_3)$, and $(\pi_2 - \pi_3)$. We could use a z statistic for each comparison, but not a single z statistic for the overall test of independence.

We can interpret the df value in a chi-squared test as the number of parameters needed to determine all the comparisons for describing the contingency table. For instance, for a 3 × 2 table for comparing three years on a binary opinion response, $df = 2$. This means we need to know only two parameters for making comparisons to figure out the third. For instance, if we know $(\pi_1 - \pi_2)$ and $(\pi_1 - \pi_3)$, then

$$(\pi_2 - \pi_3) = (\pi_1 - \pi_3) - (\pi_1 - \pi_2).$$

SAMPLE SIZE REQUIREMENTS FOR CHI-SQUARED TEST

The chi-squared test, like one- and two-sample z tests for proportions, is a large-sample test. The chi-squared distribution is the sampling distribution of the X^2 test statistic only when the sample size is relatively large. A guideline is that the expected frequency f_e should exceed 5 in each cell. Otherwise, the chi-squared distribution may poorly approximate the actual sampling distribution of the X^2 statistic.

For 2×2 contingency tables, a small-sample test of independence is ***Fisher's exact test*** (page 212). The test extends to tables of arbitrary size $r \times c$. The computations are complex and well beyond the scope of this text, but the test is available in statistical software. So, you don't need to use the chi-squared approximation when you are uncertain about its adequacy. For the GSS data on party ID and gender, Fisher's exact test also gives *P*-value = 0.002. The software R provides the results in Table 8.8 for the chi-squared test and for Fisher's exact test.

TABLE 8.8: R Software for Pearson Chi-Squared Test and Fisher's Exact Test with Table 8.1

```
> data <- matrix(c(495, 590, 272, 330, 498, 265), ncol=3, byrow=TRUE)
> chisq.test(data)

        Pearson's Chi-squared test
X-squared = 12.569, df = 2, p-value = 0.001865

> fisher.test(data)
  p-value = 0.001848
```

So, why do you often see chi-squared tests in research articles? Why not always use Fisher's exact test, since it uses an exact rather than an approximate sampling distribution and it has no sample size requirement? In practice, it's probably because methodologists have been using chi-squared tests for a long time, and only very recently has software been available for Fisher's exact test for $r \times c$ tables.

CHI-SQUARED TESTS AND TREATMENT OF CATEGORIES

In the chi-squared test, the value of X^2 does not depend on which is the response variable and which is the explanatory variable (if either). When a response variable is identified and the population conditional distributions are identical, they are said to be *homogeneous*. The chi-squared test of independence is then often referred to as a ***test of homogeneity***. For example, party ID is a response variable and gender is explanatory, so we can regard the chi-squared test applied to these data as a test of homogeneity of the conditional distributions of party ID.

Table 8.9 summarizes the five parts of the chi-squared test of independence. The test treats the classifications as nominal. That is, X^2 takes the same value if the rows or columns are reordered in any way. If either classification is ordinal or grouped interval, the chi-squared test does not use that information. In that case, even

---

TABLE 8.9: The Five Parts of the Chi-Squared Test of Independence

1. Assumptions: Two categorical variables, random sampling, $f_e \geq 5$ in all cells
   (Test treats variables as nominal scale)
2. Hypotheses: H_0: Statistical independence of variables
   H_a: Statistical dependence of variables
3. Test statistic: $X^2 = \sum \dfrac{(f_o - f_e)^2}{f_e}$, where $f_e = \dfrac{\text{(row total)(column total)}}{\text{total sample size}}$
4. *P*-value: P = right-tail probability above observed X^2 value,
   for chi-squared distribution with $df = (r-1)(c-1)$
5. Conclusion: Report *P*-value
   If decision needed, reject H_0 at α-level if $P \leq \alpha$

though the variables are categorical, it is usually better to apply stronger statistical methods designed for the higher level of measurement. Section 8.5 presents a test of independence for ordinal variables.

8.3 Residuals: Detecting the Pattern of Association

The chi-squared test of independence, like other significance tests, provides limited information. If the *P*-value is very small, strong evidence exists that the variables are associated. The chi-squared test tells us nothing, however, about the nature or strength of the association. The test does not indicate whether all cells deviate greatly from independence or perhaps only one or two of the cells do so.

RESIDUAL ANALYSIS

A cell-by-cell comparison of observed and expected frequencies reveals the nature of the evidence about the association. The difference $(f_o - f_e)$ between an observed and an expected cell frequency is called a ***residual***. A residual is positive when, as in the cell for female Democrats in Table 8.4, the observed frequency f_o exceeds the value f_e that independence predicts. The residual is negative when, as in the cell for male Democrats in Table 8.4, the observed frequency is smaller than independence predicts.

How do we know whether a residual is large enough to indicate a departure from independence that is unlikely to be due to mere chance? A standardized form of the residual that behaves like a z-score provides this information.

Standardized Residual

> The ***standardized residual*** for a cell is
>
> $$ z = \frac{f_o - f_e}{se} = \frac{f_o - f_e}{\sqrt{f_e(1 - \text{row proportion})(1 - \text{column proportion})}}. $$
>
> Here, *se* denotes the standard error of $f_o - f_e$, presuming H_0 is true. The standardized residual is the number of standard errors that $(f_o - f_e)$ falls from the value of 0 that we expect when H_0 is true.

The *se* uses the marginal proportions for the row and the column in which the cell falls. When H_0: independence is true, the standardized residuals have a large-sample standard normal distribution. They fluctuate around a mean of 0, with a standard deviation of about 1.

We use the standardized residuals in an informal manner to describe the pattern of the association among the cells. A large standardized residual provides evidence against independence in that cell. When H_0 is true, there is only about a 5% chance that any particular standardized residual exceeds 2 in absolute value. When we inspect many cells in a table, some standardized residuals could be large just by random variation. Values below -3 or above $+3$, however, are *very* convincing evidence of a true effect in that cell.

Example 8.5

Standardized Residuals for Gender and Party ID Table 8.10 displays the standardized residuals for testing independence between gender and party ID. For female Democrats, for instance, $f_o = 495$ and $f_e = 456.9$. The first row and first column

TABLE 8.10: Observed and Expected Frequencies, with Standardized Residuals in Parentheses, for Testing Independence between Party ID and Gender				
	Political Party Identification			
Gender	Democrat	Independent	Republican	Total
Female	495, 456.9 (3.3)	590, 602.6 (−1.0)	272, 297.4 (−2.5)	1357
Male	330, 368.1 (−3.3)	498, 485.4 (1.0)	265, 239.6 (2.5)	1093
Total	825	1088	537	2450

marginal proportions are $1357/2450 = 0.554$ and $825/2450 = 0.337$, respectively. Substituting into the formula, the standardized residual

$$z = \frac{f_o - f_e}{\sqrt{f_e(1 - \text{row prop.})(1 - \text{col. prop.})}} = \frac{495 - 456.9}{\sqrt{[456.9(1 - 0.554)(1 - 0.337)]}} = 3.3.$$

Since the standardized residual exceeds 3.0, we can conclude that females are Democrats more often than we would expect if the variables were truly independent.

The table also exhibits a large positive residual for male Republicans. More males identify as Republicans than the hypothesis of independence predicts. The table exhibits relatively large negative residuals for female Republicans and male Democrats. There were fewer female Republicans and male Democrats than we'd expect if party ID were independent of gender.

For each party ID, Table 8.10 contains only one nonredundant standardized residual. The one for females is the negative of the one for males. The observed counts and the expected frequencies have the same row and column totals. Thus, in a given column, if $f_o > f_e$ in one cell, the reverse must happen in the other cell. The differences $f_o - f_e$ have the same magnitude but a different sign in the two cells, implying the same pattern for their standardized residuals. ∎

Along with the X^2 statistic, statistical software[2] can provide standardized residuals. With R applied to a vector of the six cell counts in the form of a 2 × 3 table, we have

```
> data <- matrix(c(495, 590, 272, 330, 498, 265), ncol=3, byrow=TRUE)
> chisq.test(data)$stdres
          [,1]      [,2]       [,3]
[1,]  3.272365 -1.032199 -2.498557
[2,] -3.272365  1.032199  2.498557
```

STANDARDIZED RESIDUALS FOR 2 × 2 TABLES

Table 8.11 shows the standardized residuals for Table 8.7 on opinions about women's and men's roles. In this 2 × 2 table, *every* standardized residual equals either $+19.5$ or $−19.5$. The absolute value of the standardized residual is 19.5 in every cell.

For chi-squared tests with 2 × 2 tables, $df = 1$. This means that only one piece of information exists about whether an association exists. Once we find the standardized residual for one cell, other standardized residuals in the table have the same

---

[2] Also see the *Chi-Squared Test* applet at www.pearsonglobaleditions.com/Agresti.

TABLE 8.11: Standardized Residuals (in Parentheses) for Table 8.7			
Year	Agree	Disagree	Total
1977	989 (19.5)	514 (−19.5)	1503
2004	515 (−19.5)	1140 (19.5)	1655

absolute value. In fact, in 2×2 tables, each standardized residual equals the z test statistic (or its negative) for comparing the two proportions. The square of each standardized residual equals the X^2 test statistic.

8.4 Measuring Association in Contingency Tables

The main questions normally addressed in analyzing a contingency table are

- *Is there an association?* The chi-squared test of independence addresses this question. The smaller the P-value, the stronger the evidence of association.

- *How do the data differ from what independence predicts?* The standardized residuals highlight the cells that are more likely or less likely than expected under independence.

- *How strong is the association?* To summarize this, we use a statistic such as a difference of proportions, forming a confidence interval to estimate the strength of association in the population.

Analyzing the *strength* of the association reveals whether the association is important or whether it is perhaps statistically significant but practically insignificant. On page 65, we introduced the *correlation* for describing strength of association for quantitative variables. This section presents two ***measures of association*** for contingency tables.

STRONG VERSUS WEAK ASSOCIATION IN A CONTINGENCY TABLE

Let's first consider what is meant by *strong* versus *weak* association. Table 8.12 shows two hypothetical contingency tables relating religion (*fundamentalist*, *nonfundamentalist*) to opinion about legalized marriage for same-sex couples (*favor*, *oppose*). Case A, which exhibits statistical independence, represents the weakest possible association. Both religious categories have 60% in favor and 40% opposed on opinion. Opinion is not associated with race. By contrast, case B exhibits the strongest

TABLE 8.12: Cross-Classification of Opinion about Same-Sex Legalized Marriage by Religion, Showing (A) No Association and (B) Maximum Association						
	A:	Opinion		B:	Opinion	
Religion		Favor	Oppose	Total	Favor	Oppose	Total
Nonfundamentalist		360	240	600	600	0	600
Fundamentalist		240	160	400	0	400	400
Total		600	400	1000	600	400	1000

possible association. All nonfundamentalists favor, whereas all fundamentalists oppose. In this table, opinion is completely dependent on religion. For these subjects, if we know whether they are fundamentalist or not, we know their opinion.

A measure of association describes how similar a table is to the tables representing the strongest and weakest associations. It takes a range of values from one extreme to another as data range from the weakest to strongest association. It is useful for comparing associations, to determine which is stronger.

DIFFERENCE OF PROPORTIONS: MEASURING ASSOCIATION IN 2 × 2 TABLES

As discussed on page 234, many 2×2 tables compare two groups on a binary variable. In such cases, an easily interpretable measure of association is the difference between the proportions for a given response category. For example, we could measure the difference between the proportions of religious nonfundamentalists and fundamentalists who favor allowing same-sex marriage. For Table 8.12(A), this difference is

$$\frac{360}{600} - \frac{240}{400} = 0.60 - 0.60 = 0.0.$$

The population difference of proportions is 0 whenever the conditional distributions are identical, that is, when the variables are independent. The difference is 1 or -1 for the strongest possible association. For Table 8.12(B), for instance, the difference is

$$\frac{600}{600} - \frac{0}{400} = 1.0,$$

the maximum possible absolute value for the difference.

This measure falls between -1 and $+1$. In practice, we don't expect data to take these extreme values, but *the stronger the association, the larger the absolute value of the difference of proportions*. The following contingency tables illustrate the increase in this measure as the degree of association increases:

Cell Counts:	25 25 / 25 25	30 20 / 20 30	35 15 / 15 35	40 10 / 10 40	45 5 / 5 45	50 0 / 0 50
Difference of Proportions:	0	0.2	0.4	0.6	0.8	1.0

For the second table, for instance, the proportion falling in the first column is $30/(30 + 20) = 0.60$ in row 1 and $20/(20 + 30) = 0.40$ in row 2, for a difference of $0.60 - 0.40 = 0.20$.

CHI-SQUARED DOES NOT MEASURE STRENGTH OF ASSOCIATION

A large value for X^2 in the chi-squared test of independence suggests that the variables are associated. It does *not* imply, however, that the variables have a strong association. This statistic summarizes how close the observed frequencies are to the frequencies expected if the variables were independent. It merely indicates, however, how much evidence there is that the variables are dependent, not how strong that dependence is. For a given association, larger X^2 values occur for larger sample sizes. As with any significance test, large test statistic values can occur with weak effects, if the sample size is large.

For example, consider the hypothetical cases in Table 8.13 for race and an opinion response. The association in each table is very weak—the conditional distribution for whites on opinion (49% favor, 51% oppose) is nearly identical to the conditional distribution for blacks (51% favor, 49% oppose). All three tables show exactly the same degree of association, with the difference between the proportions of blacks and whites in the "yes" category being $0.51 - 0.49 = 0.02$ in each table.

TABLE 8.13: Cross-Classifications of Opinion by Race, Showing Weak but Identical Associations								
	A			**B**			**C**		
Race	Yes	No	Total	Yes	No	Total	Yes	No	Total
White	49	51	100	98	102	200	4,900	5,100	10,000
Black	51	49	100	102	98	200	5,100	4,900	10,000
	100	100	200	200	200	400	10,000	10,000	20,000
	$X^2 = 0.08$			$X^2 = 0.16$			$X^2 = 8.0$		
	P-value $= 0.78$			P-value $= 0.69$			P-value $= 0.005$		

For the sample of size 200 in case A, $X^2 = 0.08$, which has a P-value $= 0.78$. For the sample of size 400 in case B, $X^2 = 0.16$, for which $P = 0.69$. So, when the cell counts double, X^2 doubles. Similarly, for the sample size of 20,000 (100 times as large as $n = 200$) in case C, $X^2 = 8.0$ (100 times as large as $X^2 = 0.08$) and $P = 0.005$.

In summary, for a fixed percentage assignment to the cells of a contingency table, X^2 is directly proportional to the sample size—larger values occur with larger sample sizes. Like other test statistics, the larger the X^2 statistic, the smaller the P-value and the stronger the evidence against the null hypothesis. However, a small P-value can result from a weak association when the sample size is large, as case C shows.

THE ODDS RATIO*

The difference of proportions is easily interpretable. Several other measures are also available in statistical software. This subsection presents the most important one for categorical data analysis, the *odds ratio*.

For a binary response variable, recall that we use *success* to denote the outcome of interest and *failure* the other outcome. The **odds** of success are defined to be

$$\text{Odds} = \frac{\text{Probability of success}}{\text{Probability of failure}}.$$

If $P(\text{success}) = 0.75$, then $P(\text{failure}) = 1 - 0.75 = 0.25$, and the odds of success $= 0.75/0.25 = 3.0$. If $P(\text{success}) = 0.50$, then odds $= 0.50/0.50 = 1.0$. If $P(\text{success}) = 0.25$, then odds $= 0.25/0.75 = 1/3$. The odds are nonnegative, with value greater than 1.0 when a success is more likely than a failure. When odds $= 3.0$, a success is three times as likely as a failure; we expect about three successes for every failure. When odds $= 1/3$, a failure is three times as likely as a success; we expect about one success for every three failures.

The probability of an outcome relates to the odds of the outcome by

$$\text{Probability} = \frac{\text{Odds}}{\text{Odds} + 1}.$$

For instance, when odds $= 3$, probability $= 3/(3 + 1) = 0.75$.

The ratio of odds from the two rows of a 2×2 table is called the **odds ratio**. For instance, if the odds $= 4.5$ in row 1 and the odds $= 3.0$ in row 2, then the odds

ratio $= 4.5/3.0 = 1.5$. The odds of success in row 1 then equal 1.5 times the odds of success in row 2. We denote the population value of an odds ratio by the Greek letter θ (theta).

Example
8.6

Race of Murder Victims and Offenders For murders in the United States in 2013 having a single victim and single offender, Table 8.14 cross-classifies the race of the victim by the race of the offender for cases in which they were both known. We treat race of victim as the response variable. For white offenders, the proportion of victims who were white is $2509/2918 = 0.860$ and the proportion who were black is $409/2918 = 0.140$. The odds of a white victim equaled $0.860/0.140 = 6.13$. This equals $(2509/2918)/(409/2918) = 2509/409$. So, we can calculate the odds by the ratio of the counts in the two cells in row 1, without converting them to proportions.

TABLE 8.14: Cross-Classification of Race of Victim and Race of Offender

	Race of Victim		
Race of Offender	White	Black	Total
White	2509	409	2918
Black	189	2245	2434

Source: www.fbi.gov.

The value 6.13 means that for white offenders, there were 6.13 white victims for every black victim. For black offenders, the odds of a white victim were $189/2245 = 0.0842$. This means there were 0.0842 white victims for every black victim. Equivalently, since $2245/189 = 1/0.0842 = 11.88$, black offenders had 11.88 black victims for every white victim.

For these data, the odds ratio is

$$\theta = \frac{\text{Odds for white offenders}}{\text{Odds for black offenders}} = \frac{6.13}{0.0842} = 72.9.$$

The odds of a white victim for white offenders were about 73 times the odds of a white victim for black offenders. ∎

In summary,

Odds and Odds Ratio

The estimated ***odds*** for a binary response equal the number of successes divided by the number of failures.

The ***odds ratio*** is a measure of association for 2×2 contingency tables that equals the odds in row 1 divided by the odds in row 2.

PROPERTIES OF THE ODDS RATIO*

In Table 8.14, suppose we treat race of offender, rather than race of victim, as the response variable. When victims were white, the odds the race of the offender was white equaled $2509/189 = 13.28$. When victims were black, the odds the race of the offender was white equaled $409/2245 = 0.182$. The odds ratio is $13.28/0.182 = 72.9$. For each choice of the response variable, the odds ratio is 72.9. In fact,

- The odds ratio takes the same value regardless of the choice of response variable.

Since the odds ratio treats the variables symmetrically, the odds ratio is a natural measure when there is no obvious distinction between the variables, such as when both are response variables.

- The odds ratio θ equals the ratio of the products of cell counts from diagonally opposite cells.

For Table 8.13, for instance,

$$\theta = \frac{2509 \times 2245}{409 \times 189} = 72.9.$$

Because of this property, the odds ratio is also called the ***cross-product ratio***.

- The odds ratio can equal any nonnegative number.
- When the success probabilities are identical in the two rows of a 2×2 table (i.e., $\pi_1 = \pi_2$), then $\theta = 1$.

When $\pi_1 = \pi_2$, the odds are also equal. The odds of success do not depend on the row level of the table, and the variables are then independent, with $\theta = 1$. The value $\theta = 1$ for independence serves as a baseline for comparison. Odds ratios on each side of 1 reflect certain types of associations.

- When $\theta > 1$, the odds of success are *higher* in row 1 than in row 2.

For instance, when $\theta = 4$, the odds of success in row 1 are four times the odds of success in row 2.

- When $\theta < 1$, the odds of success are *lower* in row 1 than in row 2.
- Values of θ farther from 1.0 in a given direction represent stronger associations.

An odds ratio of 4 is farther from independence than an odds ratio of 2, and an odds ratio of 0.25 is farther from independence than an odds ratio of 0.50.

- Two values for θ represent the same strength of association, but in opposite directions, when one value is the reciprocal of the other.

For instance, $\theta = 4.0$ and $\theta = 1/4.0 = 0.25$ represent the same strength of association. When $\theta = 0.25$, the odds of success in row 1 are 0.25 times the odds of success in row 2. Equivalently, the odds of success in row 2 are $1/0.25 = 4.0$ times the odds of success in row 1. When the order of the rows is reversed or the order of the columns is reversed, the new value of θ is the reciprocal of the original value. This ordering of rows or columns is usually arbitrary, so whether we get 4.0 or 0.25 for the odds ratio is merely a matter of how we label the rows and columns.

In interpreting the odds ratio, be careful not to misinterpret it as a ratio of probabilities. An odds ratio of 72.9 does *not* mean that π_1 is 72.9 times π_2. Instead, $\theta = 72.9$ means that the *odds* in row 1 equal 72.9 times the odds in row 2. The odds ratio is a ratio of two odds, not a ratio of two probabilities. That is,

$$\theta = \frac{\text{Odds in row 1}}{\text{Odds in row 2}} = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}, \quad \text{not} \quad \frac{\pi_1}{\pi_2}.$$

As explained on page 194, the ratio of probabilities π_1/π_2 is itself a useful measure for comparing two groups.

The sampling distribution of the sample odds ratio $\hat{\theta}$ is highly skewed unless the sample size is extremely large, in which case the distribution is approximately normal. Exercise 8.45 shows a method of constructing confidence intervals for odds ratios.

ODDS RATIOS FOR $r \times c$ TABLES*

For contingency tables with more than two rows or more than two columns, the odds ratio describes patterns in any 2×2 subtable. We illustrate using GSS data on party ID and race, shown in Table 8.15.

TABLE 8.15: GSS Data from 2014 on Political Party Identification and Race			
	Political Party Identification		
Gender	Democrat	Independent	Republican
Black	249	108	17
White	496	828	498

Consider first the 2×2 subtable formed from the first two columns. The sample odds ratio is $(249 \times 828)/(108 \times 496) = 3.85$. The odds that a black's response was Democrat rather than Independent equal 3.85 times the odds for whites. Of those subjects who responded Democrat or Independent, blacks were more likely than whites to respond Democrat.

The sample odds ratio for the last two columns of this table is $(108 \times 498)/(17 \times 828) = 3.82$. The odds that a black's response was Independent rather than Republican equal 3.8 times the odds for whites. Of those subjects who responded Independent or Republican, blacks were much more likely than whites to respond Independent.

Finally, for the 2×2 subtable formed from the first and last columns, the sample odds ratio is $(249 \times 498)/(17 \times 496) = 14.71$. The odds that a black's response was Democrat rather than Republican equal 14.7 times the odds for whites. Of those subjects who responded Democrat or Republican, blacks were much more likely than whites to respond Democrat. This is a very strong effect, far from the independence odds ratio value of 1.0.

The odds ratio value of 14.7 for the first and last columns equals $(3.85)(3.82)$, the product of the other two odds ratios. For 2×3 tables, $df = 2$, meaning that only two independent bits of information exist about the association. Two of the odds ratios determine the third.

SUMMARY MEASURES OF ASSOCIATION FOR $r \times c$ TABLES*

Instead of studying association in 2×2 subtables, it's possible to summarize association in the entire table by a single number. One way to do this summarizes how well we can predict the value of one variable based on knowing the value of the other variable. For example, party ID and race are highly associated if race is a good predictor of party ID, that is, if knowing their race, we can make much better predictions about people's party ID than if we did not know it.

For quantitative variables, the *correlation* is such a summary measure. We'll study a similar summary measure of this type for ordinal variables (called *gamma*) in the next section. These measures describe an overall trend in the data. For nominal variables, when r or c exceeds 2 it is usually an oversimplification to describe the table with a single measure of association. In that case, too many possible patterns of association exist to describe an $r \times c$ table well by a single number. We believe you get a better feel for the association by making percentage comparisons of conditional distributions, by viewing the pattern of standardized residuals in the cells of the table, by constructing odds ratios in 2×2 subtables, and by building models such as those

presented in Chapter 15. These methods become even more highly preferred to summary measures of association when the analysis is multivariate rather than bivariate.

8.5 Association Between Ordinal Variables*

We now turn our attention to analyses of contingency tables that cross-classify ordinal variables. We introduce a popular ordinal measure of association and present related methods of inference.

Example 8.7

How Strongly Associated Are Income and Happiness? Table 8.16 cross-classifies two ordinal variables. The data, from the 2014 General Social Survey, show the relation between family income and happiness. This table has results for black Americans, and Exercise 8.13 analyzes data for white Americans.

TABLE 8.16: Family Income and Happiness for a GSS Sample, with Conditional Distributions on Happiness in Parentheses

	Happiness			
Family Income	Not Too Happy	Pretty Happy	Very Happy	Total
Below average	37 (22%)	90 (52%)	45 (26%)	172 (100.0%)
Average	25 (15%)	93 (53%)	56 (32%)	174 (100.0%)
Above average	6 (16%)	18 (49%)	13 (35%)	37 (100.0%)
Total	68	201	114	383

Let's first get a feel for the data by studying the conditional distributions on happiness. Table 8.16 shows them. For subjects with below-average family income, only 26% are very happy, whereas 35% of those at the above-average income level are very happy. Conversely, a lower percentage (16%) of the high-income group are not too happy compared to those in the low-income group (22%). The odds ratio for the four corner cells is $(37 \times 13)/(45 \times 6) = 1.78$. There seems to be a slight tendency for subjects with higher incomes to have greater happiness. ∎

Ordinal data exhibit two primary types of association between variables x and y—*positive* and *negative*. Positive association results when subjects at the high end of the scale on x tend also to be high on y, and those who are low on x tend to be low on y. For example, a positive association exists between income and happiness if those with low incomes tend to have lower happiness, and those with high incomes tend to have greater happiness. Negative association occurs when subjects classified high on x tend to be classified low on y, and those classified low on x tend to be high on y. For example, a negative association might exist between religious fundamentalism and tolerance toward homosexuality—the more fundamentalist in religious beliefs, the less tolerance toward homosexuality.

CONCORDANCE AND DISCORDANCE

Many ordinal measures of association are based on the information about the association provided by all the *pairs* of observations.

Concordant Pair, Discordant Pair	A pair of observations is ***concordant*** if the subject who is *higher* on one variable also is *higher* on the other variable. A pair of observations is ***discordant*** if the subject who is *higher* on one variable is *lower* on the other.

In Table 8.16, we regard *Not too happy* as the low end and *Very happy* as the high end of the scale on $y =$ happiness, and *Below average* as low and *Above average* as high on $x =$ family income. By convention, we construct contingency tables for ordinal variables so that the low end of the row variable is the first row and the low end of the column variable is the first column.[3]

Consider a pair of subjects, one of whom is classified (below average, not too happy), and the other of whom is classified (average, pretty happy). The first subject is one of the 37 classified in the upper left-hand cell of Table 8.16, and the second subject is one of the 92 classified in the middle cell. This pair of subjects is concordant, since the second subject is higher than the first subject both in happiness and in income. The subject who is higher on one variable is also higher on the other. Now, each of the 37 subjects classified (below average, not too happy) can pair with each of the 92 subjects classified (average, pretty happy). So, there are $37 \times 92 = 3404$ concordant pairs of subjects from these two cells.

By contrast, each of the 90 subjects in the cell (below average, pretty happy) forms a discordant pair when matched with each of the 25 subjects in the cell (average, not too happy). The 90 subjects have lower income than the other 25 subjects, yet they have greater happiness. All $90 \times 25 = 2250$ of these pairs of subjects are discordant.

Concordant pairs of observations provide evidence of positive association since, for such a pair, the subject who is higher on one variable also is higher on the other. On the other hand, the more prevalent the discordant pairs, the more evidence there is of a negative association. We let C denote the total number of concordant pairs of observations and D denote the total number of discordant pairs of observations. We leave to Exercise 8.46 a general rule for calculating C and D. Software can easily do this for us. Overall, Table 8.16 has $C = 14,804$ and $D = 11,031$. More pairs show evidence of a positive association (i.e., concordant pairs) than show evidence of a negative association (discordant pairs).

GAMMA

A positive difference for $C - D$ occurs when $C > D$. This indicates a positive association. A negative difference for $C - D$ reflects a negative association.

Larger sample sizes have larger numbers of pairs with, typically, larger absolute differences in $C - D$. Therefore, we standardize this difference to make it easier to interpret. To do this, we divide $C - D$ by the total number of pairs that are either concordant or discordant, $C + D$. This gives the measure of association called ***gamma***. Its sample formula is

$$\hat{\gamma} = \frac{C - D}{C + D}.$$

Here are some properties of gamma:

- The value of gamma falls between -1 and $+1$.
- The sign of gamma indicates whether the association is positive or negative.
- The larger the absolute value of gamma, the stronger the association.

---

[3] There is no standard, however, and some books use a different convention.

A table for which gamma equals 0.60 or −0.60 exhibits a stronger association than one for which gamma equals 0.30 or −0.30, for example. The value +1 represents the strongest positive association. This occurs when there are no discordant pairs ($D = 0$), so all the pairs reveal a positive association. Gamma equals −1 when $C = 0$, so all pairs reveal a negative association. Gamma equals 0 when $C = D$.

For Table 8.16, $C = 14{,}804$ and $D = 11{,}031$, so

$$\hat{\gamma} = \frac{14{,}804 - 11{,}031}{14{,}804 + 11{,}031} = 0.146.$$

This sample exhibits a positive association. The higher the family income, the greater the happiness tends to be. However, the sample value is closer to 0 than to 1, so the association is relatively weak.

GAMMA IS A DIFFERENCE BETWEEN TWO ORDINAL PROPORTIONS

Another interpretation for the magnitude of gamma follows from the expression

$$\hat{\gamma} = \frac{C - D}{C + D} = \frac{C}{C + D} - \frac{D}{C + D}.$$

Now, $(C + D)$ is the total number of pairs that are concordant or discordant. The ratio $C/(C + D)$ is the proportion of those pairs that are concordant, $D/(C + D)$ is the proportion of the pairs that are discordant, and $\hat{\gamma}$ is the difference between the two proportions.

For example, suppose $\hat{\gamma} = 0.60$. Then, since 0.80 and 0.20 are the two proportions that sum to 1 and have a difference of $0.80 - 0.20 = 0.60$, 80% of the pairs are concordant and 20% are discordant. Similarly, $\hat{\gamma} = -0.333$ indicates that 1/3 of the pairs are concordant and 2/3 of the pairs are discordant, since $1/3 + 2/3 = 1$ and $1/3 - 2/3 = -0.333$.

For Table 8.16, out of the $14{,}804 + 11{,}031 = 25{,}835$ pairs that are concordant or discordant, the proportion $14{,}804/25{,}835 = 0.573$ are concordant and the proportion $11{,}031/25{,}835 = 0.427$ are discordant; $\hat{\gamma} = 0.146$ is the difference between these proportions.

COMMON PROPERTIES OF ORDINAL MEASURES

Gamma is one of several ordinal measures of association. Others are **Kendall's tau-b**, **Spearman's rho-b**, and **Somers' d**. These measures are all similar in their basic purposes and characteristics. For lack of space, we do not define them, but we will list some common properties. These properties also hold for the *correlation* for quantitative variables, which we introduced in Section 3.5 and use extensively in the next chapter.

- Ordinal measures of association take values between −1 and +1. The sign tells us whether the association is positive or negative.

- If the variables are statistically independent, then the population values of ordinal measures of association equal 0.

- The stronger the association, the larger the absolute value of the measure. Values of 1.0 and −1.0 represent the strongest associations.

- With the exception of Somers' d, the ordinal measures of association named above do not distinguish between response and explanatory variables. They take the same value when variable y is the response variable as when it is the explanatory variable.

CONFIDENCE INTERVALS AND TESTS FOR ORDINAL ASSOCIATION

Confidence intervals help us gauge the strength of the association in the population. Let γ denote the population value of gamma. For sample gamma, $\hat{\gamma}$, its sampling distribution is approximately normal about γ. Its standard error se describes the variation in $\hat{\gamma}$ values around γ among samples of the given size. The formula for se is complicated, but it is reported by most software. Assuming random sampling, a confidence interval for γ has the form

$$\hat{\gamma} \pm z(se).$$

The chi-squared test of whether two categorical variables are independent treats the variables as nominal. Other tests are usually more powerful when the variables are ordinal, and here we present one using gamma. As in the chi-squared test, the null hypothesis is that the variables are statistically independent. The alternative hypothesis can take the two-sided form $H_a: \gamma \neq 0$ or a one-sided form $H_a: \gamma > 0$ or $H_a: \gamma < 0$, when we predict the direction of the association.

The test statistic has the z statistic form. It takes the difference between $\hat{\gamma}$ and the value of 0 that gamma takes when H_0: independence is true and divides by the standard error,

$$z = \frac{\hat{\gamma} - 0}{se}.$$

Under random sampling, this test statistic has approximately the standard normal distribution, when H_0 is true. Some software also reports a se and/or related P-value that holds only under H_0. The normal approximation holds better with larger n, and is adequate when each of C and D exceeds about 50.

Example 8.8	**Inference about Association between Income and Happiness** For Table 8.16 on family income and happiness, Table 8.17 shows software output. The $\hat{\gamma} = 0.145$ value has $se = 0.079$. (It is labeled as ASE, where the A stands for "asymptotic," meaning it is an approximate large-sample standard error.) With some software, if we already have the cell counts, we can enter them and find gamma and its standard error. Stata uses the command

```
tabi 37 90 45 \ 25 93 56 \ 6 18 13, gamma
```

TABLE 8.17: Part of Software Output for Analyzing Table 8.16

	Value	DF	P-Value
Pearson Chi-Square	4.1266	4	0.389
Statistic Value	ASE		P-Value
Gamma 0.1454	0.0789		0.064

A 95% confidence interval for γ in the population is

$$\hat{\gamma} \pm 1.96(se), \quad \text{or } 0.145 \pm 1.96(0.079), \quad \text{or } 0.145 \pm 0.155,$$

which equals $(-0.01, 0.30)$. We can be 95% confident that γ is no less than -0.01 and no greater than 0.30. It is plausible that essentially no association exists between

income and happiness, but it is also plausible that a moderate positive association exists. We need a larger sample size to estimate this more precisely.

For testing independence between family income and happiness, the chi-squared test of independence has $X^2 = 4.13$ with $df = 4$, for which the P-value equals 0.39. This test does not show any evidence of an association. The chi-squared test treats the variables as nominal, however, and ordinal-level methods are more powerful if there is a positive or negative trend. The ordinal test statistic using gamma is

$$z = \frac{\hat{\gamma} - 0}{se} = \frac{0.145 - 0}{0.079} = 1.84.$$

The P-value for H_a: $\gamma \neq 0$ equals 0.064. This test shows some evidence of an association. A priori, we might have predicted a positive association between family income and happiness. The test for H_a: $\gamma > 0$ has $P = 0.033$. There is relatively strong evidence of a positive association in the population. ∎

ORDINAL TESTS VERSUS PEARSON CHI-SQUARED TEST

The z test result for these data providing evidence of an association may seem surprising. The chi-squared statistic of $X^2 = 4.09$ with $df = 4$ provided no evidence ($P = 0.39$).

A test of independence based on an ordinal measure is usually preferred to the chi-squared test when both variables are ordinal. The X^2 statistic ignores the ordering of the categories, taking the same value no matter how the levels are ordered. If a positive or negative trend exists, ordinal measures are usually more powerful for detecting it. Unfortunately, the situation is not clear cut. It is possible for the chi-squared test to be more powerful even if the data are ordinal.

To explain this, we first note that the null hypothesis of independence is not equivalent to a value of 0 for population gamma. Although independence implies $\gamma = 0$, the converse is not true. Namely, γ may equal 0 even though the variables are not statistically independent. For example, Table 8.18 shows a relationship between two variables that does not have a single trend. Over the first two columns there is a positive relationship, since y increases when x increases. Over the last two columns there is a negative relationship, as y decreases when x increases. For the entire table, $C = 25(25 + 25) = 1250 = D$, so $\gamma = 0$. The proportion of concordant pairs equals the proportion of discordant pairs. However, there is not independence, because the conditional distribution on y for the low level of x is completely different from the conditional distribution on y for the high level of x.

TABLE 8.18: A Relationship for Which Ordinal Measures of Association Equal 0. The variables are dependent even though gamma equals 0.

		Level of y			
		Very Low	Low	High	Very High
Level of x	Low	25	0	0	25
	High	0	25	25	0

Thus, an ordinal measure of association may equal 0 when the variables are statistically dependent but the dependence does not have an overall positive or overall negative trend. The chi-squared test can perform better than the ordinal test when the relationship does not have a single trend. In practice, most relationships with

ordinal variables have primarily one trend, if any. So, the ordinal test is usually more powerful than the chi-squared test.

SIMILAR INFERENCE METHODS FOR OTHER ORDINAL MEASURES

The inference methods for gamma apply also to other ordinal measures of association. An alternative approach to detect trends assigns scores to the categories for each variable and uses the correlation and a z test based on it. (Section 9.5 presents a closely related test.) Some software reports this as a test of *linear-by-linear association*.

Whenever possible, it is better to choose the categories for ordinal variables finely rather than crudely. For instance, it is better to use four or five categories than only two categories. Standard errors of measures tend to be smaller with more categories, for a given sample size. Thus, the finer the categorizations, the shorter the confidence interval for a population measure of association tends to be. In addition, finer measurement makes it more valid to treat the data as quantitative and use the more powerful methods presented in the following chapter for quantitative variables.

MIXED ORDINAL–NOMINAL CONTINGENCY TABLES

For a cross-classification of an ordinal variable with a nominal variable that has only two categories, ordinal measures of association are still valid. In that case, the sign of the measure indicates which level of the nominal variable is associated with higher responses on the ordinal variable. For instance, suppose gamma $= -0.12$ for the association in a 2×3 table relating gender (female, male) to happiness (not too happy, pretty happy, very happy). Since the sign is negative, the "higher" level of gender (i.e., male) tends to occur with lower happiness. The association is weak, however.

When the nominal variable has more than two categories, it is inappropriate to use an ordinal measure such as gamma. There are specialized methods for mixed nominal–ordinal tables, but it is usually simplest to treat the ordinal variable as quantitative by assigning scores to its levels. The methods of Chapter 12, which generalize comparisons of two means to several groups, are then appropriate. Section 15.4 presents a modeling approach that does not require assigning scores to ordinal response variables. For further details about statistical methods for ordinal variables, see Agresti (2010).

8.6 Chapter Summary

This chapter introduced analyses of association for categorical variables:

- By *describing the counts* in **contingency tables** using percentage distributions, called **conditional distributions**, across the categories of the response variable. If the population conditional distributions are identical, the two variables are **statistically independent**—the probability of any particular response is the same for each level of the explanatory variable.

- By using **chi-squared** to **test H_0: independence** between the variables. The Pearson chi-squared test statistic compares each observed frequency f_o to the expected frequency f_e satisfying H_0, using

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}.$$

The test statistic has, under H_0, a large-sample chi-squared distribution. The ***degrees of freedom*** depend on the number of rows r and the number of columns c, through $df = (r - 1)(c - 1)$. The P-value is the right-tail probability above the observed value of X^2. ***Fisher's exact test*** is also applicable with small samples.

- By *describing the pattern of association* using **standardized residuals** for the cells in the table. A standardized residual reports the number of standard errors that $(f_o - f_e)$ falls from 0. A large absolute value indicates that that cell provides evidence of association in a particular direction.

- By *describing the strength of association*. For 2×2 tables the ***difference of proportions*** is useful, as is the ***odds ratio***, the ratio of odds from the two rows. Each odds measures the proportion of successes divided by the proportion of failures. When there is independence, the difference of proportions equals 0 and the odds ratio equals 1. The stronger the association, the farther the measures fall from these baseline values.

This chapter also presented methods for analyzing association between two *ordinal* categorical variables.

- Many ***ordinal measures of association*** use the numbers of **concordant pairs** (the subject who is higher on x also is higher on y) and **discordant pairs** (the subject who is higher on x is lower on y). Of the pairs that are concordant or discordant, ***gamma*** equals the difference between their proportions and falls between -1 and $+1$, with larger absolute values indicating stronger association. When the variables are independent, gamma equals 0.

The chi-squared test treats the data as nominal. When the variables are ordinal, methods that use the ordinality (such as a z test based on sample gamma) are more powerful for detecting a positive or negative association trend.

The next chapter introduces analyses of association for quantitative variables.

Exercises

Practicing the Basics

8.1. GSS surveys routinely show that in the United States, about 40% of males and 40% of females believe that a woman should be able to get an abortion if she wants it for any reason.

(a) Construct a contingency table showing the conditional distribution on whether unrestricted abortion should be legal (yes, no) by gender.

(b) Based on these results, does statistical independence seem plausible between gender and opinion about unrestricted abortion? Explain.

8.2. Whether a woman becomes pregnant in the next year is a categorical variable with categories (yes, no), and whether she and her partner use contraceptives is another categorical variable with categories (yes, no). Would you expect these variables to be statistically independent, or associated? Explain.

8.3. A March 2015 survey by the Pew Research Center (www.pewresearch.org) compared various groups of Americans in terms of their support for legalizing marijuana. In considering age groups, it estimated that legalization was supported by 68% of those of age between 18 and 34, by 51% of those of age between 35 and 69, and by 29% of those of age 70 and higher.

(a) If results for the population of adult Americans were similar to these, would age and opinion about legalizing marijuana be independent, or dependent?

(b) Display hypothetical population percentages in a contingency table for which these variables would be independent.

8.4. The World Values Survey[4] asked, "How often do you pray?" The response *never* was given by 16.5% of the 2232 respondents in the United States, by 48.4% of the 1189 respondents in Spain, and by 56.8% of the 1206 respondents

---

[4] See www.worldvaluessurvey.org/WVSOnline.jsp.

in Sweden. Show how to construct a contingency table relating the outcome on the praying question to the nation where it was asked. For this table, identify the response variable, the explanatory variable, and the conditional distributions.

8.5. Based on current estimates of how well mammograms detect breast cancer, Table 8.19 shows what to expect for 100,000 adult women over the age of 40 in terms of whether a woman has breast cancer and whether a mammogram gives a positive result (i.e., indicates that the woman has breast cancer).

(a) Construct the conditional distributions for the mammogram test result, given the true disease status. Does the mammogram appear to be a good diagnostic tool?

(b) Construct the conditional distribution of disease status, for those who have a positive test result. Use this to explain why even a good diagnostic test can have a high false positive rate when a disease is not common.

TABLE 8.19

| | | Diagnostic Test | |
		Positive	Negative
Breast	Yes	860	140
Cancer	No	11800	87120

8.6. Data posted at the FBI website (`www.fbi.gov`) indicated that of all blacks slain in 2013, 92% were slain by blacks, and of all whites slain in 2013, 93% were slain by whites. Let y denote race of victim and x denote race of murderer.

(a) Which conditional distributions do these statistics refer to, those of y at given levels of x, or those of x at given levels of y? Set up a contingency table showing these distributions.

(b) Are x and y independent or dependent? Explain.

8.7. How large a X^2 value provides a P-value of 0.05 for testing independence for the following table dimensions?

(a) 2×2 **(b)** 3×3 **(c)** 2×5 **(d)** 5×5
(e) 3×9

8.8. A sociologist uses a 2×4 contingency table to compare four groups on a binary response variable. For group i, let π_i denote the population proportion of response in the first outcome category, $i = 1, 2, 3, 4$. Explain what $df = 3$ means in the context of comparing these four population proportions.

8.9. In 2010, the GSS asked about willingness to accept cuts in the standard of living to help the environment, with categories (very willing, fairly willing, neither willing nor unwilling, not very willing, not at all willing). When this was cross tabulated with sex, $X^2 = 3.3$.

(a) What are the hypotheses for the test to which X^2 refers?

(b) Report the df value on which X^2 is based.

(c) What conclusion would you make, using a significance level of 0.05? State your conclusion in the context of this study.

8.10. Table 8.20 is based on results described in a study[5] that examined the effects of alcohol consumption and drug use on sexual behavior, for undergraduate students at a university in the southeastern United States. In this table, the columns refer to whether the student engaged in unprotected sex in the past three months, and the rows refer to whether the student reported drinking alcohol mixed with energy drinks (AmED).

(a) Construct conditional distributions that treat unprotected sex as the response variable. Interpret.

(b) Test whether the variables are statistically independent. Report the P-value, and interpret.

TABLE 8.20

| | | Unprotected Sex | |
		Yes	No
AmED	Yes	77	44
Consumption	No	194	309

8.11. Are people happier who believe in life after death? Go to the GSS website `sda.berkeley.edu/GSS` and download the contingency table for the 2014 survey relating happiness and whether you believe in life after death (variables HAPPY and POSTLIFE, with YEAR(2014) in the *selection filter*).

(a) State a research question that could be addressed with the output.

(b) Report the conditional distributions, using happiness as the response variable, and interpret.

(c) Report the X^2 value and its P-value. (You can get this by checking *Statistics*.) Interpret.

(d) Interpret the standardized residuals. (You can get them by checking *z statistic*.) Summarize what you learned from your analyses.

8.12. In the GSS, subjects who were married were asked the happiness of their marriage, the variable coded as HAPMAR.

(a) Go to `sda.berkeley.edu/GSS/` and construct a contingency table for 2014 relating HAPMAR to family income measured as (above average, average, below average), by entering FINRELA(r: 1-2; 3; 4-5) as the row variable and YEAR(2014) in the selection filter. Use a table or graph with conditional distributions to describe the association.

---

[5] By D. Snipes and E. Benotsch, *Addictive Behaviors*, vol. 38 (2013), pp. 1418–1423.

(c) By checking "Statistics," you request the chi-squared statistic. Report it and its *df* and *P*-value, and interpret.

8.13. The sample in Table 8.16 is 382 black Americans. Table 8.21 shows cell counts and standardized residuals for income and happiness for white subjects in the 2014 General Social Survey.

(a) Interpret the Pearson chi-squared statistic and its *P*-value.

(b) Interpret the standardized residuals in the four corner cells.

TABLE 8.21

```
Rows: income    Columns: happiness
            not    pretty    very
below       128      324     107
            9.14     0.95   -7.47
average      66      479     295
           -5.27     0.57    3.12
above        35      247     184
           -3.62    -1.66    4.33
Cell Contents:        Count
                      Standardized residual

Pearson Chi-Square = 114.7,
DF = 4, P-Value = 0.000
```

8.14. Table 8.22 shows SPSS analyses with the 2014 General Social Survey, for variables party ID and race.

(a) Report the expected frequency for the first cell, and show how SPSS obtained it.

(b) Test the hypothesis of independence. Report the test statistic and *P*-value, and interpret.

(c) Use the standardized residuals (labeled "Adjusted Residual" here) to describe the pattern of association.

TABLE 8.22

			party_ID			
			democr	indep	repub	Total
race	black	Count	249	108	17	374
		Expected Count	126.9	159.4	87.7	
		Adjusted Residual	14.6	-5.9	-9.5	
	white	Count	496	828	498	1822
		Expected Count	618.1	776.6	427.3	
		Adjusted Residual	-14.6	5.9	9.5	
Total		Count	745	936	515	2196

	Value	df	Asymptotic Significance
Pearson Chi-Square	230.35	2	.000

8.15. For a 2×4 cross-classification of gender and religiosity (very, moderately, slightly, not at all) for recent GSS data, the standardized residual was 3.2 for females who are very religious, −3.2 for males who are very religious,

−3.5 for females who are not at all religious, and 3.5 for males who are not at all religious. All other standardized residuals fell between −1.1 and 1.1. Interpret.

8.16. Table 8.23, from the 2014 General Social Survey, cross-classifies happiness and marital status.

(a) Software reports that $X^2 = 135.3$. Interpret.

(b) The table also shows, in parentheses, the standardized residuals. Summarize the association by indicating which marital statuses have strong evidence of (i) more, (ii) fewer people in the population in the *very happy* category than if the variables were independent.

(c) Compare the married and divorced groups by the difference in proportions in the *very happy* category.

TABLE 8.23

Marital Status	Very Happy	Pretty Happy	Not Too Happy
Married	472 (9.8)	592 (−3.9)	90 (−7.6)
Widowed	49 (−2.4)	120 (0.7)	38 (2.2)
Divorced	94 (−3.9)	233 (0.5)	84 (4.6)
Separated	12 (−3.2)	47 (0.5)	22 (3.7)
Never married	158 (−5.0)	410 (3.3)	105 (1.9)

8.17. A report by the Gallup organization in February 2015 estimated that 9% of Republicans approved of President Barack Obama's performance, whereas 79% of Democrats approved. Would you characterize the association between political party affiliation and opinion about Obama's performance as weak, or strong? Explain why.

8.18. In a recent GSS, the death penalty for subjects convicted of murder was favored by 74% of whites and 43% of blacks. It was favored by 75% of males and 63% of females. In this sample, which variable was more strongly associated with death penalty opinion—race, or gender? Explain why.

8.19. In a survey of senior high school students in Dayton, Ohio, 1449 students had used both alcohol and cigarettes, 281 had used neither, 500 had used alcohol but not cigarettes, and 46 had used cigarettes but not alcohol.

(a) Construct the 2×2 table relating alcohol use to cigarette use. Describe the strength of association using the difference between (i) users and nonusers of alcohol in the proportions who have used cigarettes, (ii) users and nonusers of cigarettes in the proportions who have used alcohol. Interpret.

(b) Describe the strength of association using the odds ratio. Interpret. Does the odds ratio value depend on your choice of response variable?

8.20. Table 8.24 cross-classifies 68,694 passengers in autos and light trucks involved in accidents in the state of Maine by whether they were wearing a seat belt and by whether they were injured or killed. Describe the association using

(a) The difference between two proportions, treating whether injured or killed as the response variable.

(b) The odds ratio.

TABLE 8.24

		Injury	
		Yes	No
Seat	Yes	2409	35,383
Belt	No	3865	27,037

Source: Thanks to Dr. Cristanna Cook, Medical Care Development, Augusta, Maine, for supplying these data.

8.21. The 2012 National Survey on Drug Use and Health (NSDUH) estimated that 23% of Americans aged 12 or over reported binge drinking in the past month, and 7% had used marijuana in the past month.

(a) Find the odds of (i) binge drinking, (ii) marijuana use. Interpret.

(b) Find the odds ratio comparing binge drinking to marijuana use. Interpret.

8.22. According to the U.S. Bureau of Justice Statistics, in 2014 the incarceration rate in the nation's prisons was 904 per 100,000 male residents, 65 per 100,000 female residents, 2805 per 100,000 black residents, and 466 per 100,000 white residents. (*Source*: www.bjs.gov.)

(a) Find the odds ratio between whether incarcerated and (i) gender, (ii) race. Interpret.

(b) According to the odds ratio, which has the stronger association with whether incarcerated, gender or race? Explain.

8.23. Refer to Table 8.1 (page 228) on political party ID and gender. Find and interpret the odds ratio for each 2×2 subtable. Explain why this analysis suggests that the last two columns show a very weak association.

8.24. According to a 2015 study by the Pew Research Center (www.people-press.org), the percentage of Americans who favor allowing gays and lesbians to marry legally was 81% for Democrats who identified themselves as liberal and 22% for Republicans who identified themselves as conservative.

(a) The odds ratio is 15.1. Explain what is wrong with the interpretation "The probability that liberal Democrats favor legalized gay marriage is 15.1 times the probability that conservative Republicans favor legalized gay marriage." Give the correct interpretation.

(b) For those born after 1980, the odds of favoring legalization equaled 2.7. Estimate the probability they favored legalization.

8.25. Table 8.25 cross-classifies happiness with family income for the subsample of the 2014 GSS that identified themselves as Jewish.

(a) Give an example of a (i) concordant pair, (ii) discordant pair.

(b) The table has 204 concordant pairs and 55 discordant pairs. Find gamma, and interpret.

(c) Show how to express gamma as a difference between two proportions.

TABLE 8.25

		HAPPY		
		Not_too	Pretty	Very
INCOME	Below	0	4	1
	Average	4	11	1
	Above	0	12	8

8.26. For the 2014 GSS, $\hat{\gamma} = 0.19$ for the relationship between job satisfaction (categories very dissatisfied, little dissatisfied, moderately satisfied, very satisfied) and family income (below average, average, above average).

(a) Would you consider this a very strong or relatively weak association? Explain.

(b) Is this a stronger or a weaker association than the one between job satisfaction and happiness, which has $\hat{\gamma} = 0.41$? Explain.

8.27. A GSS cross-classified income in thousands of dollars ($<5, 5–15, 15–25, >25$) by job satisfaction (very dissatisfied, a little satisfied, moderately satisfied, very satisfied) for black Americans. Software provides the results shown in Table 8.26.

TABLE 8.26

```
            |              jobsat
income |      1        2        3        4
       -----------------------------------
     1 |      2        4       13        3
     2 |      2        6       22        4
     3 |      0        1       15        8
     4 |      0        3       13        8
       -----------------------------------
Pearson chi2(9) =   11.5243    Pr = 0.241
         gamma =    0.3551    ASE = 0.122
```

(a) Interpret the *P*-value reported for the chi-squared test.

(b) Conduct an alternative test of independence that takes category ordering into account. Why are results so different from the chi-squared test?

(c) Construct the 95% confidence interval for gamma. Interpret.

8.28. Refer to Exercise 8.13 on happiness and income. The analysis there does not take into account the ordinality of the variables. Using software, summarize the strength of association by finding and interpreting gamma. Construct inference using it, and interpret.

Concepts and Applications

8.29. Refer to the `Students` data file (Exercise 1.11 on page 21). Using software, create and analyze descriptively and inferentially the contingency table relating opinion about abortion and **(a)** political affiliation, **(b)** religiosity.

8.30. Refer to the data file you created in Exercise 1.12. For variables chosen by your instructor, pose a research question and conduct descriptive and inferential statistical analyses. Interpret and summarize your findings in a short report.

8.31. One year the GSS asked how housework was shared between the respondent and his or her spouse. Possible responses were 1 = I do much more than my fair share, 2 = I do a bit more than my fair share, 3 = I do roughly my fair share, 4 = I do a bit less than my fair share, 5 = I do much less than my fair share. Table 8.27 shows results according to the respondent's sex. State a research question that could be addressed with this output, and prepare a 100–200-word summary of what you learn. (The *Adjusted Residual* is the standardized residual.)

8.32. Pose a research question about attitude regarding homosexual relations and political ideology. Using the most recent GSS data on variables HOMOSEX and POLVIEWS, conduct a descriptive and inferential analysis to address this question. Prepare a one-page report summarizing your analysis.

8.33. Does belief in evolution vary according to religious beliefs? Examine this using Table 8.28, for respondents to the 2014 General Social Survey. The variables are Fundamentalism/Liberalism of Respondent's Religion and response to whether human beings developed from earlier species of animals. Analyze these data. Prepare a 200–300-word report describing your analyses and providing interpretations.

8.34. For 2014 GSS data, of those identifying as Democrats, 496 classified themselves as liberal and 171 as conservative. Of those identifying as Republicans, 56 called themselves liberal and 499 conservative. Using methods of this chapter, describe the strength of association, and interpret.

TABLE 8.27

	hhwkfair				
female count	121	108	135	19	6
% within sex	31.1%	27.8%	34.7%	4.9%	1.5%
Adjusted Residual	8.0	5.9	-4.2	-7.1	-4.9
male count	18	28	148	68	29
% within sex	6.2%	9.6%	50.9%	23.4%	10.0%
Adjusted Residual	-8.0	-5.9	4.2	7.1	4.9

	Value	df	Asymp. Sig.,
Pearson Chi-Square	155.8	4	.000

	Value	Asymp. Std. Error.
Gamma	.690	.038

TABLE 8.28

Religious Preference	Evolution		Total
	True	False	
Fundamentalist	70	196	266
Moderate	239	210	449
Liberal	248	96	344
Total	557	502	1059

8.35. An abstract of an article[6] dealing with alcohol use and sexual assault among college women stated, "This study prospectively examined the relation between alcohol use and sexual assault in a sample ($n = 319$) of first-year college women. ... Over the course of their freshman year, 19.3% reported experiencing at least one sexual assault. Frequent binge drinking and frequent drinking predicted a subsequent sexual assault. Frequent binge drinking demonstrated a stronger association with sexual assault than did frequent drinking." Explain the statistical methods that were probably used in order to make such conclusions.

8.36. Shortly before a gubernatorial election, a poll asks a random sample of 50 potential voters the following questions:

Party: Do you consider yourself to be a Democrat (D), a Republican (R), or an Independent (I)?

Vote: If you were to vote today, would you vote for the Democratic candidate (D) or the Republican (R) candidate, or would you be undecided (U) about how to vote?

Plan: Do you plan on voting in the election? Yes (Y) or no (N)?

For each person interviewed, the answers to the three questions are entered in a data file, which is the `Voting`

---

[6] By E. Mouilso and S. Fischer, *Violence and Victims*, vol. 27 (2012), pp. 78–94.

data file at the text website. Using software, create a data file and conduct the following analyses:

(a) Construct the 3×3 contingency table relating party affiliation to intended vote. Report the conditional distributions on intended vote for each of the three party affiliations. Are they very different?

(b) Report the result of the test of the hypothesis that intended vote is independent of party affiliation. Provide the test statistic and the P-value, and interpret the result.

(c) Supplement the analyses in (a) and (b) to investigate the association more fully. Interpret.

8.37. (a) When the sample size is very large, we have not necessarily established an important result when we show a statistically significant association. Explain.

(b) The remarks in Sections 8.3 and 8.4 about small P-values not necessarily referring to an important effect apply for any significance test. Explain why, discussing the effect of n on standard errors and the sizes of test statistics.

8.38. Answer true or false for the following. Explain your answer.

(a) Even when the sample conditional distributions in a contingency table are only slightly different, when the sample size is very large it is possible to have a large X^2 test statistic and a very small P-value for testing H_0: independence.

(b) If the odds ratio $= 2.0$ between gender (female, male) and opinion on some issue (favor, oppose), then the odds ratio $= -2.0$ if we measure gender as (male, female).

(c) Interchanging two rows in a contingency table has no effect on the X^2 statistic.

(d) Interchanging two rows in a contingency table has no effect on gamma.

(e) If $\gamma = 0$ for two variables, then the variables are statistically independent.

8.39. The correct answer in Exercise 8.38(c) implies that if the chi-squared statistic is used for a contingency table having ordered categories in both directions, then (select the correct response(s))

(a) The statistic actually treats the variables as nominal.

(b) Information about the ordering is ignored.

(c) The test is usually not as powerful for detecting association as a test statistic based on numbers of concordant and discordant pairs.

(d) The statistic cannot differentiate between positive and negative associations.

8.40. Each subject in a sample of 100 men and 100 women is asked to indicate which of the following factors (one or more) are responsible for increases in crime committed by teenagers: A—the increasing gap in income between the rich and poor, B—the increase in the percentage of

single-parent families, C—insufficient time that parents spend with their children, D—criminal penalties given by courts are too lenient, E—increasing problems with drugs in society, F—increasing levels of violence shown on TV. To analyze whether responses differ by gender of respondent, we cross-classify the responses by gender, as Table 8.29 shows.

(a) Is it valid to apply the chi-squared test of independence to these data? Explain.

(b) Explain how this table actually provides information needed to cross-classify gender with each of six variables. Construct the contingency table relating gender to opinion about whether the increasing gap in income is responsible for increases in teenage crime.

TABLE 8.29

Gender	A	B	C	D	E	F
Men	60	81	75	63	86	62
Women	75	87	86	46	82	83

8.41. * Table 8.30 exhibits the maximum possible association between two binary variables for a sample of size n.

(a) Show that $X^2 = n$ for this table and, hence, that the maximum value of X^2 for 2×2 tables is n.

(b) The ***phi-squared*** measure of association for 2×2 contingency tables has sample value

$$\hat{\phi}^2 = \frac{X^2}{n}.$$

Explain why this measure falls between 0 and 1, with a population value of 0 corresponding to independence. (It is a special case, for 2×2 tables, of the r^2 measure introduced in the next chapter.)

TABLE 8.30

$n/2$	0
0	$n/2$

8.42. * For 2×2 tables, gamma simplifies to a measure first proposed about 1900 by the British statistician G. Udny Yule, who also introduced the odds ratio. In that special case, gamma is called ***Yule's Q***.

(a) Show that for a generic table with counts (a, b) in row 1 and (c, d) in row 2, the number of concordant pairs equals ad, the number of discordant pairs equals bc, and $Q = (ad - bc)/(ad + bc)$.

(b) Show that the absolute value of gamma equals 1 for any 2×2 table in which one of the cell frequencies is 0.

8.43. * Construct a 3×3 table for which gamma equals **(a)** 1, **(b)** -1, **(c)** 0.

8.44.* A chi-squared variable with degrees of freedom equal to df has representation $z_1^2 + \cdots + z_{df}^2$, where $z_1, \ldots, z_{df}$ are independent standard normal variates.

(a) If z is a test statistic that has a standard normal distribution, what distribution does z^2 have?

(b) Explain how to get the chi-squared values for $df = 1$ from z-scores in the standard normal table (Table A). Illustrate for the chi-squared value of 6.63 having P-value 0.01.

(c) The chi-squared statistic for testing H_0: independence between belief in an afterlife (yes, no) and happiness (not too happy, pretty happy, very happy) is X_m^2 in a 2×3 table for men and X_w^2 in a 2×3 table for women. If H_0 is true for each gender, then what is the probability distribution of $X_m^2 + X_w^2$?

8.45.* For a 2×2 table with cell counts a, b, c, d, the sample log odds ratio $\log \hat{\theta}$ has approximately a normal sampling distribution with estimated standard error

$$se = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}.$$

The antilogs of the endpoints of the confidence interval for $\log(\theta)$ are endpoints of the confidence interval for θ. For Table 8.14 on page 242, show[7] that the 95% confidence interval for the odds ratio is (60.8, 87.4). Interpret.

8.46.* In an ordinal table, to calculate the number of concordant pairs C, start at the corner of the table for the low level for each variable. Multiply that cell count by the count in every cell that is higher on both variables. Similarly, for every other cell, multiply the cell count by the counts in cells that are higher on both variables. Then C is the sum of these products. Explain how to calculate the total number of discordant pairs D.

---

[7] Stata provides this with the command `csi 2509 409 189 2245, or woolf`.

This page intentionally left blank

Chapter 9

Linear Regression and Correlation

Chapter 8 presented methods for analyzing association between categorical response and explanatory variables. This chapter presents methods for analyzing association between quantitative response and explanatory variables. The analyses are collectively called a ***regression analysis***.

Example 9.1 Regression Analysis for Crime Indices   Table 9.1 shows data from *Statistical Abstract of the United States* for the 50 states and the District of Columbia (D.C.) on

- Murder rate: The number of murders per 100,000 people in the population.

- Violent crime rate: The number of murders, forcible rapes, robberies, and aggravated assaults per 100,000 people in the population.

- Percentage of the population with income below the poverty level.

- Percentage of families headed by a single parent.

For these quantitative variables, violent crime rate and murder rate are natural response variables. We'll treat the poverty rate and percentage of single-parent families as explanatory variables for these responses as we present regression methods in this chapter. The text website contains this data file, called `Crime2`, as well as a data file `Crime` that contains 2013 violent crime and murder rates already analyzed in Chapter 3. ∎

We present three different, but related, aspects of regression analysis:

1. We investigate *whether an association exists* between the variables by testing the hypothesis of statistical independence.

2. We study the *strength of their association* using the *correlation* measure of association.

3. We estimate a *regression equation* that predicts the value of the response variable from the value of the explanatory variable.

9.1 Linear Relationships

We let *y* denote the *response* variable and let *x* denote the *explanatory* variable. We analyze how values of *y* tend to change from one subset of the population to another, as defined by values of *x*. For categorical variables, we did this by comparing the conditional distributions of *y* at the various categories of *x*, in a contingency table. For quantitative variables, a mathematical formula describes how the conditional distribution of *y* (such as *y* = crime rate) varies according to the value of *x* (such as *x* = percentage below the poverty level). Does the crime rate tend to be higher for states that have higher poverty rates?

TABLE 9.1: Statewide Data (from `Crime2` Data File at the Text Website) Used to Illustrate Regression Analyses

State	Violent Crime	Murder Rate	Poverty Rate	Single Parent	State	Violent Crime	Murder Rate	Poverty Rate	Single Parent
AK	761	9.0	9.1	14.3	MT	178	3.0	14.9	10.8
AL	780	11.6	17.4	11.5	NC	679	11.3	14.4	11.1
AR	593	10.2	20.0	10.7	ND	82	1.7	11.2	8.4
AZ	715	8.6	15.4	12.1	NE	339	3.9	10.3	9.4
CA	1078	13.1	18.2	12.5	NH	138	2.0	9.9	9.2
CO	567	5.8	9.9	12.1	NJ	627	5.3	10.9	9.6
CT	456	6.3	8.5	10.1	NM	930	8.0	17.4	13.8
DE	686	5.0	10.2	11.4	NV	875	10.4	9.8	12.4
FL	1206	8.9	17.8	10.6	NY	1074	13.38	16.4	12.7
GA	723	11.4	13.5	13.0	OH	504	6.0	13.0	11.4
HI	261	3.8	8.0	9.1	OK	635	8.4	19.9	11.1
IA	326	2.3	10.3	9.0	OR	503	4.6	11.8	11.3
ID	282	2.9	13.1	9.5	PA	418	6.8	13.2	9.6
IL	960	11.42	13.6	11.5	RI	402	3.9	11.2	10.8
IN	489	7.5	12.2	10.8	SC	1023	10.3	18.7	12.3
KS	496	6.4	13.1	9.9	SD	208	3.4	14.2	9.4
KY	463	6.6	20.4	10.6	TN	766	10.2	19.6	11.2
LA	1062	20.3	26.4	14.9	TX	762	11.9	17.4	11.8
MA	805	3.9	10.7	10.9	UT	301	3.1	10.7	10.0
MD	998	12.7	9.7	12.0	VA	372	8.3	9.7	10.3
ME	126	1.6	10.7	10.6	VT	114	3.6	10.0	11.0
MI	792	9.8	15.4	13.0	WA	515	5.2	12.1	11.7
MN	327	3.4	11.6	9.9	WI	264	4.4	12.6	10.4
MO	744	11.3	16.1	10.9	WV	208	6.9	22.2	9.4
MS	434	13.5	24.7	14.7	WY	286	3.4	13.3	10.8
					DC	2922	78.5	26.4	22.1

LINEAR FUNCTIONS: INTERPRETING THE y-INTERCEPT AND SLOPE

Any particular formula might provide a good description or a poor one of how y relates to x. This chapter introduces the simplest type of formula—a *straight line*. For it, y is said to be a ***linear function*** of x.

Linear Function	The formula $y = \alpha + \beta x$ expresses observations on y as a ***linear function*** of observations on x. The formula has a straight-line graph with ***slope*** β (beta) and ***y-intercept*** α (alpha).

Each real number x, when substituted into the formula $y = \alpha + \beta x$, yields a distinct value for y. In a graph, the horizontal axis, the ***x-axis***, lists the possible values of x. The vertical axis, the ***y-axis***, lists the possible values of y. The axes intersect at the point where $x = 0$ and $y = 0$, called the *origin*.

At $x = 0$, the equation $y = \alpha + \beta x$ simplifies to $y = \alpha + \beta x = \alpha + \beta(0) = \alpha$. Thus, the constant α in this equation is the value of y when $x = 0$. Now, points on the y-axis have $x = 0$, so the line has height α at the point of its intersection with the y-axis. Because of this, α is called the ***y-intercept***.

The ***slope*** β equals the change in y for a one-unit increase in x. That is, for two x-values that differ by 1.0 (such as $x = 0$ and $x = 1$), the y-values differ by β. Two x-values that are 10 units apart differ by 10β in their y-values. Figure 9.1 portrays the

interpretation of the *y*-intercept and slope. In the context of a regression analysis, α and β are called ***regression coefficients***.

FIGURE 9.1: Graph of the Straight Line $y = \alpha + \beta x$. The *y*-intercept is α and the slope is β.
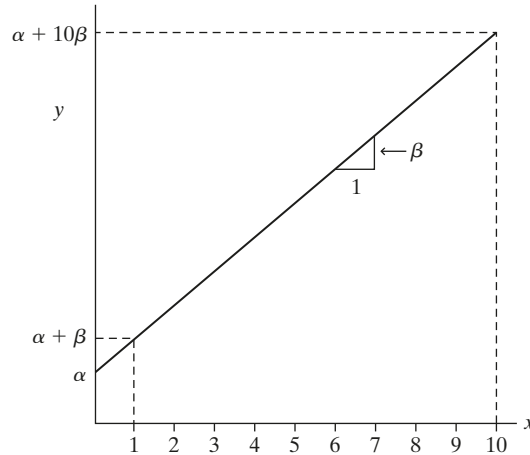
<table>
<tr><td rowspan="2">Example
9.2</td><td>**Straight Lines for Predicting Violent Crime Rate**   For the 50 states, consider $y =$ violent crime rate and $x =$ poverty rate. We'll see that a straight line cannot perfectly represent the relation between them, but the line $y = 210 + 25x$ provides a type of approximation. The *y*-intercept of 210 represents the violent crime rate at poverty rate $x = 0$ (unfortunately, there are no such states). The slope equals 25. When the percentage with income below the poverty level increases by 1, the violent crime rate increases by about 25 crimes a year per 100,000 population.</td></tr>
</table>

By contrast, if instead $x =$ percentage of the population living in urban areas, the straight line approximating the relationship is $y = 26 + 8x$. The slope of 8 is smaller than the slope of 25 when poverty rate is the explanatory variable. An increase of 1 in the percentage below the poverty level corresponds to a greater change in the violent crime rate than an increase of 1 in the percentage urban.

Figure 9.2 shows the lines relating the violent crime rate to poverty rate and to urban residence. Generally, the larger the absolute value of β, the steeper the line. When β is positive, *y increases* as *x increases*—the straight line goes upward, like these two lines. Then, large values of *y* occur with large values of *x*, and small values

FIGURE 9.2: Graphs of Lines Showing Positive Relationships $(\beta > 0)$, a Negative Relationship $(\beta < 0)$, and Independence $(\beta = 0)$
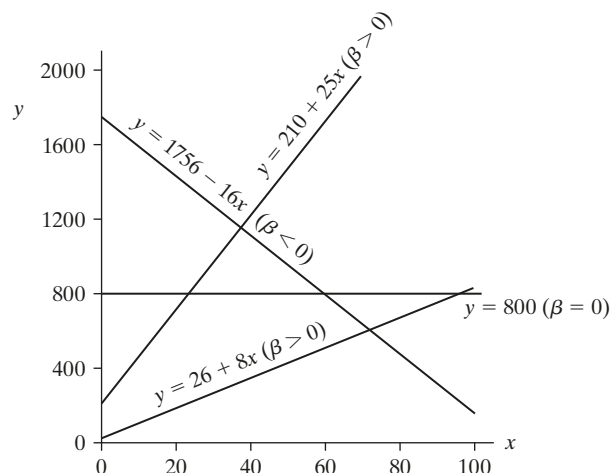
of *y* occur with small values of *x*. When a relationship between two variables follows a straight line with $\beta > 0$, the relationship is said to be ***positive***.

When β is negative, *y decreases* as *x increases*. The straight line then goes downward, and the relationship is said to be ***negative***. For instance, the equation $y = 1756 - 16x$, which has slope -16, approximates the relationship between $y =$ violent crime rate and $x =$ percentage of residents who are high school graduates. For each increase of 1.0 in the percentage who are high school graduates, the violent crime rate decreases by about 16. Figure 9.2 also shows this line. ■

When $\beta = 0$, the graph is a horizontal line. The value of *y* is constant and does not vary as *x* varies. If two variables are independent, with the value of *y* not depending on the value of *x*, a straight line with $\beta = 0$ represents their relationship. The line $y = 800$ shown in Figure 9.2 is an example of a line with $\beta = 0$.

MODELS ARE SIMPLE APPROXIMATIONS FOR REALITY

As Section 7.5 (page 205) explained, a ***model*** is a simple approximation for the relationship between variables in the population. The linear function provides a simple model for the relationship between two quantitative variables. For a given value of *x*, the model $y = \alpha + \beta x$ predicts a value for *y*. The better these predictions tend to be, the better the model.

As we shall discuss in some detail in Chapter 10, *association does not imply causation*. For example, consider the interpretation of the slope in Example 9.2 above of "When the percentage with income below the poverty level increases by 1, the violent crime rate increases by about 25 crimes a year per 100,000 population." This does not mean that if we had the ability to go to a state and increase the percentage of people living below the poverty level from 10% to 11%, we could expect the number of crimes to increase in the next year by 25 crimes per 100,000 people. It merely means that based on current data, if one state had a 10% poverty rate and one had an 11% poverty rate, we'd predict that the state with the higher poverty rate would have 25 more crimes per year per 100,000 people. But, as we'll see in Section 9.3, a *sensible* model is actually a bit more complex than the one we've presented so far, by allowing *variability* in *y*-values at each value for *x*. That model, not merely a straight line, is what we mean by a *regression model*. Before introducing the complete model, in Section 9.3, we'll next see how to find the best approximating straight line.

9.2 Least Squares Prediction Equation

Using sample data, we can estimate the equation for the simple straight-line model. The process treats α and β in the equation $y = \alpha + \beta x$ as parameters and estimates them.

A SCATTERPLOT PORTRAYS THE DATA

The first step of model fitting is to plot the data, to reveal whether a model with a straight-line trend makes sense. The data values (x, y) for any one subject form a point relative to the *x* and *y* axes. A plot of the *n* observations as *n* points is called a ***scatterplot***.

Example
9.3

Scatterplot for Murder Rate and Poverty   For Table 9.1, let x = poverty rate and y = murder rate. Figure 9.3 shows a scatterplot for the 51 observations. Each point portrays the values of poverty rate and murder rate for a given state. For Maryland, for instance, the poverty rate is $x = 9.7$, and the murder rate is $y = 12.7$. Its point $(x, y) = (9.7, 12.7)$ has coordinate 9.7 for the x-axis and 12.7 for the y-axis. This point is labeled MD in Figure 9.3.

FIGURE 9.3: Scatterplot for y = Murder Rate and x = Percentage of Residents below the Poverty Level, for 50 States and D.C.
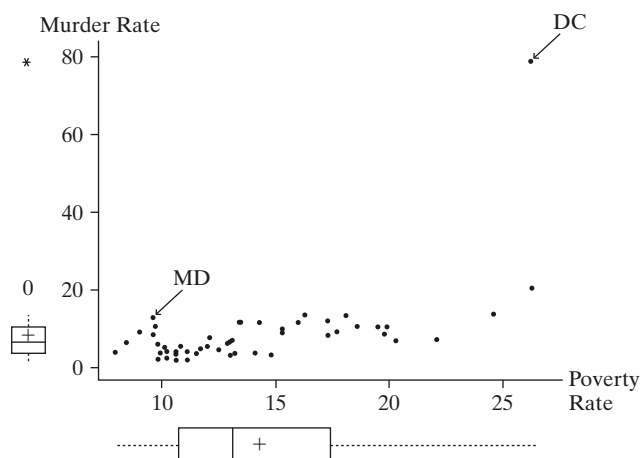
Figure 9.3 indicates that the trend of points seems to be approximated fairly well by a straight line. One point, however, is far removed from the rest. This is the point for the District of Columbia (D.C.). It had murder rate much higher than for any state. This point lies far from the overall trend. Figure 9.3 also shows box plots for these variables. They reveal that D.C. is an extreme *outlier* on murder rate. In fact, it falls 6.5 standard deviations above the mean. We shall see that outliers can have a serious impact on a regression analysis. ∎

The scatterplot provides a visual check of whether a relationship is approximately linear. When the relationship seems highly nonlinear, it is not sensible to use a straight-line model. Figure 9.4 illustrates such a case. This figure shows a negative relationship over part of the range of x-values, and a positive relationship over the rest. These cancel each other out using a straight-line model. For such data, a nonlinear model presented in Section 14.5 is more appropriate.

FIGURE 9.4: A Nonlinear Relationship, for Which It Is Inappropriate to Use a Straight-Line Model
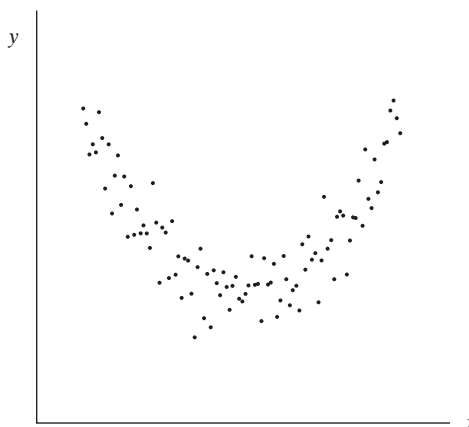
PREDICTION EQUATION

When the scatterplot suggests that the model $y = \alpha + \beta x$ may be appropriate, we use the data to estimate this line. The notation

$$\hat{y} = a + bx$$

represents a *sample* equation that estimates the linear model. In the sample equation, the y-intercept (a) estimates the y-intercept α of the model and the slope (b) estimates the slope β. The sample equation $\hat{y} = a + bx$ is called the **prediction equation**, because it provides a prediction $\hat{y}$ for the response variable at any value of x.

The prediction equation is the best straight line, falling closest to the points in the scatterplot, in a sense explained later in this section. The formulas for a and b in the prediction equation $\hat{y} = a + bx$ are

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \qquad \text{and} \qquad a = \bar{y} - b\bar{x}.$$

If an observation has both x- and y-values above their means, or both x- and y-values below their means, then $(x - \bar{x})(y - \bar{y})$ is positive. The slope estimate b tends to be positive when most observations are like this, that is, when points with large x-values also tend to have large y-values and points with small x-values tend to have small y-values.

We shall not dwell on these formulas or even illustrate how to use them, as anyone who does any serious regression modeling uses statistical software. The appendix at the end of the text provides details. Internet applets are also available.[1]

**Example
9.4**

Predicting Murder Rate from Poverty Rate   For the 51 observations on $y =$ murder rate and $x =$ poverty rate in Table 9.1, SPSS software provides the results shown in Table 9.2. Murder rate has $\bar{y} = 8.7$ and $s = 10.7$, indicating that it is probably highly skewed to the right. The box plot for murder rate in Figure 9.3 shows that the extreme outlying observation for D.C. contributes to this.

The estimates of α and β are listed under the heading[2] B. The estimated y-intercept is $a = -10.14$, listed opposite *(Constant)*. The estimate of the slope is $b = 1.32$, listed opposite the variable name of which it is the coefficient in the prediction equation, *POVERTY*. Therefore, the prediction equation is $\hat{y} = a + bx = -10.14 + 1.32x$.

TABLE 9.2: Part of SPSS Output for Fitting Linear Model to Observations from `Crime2` Data File for 50 States and D.C. on $x =$ Percentage in Poverty and $y =$ Murder Rate

Variable	Mean	Std Deviation		B	Std. Error
MURDER	8.727	10.718	(Constant)	-10.1364	4.1206
POVERTY	14.259	4.584	POVERTY	1.3230	0.2754

The slope $b = 1.32$ is positive. So, the larger the poverty rate, the larger is the predicted murder rate. The value 1.32 indicates that an increase of 1 in the percentage living below the poverty rate corresponds to an increase of 1.32 in the predicted murder rate.

---

[1] For example, the *Fit Linear Regression Model* applet at `www.pearsonglobaleditions.com/Agresti`.
[2] B is the symbol SPSS uses to denote an estimated regression coefficient. Stata uses *Coef.* as the heading, short for coefficient, R uses *Coefficients*, and SAS uses *Parameter estimate*.

Similarly, an increase of 10 in the poverty rate corresponds to a $10(1.32) =$ 13.2-unit increase in predicted murder rate. If one state has a 12% poverty rate and another has a 22% poverty rate, for example, the predicted annual number of murders per 100,000 population is 13.2 higher in the second state than the first state. This differential of 13 murders per 100,000 population translates to 130 per million or 1300 per 10 million population. If the two states each had populations of 10 million, the one with the higher poverty rate would be predicted to have 1300 more murders per year. ■

EFFECT OF OUTLIERS ON THE PREDICTION EQUATION

Figure 9.5 plots the prediction equation from Example 9.4 over the scatterplot. The diagram shows that the observation for D.C. (the sole point in the top-right corner) is a *regression outlier*—it falls quite far from the trend that the rest of the data follow. This observation seems to have a substantial effect. The line seems to be pulled up toward it and away from the center of the general trend of points.

FIGURE 9.5: Prediction Equations Relating Murder Rate and Percentage in Poverty, with and without D.C. Observation
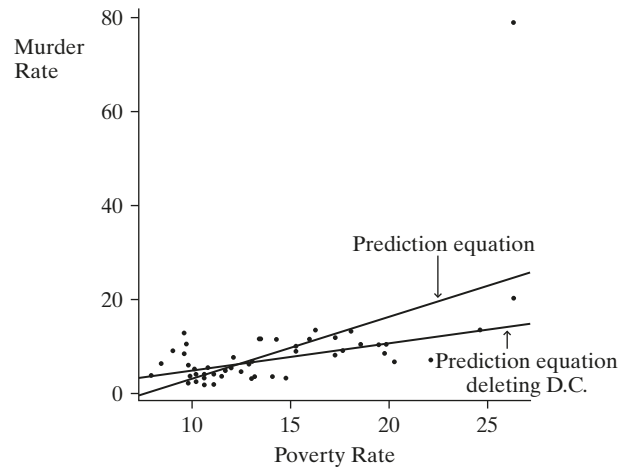
Let's now refit the line using the observations for the 50 states but not the one for D.C. Table 9.3 shows that the prediction equation is $\hat{y} = -0.86 + 0.58x$. Figure 9.5 also shows this line, which passes more directly through the 50 points. The slope is 0.58, compared to 1.32 when the observation for D.C. is included. The one outlying observation has the impact of more than doubling the slope!

An observation is called *influential* if removing it results in a large change in the prediction equation. Unless the sample size is large, an observation can have a strong influence on the slope if its *x*-value is low or high compared to the rest of the data and if it is a regression outlier.

In summary, the line for the data set including D.C. seems to distort the relationship for the 50 states. It seems wiser to use the equation based on the 50 states alone rather than to use a single equation for both the 50 states and D.C. This line for the 50 states better represents the overall trend. In reporting these results, we would note that the murder rate for D.C. falls outside this trend, being much larger than this equation predicts.

TABLE 9.3: Software Output for Fitting Linear Model to `Crime2` Data File on 50 States (but Not D.C.) on $x =$ Percentage in Poverty and $y =$ Murder Rate

	Sum of Squares	df	Mean Square		Unstandardized Coefficients	
Regression	307.342	1	307.34		B	
Residual	470.406	48	9.80		(Constant)	-.857
Total	777.749	49			poverty	.584

	murder	predict	residual
1	9.0000	4.4599	4.5401
2	11.6000	9.3091	2.2909
3	10.2000	10.8281	-0.6281
4	8.6000	8.1406	0.4594

PREDICTION ERRORS ARE CALLED RESIDUALS

For the prediction equation $\hat{y} = -0.86 + 0.58x$ for the 50 states, a comparison of the *actual* murder rates to the *predicted* values checks the goodness of the equation. For example, Massachusetts had poverty rate $x = 10.7$ and $y = 3.9$. The predicted murder rate ($\hat{y}$) at $x = 10.7$ is $\hat{y} = -0.86 + 0.58x = -0.86 + 0.58(10.7) = 5.4$. The prediction error is the difference between the actual y-value of 3.9 and the predicted value of 5.4, or $y - \hat{y} = 3.9 - 5.4 = -1.5$. The prediction equation overestimates the murder rate by 1.5. Similarly, for Louisiana, $x = 26.4$ and $\hat{y} = -0.86 + 0.58(26.4) = 14.6$. The actual murder rate is $y = 20.3$, so the prediction is too low. The prediction error is $y - \hat{y} = 20.3 - 14.6 = 5.7$. The prediction errors are called **residuals**.

Residual	For an observation, the difference between an observed value and the predicted value of the response variable, $y - \hat{y}$, is called the **residual**.

Table 9.3 shows the murder rates, the predicted values, and the residuals for the first four states in the data file. A *positive* residual results when the observed value y is *larger* than the predicted value $\hat{y}$, so $y - \hat{y} > 0$. A *negative* residual results when the observed value is *smaller* than the predicted value. The smaller the absolute value of the residual, the better is the prediction, since the predicted value is closer to the observed value.

In a scatterplot, the residual for an observation is the vertical distance between its point and the prediction line. Figure 9.6 illustrates this. For example, the observation for Louisiana is the point with (x, y) coordinates (26.4, 20.3). The prediction is represented by the point (26.4, 14.6) on the prediction line obtained by substituting $x = 26.4$ into the prediction equation $\hat{y} = -0.86 + 0.58x$. The residual is the difference between the observed and predicted points, which is the vertical distance $y - \hat{y} = 20.3 - 14.6 = 5.7$.

PREDICTION EQUATION HAS LEAST SQUARES PROPERTY

We summarize the size of the residuals by the sum of their squared values. This quantity, denoted by SSE, is

$$\text{SSE} = \sum (y - \hat{y})^2.$$

FIGURE 9.6: Prediction Equation and Residuals. A residual is a vertical distance between a data point and the prediction line.
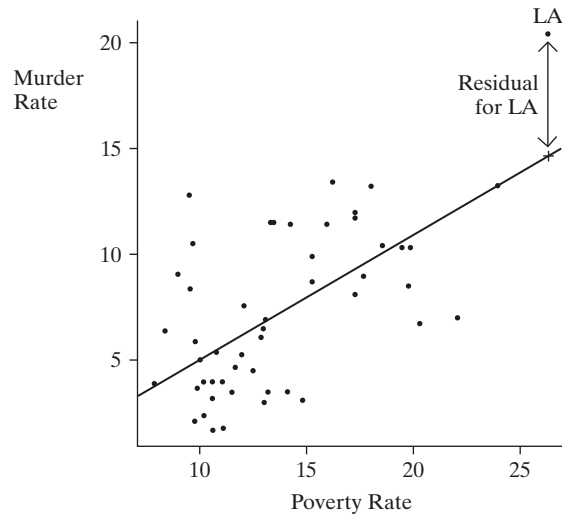
In other words, the residual is computed for every observation in the sample, each residual is squared, and then SSE is the sum of these squares. The symbol SSE is an abbreviation for **sum of squared errors**. This terminology refers to the residual being a measure of prediction error from using $\hat{y}$ to predict y.

The better the prediction equation, the smaller the residuals tend to be and, hence, the smaller SSE tends to be. Any particular equation has corresponding residuals and a value of SSE. The prediction equation specified by the formulas on page 264 for the estimates a and b of α and β has the *smallest* value of SSE out of all possible linear prediction equations.

Least Squares Estimates

> The **least squares estimates** a and b are the values that provide the prediction equation $\hat{y} = a + bx$ for which the residual sum of squares, SSE $= \sum(y - \hat{y})^2$, is a minimum.

The prediction line $\hat{y} = a + bx$ is called the **least squares line**, because it is the one with the smallest sum of squared residuals. If we square the residuals (such as those in Table 9.3) for the least squares line $\hat{y} = -0.86 + 0.58x$ and then sum them, we get

$$\text{SSE} = \sum(y - \hat{y})^2 = (4.54)^2 + (2.29)^2 + \cdots = 470.4.$$

This value is smaller than SSE for *any* other straight line predictor, such as $\hat{y} = -0.88 + 0.60x$. In this sense, the data fall closer to this line than to *any* other line. Most software (e.g., R, SPSS, Stata) calls SSE the **residual sum of squares**. It describes the variation of the data around the prediction line. Table 9.3 reports it in the *Sum of Squares* column, in the row labeled *Residual*.

Besides making the errors as small as possible in this summary sense, the least squares line

- Has some positive residuals and some negative residuals, but the sum (and mean) of the residuals equals 0.

- Passes through the point $(\bar{x}, \bar{y})$.

The first property tells us that the too-low predictions are balanced by the too-high predictions. Just as deviations of observations from their mean $\bar{y}$ satisfy $\sum(y - \bar{y}) = 0$,

so does the prediction equation satisfy $\sum(y - \hat{y}) = 0$. The second property tells us that the line passes through the center of the data.

9.3 The Linear Regression Model

For the linear model $y = \alpha + \beta x$, each value of x corresponds to a single value of y. Such a model is said to be **deterministic**. It is unrealistic in social science research, because we do not expect all subjects who have the same x-value to have the same y-value. Instead, the y-values *vary*.

For example, let $x =$ number of years of education and $y =$ annual income. The subjects having $x = 12$ years of education do not all have the same income, because income is not completely dependent upon education. Instead, a probability distribution describes annual income for individuals with $x = 12$. It is the **conditional distribution** of the y-values at $x = 12$. A separate conditional distribution applies for those with $x = 13$ years of education. Each level of education has its own conditional distribution of income. For example, the mean of the conditional distribution of income would likely be higher at higher levels of education.

A **probabilistic** model for the relationship allows for variability in y at each value of x. We now show how a linear function is the basis for a probabilistic model.

LINEAR REGRESSION FUNCTION

A probabilistic model uses $\alpha + \beta x$ to represent the *mean* of y-values, rather than y itself, as a function of x. For a given value of x, $\alpha + \beta x$ represents the mean of the conditional distribution of y for subjects having that value of x.

Expected Value of *y*	Let $E(y)$ denote the mean of a conditional distribution of y. The symbol E represents *expected value*.

We now use the equation

$$E(y) = \alpha + \beta x$$

to model the relationship between x and the mean of the conditional distribution of y. For $y =$ annual income, in dollars, and $x =$ number of years of education, suppose $E(y) = -5000 + 3000x$. For instance, those having a high school education ($x = 12$) have a mean income of $E(y) = -5000 + 3000(12) = 31{,}000$ dollars. The model states that the *mean* income is 31,000, but allows different subjects having $x = 12$ to have different incomes.

An equation of the form $E(y) = \alpha + \beta x$ that relates values of x to the mean of the conditional distribution of y is called a *regression function*.

Regression Function	A **regression function** is a mathematical function that describes how the mean of the response variable changes according to the value of an explanatory variable.

The function $E(y) = \alpha + \beta x$ is called a *linear* regression function, because it uses a straight line to relate the mean of y to the values of x. In practice, the *regression coefficients* α and β are unknown. Least squares provides the sample prediction equation $\hat{y} = a + bx$. At any particular value of x, $\hat{y} = a + bx$ *estimates* the mean of y for all subjects in the population having that value of x.

DESCRIBING VARIATION ABOUT THE REGRESSION LINE

The linear regression model has an additional parameter σ describing the standard deviation of each conditional distribution. That is, σ measures the variability of the y-values for all subjects having the same x-value. We refer to σ as the ***conditional standard deviation***.

A model also assumes a particular probability distribution for the conditional distribution of y. This is needed to make inference about the parameters. For quantitative variables, the most common assumption is that the conditional distribution of y is normal at each fixed value of x, with unknown standard deviation σ.

Example 9.5

Describing How Income Varies, for Given Education Again, suppose $E(y) = -5000 + 3000x$ describes the relationship between mean annual income and number of years of education. The slope $\beta = 3000$ implies that mean income increases 3000 for each year increase in education. Suppose also that the conditional distribution of income is normal, with $\sigma = 13,000$. According to this model, for individuals with x years of education, their incomes have a normal distribution with a mean of $E(y) = -5000 + 3000x$ and a standard deviation of $13,000$.

Those having a high school education ($x = 12$) have a mean income of $E(y) = -5000 + 3000(12) = 31,000$ dollars and a standard deviation of $13,000$ dollars. So, about 95% of the incomes fall within two standard deviations of the mean, that is, between $31,000 - 2(13,000) = 5000$ and $31,000 + 2(13,000) = 57,000$ dollars. Those with a college education ($x = 16$) have a mean annual income of $E(y) = -5000 + 3000(16) = 43,000$ dollars, with about 95% of the incomes falling between $17,000$ and $69,000$. Figure 9.7 pictures this regression model. ■

FIGURE 9.7: The Regression Model $E(y) = -5000 + 3000x$, with $\sigma = 13$, Relating the Mean of $y =$ Income (in Dollars) to $x =$ Education (in Years). The figure shows the conditional income distributions at $x = 8, 12,$ and 16 years.
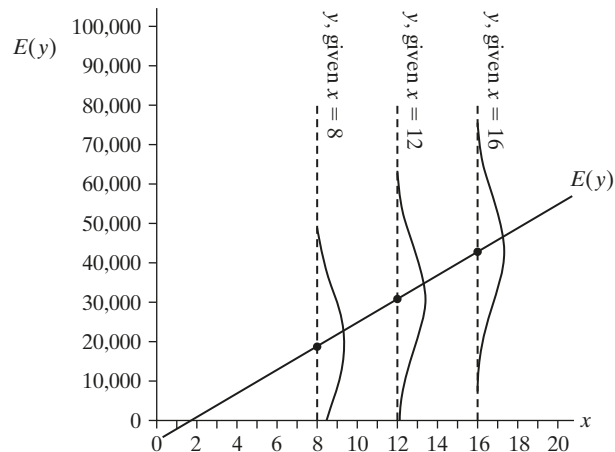
In Figure 9.7, each conditional distribution is normal, and each has the same standard deviation, $\sigma = 13$. In practice, the distributions would not be exactly normal, and the standard deviation need not be the same for each. *Any model never holds exactly in practice*. It is merely a simple approximation for reality. For sample data, we'll learn about ways to check whether a particular model is realistic. The most important assumption is that the regression equation is linear. The scatterplot helps us check whether this assumption is badly violated, as we'll discuss later in the chapter.

RESIDUAL MEAN SQUARE: ESTIMATING CONDITIONAL VARIATION

The ordinary linear regression model assumes that the standard deviation σ of the conditional distribution of y is identical at the various values of x. The estimate of σ uses $\text{SSE} = \sum(y - \hat{y})^2$, which measures sample variability about the least squares line. The estimate is

$$s = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}}.$$

If the constant variation assumption is not valid, then s summarizes the *average* variability about the line.

Example 9.6

TV Watching and Grade Point Averages A survey[3] of 50 college students in an introductory psychology class observed self-reports of $y = $ high school GPA and $x = $ weekly number of hours viewing television. The study reported $\hat{y} = 3.44 - 0.03x$. Software reports sums of squares shown in Table 9.4. This type of table is called an *ANOVA table*. Here, ANOVA is an acronym for *analysis of variability*. The residual sum of squares in using x to predict y was $\text{SSE} = 11.66$. The estimated conditional standard deviation is

$$s = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{11.66}{50-2}} = 0.49.$$

TABLE 9.4: Software Output of ANOVA Table for Sums of Squares in Fitting Regression Model to $y = $ High School GPA and $x = $ Weekly TV Watching

	Sum of Squares	df	Mean Square
Regression	3.63	1	3.63
Residual	11.66	48	.24
Total	15.29	49	

At any fixed value x of TV viewing, the model predicts that GPAs vary around a mean of $3.44 - 0.03x$ with a standard deviation of 0.49. At $x = 20$ hours a week, for instance, the conditional distribution of GPA is estimated to have a mean of $3.44 - 0.03(20) = 2.84$ and standard deviation of 0.49. ■

The term $(n - 2)$ in the denominator of s is the **degrees of freedom** (df) for the estimate. When a regression equation has p unknown parameters, then $df = n - p$. The equation $E(y) = \alpha + \beta x$ has two parameters (α and β), so $df = n - 2$. The table in the above example lists $\text{SSE} = 11.66$ and its $df = n - 2 = 50 - 2 = 48$. The ratio of these, $s^2 = 0.24$, is listed on the printout in the *Mean Square* column. Most software calls this the *residual mean square*. Its square root is the estimate of the conditional standard deviation of y, $s = \sqrt{0.24} = 0.49$. Among the names software calls this are *Root MSE* (Stata and SAS) for the square root of the mean square error, *Residual standard error* (R), and *Standard error of the estimate* (SPSS).

---

[3] https://www.iusb.edu/ugr-journal/static/2002/index.php.

CONDITIONAL VARIATION TENDS TO BE LESS THAN MARGINAL VARIATION

From pages 55 and 117, a point estimate of the population standard deviation of a variable y is

$$\sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}}.$$

This is the standard deviation of the *marginal* distribution of y, because it uses only the y-values. It ignores values of x. To emphasize that this standard deviation depends on values of y alone, the remainder of the text denotes it by s_y in a sample and σ_y in a population. It differs from the standard deviation of the *conditional* distribution of y, for a fixed value of x. To reflect its conditional form, that standard deviation is sometimes denoted by $s_{y|x}$ for the sample estimate and $\sigma_{y|x}$ for the population. For simplicity, we use s and σ.

The sum of squares $\sum(y - \bar{y})^2$ in the numerator of s_y is called the ***total sum of squares***. In Table 9.4 for the 50 student GPAs, it is 15.29. Thus, the marginal standard deviation of GPA is $s_y = \sqrt{15.29/(50 - 1)} = 0.56$. Example 9.6 showed that the conditional standard deviation is $s = 0.49$.

Typically, less spread in y-values occurs at a fixed value of x than totaled over all such values. We'll see that the stronger the association between x and y, the less the conditional variability tends to be relative to the marginal variability. For example, suppose the *marginal* distribution of college GPAs (y) at your school falls between 1.0 and 4.0, with $s_y = 0.60$. Suppose we could predict college GPA *perfectly* using x = high school GPA, with the prediction equation $\hat{y} = 0.40 + 0.90x$. Then, SSE = 0, and the conditional standard deviation would be $s = 0$. In practice, perfect prediction would not happen. However, the stronger the association in terms of less prediction error, the smaller the conditional variability would be. See Figure 9.8, which portrays a marginal distribution that is much more spread out than each conditional distribution.

FIGURE 9.8: Marginal and Conditional Distributions. The marginal distribution shows the overall variability in y-values, whereas the conditional distribution shows how y varies at a fixed value of x.
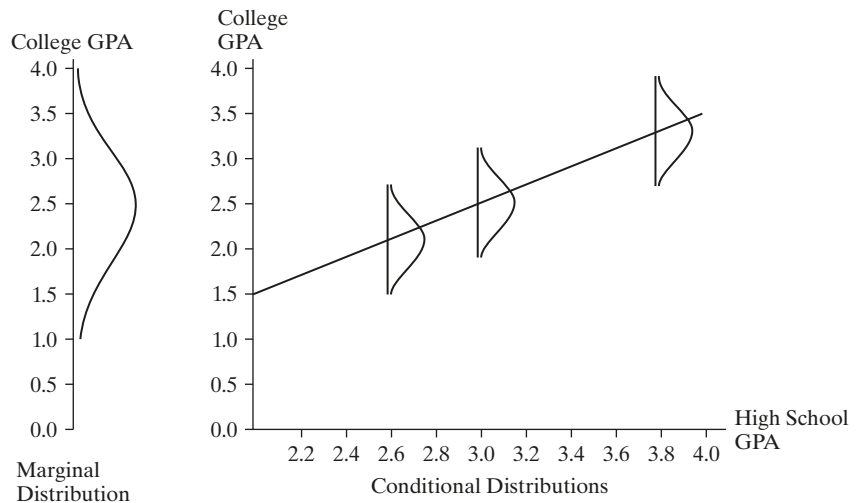
9.4 Measuring Linear Association: The Correlation

The linear regression model uses a straight line to describe the relationship. For this model, this section introduces two measures of the strength of association between two quantitative variables.

THE SLOPE AND STRENGTH OF ASSOCIATION

The slope b of the prediction equation tells us the *direction* of the association. Its sign indicates whether the prediction line slopes upward or downward as x increases, that is, whether the association is positive or negative. The slope does not, however, directly tell us the strength of the association. The reason for this is that its numerical value is intrinsically linked to the units of measurement.

For example, consider the prediction equation $\hat{y} = -0.86 + 0.58x$ for $y = $ murder rate and $x = $ percentage living below the poverty level. A one-unit increase in x corresponds to a $b = 0.58$ increase in the predicted number of murders per 100,000 people. This is equivalent to a 5.8 increase in the predicted number of murders per 1,000,000 people. So, if murder rate is the number of murders per 1,000,000 population instead of per 100,000 population, the slope is 5.8 instead of 0.58. The strength of the association is the same in each case, since the variables and data are the same. Only the units of measurement for y differed. The slope b doesn't directly indicate whether the association is strong or weak, because we can make b as large or as small as we like by an appropriate choice of units.

The slope *is* useful for comparing effects of two predictors having the same units. For instance, the prediction equation relating murder rate to percentage living in urban areas is $3.28 + 0.06x$. A one-unit increase in the percentage living in urban areas corresponds to a 0.06 predicted increase in the murder rate, whereas a one-unit increase in the percentage below the poverty level corresponds to a 0.58 predicted increase in the murder rate. An increase of 1 in percentage below the poverty level has a much greater effect on the murder rate than an increase of 1 in percentage urban.

The measures of association we now study do not depend on the units of measurement. Like the measures of association that Chapter 8 presented for categorical data, their magnitudes indicate the strength of association.

THE CORRELATION

On page 65, we introduced the ***correlation*** between quantitative variables. Its value, unlike that of the slope b, does not depend on the units of measurement.

Correlation

> The ***correlation*** between variables x and y, denoted by r, is
>
> $$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\left[\sum (x - \bar{x})^2\right]\left[\sum (y - \bar{y})^2\right]}}.$$

The formulas for the correlation and for the slope (page 264) have the same numerator, relating to the covariation of x and y. The correlation is a *standardized* version of the slope. The standardization adjusts the slope b for the fact that the standard deviations of x and y depend on their units of measurement. Let s_x and s_y denote the marginal sample standard deviations of x and y,

$$s_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \quad \text{and} \quad s_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}.$$

Here is the simple connection between the slope estimate and the sample correlation:

**Correlation Is
a Standardized Slope**

The correlation relates to the slope b of the prediction equation $\hat{y} = a + bx$ by

$$r = \left(\frac{s_x}{s_y}\right) b.$$

If the sample spreads are equal ($s_x = s_y$), then $r = b$. The correlation is the value the slope would take for units such that the variables have equal standard deviations. For example, when the variables are standardized by converting their values to z-scores, both standardized variables have standard deviations of 1.0. Because of the relationship between r and b, the correlation is also called the ***standardized regression coefficient*** for the model $E(y) = \alpha + \beta x$. In practice, it is not necessary to standardize the variables, but we can interpret the correlation as the value the slope would equal if the variables were equally spread out.

The point estimate r of the correlation was proposed by the British statistical scientist Karl Pearson in 1896, just four years before he developed the chi-squared test of independence for contingency tables. In fact, this estimate is sometimes called the ***Pearson correlation***.

**Example
9.7**

Correlation between Murder Rate and Poverty Rate   For the data in Table 9.1, the prediction equation relating $y =$ murder rate to $x =$ poverty rate is $\hat{y} = -0.86 + 0.58x$. Software tells us that $s_x = 4.29$ for poverty rate, $s_y = 3.98$ for murder rate, and the correlation $r = 0.63$. In fact,

$$r = \left(\frac{s_x}{s_y}\right) b = \left(\frac{4.29}{3.98}\right)(0.58) = 0.63.$$

We will interpret this value after studying the properties of the correlation. ∎

PROPERTIES OF THE CORRELATION

- The correlation is valid only when a straight-line model is sensible for the relationship between x and y. Since r is proportional to the slope of a linear prediction equation, it measures the *strength of the linear association*.

- $-1 \leq r \leq 1$. The correlation, unlike the slope b, must fall between -1 and $+1$. We shall see why later in the section.

- r has the same sign as the slope b. This holds because their formulas have the same numerator, relating to covariation of x and y, and positive denominators. Thus, $r > 0$ when the variables are positively related, and $r < 0$ when the variables are negatively related.

- $r = 0$ for those lines having $b = 0$. When $r = 0$, there is not a linear increasing or linear decreasing trend in the relationship.

- $r = \pm 1$ when all the sample points fall exactly on the prediction line. These correspond to *perfect* positive and negative linear associations. There is then no prediction error when we use $\hat{y} = a + bx$ to predict y.

- The larger the absolute value of r, the stronger the linear association. Variables with a correlation of -0.80 are more strongly linearly associated than variables with a correlation of 0.40. Figure 9.9 shows scatterplots having various values for r.

FIGURE 9.9: Scatterplots for Different Correlation Values
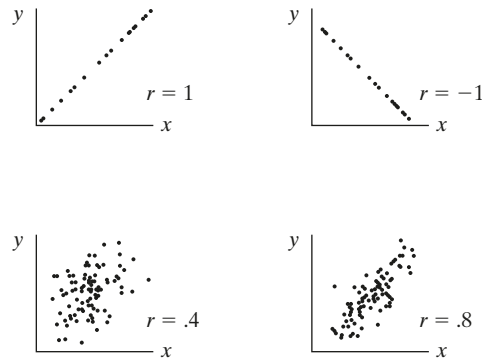
- The correlation, unlike the slope b, treats x and y symmetrically. The prediction equation using y to predict x has the same correlation as the one using x to predict y.

- The value of r does not depend on the variables' units.

For example, if y is the number of murders per 1,000,000 population instead of per 100,000 population, we obtain the same value of $r = 0.63$. Also, when murder rate predicts poverty rate, the correlation is the same as when poverty rate predicts murder rate, $r = 0.63$ in both cases.

The correlation is useful for comparing associations for variables having different units. Another potential predictor for murder rate is the mean number of years of education completed by adult residents in the state. Poverty rate and education have different units, so a one-unit change in poverty rate is not comparable to a one-unit change in education. Their slopes from the separate prediction equations are not comparable. The correlations are comparable. Suppose the correlation of murder rate with education is -0.30. Since the correlation of murder rate with poverty rate is 0.63, and since $0.63 > |-0.30|$, murder rate is more strongly associated with poverty rate than with education.

We emphasize that the correlation describes *linear* relationships. For curvilinear relationships, the best-fitting prediction line may be completely or nearly horizontal, and $r = 0$ when $b = 0$. See Figure 9.10. A low absolute value for r does not then imply that the variables are unassociated, but merely that the association is not linear.

FIGURE 9.10: Scatterplot for Which $r = 0$, Even Though There Is a Strong Curvilinear Relationship
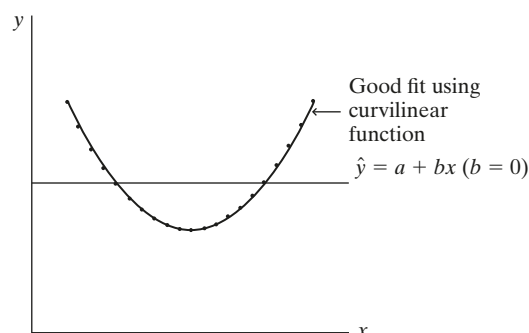
CORRELATION IMPLIES REGRESSION TOWARD THE MEAN

Another interpretation of the correlation relates to its standardized slope property. We can rewrite the equality

$$r = \left(\frac{s_x}{s_y}\right)b \quad \text{as} \quad s_x b = r s_y.$$

Now, the slope b is the change in $\hat{y}$ for a one-unit increase in x. An increase in x of s_x units has a predicted change of $s_x b$ units. (For instance, if $s_x = 10$, an increase of 10 units in x corresponds to a change in $\hat{y}$ of $10b$.) See Figure 9.11. Since $s_x b = r s_y$, an increase of s_x in x corresponds to a predicted change of r standard deviations in the y-values.

FIGURE 9.11: An Increase of s_x Units in x Corresponds to a Predicted Change of $r s_y$ Units in y
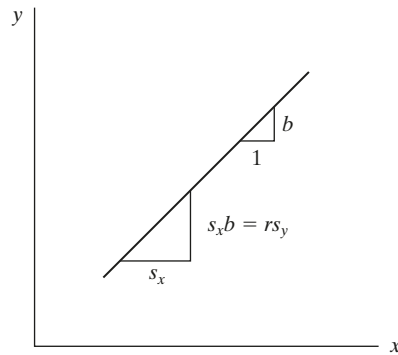
For example, let's start at the point $(\bar{x}, \bar{y})$ through which the prediction equation passes and consider the impact of x moving above $\bar{x}$ by a standard deviation. Suppose that $r = 0.5$. An increase of s_x in x, from $\bar{x}$ to $(\bar{x} + s_x)$, corresponds to a predicted increase of $0.5 s_y$ in y, from $\bar{y}$ to $(\bar{y} + 0.5 s_y)$. We predict that y is closer to the mean, in standard deviation units. This is called ***regression toward the mean***. The larger the absolute value of r, the stronger the association, in the sense that a standard deviation change in x corresponds to a greater proportion of a standard deviation change in y.

Example 9.8

Child's Height Regresses toward the Mean   The British scientist Sir Francis Galton discovered the basic ideas of regression and correlation in the 1880s. After multiplying each female height by 1.08 to account for gender differences, he noted that the correlation between $x =$ parent height (the average of father's and mother's height) and $y =$ child's height is about 0.5. From the property just discussed, a standard deviation change in parent height corresponds to half a standard deviation change in child's height.

For parents of average height, the child's height is predicted to be average. If, on the other hand, parent height is a standard deviation above average, the child is predicted to be half a standard deviation above average. If parent height is two standard deviations below average, the child is predicted to be one standard deviation below average.

Since r is less than 1, a y-value is predicted to be fewer standard deviations from its mean than x is from its mean. Tall parents tend to have tall children, but on the average not quite so tall. For instance, if you consider all fathers with height 7 feet, perhaps their sons average 6 feet 5 inches—taller than average, but not so extremely tall; if you consider all fathers with height 5 feet, perhaps their sons average 5 feet

5 inches—shorter than average, but not so extremely short. In each case, Galton pointed out the *regression toward the mean*. This is the origin of the name for regression analysis. ∎

r-SQUARED: PROPORTIONAL REDUCTION IN PREDICTION ERROR

A related measure of association summarizes how well x can predict y. If we can predict y much better by substituting x-values into the prediction equation $\hat{y} = a + bx$ than without knowing the x-values, the variables are judged to be strongly associated. This measure of association has four elements:

- A rule for predicting y without using x. We refer to this as Rule 1.

- A rule for predicting y using information on x. We refer to this as Rule 2.

- A summary measure of prediction error for each rule, E_1 for errors by rule 1 and E_2 for errors by rule 2.

- The difference in the amount of error with the two rules is $E_1 - E_2$. Converting this reduction in error to a proportion provides the definition

$$\text{Proportional reduction in error} = \frac{E_1 - E_2}{E_1}.$$

Rule 1 (Predicting y without using x): The best predictor is $\bar{y}$, the sample mean.

Rule 2 (Predicting y using x): When the relationship between x and y is linear, the prediction equation $\hat{y} = a + bx$ provides the best predictor of y.

Prediction Errors: The prediction error for each subject is the difference between the observed and predicted values of y. The prediction error using rule 1 is $y - \bar{y}$, and the prediction error using rule 2 is $y - \hat{y}$, the residual. For each predictor, some prediction errors are positive, some are negative, and the sum of the errors equals 0. We summarize the prediction errors by their sum of squared values,

$$E = \sum (\text{observed } y\text{-value} - \text{predicted } y\text{-value})^2.$$

For rule 1, the predicted values all equal $\bar{y}$. The total prediction error is

$$E_1 = \sum (y - \bar{y})^2.$$

This is the *total sum of squares* of the y-values about their mean. We denote this by TSS. For rule 2 (predicting using the $\hat{y}$-values), the total prediction error is

$$E_2 = \sum (y - \hat{y})^2.$$

We have denoted this by SSE, called the **sum of squared errors** or the **residual sum of squares**.

When x and y have a strong linear association, the prediction equation provides predictions ($\hat{y}$) that are much better than $\bar{y}$, in the sense that the sum of squared prediction errors is substantially less. Figure 9.12 shows graphical representations of the two predictors and their prediction errors. For rule 1, the same prediction ($\bar{y}$) applies for the value of y, regardless of the value of x. For rule 2, the prediction changes as x changes, and the prediction errors tend to be smaller.

FIGURE 9.12: Graphical Representation of Rule 1 and Total Sum of Squares $E_1 = \text{TSS} = \sum(y - \bar{y})^2$, Rule 2 and Residual Sum of Squares $E_2 = \text{SSE} = \sum(y - \hat{y})^2$
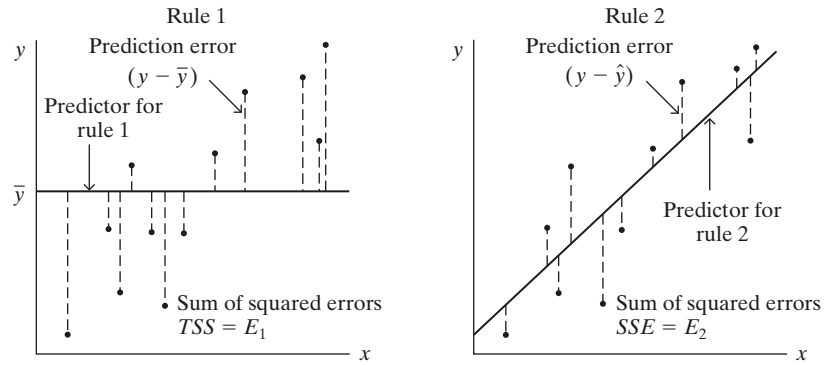
Definition of Measure: The proportional reduction in error from using the linear prediction equation instead of $\bar{y}$ to predict y is

$$r^2 = \frac{E_1 - E_2}{E_1} = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = \frac{\sum(y - \bar{y})^2 - \sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}.$$

It is called *r-squared*, or sometimes the *coefficient of determination*. The notation r^2 is used for this measure because, in fact, the proportional reduction in error equals the square of the correlation r.

Example 9.9

r^2 for Murder Rate and Poverty Rate   The correlation between poverty rate and murder rate for the 50 states is $r = 0.629$. Therefore, $r^2 = (0.629)^2 = 0.395$. For predicting murder rate, the linear prediction equation $\hat{y} = -0.86 + 0.58x$ has 39.5% less error than $\bar{y}$.

Software for regression routinely provides tables that contain the sums of squares that compose r^2. For example, part of Table 9.3 contained the ANOVA table

```
              Sum of Squares
Regression       307.342
Residual         470.406
Total            777.749
```

The sum of squared errors using the prediction equation is $\text{SSE} = \sum(y - \hat{y})^2 = 470.4$, and the total sum of squares is $\text{TSS} = \sum(y - \bar{y})^2 = 777.7$. Thus,

$$r^2 = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = \frac{777.7 - 470.4}{777.7} = \frac{307.3}{777.7} = 0.395.$$

In practice, it is unnecessary to perform this computation, since software reports r or r^2 or both.   ∎

PROPERTIES OF *r*-SQUARED

The properties of r^2 follow directly from those of the correlation r or from its definition in terms of the sums of squares.

- Since $-1 \leq r \leq 1$, r^2 falls between 0 and 1.

- The minimum possible value for SSE is 0, in which case $r^2 = \text{TSS}/\text{TSS} = 1$. For SSE = 0, all sample points must fall exactly on the prediction line. In that case, there is no error using x to predict y with the prediction equation. This condition corresponds to $r = \pm 1$.

- When the least squares slope $b = 0$, the y-intercept a equals $\bar{y}$ (because $a = \bar{y} - b\bar{x}$, which equals $\bar{y}$ when $b = 0$). Then, $\hat{y} = \bar{y}$ for all x. The two prediction rules are then identical, so SSE = TSS and $r^2 = 0$.

- Like the correlation, r^2 measures the strength of *linear* association. The closer r^2 is to 1, the stronger the linear association, in the sense that the more effective the least squares line $\hat{y} = a + bx$ is compared to $\bar{y}$ in predicting y.

- r^2 does not depend on the units of measurement, and it takes the same value when x predicts y as when y predicts x.

SUMS OF SQUARES DESCRIBE CONDITIONAL AND MARGINAL VARIABILITY

To summarize, the correlation r falls between -1 and $+1$. It indicates the direction of the association, positive or negative, through its sign. It is a standardized slope, equaling the slope when x and y are equally spread out. A one standard deviation change in x corresponds to a predicted change of r standard deviations in y. The square of the correlation has a proportional reduction in error interpretation related to predicting y using $\hat{y} = a + bx$ rather than $\bar{y}$.

The total sum of squares, TSS = $\sum(y - \bar{y})^2$, summarizes the *variability* of the observations on y, since this quantity divided by $n - 1$ is the sample variance s_y^2 of the y-values. Similarly, SSE = $\sum(y - \hat{y})^2$ summarizes the variability around the prediction equation, which refers to variability for the conditional distributions. For example, when $r^2 = 0.39$, the variability in y using x to make the predictions is 39% less than the overall variability of the y-values. Thus, the r^2 result is often expressed as "the poverty rate explains 39% of the variability in murder rate" or "39% of the variance in murder rate is explained by its linear relationship with the poverty rate." Roughly speaking, the variance of the conditional distribution of murder rate for a given poverty rate is 39% smaller than the variance of the marginal distribution of murder rate.

This interpretation has the weakness, however, that variability is summarized by the *variance*. Many statisticians find r^2 to be less useful than r because, being based on sums of squares, it uses the square of the original scale of measurement. It's easier to interpret the original scale than a squared scale. This is also the advantage of the standard deviation over the variance.

9.5 Inferences for the Slope and Correlation

We have seen that a linear regression model can represent the *form* of a relationship between two quantitative variables. We use the correlation and its square to describe the *strength* of the association. These parts of a regression analysis are descriptive. We now present inferential methods for the regression model.

A test of whether the two quantitative variables are statistically independent has the same purpose as the chi-squared test for categorical variables. A confidence interval for the slope of the regression equation or the correlation tells us about the size of the effect. These inferences enable us to judge whether the variables are associated and to estimate the direction and strength of the association.

ASSUMPTIONS FOR STATISTICAL INFERENCE

Statistical inferences for regression make the following assumptions:

- Randomization, such as a simple random sample in a survey.

- The mean of y is related to x by the linear equation $E(y) = \alpha + \beta x$.

- The conditional standard deviation σ is identical at each x-value.

- The conditional distribution of y at each value of x is normal.

The assumption about a common σ is one under which the least squares estimates are the best possible estimates of the regression coefficients.[4] The assumption about normality assures that the test statistic for a test of independence has a t sampling distribution. In practice, none of these assumptions is ever satisfied exactly. In the final section of the chapter, we'll see that the important assumptions are the first two.

TEST OF INDEPENDENCE USING SLOPE OR CORRELATION

Under the above assumptions, suppose the population mean of y is identical at each x-value. In other words, the normal conditional distribution of y is the same at each x-value. Then, the two quantitative variables are statistically independent. For the linear regression function $E(y) = \alpha + \beta x$, this means that the slope $\beta = 0$ (see Figure 9.13). The null hypothesis that the variables are statistically independent is $H_0: \beta = 0$.

FIGURE 9.13: x and y Are Statistically Independent when the Slope $\beta = 0$ in the Regression Model $E(y) = \alpha + \beta x$ with Normal Conditional Distributions Having Identical Standard Deviations
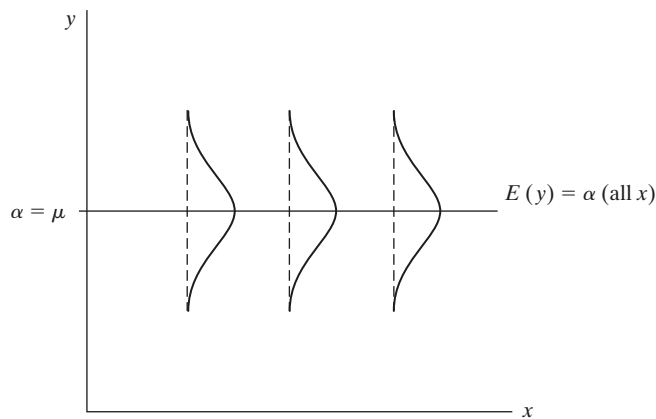
We can test independence against $H_a: \beta \neq 0$, or a one-sided alternative, $H_a: \beta > 0$ or $H_a: \beta < 0$, to predict the direction of the association. The test statistic is

$$t = \frac{b}{se},$$

where se is the standard error of the sample slope b. The form of the test statistic is the usual one for a t or z test. We take the estimate b of the parameter β, subtract the null hypothesis value ($\beta = 0$), and divide by the standard error of the estimate b. Under the assumptions, this test statistic has the t sampling distribution with $df = n - 2$. The P-value for $H_a: \beta \neq 0$ is the two-tail probability from the t distribution.

The formula for the standard error of b is

$$se = \frac{s}{\sqrt{\sum (x - \bar{x})^2}}, \quad \text{where} \quad s = \sqrt{\frac{SSE}{n - 2}}.$$

---

[4] Under the assumptions of normality with common σ, least squares estimates are special cases of *maximum likelihood* estimates, introduced in Section 5.5.

This depends on the point estimate s of the standard deviation of the conditional distributions of y. The degrees of freedom for the t test are the same as the df for s. The smaller s is, the more precisely b estimates β. A small s occurs when the data points show little variability about the prediction equation. Also, the standard error of b is inversely related to $\sum(x - \bar{x})^2$, the sum of squares of the observed x-values about their mean. This sum increases, and hence b estimates β more precisely, as the sample size n increases. (The se also decreases when the x-values are more highly spread out, but the researcher usually has no control over this except in designed experiments.)

The correlation $r = 0$ in the same situations in which the slope $b = 0$. Let ρ (Greek letter rho) denote the correlation value in the population. Then, $\rho = 0$ precisely when $\beta = 0$. In fact, a test of H_0: $\rho = 0$ using the sample value r is equivalent to the t test of H_0: $\beta = 0$ using the sample value b. The test statistic for H_0: $\rho = 0$ is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}.$$

This has the same value as the test statistic $t = b/se$. We can use either statistic to test H_0: independence, since each has the t distribution with $df = n - 2$ and yields the same P-value.

Regression for Selling Price of Homes   What affects the selling price of a house? Table 9.5 shows observations on recent home sales in Gainesville, Florida. This table shows data for eight homes. The entire file for 100 home sales is the `Houses` data file at the text website. Variables listed are selling price (in dollars), size of house (in square feet), annual property taxes (in dollars), number of bedrooms, number of bathrooms, and whether the house is newly built.

TABLE 9.5: Selling Prices and Related Factors for a Sample of Home Sales in Gainesville, Florida (`Houses` Data File)

Home	Selling Price	Size	Taxes	Bedrooms	Bathrooms	New
1	279,900	2048	3104	4	2	No
2	146,500	912	1173	2	1	No
3	237,700	1654	3076	4	2	No
4	200,000	2068	1608	3	2	No
5	159,900	1477	1454	3	3	No
6	499,900	3153	2997	3	2	Yes
7	265,500	1355	4054	3	2	No
8	289,900	2075	3002	3	2	Yes

Note: The complete `Houses` data file for 100 homes is at the text website.

For a set of variables, software can report the correlation for each pair in a **correlation matrix**. This matrix is a square table listing the variables as the rows and again as the columns. Table 9.6 shows the way software reports the correlation matrix for the variables selling price, size, taxes, and number of bedrooms. The correlation between each pair of variables appears twice. For instance, the correlation of 0.834 between selling price and size of house occurs both in the row for *price* and column for *size* and in the row for *size* and column for *price*. The correlations on the diagonal running from the upper left-hand corner to the lower right-hand corner of a correlation matrix all equal 1.0. This merely indicates that the correlation between a variable and itself is 1.0. For instance, if we know the value of y, then we can predict the value of y perfectly.

TABLE 9.6: Correlation Matrix for House Selling Price Data from HOUSES Data File				
		Correlations		
	price	size	taxes	bedrooms
price	1.00000	0.83378	0.84198	0.39396
size	0.83378	1.00000	0.81880	0.54478
taxes	0.84198	0.81880	1.00000	0.47393
bedrooms	0.39396	0.54478	0.47393	1.00000

For now, we use only the data on y = selling price and x = size of house. Since these 100 observations come from one city alone, we cannot use them to make inferences about the relationship between x and y in general. We treat them as a random sample of a conceptual population of home sales in this market in order to analyze how these variables seem to be related.

Figure 9.14 shows a scatterplot, which displays a strong positive trend. The model $E(y) = \alpha + \beta x$ seems appropriate. Some of the points at high levels of size may be outliers, however, and one point falls quite far below the overall trend. We discuss this abnormality in Section 14.4, which introduces an alternative model that does not assume constant variability around the regression line.

FIGURE 9.14: Scatterplot and Prediction Equation for y = Selling Price (in Dollars) and x = Size of House (in Square Feet)
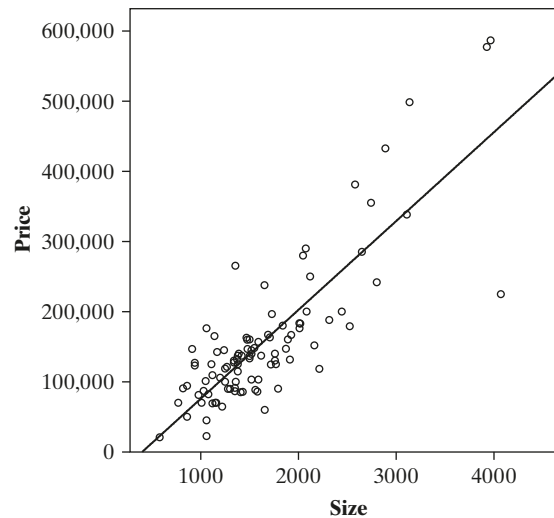
Table 9.7 shows some software output (Stata) for a regression analysis. The prediction equation is $\hat{y} = -50,926.2 + 126.6x$. The predicted selling price increases by $b = 126.6$ dollars for an increase in size of a square foot. Figure 9.14 also superimposes the prediction equation over the scatterplot.

Table 9.7 reports that the standard error of the slope estimate is $se = 8.47$. This value estimates the variability in sample slope values that would result from repeatedly selecting random samples of 100 house sales in Gainesville and calculating prediction equations. For testing independence, H_0: $\beta = 0$, the test statistic is

$$t = \frac{b}{se} = \frac{126.6}{8.47} = 14.95,$$

shown in Table 9.7. Since $n = 100$, its degrees of freedom are $df = n - 2 = 98$. This is an extremely large test statistic. The P-value, listed in Table 9.7 under the heading

TABLE 9.7: Stata Output (Edited) for Regression Analysis of $y =$ Selling Price and $x =$ Size of House from Houses Data File

```
  Source |       SS      df       MS         Number of obs =      100
- -------+----------------------------        F(1, 98)      = 223.52
   Model | 7.0573e+11    1   7.0573e+11       Prob > F      = 0.0000
Residual | 3.0942e+11   98   3.1574e+09       R-squared     = 0.6952
---------+----------------------------        Adj R-squared = 0.6921
   Total | 1.0151e+12   99   1.0254e+10       Root MSE      =  56190

   price |      Coef.   Std. Err.    t    P>|t|   [95% Conf. Interval]
    size |   126.5941   8.467517  14.95   0.000   109.7906    143.3976
   _cons |  -50926.25   14896.37  -3.42   0.001   -80487.62   -21364.89
```

$P > |t|$, is 0.000 to three decimal places. This refers to the two-sided alternative H_a: $\beta \neq 0$. It is the two-tailed probability of a t statistic at least as large in absolute value as the absolute value of the observed t, $|t| = 14.95$, presuming H_0 is true.

We get the same result if we conduct the test using the correlation. The correlation of $r = 0.834$ for the house selling price data has

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{0.834}{\sqrt{(1 - 0.695)/98}} = 14.95.$$

Table 9.8 shows some R output for the same analysis. The two-sided P-value, listed under the heading $Pr(> |t|)$, is 0 to many decimal places.

TABLE 9.8: R Output for Regression Analysis of $y =$ Selling Price and $x =$ Size of House from Houses Data File

```
> fit <- lm(price ~ size)
> summary(fit)


            Estimate Std. Error t value Pr(>|t|)
(Intercept) -50926.255  14896.373   -3.419 0.000918
size            126.594      8.468   14.951  < 2e-16
---
Residual standard error: 56190 on 98 degrees of freedom
Multiple R-squared:  0.6952,    Adjusted R-squared:  0.6921

> cor.test(price,size)

t = 14.951, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:  0.7621910 0.8852286
sample estimates:  cor  0.8337848
```

Both the Stata and R outputs also contain a standard error and t test for the y-intercept. We won't use this information, since rarely is there any reason to test the hypothesis that a y-intercept equals 0. For this example, the y-intercept does not have any interpretation, since houses of size $x = 0$ do not exist.

In summary, the evidence is extremely strong that size of house has a positive effect on selling price. On the average, selling price increases as size increases. This

is no surprise. Indeed, we would be shocked if these variables were independent. For these data, estimating the size of the effect is more relevant than testing whether it exists. ∎

CONFIDENCE INTERVAL FOR THE SLOPE AND CORRELATION

A small P-value for H_0: $\beta = 0$ suggests that the regression line has a nonzero slope. We should be more concerned with the size of the slope β than in knowing merely that it is not 0. A confidence interval for β has the formula

$$b \pm t(se).$$

The t-score is the value, with $df = n - 2$, for the desired confidence level. The form of the interval is similar to the confidence interval for a mean (Section 5.3), namely, take the estimate of the parameter and add and subtract a t multiple of the standard error. The se is the same as se in the test about β.

Constructing a confidence interval for the correlation ρ is more complicated than for the slope β. The reason is that the sampling distribution of r is not symmetric except when $\rho = 0$. The lack of symmetry is caused by the restricted range $[-1, 1]$ for r values. If ρ is close to 1.0, for instance, the sample r cannot fall much above ρ, but it can fall well below ρ. The sampling distribution of r is then skewed to the left. Exercise 9.64 shows how to construct confidence intervals for correlations. This is available with software.

**Example
9.11**

Estimating the Slope and Correlation for House Selling Prices For the data on $x = $ size of house and $y = $ selling price, $b = 126.6$ and $se = 8.47$. The parameter β refers to the change in the mean selling price (in dollars) for each 1-square-foot increase in size. For a 95% confidence interval, we use the $t_{.025}$ value for $df = n - 2 = 98$, which is $t_{.025} = 1.984$. The interval is

$$b \pm t_{.025}(se) = 126.6 \pm 1.984(8.47)$$

$$= 126.6 \pm 16.8, \quad \text{or} \quad (110, 143).$$

We can be 95% confident that β falls between 110 and 143. The mean selling price increases by between \$110 and \$143 for a 1-square-foot increase in house size. ∎

In practice, we make inferences about the change in $E(y)$ for an increase in x that is a relevant portion of the actual range of x-values. If a one-unit increase in x is too small or too large in practical terms, the confidence interval for β can be adjusted to refer to a different change in x. For Table 9.5, $x = $ size of house has $\bar{x} = 1629$ and $s_x = 669$. A change of 1 square foot in size is small. Let's estimate the effect of a 100-square-foot increase in size. The change in the mean of y is 100β. The 95% confidence interval for β is (110, 143), so the 95% confidence interval for 100β has endpoints $100(110) = 11,100$ and $100(143) = 14,300$. We infer that the mean selling price increases by at least \$11,100 and at most \$14,300 for a 100-square-foot increase in house size. For example, assuming that the linear regression model is valid, we conclude that the mean is between \$11,100 and \$14,300 higher for houses of 1700 square feet than for houses of 1600 square feet.

For the house selling price data, we found that the correlation between selling price and size is 0.834. The R output in Table 9.8 tells us that a 95% confidence interval for the population correlation is (0.762, 0.885).

SUMS OF SQUARES IN SOFTWARE OUTPUT

How do we interpret the sums of squares (SS) output in tables such as Table 9.7? In that table, the residual sum of squares (SSE $= 3.0942 \times 10^{11}$) is a huge number because the y-values are very large and their deviations are squared. The estimated conditional standard deviation of y is

$$s = \sqrt{\text{SSE}/(n-2)} = 56{,}190,$$

labeled as *Root MSE* by Stata and *Residual standard error* by R. The sum of squares table also reports the total sum of squares, TSS $= \sum (y - \bar{y})^2 = 1.0151 \times 10^{12}$. From this value and SSE,

$$r^2 = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = 0.695.$$

This is the proportional reduction in error in predicting the selling price using the linear prediction equation instead of the sample mean selling price. A strong association exists between these variables.

   The total sum of squares TSS partitions into two parts, the sum of squared errors, SSE $= 3.0942 \times 10^{11}$, and the difference TSS $-$ SSE $= 7.0573 \times 10^{11}$. This difference is the numerator of the r^2 measure. Software calls this the ***regression sum of squares*** (e.g., SPSS) or the ***model sum of squares*** (e.g., Stata, SAS). It represents the amount of the total variation TSS in y that is explained by x in using the least squares line. The ratio of this sum of squares to TSS equals r^2.

   Tables of sums of squares have an associated list of degrees of freedom values. The *df* for the total sum of squares TSS $= \sum (y - \bar{y})^2$ is $n - 1 = 99$, since TSS refers to variability in the *marginal* distribution of y, which has sample variance $s_y^2 = \text{TSS}/(n-1)$. The degrees of freedom for SSE are $n - 2 = 98$, since it refers to variability in the *conditional* distribution of y, which has variance estimate $s^2 = \text{SSE}/(n-2)$ for a model having two parameters. The regression (model) sum of squares has *df* equal to the number of explanatory variables in the regression model, in this case 1. The sum of *df* for the regression sum of squares and *df* for the residual sum of squared errors is $df = n - 1$ for the total sum of squares, in this case $1 + 98 = 99$.

9.6 Model Assumptions and Violations

We end this chapter by reconsidering the assumptions underlying linear regression analysis. We discuss the effects of violating these assumptions and the effects of *influential* observations. Finally, we show an alternate way to express the model.

WHICH ASSUMPTIONS ARE IMPORTANT?

The linear regression model assumes that the relationship between x and the mean of y follows a straight line. The actual form is unknown. It is almost certainly not *exactly* linear. Nevertheless, a linear function often provides a decent approximation for the actual form. Figure 9.15 illustrates a straight line falling close to an actual curvilinear relationship.

   The inferences introduced in the previous section are appropriate for detecting positive or negative linear associations. Suppose that instead the true relationship were U-shaped, such as in Figure 9.4. Then, the variables would be statistically dependent, since the mean of y would change as the value of x changes. The t test of H_0: $\beta = 0$ might not detect it, though, because the slope b of the least squares line would be close to 0. In other words, a small P-value would probably not occur even though

FIGURE 9.15: A Linear Regression Equation as an Approximation for a Nonlinear Relationship
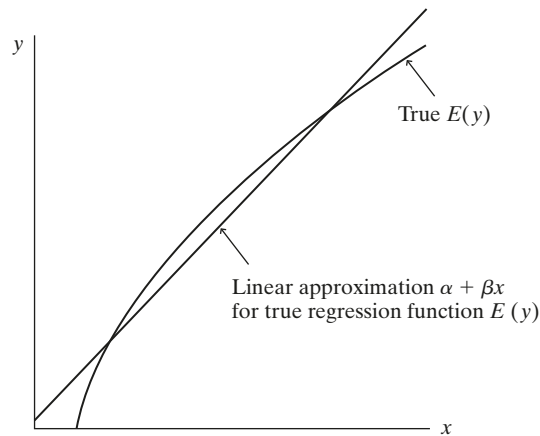
an association exists. In summary, $\beta = 0$ need not correspond to independence if the assumption of a linear regression model is violated. For this reason, you should always construct a scatterplot to check this fundamental assumption.

The least squares line and r and r^2 are valid descriptive statistics no matter what the shape of the conditional distribution of y-values for each x-value. However, the statistical inferences in Section 9.5 also assume that the conditional distributions of y are (1) normal, with (2) identical standard deviation σ for each x-value. These assumptions are also not *exactly* satisfied in practice. For large samples, the normality assumption is relatively unimportant, because an extended Central Limit Theorem implies that sample slopes and correlations have approximately normal sampling distributions. If the assumption about common σ is violated, other estimates may be more efficient than least squares (i.e., having smaller se values), but ordinary inference methods are still approximately valid.

The random sample and straight-line assumptions are very important. If the true relationship deviates greatly from a straight line, for instance, it does not make sense to use a slope or a correlation to describe it. Chapter 14 discusses ways of checking the assumptions and modifying the analysis, if necessary.

INFLUENTIAL OBSERVATIONS

The least squares method has a long history and is the standard way to fit prediction equations to data. A disadvantage of least squares, however, is that individual observations can unduly influence the results. A single observation can have a large effect if it is a *regression outlier*—having x-value relatively large or relatively small and falling quite far from the trend that the rest of the data follow.

Figure 9.16 illustrates this. The figure plots observations for several African and Asian nations on $y = $ crude birth rate (number of births per 1000 population size) and $x = $ number of televisions per 100 people. We added to the figure an observation on these variables for the United States, which is the outlier that is much lower than the other countries in birth rate but much higher on number of televisions. Figure 9.16 shows the prediction equations both without and with the U.S. observation. The prediction equation changes from $\hat{y} = 29.8 - 0.024x$ to $\hat{y} = 31.2 - 0.195x$. Adding only a single point to the data set causes the prediction line to tilt dramatically downward.

When a scatterplot shows a severe regression outlier, you should investigate the reasons for it. An observation may have been incorrectly recorded. If the observation is correct, perhaps that observation is fundamentally different from the others in some way, such as the U.S. observation in Figure 9.16. It may suggest an additional

FIGURE 9.16: Prediction Equations for $y =$ Birth Rate and $x =$ Television Ownership, with and without Observation for the United States
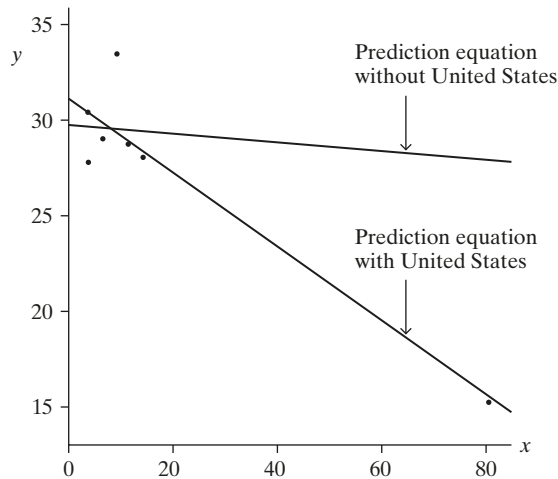
predictor for the model, using methods of Chapter 11. It is often worthwhile to refit the model without one or two extreme regression outliers to see if those observations have a large effect on the fit. We did this following Example 9.4 (page 264) with the D.C. observation for the murder rates. The slope of the prediction equation relating murder rate to poverty rate more than doubled when we included the observation for D.C.

Observations that have a large influence on the model parameter estimates can also have a large impact on the correlation. For instance, for the data in Figure 9.16, the correlation is -0.935 when the United States is included and -0.051 when it is deleted from the data set. One point can make quite a difference, especially when the sample size is small.

FACTORS INFLUENCING THE CORRELATION

Besides being influenced by outliers, the correlation depends on the range of x-values sampled. When a sample has a much narrower range of variation in x than the population, the sample correlation tends to underestimate drastically (in absolute value) the population correlation.

Figure 9.17 shows a scatterplot of 500 points that has a correlation of $r = 0.71$. Suppose, instead, we had only sampled the middle half of the points, roughly between

FIGURE 9.17: The Correlation Is Affected by the Range of x-Values. The correlation decreases from 0.71 to 0.33 using only points with x between 43 and 57.
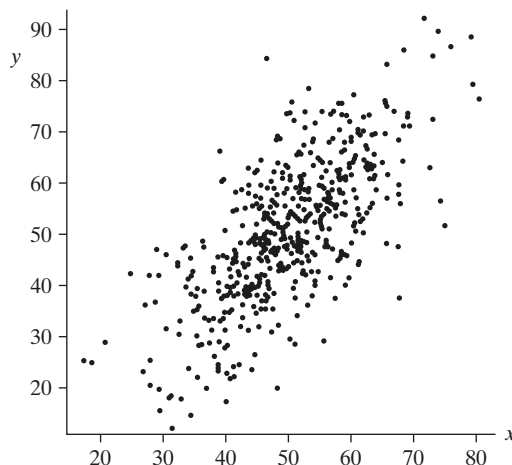
x-values of 43 and 57. Then the correlation equals only $r = 0.33$, considerably lower. For the relation between housing price and size of house, portrayed in Figure 9.14, $r = 0.834$. If we sampled only those sales in which house size is between 1300 and 2000 square feet, which include 44 of the 100 observations, r decreases to 0.254.

The correlation is most appropriate as a summary measure of association when the sample (x, y)-values are a random sample of the population. This way, there is a representative sample of the x variation as well as the y variation.

<table>
<tr><td>**Example**
9.12</td><td>**Does the SAT Predict College GPA?**   Consider the association between $x =$ score on the SAT college entrance exam and $y =$ college GPA at end of second year of college. The strength of the correlation depends on the variability in SAT scores in the sample. If we study the association only for students at Harvard University, the correlation will probably be weak, because the sample SAT scores will be concentrated very narrowly at the upper end of the scale. By contrast, if we could randomly sample from the population of *all* high school students who take the SAT and then place those students in the Harvard environment, students with poor SAT scores would tend to have low GPAs at Harvard. We would then observe a much stronger correlation.   ■</td></tr>
</table>

Other aspects of regression, such as fitting a prediction equation to the data and making inferences about the slope, remain valid when we randomly sample y within a restricted range of x-values. We simply limit our predictions to that range. The slope of the prediction equation is not affected by a restriction in the range of x. For Figure 9.17, for instance, the sample slope equals 0.97 for the full data and 0.96 for the restricted middle set. The correlation makes most sense, however, when both x and y are random, rather than only y.

EXTRAPOLATION IS DANGEROUS

It is dangerous to apply a prediction equation to values of x outside the range of observed values. The relationship might be far from linear outside that range. We may get poor or even absurd predictions by extrapolating beyond the observed range.

To illustrate, the prediction equation $\hat{y} = -0.86 + 0.58x$ in Section 9.2 relating $x =$ poverty rate to $y =$ murder rate was based on observed poverty rates between 8.0 and 26.4. It is not valid to extrapolate much below or above this range. The predicted murder rate for a poverty rate of $x = 0\%$ is $\hat{y} = -0.86$. This is an impossible value for murder rate, which cannot be negative.

Here is another type of inappropriate extrapolation: x being positively correlated with y and y being positively correlated with z does not imply that x is positively correlated with z. For example,[5] in the United States wealthier people tend to reside in wealthier states and wealthier states tend to have a higher percentage favoring the Democratic candidate in presidential elections, yet wealthier people tend to be *less* likely to vote Democratic.

REGRESSION MODEL WITH ERROR TERMS*

Recall that at each fixed value of x, the regression model permits values of y to fluctuate around their mean, $E(y) = \alpha + \beta x$. Any one observation may fall above that

---

[5] See *Red State, Blue State, Rich State, Poor State* by A. Gelman (Princeton University Press, 2008).

mean (i.e., above the regression line) or below that mean (i.e., below the regression line). The standard deviation σ summarizes the typical sizes of the deviations from the mean.

An alternative formulation for the model expresses each observation on y, rather than the mean $E(y)$ of the values, in terms of x. We've seen that the deterministic model $y = \alpha + \beta x$ is unrealistic, because of not allowing variability of y-values. To allow variability, we include a term for the deviation of the observation y from the mean,

$$y = \alpha + \beta x + \varepsilon.$$

The term denoted by ε (the Greek letter epsilon) represents the deviation of y from the mean, $\alpha + \beta x$. Each observation has its own value for ε.

If ε is positive, then $\alpha + \beta x + \varepsilon$ is larger than $\alpha + \beta x$, and the observation falls above the mean. See Figure 9.18. If ε is negative, the observation falls below the mean. When $\varepsilon = 0$, the observation falls exactly at the mean. The mean of the ε-values is 0.

FIGURE 9.18: Positive and Negative ε-Values Correspond to Observations above and below the Mean of the Conditional Distribution
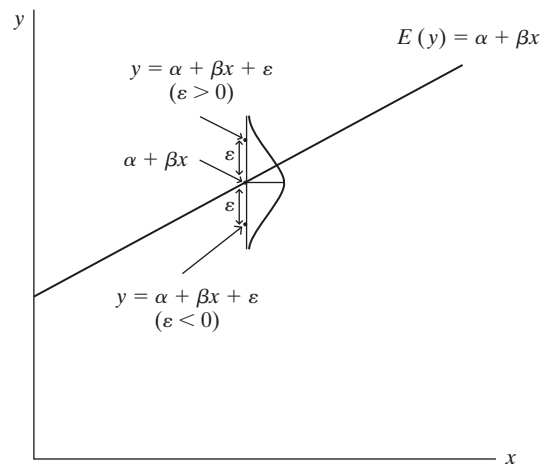
For each x, variability in the y-values corresponds to variability in ε. The ε term is called the ***error term***, since it represents the error that results from using the mean value $(\alpha + \beta x)$ of y at a certain value of x to predict the individual observation.

In practice, we do not know the n values for ε, just like we do not know the parameter values and the true mean $\alpha + \beta x$. For the sample data and their prediction equation, let e be such that

$$y = a + bx + e.$$

That is, $y = \hat{y} + e$, so $e = y - \hat{y}$. Then e is the *residual*, the difference between the observed and predicted values of y, which we *can* observe. Since $y = \alpha + \beta x + \varepsilon$, the residual e estimates ε. We can interpret ε as a ***population residual***. Thus, ε is the difference between the observation y and the mean $\alpha + \beta x$ of all possible observations on y at that value of x. Graphically, ε is the vertical distance between the observed point and the true regression line.

In summary, we can express the regression model either as

$$E(y) = \alpha + \beta x \quad \text{or as} \quad y = \alpha + \beta x + \varepsilon.$$

We use the first equation in later chapters, because it connects better with regression models for response variables assumed to have distributions other than the normal. Models for discrete quantitative variables and models for categorical variables are expressed in terms of their means, not in terms of y itself.

MODELS, REALITY, AND ALTERNATIVE APPROACHES

We emphasize again that the regression model *approximates* the true relationship. No sensible researcher expects a relationship to be exactly linear, with exactly normal conditional distributions at each x and with exactly the same standard deviation of y-values at each x-value. Models merely approximate reality.

If the model seems too simple to be adequate, the scatterplot or other diagnostics may suggest improvement by using other models introduced later in this text. Such models can be fitted, rechecked, and perhaps modified further. Model building is an iterative process. Its goals are to find a realistic model that is adequate for describing the relationship and making predictions but that is still simple enough to interpret easily. Chapters 11–15 extend the basic regression model so that it applies to situations in which the assumptions of this chapter are too simplistic.

9.7 Chapter Summary

Chapters 7–9 have dealt with the detection and description of *association between two variables*. Chapter 7 showed how to compare means or proportions for two groups. When the variables are statistically independent, the population means or proportions are identical for the two groups. Chapter 8 dealt with *association between two categorical variables*. Measures of association such as the difference of proportions, the odds ratio, and gamma describe the strength of association. The chi-squared statistic for nominal data or a z statistic based on sample gamma for ordinal data tests the hypothesis of independence.

This chapter dealt with *association between quantitative variables*. A new element studied here was a regression model to describe the *form* of the relationship between the explanatory variable x and the mean $E(y)$ of the response variable. The major aspects of the analysis are as follows:

- The **linear regression equation** $E(y) = \alpha + \beta x$ describes the *form* of the relationship. This equation is appropriate when a straight line approximates the relationship between x and the mean of y.

- A **scatterplot** views the data and checks whether the relationship is approximately linear. If it is, the **least squares** estimates of the y-intercept α and the slope β provide the prediction equation $\hat{y} = a + bx$ closest to the data, minimizing the sum of squared residuals.

- The **correlation r** and its square describe the *strength* of the linear association. The correlation is a standardized slope, having the same sign as the slope but falling between -1 and $+1$. Its square, r^2, gives the proportional reduction in variability about the prediction equation compared to the variability about $\bar{y}$.

- For inference about the relationship, a t test using the slope or correlation tests the **null hypothesis of independence**, namely, that the population slope and correlation equal 0. A confidence interval estimates the size of the effect.

Table 9.9 summarizes the methods studied in the past three chapters.

Chapter 11 introduces the **multiple regression** model, a generalization that permits *several* explanatory variables in the model. Chapter 12 shows how to include categorical predictors in a regression model. Chapter 13 includes both categorical and quantitative predictors. Chapter 14 introduces models for more complex relationships, such as nonlinear ones. Finally, Chapter 15 presents regression models for

TABLE 9.9: Summary of Tests of Independence and Measures of Association

	Measurement Levels Of Variables		
	Nominal	Ordinal	Interval
Null hypothesis	H_0: independence	H_0: independence	H_0: independence ($\beta = 0$)
Test statistic	$X^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$	$z = \frac{\hat{\gamma}}{se}$	$t = \frac{b}{se} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$, $df = n - 2$
Measure of association	$\hat{\pi}_2 - \hat{\pi}_1$	$\hat{\gamma} = \frac{C-D}{C+D}$	$r = b\left(\frac{s_x}{s_y}\right)$
	Odds ratio		$r^2 = \frac{TSS-SSE}{TSS}$

categorical response variables. Before discussing these multivariate models, how-ever, we introduce in the next chapter some new concepts that help us to understand and interpret multivariate relationships.

Exercises

Practicing the Basics

9.1. For the following variables in a regression analysis, which variable more naturally plays the role of x (explanatory variable) and which plays the role of y (response variable)?

(a) College grade point average (GPA) and high school GPA.

(b) Number of children and mother's education level.

(c) Annual income and number of years of education.

(d) Annual income of homeowner and assessed value of home.

9.2. Sketch plots of the following prediction equations, for values of x between 0 and 10:
(a) $\hat{y} = 7 + x$, **(b)** $\hat{y} = 7 - x$, **(c)** $\hat{y} = 7$, **(d)** $\hat{y} = x$.

9.3. Anthropologists often try to reconstruct information using partial human remains at burial sites. For instance, after finding a femur (thighbone), they may want to predict how tall an individual was. An equation they use to do this is $\hat{y} = 61.4 + 2.4x$, where $\hat{y}$ is the predicted height and x is the length of the femur, both in centimeters.

(a) Identify the y-intercept and slope of the equation. Interpret the slope.

(b) A femur found at a particular site has length of 50 cm. What is the predicted height of the person who had that femur?

9.4. The OECD (Organization for Economic Cooperation and Development) consists of advanced, industrialized countries. For these nations, a recent prediction equation[6] relating y = child poverty rate to x = social

expenditure as a percentage of gross domestic product is $\hat{y} = 22 - 1.3x$. The y-values ranged from 2.8% (Finland) to 21.9% (the United States). The x-values ranged from 2% (the United States) to 16% (Denmark).

(a) Interpret the y-intercept and the slope.

(b) Find the predicted poverty rates for the United States and for Denmark.

(c) The correlation is -0.79. Interpret.

9.5. Look at Figure 2 at http://ajph.aphapublica tions.org/doi/pdf/10.2105/AJPH.93.4.652, a scatterplot for U.S. states with correlation 0.53 between x = child poverty rate and y = child mortality rate. Approximate the y-intercept and slope of the prediction equation shown there.

9.6. The Firearms data file at the text website shows U.S. statewide data on x = percentage of people who report owning a gun and y = firearm death rate (annual number of deaths per 100,000 population), from www.cdc.gov.

(a) Find the prediction equation, and interpret.

(b) The correlation is 0.70. Identify an outlier, and show that $r = 0.78$ when you remove this state from the data file.

9.7. Access the UN data file (shown in Table 3.9) at the text website from 42 countries. Let y = carbon dioxide emissions (metric tons per capita) and x = gross domestic product (per capita GDP, in dollars).

(a) Find the prediction equation, and interpret the coefficients.

---

[6] *Source*: Figure 8H at www.stateofworkingamerica.org.

(b) For the United States, $x = 50.9$ and $y = 17.0$. Find the predicted CO2 value. Find the residual, and interpret.

9.8. For the 2014 GSS, a prediction equation relates $y =$ highest year of school completed to $x =$ father's highest year of school completed.

(a) Which equation is more realistic for the result: $\hat{y} = 9.57 + 0.35x$, or $\hat{y} = 0.35 + 9.57x$? Why?

(b) Suppose the prediction equation had been $\hat{y} = x$. Identify the y-intercept and slope, and interpret the slope.

9.9. Access the data file `Crime2` shown in Table 9.1. Let $y =$ violent crime rate and $x =$ poverty rate.

(a) Using software, show that the prediction equation is $\hat{y} = 209.9 + 25.5x$. Interpret the y-intercept and the slope.

(b) Find the predicted violent crime rate and the residual for Massachusetts, which had $x = 10.7$ and $y = 805$. Interpret.

(c) Two states differ by 10.0 in their poverty rates. Find the difference in their predicted violent crime rates.

(d) From the prediction equation, can you tell the sign of the correlation between these variables? How?

9.10. In the 2000 Presidential election in the United States, the Democratic candidate was Al Gore and the Republican candidate was George W. Bush. In Palm Beach County, Florida, initial election returns reported 3407 votes for the Reform party candidate, Pat Buchanan. Some political analysts thought that most of these votes may have actually been intended for Gore (whose name was next to Buchanan's on the ballot) but wrongly cast for Buchanan because of the design of the "butterfly ballot" used in that county, which some voters found confusing. For the 67 counties in Florida, Figure 9.19 is a scatterplot

of the county wide vote for the Reform party candidates in 2000 (Buchanan) and in 1996 (Perot).

(a) The top point is for Palm Beach county. What does it suggest?

(b) The prediction equation fitted to all but the observation for Palm Beach county is $\hat{y} = 45.7 + 0.02414x$. In Palm Beach county, $x = 30,739$. Find the predicted Buchanan vote and the residual, and interpret.

(c) Why is the top point, but not each of the two rightmost points, considered a regression outlier? (*Note*: Statistical analyses predicted that fewer than 900 of the 3407 votes were truly intended for Buchanan. Bush won the state by 537 votes and, with it, the Electoral College and the election. Other factors that played a role were 110,000 disqualified "overvote" ballots in which people mistakenly voted for more than one presidential candidate—with Gore marked on 84,197 ballots and Bush on 37,731—often because of confusion from names being listed on more than one page of the ballot, and 61,000 "undervotes" caused by factors such as "hanging chads" from manual punch-card machines.)

9.11. Figure 9.20 is a scatterplot relating $y =$ percentage of people using cell phones and $x =$ per capita gross domestic product (GDP) for some nations listed in the *Human Development Report*.

(a) Give the approximate x- and y-coordinates for the nation that has the highest (i) cell phone use, (ii) GDP.

(b) The least squares prediction equation is $\hat{y} = -0.13 + 2.62x$. For one nation, $x = 34.3$ and $y = 45.1$. Find the predicted cell-phone use and the residual. Interpret the residual.

(c) Is the correlation positive, or negative? Explain what it means for the correlation to have this sign.
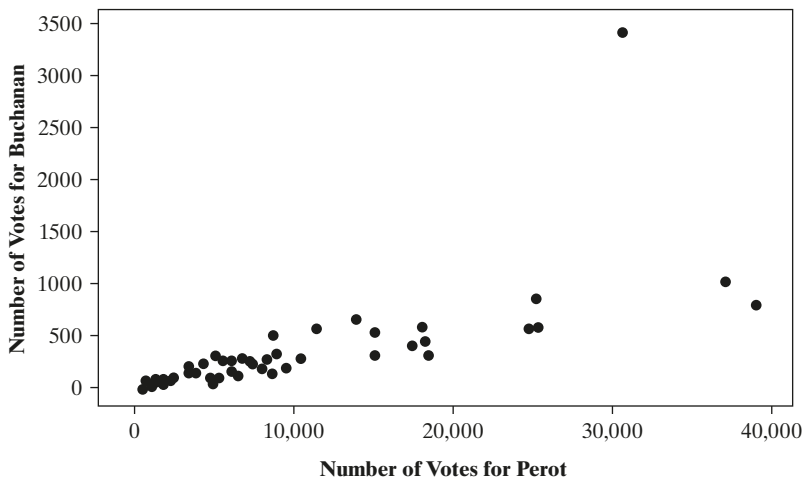
FIGURE 9.19: Scatterplot of Florida Countywide Vote for Reform Party Candidates Pat Buchanan in 2000 and Ross Perot in 1996
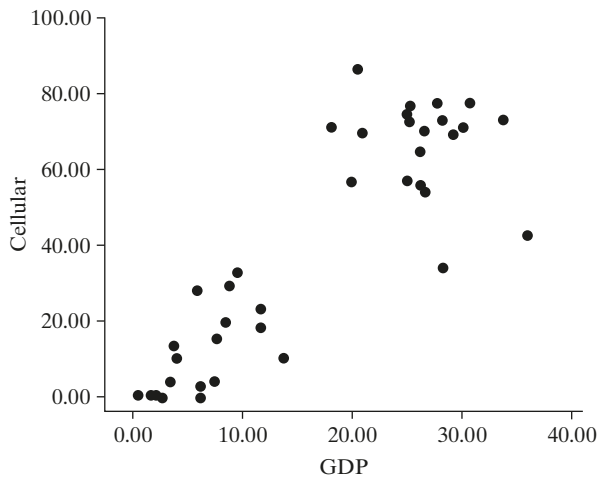
FIGURE 9.20: Scatterplot of Percentage Using Cell Phones and Per Capita GNP

9.12. For nations listed in the *Human Development Report*, the correlation with percentage of people using the Internet is 0.888 for per capita gross domestic product (GDP, a summary description of a nation's wealth), 0.818 for percentage using cell phones, 0.669 for literacy rate, −0.551 for fertility rate (the mean number of children per adult woman), and 0.680 for per capita emissions of carbon dioxide.

(a) Explain how to interpret the sign of the correlation between Internet use and (i) GDP, (ii) fertility rate.

(b) Which variable has the (i) strongest, (ii) weakest linear association with Internet use?

9.13. A report summarizing the results of a study on the relationship between a verbal aptitude test x and a mathematics aptitude test y states that $\bar{x} = 480$, $\bar{y} = 500$, $s_x = 80$, $s_y = 120$, and $r = 0.60$. Using the formulas for the correlation and for the least squares estimates, find the prediction equation.

9.14. Table 9.15 in Exercise 9.39 shows countywide data for several variables in Florida. For those counties, Table 9.10 shows part of the output for the regression analysis relating y = median income (in thousands of dollars) to x = percentage of residents with at least a high school education.

(a) Report the prediction equation, and interpret the slope.

(b) County A has 10% more of its residents than county B with at least a high school education. Find their difference in predicted median incomes.

(c) Find the correlation. Interpret using (i) the sign, (ii) the magnitude, (iii) the standardized slope.

(d) Find r^2. Explain how to interpret it.

TABLE 9.10

Variable	Mean	Std Dev	Variable	Parameter Estimate
INCOME	24.51	4.69	(Constant)	-4.63
EDUC	69.49	8.86	EDUC	0.42

9.15. The Internet site www.pearsonglobaleditions .com/Agresti has an *Explore Linear Regression* applet. To show the impact of an outlier, use the *Draw Own* option to put 10 points on the scatterplot that have a correlation close to +1, and then add a single point that changes the correlation to a negative value. Explain why this single observation has so much influence. Download the scatterplot and print it as part of your solution.

9.16. For the student survey data set described in Exercise 1.11, the sample correlation between y = political ideology (scored 1 to 7, with higher values representing more conservatism) and x = number of times a week reading a newspaper is $r = -0.066$.

(a) Would you conclude that the sample association is strong, or weak?

(b) Interpret the square of the correlation.

(c) When y is predicted using x = religiosity (how often attend religious services, scored 0, 1, 2, 3), the sample correlation is $r = 0.580$. Which of these two explanatory variables seems to have a stronger linear relationship with y? Explain.

9.17. For the study in Example 9.6 (page 270) of y = high school GPA and x = weekly number of hours viewing television, $\hat{y} = 3.44 - 0.03x$.

(a) The study reported that r-squared = 0.237. Interpret.

(b) Report and interpret the correlation.

(c) Suppose you found the correlation only for those students having TV watching of no more than three hours per week. Would you expect the correlation to be stronger, or weaker, than for all students? Why?

9.18. For students who take Statistics 101 at Lake Wobegon College in Minnesota, both x = midterm exam score and y = final exam score have mean = 75 and standard deviation = 10.

(a) The prediction equation is $\hat{y} = 30 + 0.60x$. Find the predicted final exam score for a student who has (i) midterm score = 100, (ii) midterm score = 50. Note that the predicted final exam score regresses from the midterm score toward the mean.

(b) Show that the correlation equals 0.60.

(c) If instead, $\hat{y} = x$, show that $r = 1.0$.

(d) If instead, $\hat{y} = 75$ (i.e., slope = 0), show that $r = 0.0$.

9.19. The prediction equation relating x = years of education and y = annual income (in dollars) is $\hat{y} = -20,000 + 4000x$, and the correlation equals 0.50. The standard deviations were 2.0 for x and 16,000 for y.

(a) Show how to find the correlation from the slope.

(b) Results were translated to units of euros, at a time when the exchange rate was $1.25 per euro. Find the prediction equation and the correlation.

9.20. Access the Houses data file (shown partly in Table 9.5) at the text website. For the response variable *taxes* and explanatory variable *size*, using software,

(a) Graphically portray the association, and describe it.

(b) Find and interpret the prediction equation.

(c) Find and interpret the correlation and r^2.

9.21. For 2014 GSS data, the correlation matrix for subject's education (EDUC), mother's education (MAEDUC), and father's education (PAEDUC) is

	EDUC	PAEDUC	MAEDUC
EDUC	1.00	.46	.45
PAEDUC	.46	1.00	.69
MAEDUC	.45	.69	1.00

Interpret this matrix, identifying the pair of variables with the strongest association and giving the implication of the sign of each correlation.

9.22. In the UN *Human Development Report*, one variable measured was x = percentage of adults who use contraceptive methods. Table 9.11 shows part of a regression analysis using y = fertility (mean number of children per adult woman), for 22 nations listed in that report. For those nations, x had a mean of 60.0 and standard deviation of 20.6.

(a) State a research question that could be addressed with this output.

(b) Report the prediction equation and find the predicted fertility when (i) $x = 0$, (ii) $x = 100$. Show how the difference between these can be obtained using the slope.

(c) Find and interpret r and r^2.

(d) What do your analyses say about the question in (a)?

TABLE 9.11

Predictor	Coef.	Std. Error	t	P>|t|
Constant	6.6633	0.4771	13.97	0.000
CONTRA	-0.064843	0.007539	-8.60	0.000

Source	Sum of Squares	df
Regression	37.505	1
Residual	10.138	20
Total	47.644	21

9.23. For data on several nations, we want to describe whether the percentage of people using the Internet is more strongly associated with per capita GDP or with the fertility rate.

(a) Can we compare the slopes when GDP and fertility each predict Internet use in separate regression equations? Why or why not?

(b) Let x = GDP (thousands of dollars per capita). For recent data on 39 nations from the UN, for y = percentage using cell phones, $\hat{y} = -0.13 + 2.62x$, whereas for y percentage using the Internet, $\hat{y} = -3.61 + 1.55x$. Why does it make sense to compare these slopes, thus concluding that a one-unit increase in GDP has a slightly greater impact on the percentage using cell phones than on the percentage using the Internet?

9.24. For the Houses data file (shown partly in Table 9.5), Table 9.12 shows a regression analysis relating selling price to number of bedrooms.

(a) Report the prediction equation, and interpret the slope.

(b) Report r^2, and interpret its value.

(c) Report the correlation and its confidence interval, and interpret.

(d) Interpret the value labeled *Root MSE*.

TABLE 9.12

	Root MSE 93547	R-Square 0.1552		
		correlation and 95% limits		
price	bedrooms	0.394	0.214	0.548

Variable	Parameter Estimate	Standard Error	t	Sig.
Intercept	-28412	44303	-0.64	0.5228
bedrooms	61248	14435	4.24	<.0001

9.25. Refer to Table 9.1 and the `Crime2` data file at the text website. For all 51 observations, use software to analyze the relationship between $y =$ murder rate and $x =$ poverty rate.

(a) Construct a scatterplot. Does there seem to be a positive, or a negative, relationship?

(b) Report the prediction equation, and find the predicted murder rate and the residual for D.C. Interpret.

(c) Based on the scatterplot, would you regard D.C. as a regression outlier? Refit the model without it, and note the effect on the slope.

9.26. Refer to the `Florida` data file at the text website, shown partly in Table 9.15 for Exercise 9.39, giving countywide data for several variables in Florida. For those data, use software to analyze $y =$ crime rate and $x =$ percentage living in an urban environment.

(a) Construct a box plot for y. Interpret.

(b) Find the prediction equation. Interpret the y-intercept and slope.

(c) Using the slope, find the difference in predicted crime rates between counties that are 100% urban and counties that are 0% urban. Interpret.

(d) Report and interpret the correlation and r^2.

9.27. Refer to Table 3.9 on page 65. This exercise uses $y =$ fertility rate and $x =$ gender inequality index. Table 9.13 shows part of an SPSS output for a regression analysis.

(a) State a research question that could be addressed with this printout.

(b) Report the prediction equation, and interpret.

(c) Report r and r^2, and interpret.

(d) What do your analyses suggest about the question posed in (a)?

TABLE 9.13 Fertility Rate Regressed on Gender Inequality Index

R .598	R Square	0.357		
	B	Std. Error	t	Sig.
(Constant)	1.378	0.172	8.027	0.000
GII	2.734	0.580	4.717	0.000

9.28. Refer to the previous exercise. Now let the human development index (HDI) be the explanatory variable for predicting fertility. Using software with the UN data file at the text website,

(a) Construct a scatterplot. Do any observations stand out as potential regression outliers?

(b) Fit the model, and interpret the parameter estimates and correlation.

(c) Redo the analyses without the observation that may be a regression outlier. Compare results to (b).

9.29. For 2271 observations from the 2014 GSS on $y =$ number of years of education (EDUC) and $x =$ number of years of mother's education (MAEDUC), $\hat{y} = 9.86 + 0.345x$, with $se = 0.0147$ for the slope.

(a) Test the null hypothesis that these variables are independent, and interpret.

(b) Find a 95% confidence interval for the population slope. Interpret.

(c) The correlation equals 0.441. Explain "regression toward the mean" in terms of these variables.

9.30. A study was conducted using 49 Catholic female undergraduates at Texas A&M University. The variables measured refer to the parents of these students. The response variable is the number of children that the parents have. One of the explanatory variables is the mother's educational level, measured as the number of years of formal education. For these data, $\bar{x} = 9.88, s_x = 3.77, \bar{y} = 3.35, s_y = 2.19$, the prediction equation is $\hat{y} = 5.40 - 0.207x$, the standard error of the slope estimate is 0.079, and SSE = 201.95.

(a) Find the correlation and interpret its value.

(b) Test the null hypothesis that mean number of children is independent of mother's educational level, and report and interpret the P-value.

(c) Sketch a potential scatterplot such that the analyses you conducted in (a) and (b) would be inappropriate.

9.31. Is political ideology associated with income? When GSS data for 1478 cases in 2014 were used to regress $y =$ political views (POLVIEWS, using scores 1–7 with 1 = extremely liberal and 7 = extremely conservative) on $x =$ respondent's income (RINCOME, using scores 1–12 for the 12 income categories), we get the results shown in Table 9.14.

(a) Construct a confidence interval to make an inference about the size of the slope effect of x on y. In practical terms, would you characterize this effect as major, or minor? Why?

(b) SPSS and Stata report the correlation under the misleading heading of "Beta." How would you interpret this value?

TABLE 9.14 Political Views Regressed on Income

	B	Std. Error	Beta	t	Sig.
R Square	0.0032				
Constant	3.6950	0.1395		26.48	0.000
RINCOME	0.0282	0.0129	0.0569	2.19	0.029

9.32. Refer to the previous exercise. When we regress political ideology in 2014 on x = number of hours spent in the home on religious activity in the past month (RELHRS1), we obtain

```
           B    Std. Error  Beta    t      Sig.
Constant 4.0115   0.0422           95.10  0.0000
RELHRS1  0.0064   0.0020    0.087   3.20  0.0015
```

(a) Report and interpret the *P*-value for testing the hypothesis that these variables are independent.

(b) Use these results to illustrate that statistical significance does not imply practical significance.

9.33. For the OECD data file for OECD nations in Table 3.13 on page 70, use software to construct a scatterplot relating x = carbon dioxide emissions and y = prison population.

(a) Based on this plot, identify a point that has a large influence in determining the correlation. Show how the correlation changes if you remove this observation from the data set.

(b) Suppose you constructed this plot using UN data for *all* nations, rather than only the highly economically advanced nations that form the OECD. Would you expect the correlation to be weaker, about the same size, or stronger? Why?

Concepts and Applications

9.34. For the Students data file (Exercise 1.11 on page 21), conduct regression analyses relating (i) y = political ideology and x = religiosity, (ii) y = high school GPA and x = hours of TV watching. Prepare a report

(a) Using graphical ways of portraying the individual variables and their relationship.

(b) Interpreting descriptive statistics for summarizing the individual variables and their relationship.

(c) Summarizing and interpreting results of inferential analyses.

9.35. Refer to the data file you created in Exercise 1.12. For variables chosen by your instructor, pose a research question and conduct a regression and correlation analysis. Report both descriptive and inferential statistical analyses, interpreting and summarizing your findings.

9.36. Pose a research question about job satisfaction and educational attainment. Using the most recent GSS data on SATJOB and EDUC with the multiple regression option at sda.berkeley.edu/GSS, with scores (1, 2, 3, 4) for (very satisfied, ..., very dissatisfied), conduct a descriptive and inferential analysis to address this question. Prepare a one-page report summarizing your analysis.

9.37. Refer to the UN data file shown in Table 3.9 on page 65. Pose a research question relating to the association between fertility and the gender inequality index. Using software, analyze data in that file to address this question, and summarize your analyses.

9.38. The Guns_suicide data file at the text website shows statewide data[7] on y = suicides (per 100,000 people) and x = percentage of residents who own a firearm. Conduct a regression and correlation analysis to analyze the association.

9.39. Table 9.15 shows a small excerpt of the Florida data file at the text website. That file has data from all 67 Florida counties on crime rate (number of crimes per 1000 residents), median income (in thousands of dollars), percentage of residents with at least a high school education (of those aged at least 25), and the percentage of the county's residents living in an urban environment. Using crime rate as the response variable and percentage urban as the predictor, analyze these data. In your report, provide interpretations of all the analyses.

9.40. Refer to Table 9.1 (page 260), available in the Crime2 data file at the text website. Pose a research question about the relationship between the murder rate and the percentage of single-parent families. Using software, conduct analyses to address this question. Write a report showing your analyses and providing interpretations.

TABLE 9.15

County	Crime Rate	Median Income	High School	Percentage Urban	County	Crime Rate	Median Income	High School	Percentage Urban
ALACHUA	104	22.1	82.7	73.2	LAKE	42	23.4	70.6	43.2
BAKER	20	25.8	64.1	21.5	LEE	59	28.4	76.9	86.1

Source: Dr. Larry Winner, University of Florida. Complete data are in Florida data file at the text website.

[7] *Source*: *The Economist*, January 31, 2015.

9.41. Refer to the UN data for several nations shown in Table 3.9 (page 65) and given at the text website. Using software, obtain the correlation matrix. Which pairs of variables are highly correlated? Describe the nature of those correlations, and inspect scatterplots to analyze whether any of them may be misleading. Explain how your software handled the missing values.

9.42. A study,[8] after pointing out that diets high in fats and sugars (bad for our health) are more affordable than diets high in fruit and vegetables (good for our health), reported, "Every extra 100 g of fats and sweets eaten decreased diet costs by 0.05 to 0.4 Euros, whereas every extra 100 g of fruit and vegetables eaten increased diet costs by 0.18 to 0.29 Euros." Indicate the parameters to which these interpretations refer and the statistical inference that was performed to give this summary.

9.43. The headline of an article in the *Gainesville Sun* newspaper stated, "Height can yield a taller paycheck." It described an analysis of four large studies in the United States and Britain by a University of Florida professor on subjects' height and salaries. The article reported that for each gender "an inch is worth about $789 a year in salary. So, a person who is 6 feet tall will earn about $5,523 more a year than a person who is 5 foot 5."

(a) For the interpretation in quotes, identify the response variable and explanatory variable, and state the slope of the prediction equation, when height is measured in inches and salary in dollars.

(b) Explain how the value $5,523 relates to the slope.

9.44. A recent survey reported that the mean total earnings that a worker in Country X can expect to earn between ages 25 and 64 is $1 million for those with only a high school education and $3 million for those with a doctoral degree.

(a) Assuming eight years for a doctoral degree and a straight-line regression of $y =$ total earnings on $x =$ number years of education, what is the slope?

(b) If y instead measures earnings per year (rather than for 40 years), then what is the slope?

9.45. Explain why conditional variability can be much less than marginal variability, using the relationship between $y =$ weight and $x =$ age for a sample of boys of ages 2–12, for which perhaps $\sigma_y = 30$ but the conditional $\sigma = 10$.

9.46. For counties in a particular state, crime rate (number of crimes per thousand residents) in the past two years varies around a mean $= 50$ with standard deviation $= 20$. The crime rate last year has correlation 0.50 with the crime rate this year. Last year, the crime rate was 100 for the

county having the highest rate. Predict the crime rate in that county this year. Explain your reasoning.

9.47. Annual income, in dollars, is an explanatory variable in a regression analysis. For a British version of the report on the analysis, all responses are converted to British pounds sterling (1 pound equals about 1.33 dollars, as of 2016).

(a) How, if at all, does the slope of the prediction equation change?

(b) How, if at all, does the correlation change?

9.48. A magazine article[9] reported results of a study suggesting that "each extra year of schooling makes someone 10% less likely to describe himself as religious." Explain how the 10% could be a slope estimate from a regression analysis, but the prediction equation obtained would apply over a restricted range of education levels.

9.49. State the assumptions in fitting and making inferences with the model $E(y) = \alpha + \beta x$. Which assumptions are most critical? In view of these assumptions, indicate why the model might not be adequate for

(a) $x =$ income, $y =$ charitable contributions within the previous year. (*Hint*: Would poor people show as much variation as wealthy people?)

(b) $x =$ age, $y =$ annual medical expenses.

9.50. For a total of 100 salesmen in a firm, the 10 who perform poorest in the first quarter are enrolled in a special training program. The overall sales mean is 70 in both the first and the last quarter, but the mean for the specially trained employees increases from 50 to 60. Use the concept of regression toward the mean to explain why this is not sufficient evidence to imply that the training program was successful.

9.51. A study by the Readership Institute[10] at Northwestern University used survey data to analyze how reader behavior was influenced by the Iraq war. The response variable was a Reader Behavior Score (RBS), a combined measure summarizing newspaper use frequency, time spent with the newspaper, and how much was read. Comparing RBS scores before the war and during the war, the study noted that there was a significant increase in reading by light readers (mean RBS changing from 2.05 to 2.32, $P < 0.001$) but a significant decrease in reading by heavy readers (mean RBS changing from 5.87 to 5.66, $P < 0.001$). Would you conclude that the Iraq war caused a change in reader behavior, or could there be some other explanation?

9.52. Refer to Exercise 9.39. For these counties, the correlation between high school education rate and median income equals 0.79. Suppose we also have data at the

---

[8] E. Frazao and E. Golan, *Evidence-Based Healthcare and Public Health*, vol. 9 (2005), pp. 104–107.

[9] *The Economist*, October 11, 2014.

[10] www.readership.org/consumers/data/FINAL_war_study.pdf.

individual level as well as aggregated for a county. Sketch a scatterplot to show that at the individual level, the correlation could be much weaker. (*Hint*: Show that lots of variability could exist for individuals, yet the summary values for counties could fall close to a straight line. Thus, it is misleading to extend results from the aggregate level to individuals. Making predictions about individuals based on the behavior of aggregate groups is known as the *ecological fallacy*.)

9.53. For which student body do you think the correlation between high school GPA and college GPA would be higher: Yale University or the University of Bridgeport, Connecticut? Explain why.

9.54. Explain why the correlation between x = number of years of education and y = annual income is likely to be smaller if we use a random sample of adults who have a college degree than if we use a random sample of all adults.

9.55. Explain carefully the interpretations of the standard deviations **(a)** s_y, **(b)** s_x, **(c)** s = square root of residual MS, **(d)** *se* for b.

9.56. Estimating a single mean μ corresponds to estimating the parameter in the simple model, $E(y) = \mu$. Use this fact to explain why the estimate s_y of the standard deviation of the marginal distribution has $df = n - 1$.

9.57. The statistician George Box, who had an illustrious academic career at the University of Wisconsin, is often quoted as saying, "All models are wrong, but some models are useful." Why do you think that, in practice, **(a)** all models are wrong, **(b)** some models are *not* useful?

9.58. The variables y = annual income (thousands of dollars), x_1 = number of years of education, and x_2 = number of years of experience in job are measured for all the employees having city-funded jobs, in Knoxville, Tennessee. The following prediction equations and correlations apply:

  i. $\hat{y} = 10 + 1.0x_1$, $r = 0.30$.
  ii. $\hat{y} = 14 + 0.4x_2$, $r = 0.60$.

The correlation is -0.40 between x_1 and x_2. Which of the following statements are true?

(a) The strongest sample association is between y and x_2.

(b) The weakest sample association is between x_1 and x_2.

(c) The prediction equation using x_2 to predict x_1 has negative slope.

(d) A standard deviation increase in education corresponds to a predicted increase of 0.3 standard deviations in income.

(e) There is a 30% reduction in error in using education, instead of $\bar{y}$, to predict income.

(f) Each additional year on the job corresponds to a $400 increase in predicted income.

(g) When x_1 is the predictor of y, the sum of squared residuals (SSE) is larger than when x_2 is the predictor of y.

(h) The predicted mean income for employees having 20 years of experience is $4000 higher than the predicted mean income for employees having 10 years of experience.

(i) If $s = 8$ for the model using x_1 to predict y, then it is not unusual to observe an income of $70,000 for an employee who has 10 years of education.

(j) It is possible that $s_y = 12.0$ and $s_{x_1} = 3.6$.

(k) It is possible that $\bar{y} = 20$ and $\bar{x}_1 = 13$.

Select the best response(s) in Exercises 9.59–9.61. (More than one response may be correct.)

9.59. One can interpret $r = 0.60$ as follows:

(a) A 60% reduction in error occurs in using x to predict y.

(b) A 9% reduction in error occurs in using x to predict y compared to using $\bar{y}$ to predict y.

(c) 9% of the time $\hat{y} = y$.

(d) y changes 0.60 units for every one-unit increase in x.

(e) When x predicts y, the average residual is 0.6.

(f) x changes exactly 0.60 standard deviations when y changes one standard deviation.

9.60. The correlation is inappropriate as a measure of association between two quantitative variables

(a) When different people measure the variables using different units.

(b) When the relationship is highly nonlinear.

(c) When the data points fall exactly on a straight line.

(d) When the slope of the prediction equation is 0 using nearly all the data, but a couple of outliers are extremely high on y at the high end of the x scale.

(e) When y tends to decrease as x increases.

(f) When we have data for the entire population rather than a sample.

(g) When the sample has a much narrower range of x-values than does the population.

9.61. The slope of the least squares prediction equation and the correlation are similar in the sense that

(a) They do not depend on the units of measurement.

(b) They both must fall between -1 and $+1$.

(c) They both have the same sign.

(d) They both equal 1 when there is the strongest association.

(e) Their squares both have proportional reduction in error interpretations.

(f) They have the same t statistic value for testing H_0: independence.

(g) They both can be strongly affected by severe outliers.

9.62.* A study by the National Highway Traffic Safety Administration estimated that 73% of people wear seat belts, that failure to wear seat belts led to 9200 deaths in

the previous year, and that that value would decrease by 270 for every 1 percentage point gain in seat belt usage. Let $\hat{y}$ = predicted number of deaths in a year and x = percentage of people who wear seat belts. Find the prediction equation that yields these results.

9.63.* Observations on both x and y are standardized, having estimated means of 0 and standard deviations of 1 (see Section 4.3). Show that the prediction equation has the form $\hat{y} = rx$, where r is the sample correlation between x and y. That is, for the standardized variables, the y-intercept equals 0 and the slope is the same as the correlation.

9.64.* A confidence interval for a population correlation ρ requires a transformation of r, $T = (1/2)\log_e[(1 + r)/(1 - r)]$, for which the sampling distribution is approximately normal, with standard error $1/\sqrt{n-3}$. Once we get the endpoints of the interval for the population value of T, we substitute each endpoint in the inverse transformation $\rho = (e^{2T} - 1)/(e^{2T} + 1)$ to get the endpoints of the confidence interval for ρ. For $r = 0.8338$ for the data on house selling price and size of house (Table 9.5), show that $T = 1.20$ with standard error 0.1015, a 95% confidence interval for population T is (1.00, 1.40), and the corresponding confidence interval for ρ is (0.76, 0.89). (Unless $r = 0$, the confidence interval for ρ is not symmetric about r, because of the nonsymmetry of its sampling distribution.)

9.65.* Refer to the previous exercise. Let ρ_1 and ρ_2 denote the population correlation values between two variables for two separate populations. Let r_1 and r_2 denote sample values for independent random samples from the populations. To test H_0: $\rho_1 = \rho_2$, the test statistic is

$$z = \frac{T_2 - T_1}{s_{T_2 - T_1}} \quad \text{with} \quad s_{T_2 - T_1} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}},$$

where T_1 and T_2 are the transformed values of r_1 and r_2. If H_0 is true, this test statistic has approximately the standard normal distribution. In Table 9.5, the correlation between housing price and size of home is $r_1 = 0.96$ for the 11 new homes and $r_2 = 0.76$ for the 89 older homes. Find the P-value for testing H_0: $\rho_1 = \rho_2$ against H_a: $\rho_1 \neq \rho_2$. Interpret.

9.66.* Refer to the formula $a = \bar{y} - b\bar{x}$ for the y-intercept.
(a) Show that substituting $x = \bar{x}$ into the prediction equation $\hat{y} = a + bx$ yields the predicted y-value of $\hat{y} = \bar{y}$. Show that this means that the least squares prediction equation passes through the point with coordinates $(\bar{x}, \bar{y})$, the center of gravity of the data.
(b) Show that an alternative way of expressing the regression model is as $(\hat{y} - \bar{y}) = b(x - \bar{x})$.
(c) Let y = final exam score and x = midterm exam score. Suppose the correlation is 0.70 and the standard deviation is the same for each set of scores. Show that

$(\hat{y} - \bar{y}) = 0.70(x - \bar{x})$; that is, the predicted difference between your final exam grade and the class mean is 70% of the difference between your midterm exam score and the class mean, so your score is predicted to regress toward the mean.

9.67.* The formula for the correlation can be expressed as

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\left[\sum(x - \bar{x})^2\right]\left[\sum(y - \bar{y})^2\right]}} = \frac{1}{n-1}\sum\left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right).$$

(a) Using the first formula, explain why the correlation has the same value when x predicts y as when y predicts x.
(b) By the second formula, the correlation is approximately the average product of the z-score for x times the z-score for y. Use this to explain why the correlation does not depend on the units of measurement. (*Note*: For the population, page 92 showed an analogous formula.)

9.68.* From the formulas for b (page 264) and r (page 272), show that $r = bs_x/s_y$.

9.69.* Suppose that the linear regression model $E(y) = \alpha + \beta x$ with normality and constant standard deviation σ is truly appropriate. Then, the interval of numbers

$$\hat{y} \pm t_{.025}s\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

predicts where a new observation on y will fall at that value of x. This interval, which for large n is roughly $\hat{y} \pm 2s$, is a 95% **prediction interval** for y. To make an inference about the *mean* of y (rather than a single value of y) at that value of x, we use the **confidence interval**

$$\hat{y} \pm t_{.025}s\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x - \bar{x})^2}}.$$

For large n, near $\bar{x}$ this is roughly $\hat{y} \pm 2s/\sqrt{n}$. The t-value in these intervals is based on $df = n - 2$. Most software has options for calculating these formulas. Refer to the `Houses` data file at the text website.

(a) Using software, find a 95% prediction interval for selling price at house size $x = 2000$.
(b) Using software, find a 95% confidence interval for the mean selling price at house size $x = 2000$.
(c) Explain intuitively why a prediction interval for a single observation is much wider than a confidence interval for the mean.
(d) Explain how prediction intervals would likely be in error if, in fact, (i) the variability in housing prices tends to increase as house size increases, (ii) the response variable is highly discrete, such as y = number of children in Exercise 9.30.

Chapter 10

INTRODUCTION TO MULTIVARIATE RELATIONSHIPS

Chapters 7–9 introduced methods for analyzing the association between two variables. In most social science research, these analyses are but the first step. Subsequent steps use *multivariate* methods to include other variables in the analysis that might influence that association.

For instance, Examples 8.1 and 8.3 showed that political party identification in the United States is associated with gender, with men somewhat more likely than women to be Republicans. To analyze why this is so, we could analyze whether differences between men and women in political ideology (measured on a conservative–liberal scale) could explain the association. For example, perhaps men tend to be more conservative than women, and being conservative tends to be associated with being Republican. If we compare men to women just for those classified as liberal in political ideology, and then again just for those classified as conservative, is it still true that men are more likely than women to be Republicans? Or, could the difference between men and women on political party ID be explained by some other factor, such as income or religion?

Several types of research questions require adding variables to the analysis. These questions often involve notions of *causal* connections among the variables. This chapter discusses causation and outlines methods for testing causal hypotheses. We present the notion of *statistical control*, a fundamental tool for studying how an association changes or possibly even disappears after we remove the influence of other variables. We also show the types of multivariate relationships that statistical control can reveal.

10.1 Association and Causality

Causality is central to the scientific endeavor. Most people are familiar with this concept, at least in an informal sense. We know, for instance, that being exposed to a virus can cause the flu and that smoking can cause lung cancer. But how can we judge whether there is a causal relationship between two social science variables? For instance, what causes juvenile delinquency? Being poor? Coming from a single-parent home? A lack of moral and religious training? Genetic factors? A combination of these and other factors? We now look at some guidelines that help us assess a hypothesis of the form "*x* causes *y*."

Causal relationships usually have an asymmetry, with one variable having an influence on the other, but not vice versa. An arrow drawn between two variables *x* and *y*, pointing to the response variable, denotes a causal association between the variables. Thus,

$$x \rightarrow y$$

specifies that *x* is an explanatory variable having a causal influence on *y*. For example, suppose we suspect that being a Boy Scout has a causal effect on being a juvenile delinquent, scouts being less likely to be delinquents. We are hypothesizing that

$S \rightarrow D$, where S (for Scouting) and D (for Delinquency) denote the binary variables "whether a Boy Scout (yes, no)" and "whether a juvenile delinquent (yes, no)."

If we suspect that one variable is causally explained by another, how do we analyze whether it actually is? A relationship must satisfy three criteria to be considered a causal one. These criteria, which we'll discuss below, are

- association between the variables,
- an appropriate time order, and
- the elimination of alternative explanations.

If all three are met, then the evidence supports the hypothesized causal relationship. If one or more criteria are not met, then we conclude that there is not a causal relationship.

ASSOCIATION IS REQUIRED, BUT NOT SUFFICIENT, FOR CAUSALITY

The first criterion for causality is ***association***. We must show that x and y are associated. If $x \rightarrow y$, then as x changes, the distribution of y should change in some way. If scouting causes lower delinquency rates, for example, then the population proportion of delinquents should be higher for nonscouts than for scouts. For sample data, a statistical test, such as chi-squared for categorical data or a t test for the regression slope or for a comparison of means for quantitative data, analyzes whether this criterion is satisfied.

Association by itself cannot establish causality.

Association does not imply causation.

The remainder of this section explains why.

CAUSALITY REQUIRES APPROPRIATE TIME ORDER

The second criterion for causality is that the two variables have the appropriate ***time order***, with the cause preceding the effect. Sometimes this is just a matter of logic. For instance, race, age, and gender exist prior to current attitudes or achievements, so any causal association must treat them as causes rather than effects.

In other cases, the causal direction is not as obvious. Consider scouting and delinquency. It is logically possible that scouting reduces delinquency tendencies. On the other hand, it is also possible that delinquent boys avoid scouting but nondelinquent boys do not. Thus, the time order is not clear, and both possibilities, $S \rightarrow D$ and $D \rightarrow S$, are plausible. Just showing that an association exists does not solve this dilemma, because a lower proportion of delinquents among scout members is consistent with both explanations.

It is difficult to study cause and effect when two variables do not have a time order but are measured together over time. The variables may be associated merely because they both have a time trend. For example, for recent annual data there is a correlation of 0.993 between $y =$ divorce rate in Maine and $x =$ per capita consumption of margarine.[1] They both have a decreasing trend over time, so they have a strong positive correlation, with higher divorce rates occurring in years that have higher consumption of margarine. Each variable would be strongly negatively correlated with all variables that have a positive time trend, such as percentage of people who

---

[1] See www.tylervigen.com/spurious-correlations.

use smart phones, percentage of people who belong to an Internet social network such as Facebook, and annual average worldwide temperature.

ALTERNATIVE EXPLANATION MAY INVALIDATE CAUSAL RELATION

When two variables are associated and have the proper time order to satisfy a casual relation, this is still insufficient to imply causality. The association may have an ***alternative explanation***.

For example, airline pilots turn on the "fasten seat belt" sign just before their planes encounter turbulence. We observe an association, greater turbulence occurring when the sign is on than when it is off. There is usually also the appropriate time order, the sign coming on, followed by turbulence shortly afterward. But this does not imply that turning on the sign causes turbulence.

An alternative explanation for an association is responsible for rejecting many hypotheses of causal relationships. Many alternative explanations involve an additional variable z or a set of variables. For example, there may be a variable z that causes both x and y.

With observational data, it is easy to find associations, but those associations are often explained by other variables that may not have been measured in a study. For example, some medical studies have found associations between coffee drinking and various responses, such as the likelihood of a heart attack. But after taking into account other variables associated with the extent of coffee drinking, such as country of residence, occupation, and levels of stress, such associations have disappeared or weakened considerably.

This criterion for causality of eliminating an alternative explanation is the most difficult to achieve. We may think we've found a causal relationship, but we may merely have not thought of a particular reason that can explain the association. Because of this, *with observational studies we can never prove that one variable is a cause of another*. We can disprove causal hypotheses, however, by showing that empirical evidence contradicts at least one of these three criteria.

ASSOCIATION, CAUSALITY, AND ANECDOTAL EVIDENCE

The association between smoking and lung cancer is regarded as having a causal link. The association is moderately strong, there is the proper time order (lung cancer following a period of smoking), and no alternative explanation has been found to explain the relationship. In addition, the causal link has been bolstered by biological theories that explain how smoking could cause lung cancer.

Sometimes you hear people give anecdotal evidence to attempt to disprove causal relationships. "My Uncle John is 85 years old, he still smokes a pack of cigarettes a day, and he's as healthy as a horse." An association does not need to be perfect, however, to be causal. Not all people who smoke two packs of cigarettes a day will get lung cancer, but a much higher percentage of them will do so than people who are nonsmokers. Perhaps Uncle John is still in fine health, but that should not encourage us to tempt the fates by smoking a pack each day. Anecdotal evidence is not enough to disprove causality unless it can deflate one of the three criteria for causality.

ESTABLISHING CAUSALITY WITH RANDOMIZED EXPERIMENTS

As mentioned in Section 2.2 (page 26), a randomized experiment is the ideal way to compare two groups. This approach, by which we randomly select the subjects

for each group and then observe the response, provides the gold standard for establishing causality. For instance, does a new drug have a beneficial effect in treating a disease? We could randomly assign subjects suffering from the disease to receive either the drug or a placebo. Then, to analyze whether the drug assignment may have a causal influence on the response outcome, we would observe whether the proportion successfully treated was significantly higher for the drug group.

In a randomized experiment, suppose that we observe an association between the group variable and the response variable, such as a statistically significant difference between two proportions for a categorical response or between two means for a quantitative response. With such an experiment, we do not expect another variable to provide an alternative explanation for the association. With randomized assignment to groups, the two groups should have about the same distributions for variables not observed but which may be associated with the response variable. So, the association is not because of an association between the group variable and an alternative variable. In addition, when a research study is experimental, the time order is fixed. The outcome for a subject is observed *after* the group is assigned, so the time order is certain. Because of these factors, *it is easier to assess causality with randomized experiments than with observational studies.*

In most social research, unfortunately, randomized experiments are not possible. If we want to investigate the effect of level of education on political ideology, we cannot randomly assign children to different levels of attained education and then later ask them about their ideology. For each person sampled, we can merely observe their actual attained education and political ideology, and data are missing for that subject about what the political ideology would have been had they attained a different level of education. In the next section, we present a way that we can attempt to adjust for different groups differing in their distributions of other variables that could be associated with the response variable.

10.2 Controlling for Other Variables

A fundamental component to evaluating whether x could cause y is searching for an alternative explanation. We do this by studying whether the association between x and y remains when we remove the effects of other variables on this association. In a multivariate analysis, a variable is said to be ***controlled*** when its influence is removed.

A laboratory experiment controls variables that could affect the results by holding their values constant. For instance, an experiment in chemistry or physics might control temperature and atmospheric pressure by holding them constant in a laboratory environment during the course of the experiment. A lab experiment investigating the effect of different doses of a carcinogen on mice might control the age and diet of the mice.

Randomized experiments cannot strictly control other variables. But in their randomization, we expect groups on which we perform randomization to have similar distributions on the other variables. So, randomized experiments inherently control other variables in a probabilistic sense.

STATISTICAL CONTROL IN SOCIAL RESEARCH

Unlike laboratory sciences, social research is usually observational rather than experimental. We cannot fix values of variables we might like to control, such as intelligence or education, before obtaining data on the variables of interest. But we can approximate an experimental type of control by grouping together observations

with equal, or similar, values on the control variables. Socioeconomic status or a related variable such as education or income is often a prime candidate for control in social research. To control education, for instance, we could group the sample results into those subjects with less than a high school education, those with a high school education but no college education, those with some college education, and those with at least one college degree. This is ***statistical control***, rather than experimental control.

The following example, although artificial, illustrates statistical control in a social science setting, by holding a key variable constant.

<table>
<tr><td>

**Example
10.1**

</td><td>

Causal Effect of Height on Math Achievement? Do tall students tend to be better than short students in learning math skills? We might think so looking at a random sample of students from Lake Wobegon school district who take a math achievement test. The correlation is 0.81 between height and math test score. Taller students tend to have higher scores.

Is being tall a causal influence on math achievement? Perhaps an alternative explanation for this association is that the sample has students of various ages. As age increases, both height and math test score would tend to increase. Older students tend to be taller, and older students tend to have stronger math knowledge.

We can remove the effects of age from the association by *statistical control*, studying the association between height and math test score for students of the same age. That is, we control for age by analyzing the association separately at each age level. Then, variation in age cannot jointly cause variation in both height and test score.

In fact, the achievement test was administered to students from grades 2, 5, and 8 at Lake Wobegon, so the sample contained considerable variability in the students' ages. Figure 10.1 shows a scatterplot of the observations, with labels indicating the grade for each student. The overall pattern of points shows a strong positive correlation, with higher math scores at higher heights. View the points within a fixed grade level (for which age is approximately constant), however, and you see random variation, with no particular pattern of increase or decrease. The correlation between height and math test score is close to zero for students of about the same age. Height does not have a causal effect on math test score, because the association disappears when age is held constant. ∎

</td></tr>
</table>

FIGURE 10.1: Scatterplot Showing Relationship between Height and Math Achievement Test Score, with Observations Labeled by Grade Level. Students at a particular grade level have about the same age and show a lack of association between height and test score.
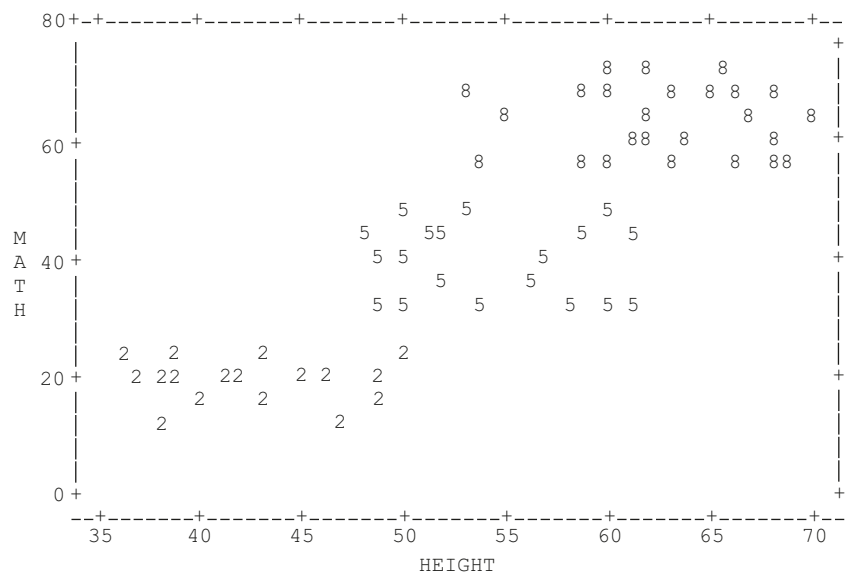
In summary, we control a variable by holding its value constant, or nearly so. We can then study the relationship between x and y for cases with equal, or similar, values of that variable. The variable controlled is called a ***control variable***. In holding the control variable constant, we remove the influence of that variable on the association between x and y.

STATISTICAL CONTROL FOR VARIABLE TYPES IN AN ASSOCIATION

The scatterplot in Figure 10.1 describes the association between two quantitative variables, controlling for a third variable. We can describe association between a quantitative variable and a categorical variable by comparing means. For example, at your school suppose the mean salary for male faculty is higher than the mean salary for female faculty. Suppose that the percentage of professors who are female is lowest for full professors and is considerably higher for instructors and assistant professors, perhaps because relatively few female faculty were hired until recent years. Then, this difference in mean salaries would diminish and could even disappear if we control for academic rank.

To study the association between two categorical variables, while controlling for a third variable, we form contingency tables relating those variables separately for subjects at each level of that control variable. The separate tables that display the relationships within the fixed levels of the control variable are called ***partial tables***.

Example 10.2

Partial Tables for Control with Categorical Variables Table 10.1 is a hypothetical table relating scouting to delinquency. The percentage of delinquents among scout members is lower than among nonscouts. This table is ***bivariate***, meaning that it contains data only on *two* variables. All other variables are ignored. None is controlled.

TABLE 10.1: Contingency Table Relating Scouting and Delinquency. Percentages refer to conditional distribution of delinquency, given whether a boy scout.					
		Delinquency			
		Yes		No	Total
Boy Scout	Yes	36	(9%)	364 (91%)	400
	No	60	(15%)	340 (85%)	400

In seeking a possible explanation for the association, we could control for church attendance. Perhaps boys who attend church are more likely than nonattenders to be scouts, and perhaps boys who attend church are less likely to be delinquent. Then, the difference in delinquency rates between scouts and nonscouts might be due to variation in church attendance.

To control for church attendance, we examine the association between scouting and delinquency within partial tables formed by various levels of church attendance. Table 10.2 shows partial tables for three levels: Low = no more than once a year, Medium = more than once a year but less than once a week, and High = at least once a week. Adding these three partial tables together produces the bivariate table (Table 10.1), which ignores church attendance. For instance, the number of Boy Scouts who are delinquents is $36 = 10 + 18 + 8$.

TABLE 10.2: Contingency Table Relating Scouting and Delinquency, Controlling for Church Attendance						
	Church Attendance					
	Low		Medium		High	
Delinquency	Yes	No	Yes	No	Yes	No
Scout	10 (20%)	40 (80%)	18 (12%)	132 (88%)	8 (4%)	192 (96%)
Not scout	40 (20%)	160 (80%)	18 (12%)	132 (88%)	2 (4%)	48 (96%)

In each partial table, the percentage of delinquents is the same for scouts as for nonscouts. Controlling for church attendance, no association appears between scouting and delinquency. These data provide an alternative explanation for the association between scouting and delinquency, making us skeptical of any causal links. The alternative explanation is that both these variables are associated with church attendance. Youngsters who attend church are less likely to be delinquents and more likely to be scouts. For a fixed level of church attendance, scouting has no association with delinquency. Since the association can be explained by church attendance, no causal link exists between scouting and delinquency. ■

Some examples in this chapter, like this one, use artificial data in order to make it simpler to explain the concepts. In practice, some distortion occurs because of sampling variation. Even if an association between two variables truly disappears under a control, *sample* partial tables would not look exactly like those in Table 10.2. Because of sampling variation, they would not show a *complete* lack of association. Moreover, few associations disappear *completely* under a control. There may be *some* causal connection between two variables, but not as strong as the bivariate table suggests. Moreover, in practice we need to control for several variables, and we'll see in the next chapter that statistical control then involves further assumptions.

BE WARY OF LURKING VARIABLES

It is not always obvious which variables require control in a study. Knowing about the theory and previous research in a field of study helps a researcher to know which variables to control. A potential pitfall of almost all social science research is the possibility that the study did not include an important variable. If you fail to control for a variable that strongly influences the association between the variables of primary interest, you will obtain misleading results.

A variable that is *not* measured in a study (or perhaps even known about to the researchers) but that influences the association under study is called a ***lurking variable***. In analyzing the positive correlation between height and math achievement in Example 10.1 (page 303), we observed that the correlation could be due to a lurking variable, the age of the student.

When you read about a study that reports an association, try to think of a lurking variable that could be responsible. For example, suppose a study reports a positive correlation between individuals' college GPA and their income later in life. Is doing well in school responsible for the higher income? An alternative explanation is that both high GPA and high income could be caused by a lurking variable such as IQ or an individual's tendency to work hard.

10.3 Types of Multivariate Relationships

Section 10.2 showed that an association may change dramatically when we control for another variable. This section describes types of multivariate relationships that often occur in social science research. For a response variable y, there may be several explanatory variables and control variables, and we denote them by $x_1, x_2, \ldots$.

SPURIOUS ASSOCIATIONS

An association between y and x_1 is said to be ***spurious*** if both variables are dependent on a third variable x_2, and their association disappears when x_2 is controlled. Such an association results from the relationship of y and x_1 with the control variable x_2, rather than indicating a causal connection. Showing that the association between two variables may be spurious provides an alternative explanation to a causal connection between them.

Example 10.3

Examples of Spurious Associations Table 10.1 (page 304) displayed an association between scouting and delinquency. Controlling for church attendance, the partial tables in Table 10.2 (page 305) showed no association. This is consistent with spuriousness. Table 10.2 shows that as church attendance increases, the percentage of delinquents decreases (compare percentages across the partial tables) and the percentage of scout members increases. By the nature of these two associations, it is not surprising that Table 10.1 exhibits lower overall delinquency rates for scouts than nonscouts.

The association between height and mathematics achievement test score in Example 10.1 disappears at fixed levels of age. That association is spurious, with age being a common cause of both height and math achievement. Figure 10.2 graphically depicts this spurious association, using $x_1 =$ height and $y =$ math test score. They are associated only because they both depend on a common cause, $x_2 =$ age. As x_2 changes, it produces changes simultaneously in x_1 and y, so that x_1 and y are associated. In fact, they are associated only because of their common dependence on the third variable (age). ∎

FIGURE 10.2: Graphical Depiction of a Spurious Association between y and x_1. The association disappears when we control for x_2, which causally affects both x_1 and y.
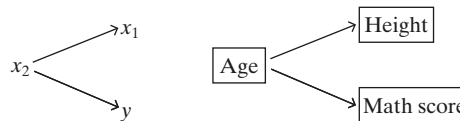
Example 10.4

Do Fewer Vacations Cause Increased Risk of Death? When an association is observed between two variables, later studies often attempt to determine whether that association might be spurious, by controlling for variables that could be a common cause. For example, some studies have observed an association between frequency of vacationing and quality of health. In particular, a study using a 20-year follow-up of women participants in the Framingham Heart Study found[2] that less frequent vacationing was associated with greater frequency of deaths from heart attacks.

---

[2] E. D. Eaker et al., *American Journal of Epidemiology*, vol. 135 (1992), pp. 835–864.

A later study[3] questioned whether this could be a spurious association, explained by the effects of socioeconomic status (SES). For example, perhaps higher SES is responsible both for lower mortality and for more frequent vacations. But after controlling for education, family income, and other potentially important variables with a much larger data set, this study also observed higher risk of heart disease and related death for those who took less vacation time. Perhaps the association is not spurious, unless researchers find another variable to control such that the association disappears. ∎

CHAIN RELATIONSHIPS AND INTERVENING (MEDIATOR) VARIABLES

Another way that an association can disappear when we control for a third variable is with a *chain* of causation, in which x_1 affects x_2, which in turn affects y. Figure 10.3 depicts the chain. Here, x_1 is an *indirect*, rather than direct, cause of y. Variable x_2 is called an **intervening** variable or a **mediator** variable.

FIGURE 10.3: A Chain Relationship, in Which x_1 Indirectly Affects y through an Intervening Variable x_2, Which Has a Mediating Effect

$$x_1 \longrightarrow x_2 \longrightarrow y$$

Example 10.5

Is Education Responsible for a Long Life?   A *New York Times* article[4] summarized research studies dealing with human longevity. It noted that consistently across studies in many nations, life span was positively associated with educational attainment. Many researchers believe education is the most important variable in explaining how long a person lives. Is having more education responsible for having a longer life?

Establishing causal connections is difficult. In some societies, perhaps the causation could go in the other direction, with sick children not going to school or dropping out early because they were ill. Many researchers believe there could be a chain of causation, perhaps with income as an intervening variable. For example, perhaps having more education leads to greater wealth, which then (possibly for a variety of reasons, such as access to better health care) leads to living longer. Figure 10.4 depicts this causal chain model.

FIGURE 10.4: Example of a Chain Relationship. Income is an intervening variable (also called a *mediator* variable), and the association between education and life span disappears when it is controlled.
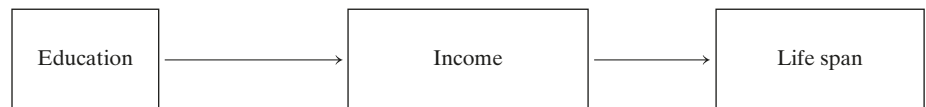
This model is supported if the association between education and life span disappears after controlling for income; that is, if within fixed levels of income (the intervening variable), no significant association occurs. If this happens, education does not directly affect life span, but it is an indirect cause through income. ∎

For both spurious relationships and chain relationships, an association between y and x_1 disappears when we control for a third variable, x_2. The difference between

---

[3] B. B. Gump and K. A. Matthews, *Psychosomatic Medicine*, vol. 62 (2000), pp. 608–612.
[4] Written by G. Kolata, January 3, 2007.

the two is in the causal order among the variables. For a spurious association, x_2 is causally prior to both x_1 and y, as in Figure 10.2. In a chain association, x_2 intervenes between the two, as in Figures 10.3 and 10.4.

To illustrate, a study[5] of mortality rates in the United States found that states that had more income inequality tended to have higher age-adjusted mortality rates. However, this association disappeared after controlling for the percentage of a state's residents who had at least a high school education. Might this reflect a chain relationship, or a spurious relationship? Greater education could tend to result in less income inequality, which could in turn tend to result in lower mortality rates. Thus, the chain relationship

$$\text{Education} \longrightarrow \text{Income inequality} \longrightarrow \text{Mortality rate}$$

is plausible. For the relationship to be spurious, education would need to have a causal effect on both income inequality and mortality. This is also plausible. Just from viewing the association patterns, we do not know which provides a better explanation.

MULTIPLE CAUSES

Response variables in social science research almost always have more than one cause. For instance, a variety of factors likely have causal influences on responses such as $y =$ juvenile delinquency or $y =$ length of life. Figure 10.5 depicts x_1 and x_2 as separate causes of y. We say that y has ***multiple causes***.

FIGURE 10.5: Graphical Depiction of Multiple Causes of y
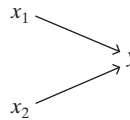
Sometimes variables that are separate causes of y are themselves statistically independent. That is, they are *independent causes*. For instance, $x_1 =$ gender and $x_2 =$ race are essentially statistically independent. If they both have effects on juvenile delinquency, with delinquency rates varying both according to gender and race, they are likely to be independent causes.

In the social sciences, most explanatory variables are associated. Both being poor and being from a single-parent family may cause delinquency, but those factors are themselves probably associated. Because of complex association linkages, when we control for a variable x_2 or a set of variables $x_2, x_3, \ldots,$ the $x_1 y$ association usually changes somewhat. Often the association decreases somewhat, although usually it does not completely disappear as in a spurious or chain relationship. Sometimes this is because x_1 has direct effects on y and also indirect effects through other variables. Figure 10.6 illustrates this. For instance, perhaps being from a single-parent family has direct effects on delinquency but also indirect effects through being more likely to be poor. Most response variables have many causes, both direct and indirect.

FIGURE 10.6: Graphical Depiction of Direct and Indirect Effects of x_1 on y

---
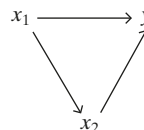
[5] A. Muller, *BMJ*, vol. 324 (2002), pp. 23–25.

SUPPRESSOR VARIABLES

In examples so far, an association disappears or weakens when we control for another variable. By contrast, occasionally two variables show no association until a third variable is controlled. That control variable is called a ***suppressor variable***.

Age Suppresses the Association between Education and Income   Is educational level positively related with income? Table 10.3 shows such a relationship, measured as binary variables, controlling for age. In each partial table, the percentage of subjects at the high level of income is greater when education is high than when education is low.

TABLE 10.3: Partial Tables Relating Education and Income, Controlling for Age								
		Age $=$ Low				Age $=$ High		
Income:		High	Low	% High		High	Low	% High
Education	High	125	225	35.7		125	25	83.3
	Low	25	125	16.7		225	125	64.3

Suppose now that we ignore age, adding these two partial tables together. The bivariate table for education and income is the first panel of Table 10.4. Every count equals 250. Both when education is high and when education is low, the percentage having a high income is 50%. For the bivariate table, no association exists between education and income.

TABLE 10.4: Bivariate Tables Relating Education, Income, and Age								
	Income			Income			Education	
Education	High	Low	Age	High	Low	Age	High	Low
High	250	250	High	350	150	High	150	350
Low	250	250	Low	150	350	Low	350	150

A look at the other two bivariate tables in Table 10.4 reveals how this could happen. Age is positively associated with income but negatively associated with education. Older subjects tend to have higher income, but they tend to have lower education. Thus, when we ignore rather than control age, we give an inadvertent boost to the relative numbers of people at high incomes with low educational levels and at low incomes with high educational levels. Because of the potential for a suppressor effect, it can be informative to control for a variable even when the bivariate analysis does not show an association with y. This is especially true when there is a theoretical reason for a potential suppression effect.   ∎

STATISTICAL INTERACTION

Often the effect of an explanatory variable on a response variable changes according to the level of another explanatory variable or control variable. When the true effect of x_1 on y changes at different levels of x_2, the relationship is said to exhibit *statistical interaction*.

Statistical Interaction	***Statistical interaction*** exists between x_1 and x_2 in their effects on y when the true effect of one predictor on y changes as the value of the other predictor changes.

Example 10.7

Interaction between Education and Gender in Predicting Income   Consider the relationship between $y =$ annual income (in thousands of dollars) and $x_1 =$ number of years of education, by $x_2 =$ gender. Many studies in the United States have found that the slope for a regression equation relating y to x_1 is larger for men than for women. Suppose that in the population, the regression equations are $E(y) = -10 + 5x_1$ for men and $E(y) = -5 + 3x_1$ for women. On the average, income for men increases by \$5000 for every year of education, whereas for women it increases by \$3000 for every year of education. That is, the effect of education on income varies according to gender, with the effect being greater for men than for women. So, there is interaction between education and gender in their effects on income.   ∎

Example 10.8

Interaction between SES and Age in Predicting Health   Some studies[6] have noted that quality of health (measured by self-rated and health indexes) tends to be positively associated with SES (measured by years of education and annual household income), and that the association strengthens with age. For example, the gap in health between low SES and high SES levels tends to be larger at older ages. Thus, there is interaction between SES and age in their effects on health.   ∎

ANALYZING AND DEPICTING INTERACTION

To assess whether a sample shows evidence of interaction, we can compare the effect of x_1 on y at different levels of x_2. When the sample effect is similar at each level of x_2, it's simpler to use statistical analyses that assume an absence of interaction. The interaction is worth noting when the variability in effects is large. For instance, perhaps the association is positive at one level of x_2 and negative at another, or strong at one level and weak or nonexistent at another.

Figure 10.7 depicts a three-variable relationship having statistical interaction. Here, x_2 affects the relationship between x_1 and y. When this happens, then likewise x_1 affects the relationship between x_2 and y.

FIGURE 10.7: Graphical Depiction of Statistical Interaction. The effect of one explanatory variable on y depends on the level of the other explanatory variable.
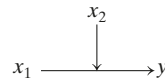
Suppose *no* interaction occurs between x_1 and x_2 in their effects on y. This does not mean that x_1 and x_2 have no association. There can be a lack of statistical interaction even when all the variables are associated. For instance, Tables 10.2 (page 305) and 10.3 (page 309) showed no interaction—in each case the association was similar in each partial table. However, in each case the explanatory variables

---

[6] For example, see S. G. Prus, *Canadian Journal on Aging*, vol. 23 (2004), Supplement, pp. S145–S153.

were associated, with each other and with the response. In Table 10.3, for instance, age was negatively associated with education and positively associated with income.

SUMMARY OF MULTIVARIATE RELATIONSHIPS

In summary,

- For spurious relationships (i.e., x_2 affects both x_1 and y) and chain relationships (i.e., x_2 intervenes between x_1 and y), the $x_1 y$ association disappears when we control for x_2.

- For multiple causes, an association may change under a control but does not disappear.

- When there is a suppressor variable, an association appears only under the control.

- When there is statistical interaction, an association has different strengths and/or directions at different values of a control variable.

This does not exhaust the possible association structures. It is even possible that, after controlling for a variable, each association in a partial table has the *opposite* direction as the bivariate association. This is called **Simpson's paradox** and is illustrated in Exercises 10.14, 10.29, and 10.30.

CONFOUNDING AND OMITTED VARIABLE BIAS

When two explanatory variables both have effects on a response variable but are also associated with each other, there is said to be **confounding**. It is difficult to determine whether either of them truly causes the response, because a variable's effect could be at least partly due to its association with the other variable. We usually observe a different effect on *y* for a variable when we control for the other variable than when we ignore it.

In analyzing the effect of an explanatory variable of key interest, if our study neglects to observe a confounding variable that explains a major part of that effect, our results and conclusions will be biased. Such bias is called **omitted variable bias**.

Confounding and omitted variable bias are constant worries in social science research. They are the main reason it is difficult to study many issues of importance, such as what causes crime or what causes the economy to improve or what causes students to succeed in school.

10.4 Inferential Issues in Statistical Control

To conduct research well, you must select the key variables, determine which variables to control, choose an appropriate model, and analyze the data and interpret the results properly. So far this chapter has ignored inferential matters, to avoid confusing them with the new concepts presented. We now discuss some inferential issues in studying associations while controlling other variables.

EFFECTS OF SMALLER SAMPLE SIZE IN PARTIAL ANALYSES

Suppose we control for x_2 in studying the $x_1 y$ association. The sample size at a fixed level of x_2 may be much smaller than in the full data set. Even if no reduction in

association occurs relative to the full data, standard errors of parameter estimators tend to be larger. Thus, confidence intervals for those parameters at fixed levels of x_2 tend to be wider, and test statistic values tend to be smaller.

For categorical data, for example, we could compute the Pearson X^2 statistic within a particular partial table to test whether the variables are independent at that level of x_2. This X^2-value may be small relative to the X^2-value for the bivariate x_1y table. This could be due partly to a weaker association, but it could also reflect the reduction in sample size. Section 8.4 showed that larger sample sizes tend to produce larger X^2-values, for a particular degree of association.

EFFECTS OF CATEGORIZATION IN CONTROLLING A VARIABLE

Categorical control variables (e.g., gender) have the categories as the natural values held constant in partial tables. For ordinal control variables, you should avoid overly crude categorizations. The greater the number of control levels, the more nearly constant the control variable is within each partial table. It is usually advisable to use at least three or four partial tables.

On the other hand, it is preferable not to use more partial tables than needed, because then each one may have a small sample size. Separate estimates may have large standard errors, resulting in imprecise inferences within the partial tables and comparisons of associations between tables. Fortunately, the model-building methods presented in the rest of the text allow us to attempt statistical control and assess patterns of association and interaction without necessarily performing separate analyses at the various combinations of levels of the control variables.

COMPARING AND POOLING MEASURES

It is often useful to compare parameter values describing the effect of an explanatory variable on a response variable at different levels of a control variable. You can construct a confidence interval for a difference between two parameter values in the same way as Chapter 7 showed for a difference of proportions or a difference of means. Suppose that the two sample estimates are based on independent random samples, with standard errors se_1 and se_2. Then, Section 7.1 noted that the standard error for the difference between the estimates is $\sqrt{(se_1)^2 + (se_2)^2}$. For large random samples, most estimates have approximately normal sampling distributions. Then, a confidence interval for the difference between the parameters is

$$(\text{Estimate}_2 - \text{Estimate}_1) \pm z\sqrt{(se_1)^2 + (se_2)^2}.$$

If the interval does not include 0, the evidence suggests that the parameter values differ.

Example 10.9

Comparing Happiness Associations for Men and Women Is there a difference between men and women in the association between happiness and marital happiness? For recent data from the GSS, the sample value of gamma for a 3×3 table relating these two ordinal variables is 0.674 ($se = 0.0614$, $n = 326$) for males and 0.689 ($se = 0.0599$, $n = 350$) for females.

A 95% confidence interval for the difference between the population values of gamma is

$$(0.689 - 0.674) \pm 1.96\sqrt{(0.0614)^2 + (0.0599)^2}, \quad \text{or} \quad 0.015 \pm 0.168,$$

which is $(-0.153, 0.183)$. It is plausible that the population gamma values are identical. If they are not identical, they seem not to be very different. ■

When the association between two variables is similar in the partial analyses, we can form a measure that summarizes the strength of the association, conditional on the control variable. This is referred to as a measure of *partial association*. The rest of the text shows how to do this in various contexts, using models that handle all the variables at once.

10.5 Chapter Summary
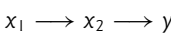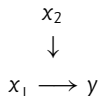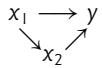
We use a multivariate analysis to study effects of multiple explanatory variables on a response variable. To demonstrate a causal relationship, we must show *association* between variables, ensure proper *time order*, and *eliminate alternative explanations* for the association. This is possible for randomized experiments, but eliminating alternative explanations is a challenge for observational studies.

To consider alternative explanations in observational studies, we introduce *control variables*. We perform statistical control by analyzing associations while keeping the values of control variables essentially constant. This helps us to detect

- *Spuriousness*, in which x_2 jointly affects both y and x_1.

- *Chain relationships*, in which x_2 is an *intervening variable* (also called a *mediator variable*), so that x_1 affects y indirectly through its effects on x_2.

- *Suppressor variables*, in which the $x_1 y$ association appears only after controlling for x_2.

- *Statistical interaction*, in which the effect of x_1 on y varies according to the value of x_2.

Table 10.5 summarizes some possible relationships. The remainder of this text presents statistical methods for multivariate relationships. As you learn about these methods, be careful not to overextend your conclusions: Realize the limitations in making causal inferences with inferential statistical analyses, and keep in mind that any inferences you make must usually be tentative because of assumptions that may be violated or lurking variables that you did not include in your analyses. For further discussion of these points in the context of regression modeling, see Berk (2004), Freedman (2005), Morgan and Winship (2007), and Pedhazur (1997).

TABLE 10.5: Some Three-Variable Relationships

Graph	Name of Relationship	Controlling for x_2
$x_2 \nearrow x_1$ $\searrow y$	Spurious $x_1 y$ association	Association between x_1 and y disappears.
$x_1 \longrightarrow x_2 \longrightarrow y$	Chain relationship; x_2 intervenes; x_1 indirectly causes y	Association between x_1 and y disappears.
x_2 $\downarrow$ $x_1 \longrightarrow y$	Interaction	Association between x_1 and y varies according to level of x_2.
$x_2 \searrow y$ $x_1 \nearrow$	Multiple causes	Association between x_1 and y does not change.
$x_1 \longrightarrow y$ $\searrow x_2 \nearrow$	Both direct and indirect effects of x_1 on y	Association between x_1 and y changes, but does not disappear.

Exercises

Practicing the Basics

10.1. State the three criteria for a causal relationship. For each, describe a relationship between two variables that is not causal because that criterion would be violated.

10.2. A young child wonders what causes women to have babies. For each woman who lives on her block, she observes whether her hair is gray and whether she has young children. The four women with gray hair do not have young children, whereas all five women not having gray hair have young children. Noticing this association, the child concludes that not having gray hair is what causes women to have children.

(a) Form the contingency table displaying the data.

(b) Use this example to explain why association does not imply causation.

10.3. For all fires in Chicago last year, data are available on x = number of firefighters at the fire and y = cost of damages due to the fire. The correlation is positive.

(a) Does this mean that having more firefighters at a fire causes the damage to be worse? Explain.

(b) Identify a third variable that could be a common cause of x and y. Construct a hypothetical scatterplot (like Figure 10.1 on page 303), identifying points according to their value on the third variable, to illustrate your argument.

10.4. Cities in the United States have a positive correlation between y = crime rate and x = size of police force. Does this imply that x causes y? Explain.

10.5. An association exists between college GPA and whether a college student has ever used marijuana. Explain how

(a) The direction of a causal arrow might go in either direction.

(b) A third variable might be responsible for the association.

10.6. Explain what it means to *control* for a variable, using an example to illustrate.

10.7. Explain what is meant by a *spurious* association, drawing a scatter diagram to illustrate.

(a) Illustrate using x_1 = shoe size, x_2 = age, and y = number of books one has ever read, for children from schools in Winnipeg, Canada.

(b) Illustrate using x_1 = height, x_2 = gender, and y = annual income, for a random sample of adults. Suppose that, overall, men tend to be taller and have higher income, on the average, than females.

10.8. Figure 9.16 on page 286 showed a negative correlation between birth rate and television ownership. Identify a variable to help explain how this association could be spurious.

10.9. An Associated Press story quoted a study at the University of California at San Diego that reported, based on a nationwide survey, that those who averaged at least 8 hours of sleep a night were 12 percent more likely to die within six years than those who averaged 6.5 to 7.5 hours of sleep a night.

(a) Explain how the subject's age could be positively associated both with time spent sleeping and with an increased death rate, and hence could explain the association between sleeping and the death rate.

(b) If the association disappears when we control for subject's age, do you think age is more likely to be a common cause, or a mediator variable?

10.10. A study found that children who eat breakfast get better math grades than those who do not eat breakfast. This result was based on the association between x = whether eat breakfast (yes, no) and y = grade in last math course taken. How might this result be spurious, and how could you check for that possibility?

10.11. Suppose race is related to frequency of juvenile arrests, with black juveniles more likely to be arrested than white juveniles. A possible chain relationship explanation is that (1) race affects family income, with blacks tending to have lower family incomes than whites, and (2) being poor increases the chance of being arrested as a juvenile. Show a figure to portray the chain relationship. To support this explanation, what would need to happen to the difference between the arrest rates for whites and blacks, after controlling for family income?

10.12. A study at your university finds that of those who applied to its graduate school last year, the percentage admitted was *higher* for the male applicants than for the female applicants. However, for each department that received applications, the percentage admitted was *lower* for the male applicants than for the female applicants. How could this possibly happen? In your answer, explain what plays the role of the response variable, the explanatory variable, the control variable, the bivariate table, and the partial tables. (Exercise 15.12 shows data that have similar behavior.)

10.13. Table 10.6 relates occupational level (white collar, blue collar) and political party choice, controlling for income.

(a) Construct the bivariate table between occupational level and political party, ignoring income. Is there an association? If so, describe it.

(b) Do the partial tables display an association? Interpret them.

(c) Using the nature of the association between income and each of the other variables, explain why the bivariate table has such different association than the partial tables.

TABLE 10.6

Party	High Income		Medium High Income		Medium Low Income		Low Income	
	White Collar	Blue Collar	White Collar	Blue Collar	White Collar	Blue Collar	White Collar	Blue Collar
Democrat	45	5	100	25	75	300	45	405
Republican	405	45	300	75	25	100	5	45

(d) Construct a chain diagram that might explain the relationships, identifying the intervening (i.e., mediator) variable.

(e) Show that the data are also consistent with a spurious association, and draw the corresponding diagram. Which diagram seems more appropriate? Why?

10.14. In murder trials[7] in 20 Florida counties in two years, the death penalty was given in 19 out of 151 cases in which a white killed a white, in 0 out of 9 cases in which a white killed a black, in 11 out of 63 cases in which a black killed a white, and in 6 out of 103 cases in which a black killed a black.

(a) Construct partial tables relating D = defendant's race and P = the death penalty verdict, controlling for V = victim's race. In those tables, compare the proportions of white and black defendants who received the death penalty.

(b) Construct the bivariate table, ignoring victim's race. Describe the association, and compare to (a).

(c) *Simpson's paradox* states that the associations in partial tables can all have a different direction than the association in the bivariate table. Show that these data satisfy Simpson's paradox, with white defendants having a lower or higher chance of the death penalty than black defendants according to whether we control victim's race.

(d) By describing how V is associated with D (whites tending to kill whites and blacks tending to kill blacks) and how V is associated with P (killing a white more likely to lead to the death penalty), explain why the partial association differs as it does from the bivariate association.

(e) For these variables, indicate whether each of the following diagrams seems to provide a reasonable model. Give your reasoning.

i. $V \underset{D}{\overset{P}{\nearrow}}$ ii. $D \longrightarrow V \longrightarrow P$ iii. $\overset{D}{\underset{V}{\searrow}} P$ iv. $D \underset{V}{\overset{}{\searrow}} P$
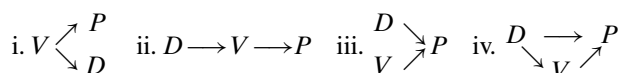
10.15. For the data on house sales shown partly in Table 9.5 on page 280, the number of bedrooms has a moderately strong positive correlation with selling price. Controlling for size of home, however, this association diminishes greatly.

(a) Explain how this could happen, illustrating with a diagram showing potential direct and indirect effects of number of bedrooms on selling price.

(b) Explain what it means to say that there is *confounding* in the effects of size of home and number of bedrooms in their effects on the selling price.

10.16. For the `Florida` data file (shown partly on page 295), giving countywide data in Florida for several variables, a moderate positive correlation ($r = 0.47$) exists between crime rate and the percentage who are high school graduates. The percentage living in urban areas is also strongly correlated with crime rate ($r = 0.68$) and with high school graduation rate ($r = 0.79$).

(a) Explain why the association between crime rate and high school graduation rate could disappear, or even change direction, when we control for the percentage living in urban areas.

(b) Under the control in (a), if the association disappears, which type of relationship is more plausible—a spurious relationship or a chain relationship? Explain.

10.17. Opposition to the legal availability of abortion is stronger among the very religious than the nonreligious, and it is also stronger among those with conservative sexual attitudes than those with more permissive attitudes. Draw a three-variable diagram of how these variables might be related, treating abortion attitude as the response variable. (*Note*: More than one diagram is plausible.)

10.18. Table 10.7 lists the mean salary, in thousands of dollars, of faculty on nine-month contracts in U.S. institutions of higher education in 2013–2014, by gender and academic rank.

(a) Suppose that gender is the explanatory variable. Identify the response variable and the control variable.

(b) Describe the bivariate relationship between gender and salary.

(c) Describe the relationship between gender and salary, controlling for academic rank.

(d) A hypothesis of interest for these variables is "Controlling for academic rank, annual salary and gender are independent." Draw a causal diagram that is consistent with this hypothesis. Refer to your interpretation in

---

[7] Data from M. Radelet, *American Sociological Review*, vol. 46 (1981), pp. 918–927.

part (c), and comment on whether the hypothesis seems plausible.

(e) Is it possible that the overall difference between mean income of men and women could be larger than the difference for each academic rank? (It nearly is.) Explain how or how not.

TABLE 10.7

Gender	Professor	Associate	Assistant	Instructor	Overall
Men	115.5	81.2	68.5	59.6	85.5
Women	98.1	75.4	63.6	56.9	70.4

Source: National Center for Education Statistics, *Digest of Education Statistics, 2015*, Table 316.10.

10.19. Table 10.8 relates y = exam score (1 = below median, 2 = above median) to gender, controlling for the subject of the exam (Math, Verbal). Show that subject of exam is a suppressor variable.

TABLE 10.8

	Math		Verbal	
Gender	$y=1$	$y=2$	$y=1$	$y=2$
Females	100	50	50	100
Males	50	100	100	50

10.20. When we analyze data for the census tracts in the greater Los Angeles area, we find no significant correlation between median tax bill and median lot size. Yet a considerable positive correlation occurs when we control for the percentage of the tract used for business. Explain how the percentage of the tract used for businesses could be a suppressor variable, if it is positively correlated with median tax bill and negatively correlated with median lot size.

10.21. According to the U.S. Census Bureau, in 2013 the population median income was estimated to be \$29,127 for white females, \$26,006 for black females, \$41,086 for white males, and \$30,394 for black males. Compare the difference in median incomes between males and females for **(a)** white subjects, **(b)** black subjects. If these are close estimates of the population medians, explain why there is interaction and describe its nature.

10.22. For lower-level managerial employees of a fast-food chain, the prediction equation relating y = annual income (thousands of dollars) to x_1 = number of years of experience on the job is $\hat{y} = 14.2 + 1.1x_1$ for males and $\hat{y} = 14.2 + 0.8x_1$ for females. Explain how these equations show evidence of statistical interaction.

10.23. A study of the association between whether a smoker (yes, no) and whether have had some form of cancer (yes, no) has odds ratio 1.1 for subjects of age less than 30, 2.4 for subjects of age 30 to 50, and 4.3 for subjects of age over 50.

(a) Identify the response variable, explanatory variable, and control variable.

(b) Does the study show evidence of interaction? Explain.

10.24. A study of students at Oregon State University found an association between frequency of church attendance and favorability toward the legalization of marijuana. Both variables were measured in ordered categories. Controlling for gender, the gamma measures for the two partial tables were

Males: gamma = −0.287, standard error = 0.081.
Females: gamma = −0.581, standard error = 0.091.

(a) These results show a slight degree of _____, since the association is somewhat stronger for females than males.

(b) Construct and interpret a 95% confidence interval for the difference between the population gamma values.

Concepts and Applications

10.25. Refer to the `Students` data file (Exercise 1.11 on page 21). Construct partial tables relating opinion about abortion to opinion about life after death, controlling for attendance at religious services, measured using the two categories (Never or occasionally, Most weeks or every week). Prepare a report **(a)** posing and interpreting a possible arrow diagram, before you analyze the data, for relationships among the variables, **(b)** interpreting the sample associations in the bivariate table and the partial tables, **(c)** revising, if necessary, your arrow diagram based on the evidence in the sample data.

10.26. For the `Students` data (Exercise 1.11), are there any pairs of variables for which you expect the association to disappear under control for a third variable? Explain.

10.27. Using the most recent General Social Survey, construct a contingency table relating gender (GSS variable SEX) and party identification (PARTYID). Is there still a gender gap? Control for political ideology (the GSS variable POLVIEWS) by forming partial tables for the most conservative and the most liberal subjects. Does the association seem to persist for these subjects?

10.28. Suppose that x_1 = father's education is positively associated with y = son's income at age 40. However, for the regression analysis conducted separately at fixed levels of x_2 = son's education, the correlation does not differ significantly from zero. Do you think this is more likely to reflect a chain relationship, or a spurious relationship? Explain.

10.29. Table 10.9 shows the mean number of children in Canadian families, classified by whether the family was English speaking or French speaking and by whether the family lived in Quebec or in another province. Let y = number of children in family, x_1 = primary language of family, and x_2 = province (Quebec, others).

(a) Describe the association between y and x_1, based on the overall means in this table.

(b) Describe the association between y and x_1, controlling for x_2.

(c) Explain how it is possible that for each level of province the mean is higher for French-speaking families, yet overall the mean is higher for English-speaking families. (This illustrates *Simpson's paradox*. See Exercise 10.14.)

TABLE 10.9

Province	English	French
Quebec	1.64	1.80
Other	1.97	2.14
Overall	1.95	1.85

10.30. Eighth-grade math scores on the National Assessment of Educational Progress had means of 277 in Nebraska and 271 in New Jersey. For white students, the means were 281 in Nebraska and 283 in New Jersey. For black students, the means were 236 in Nebraska and 242 in New Jersey. For other nonwhite students, the means were 259 in Nebraska and 260 in New Jersey.[8]

(a) Identify the group variable specifying the two states as the explanatory variable. What is the response variable and the control variable?

(b) Explain how it is possible for New Jersey to have the higher mean for each race yet for Nebraska to have the higher mean when the data are combined. (This illustrates *Simpson's paradox*.).

10.31. Example 7.1 (page 194) discussed a study that found that prayer did not reduce the incidence of complications for coronary surgery patients.

(a) Just as association does not imply causality, so does a lack of association not imply a lack of causality, because there may be an alternative explanation. Illustrate this using this study.

(b) A summary of this study in *Time Magazine* (December 4, 2006, p. 87) noted that "… the prayers said by strangers were provided by the clergy and were all identical. Maybe that prevented them from being truly heartfelt. In short, the possible confounding factors in this study made it extraordinarily limited." Explain what "possible confounding" means in the context of this study.

10.32. A study observes that subjects who say they exercise regularly reported only half as many serious illnesses per year, on the average, as those who say they do not exercise regularly. The results section in the article states, "We next analyzed whether age was a confounding variable affecting this association." Explain what this sentence means and how age could potentially explain the association between exercising and illnesses.

10.33. A research study funded by Wobegon Springs Mineral Water, Inc., discovers that the probability that a newborn child has a birth defect is lower for families that regularly buy bottled water than for families that do not. Does this association reflect a causal link between drinking bottled water and a reduction in birth defects? Why or why not?

10.34. The percentage of women who get breast cancer is higher now than a century ago. Suppose that cancer incidence tends to increase with age, and suppose that women tend to live longer now than a century ago. How might a comparison of breast cancer rates now with 100 years ago show different results from these if we control for the age of the woman?

10.35. The crude death rate is the number of deaths in a year, per size of the population, multiplied by 1000. According to the U.S. Bureau of the Census, recently Mexico had a crude death rate of 4.6 (i.e., 4.6 deaths per 1000 population) while the United States had a crude death rate of 8.4. Could the overall death rate be higher in the United States even if the United States has a lower death rate than Mexico for people of each specific age? Explain.

10.36. In the United States, the median age of residents is lowest in Utah. At each age level, the death rate from heart disease is higher in Utah than in Colorado; yet overall, the death rate from heart disease is lower in Utah than Colorado. Are there any contradictions here, or is this possible? Explain.

10.37. A study of the relationship between student's high school GPA and mother's employment (yes, no) suspects an interaction with the gender of a student. Controlling gender, Table 10.10 shows results.

(a) Describe the relationship between mother's employment and GPA for females and for males. Does this sample show evidence of statistical interaction? Explain.

(b) A journal article written about the study states, "Having a mother who is employed outside the home seems to have positive effects on daughter's achievement in high school, but no substantive effect on son's achievement." Explain how Table 10.10 suggests this interpretation.

TABLE 10.10 Mean GPA by Mother's Employment, Controlling for Gender

Gender	Mother Employed	Mother Not Employed
Females	2.94	2.71
Males	2.72	2.74

---

[8] H. Wainer and L. Brown, *American Statistician*, vol. 58 (2004), p. 119

10.38. Give an example of three variables for which the effect of x_1 on y would be

(a) Spurious, disappearing when x_2 is controlled.

(b) Part of a chain relationship, disappearing when a mediator variable x_2 is controlled.

(c) Weakened, but not eliminated, when x_2 is controlled.

(d) Unaffected by controlling x_2.

(e) Different at different levels of x_2 (i.e., showing interaction).

(f) Confounded with the effect of x_2.

10.39. Exercise 7.17 on page 219 mentioned a study of compulsive buying behavior that conducted a national telephone survey. The study found that lower-income subjects
were more likely to be compulsive buyers. They reported, "Compulsive buyers did not differ significantly from other respondents in mean total credit card balances, but the compulsive buyers' lower income was a confounding factor." Explain what it means to say that income was a confounding factor, and explain why a comparison of the mean total credit card balances between compulsive and noncompulsive buyers could change depending on whether income is controlled.

10.40. A study[9] reported a correlation of 0.68 between scores on an index of depression and scores on an index that measures the amount of saturated fat intake. True or false: You can conclude that if you increase your saturated fat intake by a standard deviation, your degree of depression will increase by more than half a standard deviation.

10.41. For recent U.S. presidential elections, in each state wealthier voters tend to be more likely to vote Republican, yet states that are wealthier in an aggregate sense are more likely to have more Democrat than Republican votes (Gelman and Hill 2007, Section 14.2). Sketch a plot that illustrates how this instance of Simpson's paradox could occur.

Select the best response(s) in Exercises 10.42–10.45. (More than one response may be correct.)

10.42. For all court trials about homicides in Florida in a certain period, the difference between the proportions of whites and blacks receiving the death penalty was 0.026 when the victim was black and -0.077 when the victim was white.[10] This shows evidence of (**a**) a spurious association, (**b**) statistical interaction, (**c**) a chain relationship, (**d**) all of these.

10.43. Statistical interaction refers to which of the following?

(a) Association exists between two variables.

(b) The effect of an explanatory variable on a response variable changes greatly over the levels of a control variable.

(c) The partial association is the same at each level of the control variable, but it is different from the overall bivariate association, ignoring the control variable.

(d) For a collection of three variables, each pair of variables is associated.

(e) All of the above.

10.44. Example 9.10 (page 280) used a data set on house sales to regress y = selling price of home (in dollars) to x = size of house (in square feet). The prediction equation was $\hat{y} = -50{,}926 + 126.6x$. Now, we regard size of house as x_1 and also consider x_2 = whether the house is new (yes or no). The prediction equation relating $\hat{y}$ to x_1 has slope 161 for new homes and 109 for older homes. This gives evidence:

(a) of interaction between x_1 and x_2 in their effects on y.

(b) of a spurious association between selling price and size.

(c) of a chain relationship, whereby whether new affects size which affects selling price.

(d) that size of house does not have a causal effect on price.

10.45. Consider the relationship between y = political party preference (Democrat, Republican) and x_1 = race (Black, White) and x_2 = gender. There is an association between y and both x_1 and x_2, with the Democrat preference being more likely for blacks than whites and for women than men.

(a) x_1 and x_2 are probably independent causes of y.

(b) The association between y and x_1 is probably spurious, controlling for x_2.

(c) Since both variables affect y, there is probably interaction.

(d) The variables probably satisfy a chain relationship.

(e) Race is probably a suppressor variable.

(f) None of the above.

[9] In *Behavior Modification*, vol. 29 (2005), p. 677.
[10] M. L. Radelet and G. L. Pierce, *Florida Law Review*, vol. 43 (1991).

MULTIPLE REGRESSION AND CORRELATION

Chapter 9 introduced regression modeling of the relationship between two quantitative variables. Multivariate relationships require more complex models, containing several explanatory variables. Some of these may be predictors of theoretical interest, and some may be control variables.

This chapter extends the regression model to a **multiple regression model** that can have multiple explanatory variables. Such a model provides better predictions of y than does a model with a single explanatory variable. The model also can analyze relationships between variables while controlling for other variables. This is important because Chapter 10 showed that after controlling for a variable, an association can appear quite different from when the variable is ignored.

After defining the multiple regression model and showing how to interpret its parameters, we present correlation and r-squared measures that describe association between y and a set of explanatory variables, and we present inference procedures for the model parameters. We then show how to allow *statistical interaction* in the model, whereby the effect of an explanatory variable changes according to the value of another explanatory variable. A significance test can analyze whether a complex model, such as one permitting interaction, provides a better fit than a simpler model. The final two sections introduce correlation-type measures that summarize the association between the response variable and an explanatory variable while controlling other variables.

11.1 The Multiple Regression Model

Chapter 9 modeled the relationship between the explanatory variable x and the mean of the response variable y by the straight-line (linear) equation $E(y) = \alpha + \beta x$. We refer to this model containing a *single* predictor as a **bivariate model**, because it contains only two variables.

THE MULTIPLE REGRESSION FUNCTION

With two explanatory variables, denoted by x_1 and x_2, the bivariate regression function generalizes to the **multiple regression function**

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2.$$

In this equation, α, β_1, and β_2 are parameters discussed below. For particular values of x_1 and x_2, the equation specifies the population mean of y for all subjects with those values of x_1 and x_2. With additional explanatory variables, each has a βx term, such as $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ with four explanatory variables.

The multiple regression function is more difficult to portray graphically than the bivariate regression function. With two explanatory variables, the x_1 and x_2 axes are perpendicular but lie in a horizontal plane and the y axis is vertical and perpendicular

to both the x_1 and x_2 axes. The equation $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ traces a plane (a flat surface) cutting through three-dimensional space, as Figure 11.1 portrays.

FIGURE 11.1: Graphical Depiction of a Multiple Regression Function with Two Explanatory Variables
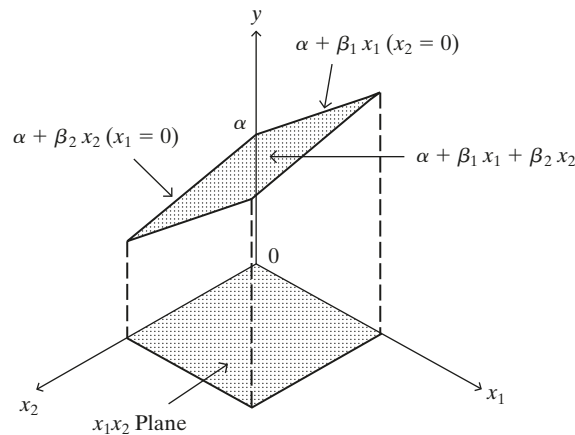
The simplest interpretation treats all but one explanatory variable as control variables and fixes them at particular levels. This leaves an equation relating the mean of y to the remaining explanatory variables.

Example 11.1

Do Higher Levels of Education Cause Higher Crime Rates? The `Florida` data file at the text website, shown partly in Table 9.15 on page 295, contains data for the 67 counties in the state of Florida on $y =$ crime rate (annual number of crimes per 1000 population), $x_1 =$ education (percentage of adult residents having at least a high school education), and $x_2 =$ urbanization (percentage living in an urban environment). The bivariate relationship between crime rate and education is approximated by $E(y) = -51.3 + 1.5x_1$. Surprisingly, the association is moderately *positive*, the correlation being $r = 0.47$. As the percentage of county residents having at least a high school education increases, so does the crime rate.

A closer look at the data reveals strong positive associations between crime rate and urbanization ($r = 0.68$) and between education and urbanization ($r = 0.79$). This suggests that the association between crime rate and education may be spurious. Perhaps urbanization is a common causal factor. See Figure 11.2. As urbanization increases, both crime rate and education increase, resulting in a positive correlation between crime rate and education.

FIGURE 11.2: The Positive Association between Crime Rate and Education May Be Spurious, Explained by the Effects of Urbanization on Each
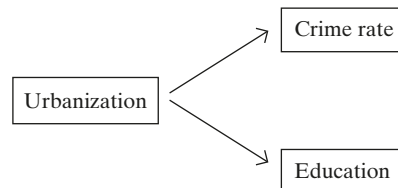
The relation between crime rate and both explanatory variables considered together is approximated by the multiple regression function

$$E(y) = 58.9 - 0.6x_1 + 0.7x_2.$$

For instance, the expected crime rate for a county at the mean levels of education ($\bar{x}_1 = 70$) and urbanization ($\bar{x}_2 = 50$) is $E(y) = 58.9 - 0.6(70) + 0.7(50) = 52$ annual crimes per 1000 population.

Let's study the effect of x_1, controlling for x_2. We first set x_2 at its mean level of 50. Then, the relationship between crime rate and education is

$$E(y) = 58.9 - 0.6x_1 + 0.7(50) = 58.9 - 0.6x_1 + 35.0 = 93.9 - 0.6x_1.$$

Figure 11.3 plots this line. Controlling for x_2 by fixing it at 50, the relationship between crime rate and education is negative, rather than positive. The slope decreased and changed sign from 1.5 in the bivariate relationship to -0.6. At this fixed level of urbanization, a negative relationship exists between education and crime rate. We use the term *partial* regression equation to distinguish the equation $E(y) = 93.9 - 0.6x_1$ from the regression equation $E(y) = -51.3 + 1.5x_1$ for the *bivariate* relationship between y and x_1. The *partial* regression equation refers to *part* of the potential observations, in this case counties having $x_2 = 50$.

FIGURE 11.3: Partial Relationships between $E(y)$ and x_1 for the Multiple Regression Equation $E(y) = 58.9 - 0.6x_1 + 0.7x_2$. These partial regression equations fix x_2 to equal 50 or 40.
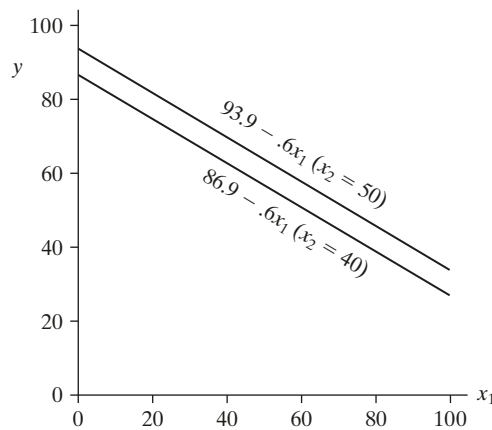
Next we fix x_2 at a different level, say $x_2 = 40$ instead of 50. Then, you can check that $E(y) = 86.9 - 0.6x_1$. Thus, decreasing x_2 by 10 units shifts the partial line relating y to x_1 downward by $10\beta_2 = 7.0$ units (see Figure 11.3). The slope of -0.6 for the partial relationship remains the same, so the line is parallel to the one with $x_2 = 40$. Setting x_2 at a variety of values yields a collection of parallel lines, each having slope $\beta_1 = -0.6$.

Similarly, setting x_1 at a variety of values yields a collection of parallel lines, each having slope 0.7, relating the mean of y to x_2. In other words, controlling for education, the slope of the partial relationship between crime rate and urbanization is $\beta_2 = 0.7$.

In summary, education has an overall positive effect on crime rate, but it has a negative effect when controlling for urbanization. The partial association has the opposite direction from the bivariate association. This is called ***Simpson's paradox***. Figure 11.4 illustrates how this happens. It shows the scatterplot relating crime rate to education, portraying the overall positive association between these variables. The diagram circles the 19 counties that are highest in urbanization. That subset of points for which urbanization is nearly constant has a negative trend between crime rate and education. The high positive association between education and urbanization is reflected by the fact that most of the highlighted observations that are highest on urbanization also have high values on education. ■

FIGURE 11.4: Scatterplot
Relating Crime Rate and
Education. The circled
points are the counties
highest in urbanization. A
regression line fitting the
circled points has negative
slope, even though the
regression line passing
through *all* the points has
positive slope (Simpson's
paradox).
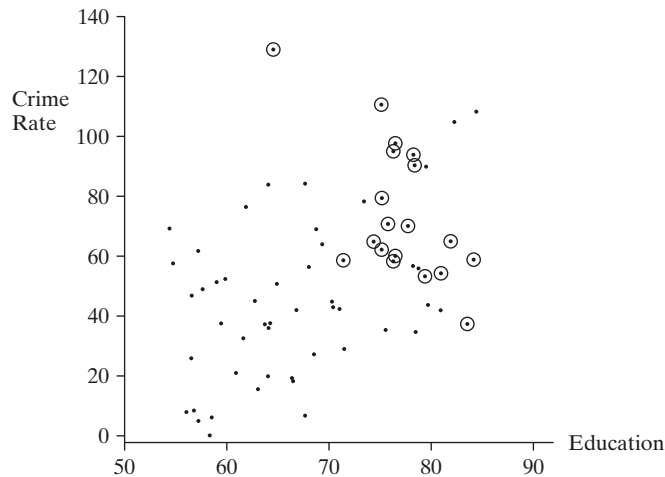
INTERPRETATION OF REGRESSION COEFFICIENTS

We have seen that for a fixed value of x_2, the equation $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ simplifies to a straight-line equation in x_1 with slope β_1. The slope is the same for each fixed value of x_2. When we fix the value of x_2, we are holding it constant: We are *controlling* for x_2. That's the basis of the major difference between the interpretation of slopes in multiple regression and in bivariate regression:

- In *multiple regression*, a slope describes the effect of an explanatory variable while *controlling* effects of the other explanatory variables in the model.

- *Bivariate regression* has only a single explanatory variable. So, a slope in bivariate regression describes the effect of that variable while *ignoring* all other possible explanatory variables.

The parameter β_1 measures the *partial effect* of x_1 on y, that is, the effect of a one-unit increase in x_1, holding x_2 constant. The partial effect of x_2 on y, holding x_1 constant, has slope β_2. Similarly, for the multiple regression model with *several* explanatory variables, the beta coefficient of a particular explanatory variable describes the change in the mean of y for a one-unit increase in that variable, controlling for the other variables in the model. The parameter α represents the mean of y when each explanatory variable equals 0.

The parameters $\beta_1, \beta_2, \ldots$ are called ***partial regression coefficients***. The adjective *partial* distinguishes these parameters from the regression coefficient β in the *bivariate* model $E(y) = \alpha + \beta x$, which *ignores* rather than *controls* effects of other explanatory variables.

A partial slope in a multiple regression model usually differs from the slope in the bivariate model for that explanatory variable, but it need not. With two explanatory variables, the partial slopes and bivariate slopes are equal if the correlation between x_1 and x_2 equals 0. When x_1 and x_2 are independent causes of y, the effect of x_1 on y does not change when we control for x_2.

LIMITATIONS OF THIS MULTIPLE REGRESSION MODEL

In interpreting partial regression coefficients in observational studies, we need to be cautious not to regard the estimated effects as implying causal relations. For example, for a sample of college students, suppose y = math achievement test score (scale of

0 to 100) and the explanatory variables are $x_1 =$ number of years of math education, $x_2 =$ mother's number of years of math education, and $x_3 =$ GPA, and we fit the multiple regression model

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Suppose that the estimate of β_1 is 5. In interpreting the effect of x_1, we might say, "A one-year increase in math education corresponds to an increase in the predicted math achievement test score of 5, controlling for the mother's math education and GPA." However, this does not imply that if a student attains another year of math education, her or his math achievement test score is predicted to change by 5. To validly make such a conclusion, we'd need to conduct an experiment that adds a year of math education for each student and then observes the results. Otherwise, a higher mean test score at a higher math education level could at least partly reflect the correlation of several other variables with both math test score and math education level, such as the student's IQ, father's number of years of math education, and number of years of science courses.

What the above interpretation actually means is this: "The difference between the estimated mean math achievement test score of a subpopulation of students having a certain number of years of math education and a subpopulation having one fewer year of math education equals 5, when we control (keep constant) the mother's math education and GPA." However, we need to be cautious even with this interpretation. It is unnatural and even inconsistent with the data for some observational studies to envision increasing one explanatory variable while keeping all the others fixed. For example, x_1 and x_2 are likely to be positively correlated, so increases in x_1 naturally tend to occur with increases in x_2. In some data sets, one might not even observe a one-unit range in an explanatory variable when the other explanatory variables are all held constant. As an extreme example, suppose $y =$ height, $x_1 =$ length of left leg, and $x_2 =$ length of right leg. The correlation between x_1 and x_2 is extremely close to 1. It does not make much sense to imagine how y changes as x_1 changes while x_2 is controlled.

Because of this limitation, some methodologists prefer to use more cautious wording than "controlling." In interpreting an estimate of 5 for β_1, they would say, "The difference between the estimated mean math achievement test score of a subpopulation of students having a certain number of years of math education and a subpopulation having one fewer year equals 5, when both subpopulations have the same estimated value for $\beta_2 x_{i2} + \beta_3 x_{i3}$." More concisely, "The effect of the number of years of math education on the estimated mean math achievement test score equals 5, *adjusting* for student's age and mother's math education." In the rest of the text, we will use the simpler "controlling" wording, but we should keep in mind its limitations.

Finally, this multiple regression model also has a structural limitation. It assumes that the slope of the partial relationship between y and each explanatory variable is identical for *all* combinations of values of the other explanatory variables. This means that the model is appropriate when there is a lack of *statistical interaction*, in the sense explained in Section 10.3 (page 306). If the true partial slope between y and x_1 is very different at $x_2 = 50$ than at $x_2 = 40$, for example, we need a more complex model. Section 11.4 will show this model and Section 11.5 will show how to analyze whether it fits significantly better.

PREDICTION EQUATION AND RESIDUALS

Corresponding to the multiple regression equation, software finds a prediction equation by estimating the model parameters using sample data. In general, we let p denote the number of explanatory variables.

Notation for Prediction Equation

> The prediction equation that estimates the multiple regression equation
> $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ is denoted by
> $\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p.$

We use statistical software to find the prediction equation. The calculation formulas are complex and are not shown in this text.

To get the predicted value of y for a subject, we substitute the x-values for that subject into the prediction equation. Like the bivariate model, the multiple regression model has **residuals** that measure prediction errors. For a subject with predicted response $\hat{y}$ and observed response y, the residual is $y - \hat{y}$. The next section shows an example.

The **sum of squared errors** (SSE),

$$SSE = \sum (y - \hat{y})^2,$$

summarizes the closeness of fit of the prediction equation to the response data. Most software calls SSE the **residual sum of squares**. The formula for SSE is the same as in Chapter 9. The only difference is that the predicted value $\hat{y}$ results from using *several* explanatory variables instead of just a single one.

The parameter estimates in the prediction equation satisfy the **least squares** criterion: The prediction equation has the *smallest* SSE value of all possible equations of the form $\hat{y} = a + b_1 x_1 + \cdots + b_p x_p$.

Example 11.2

Multiple Regression for Mental Health Study   A study in Alachua County, Florida, investigated the relationship between certain mental health indices and several explanatory variables. Primary interest focused on an index of mental impairment, which incorporates various dimensions of psychiatric symptoms, including aspects of anxiety and depression. This measure, which is the response variable y, ranged from 17 to 41 in the sample. Higher scores indicate greater psychiatric impairment.

The two explanatory variables used here are x_1 = life events score and x_2 = socioeconomic status (SES). The life events score is a composite measure of both the number and severity of major life events the subject experienced within the past three years. These events range from severe personal disruptions, such as a death in the family, a jail sentence, or an extramarital affair, to less severe events, such as getting a new job, the birth of a child, moving within the same city, or having a child marry. This measure ranged from 3 to 97 in the sample. A high score represents a greater number and/or greater severity of these life events. The SES score is a composite index based on occupation, income, and education. Measured on a standard scale, it ranged from 0 to 100. The higher the score, the higher the status.

Table 11.1 shows data[1] on the three variables for a random sample of 40 adults in the county. Table 11.2 summarizes the sample means and standard deviations of the three variables. ∎

SCATTERPLOT MATRIX FOR BIVARIATE RELATIONSHIPS

Plots of the data provide an informal check of whether the relationships are linear. Software can construct scatterplots on a single diagram for each pair of the variables.

---

[1]These data are based on a much larger survey. Thanks to Dr. Charles Holzer for permission to use the study as the basis of this example.

TABLE 11.1: Mental Data File from the Text Website with y = Mental Impairment, x_1 = Life Events, and x_2 = Socioeconomic Status

y	x_1	x_2	y	x_1	x_2	y	x_1	x_2
17	46	84	26	50	40	30	44	53
19	39	97	26	48	52	31	35	38
20	27	24	26	45	61	31	95	29
20	3	85	27	21	45	31	63	53
20	10	15	27	55	88	31	42	7
21	44	55	27	45	56	32	38	32
21	37	78	27	60	70	33	45	55
22	35	91	28	97	89	34	70	58
22	78	60	28	37	50	34	57	16
23	32	74	28	30	90	34	40	29
24	33	67	28	13	56	41	49	3
24	18	39	28	40	56	41	89	75
25	81	87	29	5	40			
26	22	95	30	59	72			

TABLE 11.2: Sample Means and Standard Deviations of Mental Impairment, Life Events, and Socioeconomic Status (SES)

Variable	Mean	Standard Deviation
Mental impairment	27.30	5.46
Life events	44.42	22.62
SES	56.60	25.28

Figure 11.5 shows the plots for the variables from Table 11.1. This type of plot is called a ***scatterplot matrix***. Like a correlation matrix, it shows each pair of variables twice. In one plot, a variable is on the *y*-axis and in the other it is on the *x*-axis. Mental impairment (the response variable) is on the *y*-axis for the plots in the first row of Figure 11.5, so these are the plots of interest to us. The plots show no evidence of nonlinearity, and models with linear effects seem appropriate. The plots suggest that life events have a mild positive effect and SES has a mild negative effect on mental impairment.

FIGURE 11.5: Scatterplot Matrix: Scatterplots for Pairs of Variables from Table 11.1
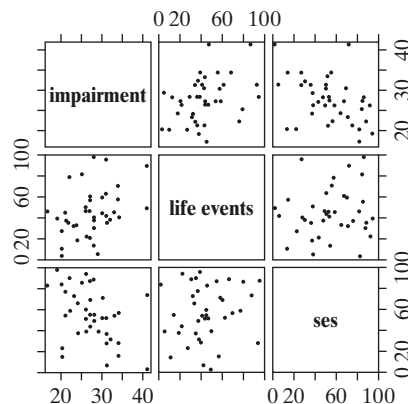
PARTIAL PLOTS FOR PARTIAL RELATIONSHIPS

The multiple regression model states that each explanatory variable has a linear effect with common slope, controlling for the other predictors. Although the regression formula is relatively simple, this itself is quite a strong assumption. To check it, we can compare the fit of the model to the fit of a more complex model, as we'll explain in Section 11.4. We can also plot y versus each predictor, for subsets of points that are nearly constant on the other predictors. With a single control variable, for example, we could sort the observations into four groups using the quartiles as boundaries, and then either construct four separate scatterplots or mark the observations on a single scatterplot according to their group.

With several control variables, however, keeping them all nearly constant can reduce the sample to relatively few observations and is impractical. A more informative single picture is provided by the ***partial regression plot*** (also called *added-variable plot*). It displays the relationship between the response variable and an explanatory variable after removing the effects of the other predictors in the multiple regression model. It does this by plotting the residuals from models using these two variables as responses and the other explanatory variables as predictors.

Here is how software constructs the partial regression plot for the effect of x_1 when the multiple regression model also has explanatory variables x_2 and x_3. It finds the residuals from the models (i) using x_2 and x_3 to predict y and (ii) using x_2 and x_3 to predict x_1. Then it plots the residuals from the first analysis (on the y-axis) against the residuals from the second analysis. For these residuals, the effects of x_2 and x_3 are removed. The least squares slope for the points in this plot is necessarily the same as the estimated partial slope b_1 for the multiple regression model.

Figure 11.6 shows a partial regression plot for y = mental impairment and x_1 = life events, controlling for x_2 = SES. It plots the residuals on the y-axis from $\hat{y} = 32.2 - 0.086x_2$ against the residuals on the x-axis from $\hat{x}_1 = 38.2 + 0.110x_2$. Both axes have negative and positive values, because they refer to residuals. Recall that residuals (prediction errors) can be positive or negative, and have a mean of 0. Figure 11.6 suggests that the partial effect of life events is approximately linear and is positive.

FIGURE 11.6: Partial Regression Plot for Mental Impairment and Life Events, Controlling for SES. This plots the residuals from regressing mental impairment on SES against the residuals from regressing life events on SES.
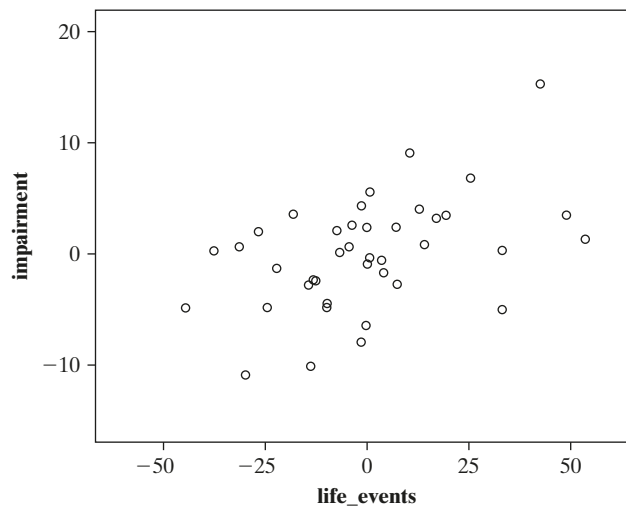
Figure 11.7 shows the partial regression plot for SES. It shows that its partial effect is also approximately linear but is negative.

FIGURE 11.7: Partial Regression Plot for Mental Impairment and SES, Controlling for Life Events. This plots the residuals from regressing mental impairment on life events against the residuals from regressing SES on life events.
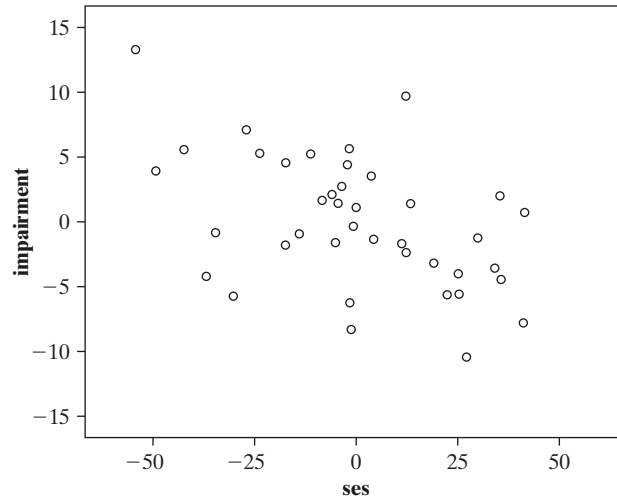
SOFTWARE OUTPUT FOR MENTAL IMPAIRMENT EXAMPLE

Tables 11.3 and 11.4 are Stata outputs of the coefficients table for the bivariate relationships between mental impairment and the separate explanatory variables. The estimated regression coefficients fall in the column labeled *Coef.* The prediction equations are

$$\hat{y} = 23.31 + 0.090x_1 \text{ and } \hat{y} = 32.17 - 0.086x_2.$$

In the sample, mental impairment is positively related to life events, since the coefficient of x_1 (0.090) is positive. The greater the number and severity of life events in the previous three years, the higher the mental impairment (i.e., the poorer the mental health) tends to be. Mental impairment is negatively related to socioeconomic status. The correlations between mental impairment and the explanatory variables are modest, 0.372 for life events and -0.399 for SES (the appropriate square roots of the r^2-values reported).

TABLE 11.3: Bivariate Regression Analysis for y = Mental Impairment and x_1 = Life Events from `Mental` Data File

impair	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
				R-squared = 0.1385		
life	.0898257	.0363349	2.47	0.018	.0163	.1634
_cons	23.30949	1.806751	12.90	0.000	19.65	26.97

TABLE 11.4: Bivariate Regression Analysis for y = Mental Impairment and x_2 = Socioeconomic Status (SES) from `Mental` Data File

impair	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
				R-squared = 0.1589		
ses	-.086078	.0321317	-2.68	0.011	-.1511	-.0210
_cons	32.17201	1.987649	16.19	0.000	28.148	36.196

Table 11.5 shows output for the multiple regression model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$. The prediction equation is

$$\hat{y} = a + b_1 x_1 + b_2 x_2 = 28.230 + 0.103x_1 - 0.097x_2.$$

TABLE 11.5: Fit of Multiple Regression Model for y = Mental Impairment, x_1 = Life Events, and x_2 = Socioeconomic Status from `Mental` Data File

```
                                       R-squared = 0.3392
impair |      Coef.   Std. Err.        t    P>|t|   [95% Conf. Interval]
  life |   .1032595   .0324995      3.18    0.003      .0374      .1691
   ses |  -.0974755   .0290848     -3.35    0.002     -.1564     -.0385
 _cons |   28.22981   2.174222     12.98    0.000      23.82      32.64
```

Controlling for SES, the sample relationship between mental impairment and life events is positive, since the coefficient of life events ($b_1 = 0.103$) is positive. The estimated mean of mental impairment increases by about 0.1 for every one-unit increase in the life events score, controlling for SES. Since $b_2 = -0.097$, a negative association exists between mental impairment and SES, controlling for life events. For example, over the 100-unit range of potential SES values (from a minimum of 0 to a maximum of 100), the estimated mean mental impairment changes by $100(-0.097) = -9.7$. Since mental impairment ranges only from 17 to 41 with a standard deviation of 5.5, a decrease of 9.7 points in the mean is noteworthy.

From Table 11.1, the first subject in the sample had $y = 17$, $x_1 = 46$, and $x_2 = 84$. This subject's predicted mental impairment is

$$\hat{y} = 28.230 + 0.103(46) - 0.097(84) = 24.8.$$

The prediction error (residual) is $y - \hat{y} = 17 - 24.8 = -7.8$.

Table 11.6 summarizes some results of the regression analyses. It shows standard errors in parentheses below the parameter estimates. The partial slopes for the multiple regression model are similar to the slopes for the bivariate models. In each case, the introduction of the second explanatory variable does little to alter the effect of the other one. This suggests that these explanatory variables may have nearly independent sample effects on y. In fact, the sample correlation between x_1 and x_2 is only 0.123. The next section shows how to interpret the R^2-value listed for the multiple regression model.

TABLE 11.6: Summary of Regression Models for Mental Impairment

	Explanatory Variables in Regression Model		
Effect	Multiple	Life Events	SES
Intercept	28.230	23.309	32.172
Life events	0.103	0.090	—
	(0.032)	(0.036)	
SES	−0.097	—	−0.086
	(0.029)		(0.032)
R^2	0.339	0.138	0.159

11.2 Multiple Correlation and R^2

The correlation r and its square describe strength of linear association for bivariate relationships. This section presents analogous measures for the multiple regression model. They describe the strength of association between y and the set of explanatory variables acting together as predictors in the model.

THE MULTIPLE CORRELATION R

The explanatory variables collectively are strongly associated with y if the observed y-values correlate highly with the $\hat{y}$-values from the prediction equation. The correlation between the observed and predicted values summarizes this association.

Multiple Correlation

> The sample ***multiple correlation*** for a regression model, denoted by R, is the correlation between the observed y-values and the predicted $\hat{y}$-values.

For each subject, the prediction equation provides a predicted value $\hat{y}$. So, each subject has a y-value and a $\hat{y}$-value. For the first three subjects in Table 11.1, the observed and predicted y-values are

y	$\hat{y}$
17	24.8
19	22.8
20	28.7

The sample correlation computed between all 40 of the y- and $\hat{y}$-values is R, the multiple correlation. The larger the value of R, the better the predictions of y by the set of explanatory variables.

The predicted values cannot correlate negatively with the observed values. The predictions must be at least as good as the sample mean $\bar{y}$, which is the prediction when all partial slopes = 0, and $\bar{y}$ has zero correlation with y. So, R always falls between 0 and 1. In this respect, the correlation between y and $\hat{y}$ differs from the correlation between y and an explanatory variable x, which falls between -1 and $+1$.

R^2: THE COEFFICIENT OF MULTIPLE DETERMINATION

Another measure uses the *proportional reduction in error* concept, generalizing r^2 for bivariate models. This measure summarizes the relative improvement in predictions using the prediction equation instead of $\bar{y}$. It has the following elements:

Rule 1 (Predict y without using $x_1, \ldots, x_p$): The best predictor is then the sample mean, $\bar{y}$.

Rule 2 (Predict y using $x_1, \ldots, x_p$): The best predictor is the prediction equation $\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$.

Prediction Errors: The prediction error for a subject is the difference between the observed and predicted values of y. With rule 1, the error is $y - \bar{y}$. With rule 2, it is the residual $y - \hat{y}$. In either case, we summarize the error by the sum of the squared prediction errors. For rule 1, this is TSS $= \sum (y - \bar{y})^2$, the *total sum of squares*. For rule 2, it is SSE $= \sum (y - \hat{y})^2$, the sum of squared errors using the prediction equation, which is the *residual sum of squares*.

Definition of Measure: The proportional reduction in error from using the prediction equation $\hat{y} = a + b_1 x_1 + \cdots + b_p x_p$ instead of $\bar{y}$ to predict y is ***R-squared***, also called the ***coefficient of multiple determination***.

R-Squared: The Coefficient of Multiple Determination

$$R^2 = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = \frac{\sum (y - \bar{y})^2 - \sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

R^2 measures the proportion of the total variation in y that is explained by the predictive power of all the explanatory variables, through the multiple regression

model. The symbol reflects that R^2 *is the square of the multiple correlation R*. The uppercase notation R^2 distinguishes this measure from r^2 for the bivariate model. Their formulas are identical, and r^2 is the special case of R^2 applied to a regression model with one explanatory variable. For the multiple regression model to be useful for prediction, it should provide improved predictions relative not only to $\bar{y}$ but also to the separate bivariate models for y and each explanatory variable.

Example 11.3

Multiple Correlation and R^2 for Mental Impairment   For the data on $y =$ mental impairment, $x_1 =$ life events, and $x_2 =$ socioeconomic status in Table 11.1, Table 11.5 showed some output. Software (SPSS) also reports ANOVA tables with sums of squares and shows R and R^2. See Table 11.7.

TABLE 11.7: ANOVA Table and Model Summary for Regression of Mental Impairment on Life Events and Socioeconomic Status from `Mental` Data File

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	394.238	2	197.119	9.495	.000
Residual	768.162	37	20.761		
Total	1162.400	39			

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.582	.339	.303	4.556

Predictors: (Constant), SES, LIFE
Dependent Variable: IMPAIR

From the *Sum of Squares* column, the total sum of squares is TSS $= \sum(y-\bar{y})^2 =$ 1162.4, and the residual sum of squares from using the prediction equation to predict y is SSE $= \sum(y-\hat{y})^2 = 768.2$. Thus,

$$R^2 = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = \frac{1162.4 - 768.2}{1162.4} = 0.339.$$

Using life events and SES together to predict mental impairment provides a 33.9% reduction in the prediction error relative to using only $\bar{y}$. The multiple regression model provides a substantially larger reduction in error than either bivariate model (Table 11.6 reported r^2-values of 0.138 and 0.159 for them). It is more useful than those models for predictive purposes.

The multiple correlation between mental impairment and the two explanatory variables is $R = +\sqrt{0.339} = 0.582$. This equals the correlation between the observed y- and predicted $\hat{y}$-values for the model. ∎

PROPERTIES OF R AND R^2

The properties of R^2 are similar to those of r^2 for bivariate models.

- R^2 falls between 0 and 1.

- The larger the value of R^2, the better the set of explanatory variables $(x_1, \ldots, x_p)$ collectively predicts y.

- $R^2 = 1$ only when all the residuals are 0, that is, when all $y = \hat{y}$, so that predictions are perfect and SSE $= 0$.

- $R^2 = 0$ when the predictions do not vary as any of the x-values vary. In that case, $b_1 = b_2 = \cdots = b_p = 0$, and $\hat{y}$ is identical to $\bar{y}$, since the explanatory variables do not add any predictive power. The correlation is then 0 between y and each explanatory variable.

- R^2 cannot decrease when we add an explanatory variable to the model. It is impossible to explain *less* variation in y by adding explanatory variables to a regression model.

- R^2 for the multiple regression model is at least as large as the r^2-values for the separate bivariate models. That is, R^2 for the multiple regression model is at least as large as $r^2_{yx_1}$ for y as a linear function of x_1, $r^2_{yx_2}$ for y as a linear function of x_2, and so forth.

- R^2 tends to overestimate the population value, because the sample data fall closer to the sample prediction equation than to the true population regression equation. Most software also reports a less biased estimate, called ***adjusted*** R^2. Exercise 11.61 shows its formula. For the mental impairment example, Table 11.7 reports its value of 0.303, compared to ordinary $R^2 = 0.339$.

Properties of the multiple correlation R follow directly from the ones for R^2, since R is the positive square root of R^2. For instance, R for the model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ is at least as large as R for the model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$.

The numerator of R^2, TSS $-$ SSE, summarizes the variation in y explained by the multiple regression model. This difference, which equals $\sum(\hat{y} - \bar{y})^2$, is called the ***regression sum of squares***. The ANOVA table in Table 11.7 lists the regression sum of squares as 394.2. (Some software, such as Stata and SAS, labels this the *Model* sum of squares.) The total sum of squares TSS of the y-values about $\bar{y}$ partitions into the variation explained by the regression model (regression sum of squares) plus the variation not explained by the model (the residual sum of squares, SSE).

MULTICOLLINEARITY WITH MANY EXPLANATORY VARIABLES

When a study has many explanatory variables but the correlations among them are strong, once you have included a few of them in the model, R^2 usually doesn't increase much more when you add additional ones. For example, for the `Houses` data file at the text website (introduced in Example 9.10 on page 280), r^2 is 0.71 with the house's tax assessment as a predictor of selling price. Then, R^2 increases to 0.77 when we add house size as a second predictor. But then it increases only to 0.79 when we add number of bathrooms, number of bedrooms, and whether the house is new as additional predictors.

When R^2 does not increase much, this does not mean that the additional variables are uncorrelated with y. It means merely that they don't add much new power for predicting y, given the values of the explanatory variables already in the model. These other variables may have small associations with y, given the variables already in the model. This often happens in social science research when the explanatory variables are highly correlated, no one having much unique explanatory power. Section 14.3 discusses this condition, called ***multicollinearity***.

Figure 11.8, which portrays the portion of the total variability in y explained by each of three explanatory variables, shows a common occurrence. The size of the set for an explanatory variable in this figure represents the size of its r^2-value in predicting y. The amount a set for an explanatory variable overlaps with the set for another explanatory variable represents its association with that predictor. The part of the set for an explanatory variable that does not overlap with other sets represents the

part of the variability in y explained uniquely by that explanatory variable. In Figure 11.8, all three explanatory variables have moderate associations with y, and together they explain considerable variation. Once x_1 and x_2 are in the model, however, x_3 explains little additional variation in y, because of its strong correlations with x_1 and x_2. Because of this overlap, R^2 increases only slightly when x_3 is added to a model already containing x_1 and x_2.

FIGURE 11.8: R^2 Does Not Increase Much when x_3 Is Added to the Model Already Containing x_1 and x_2
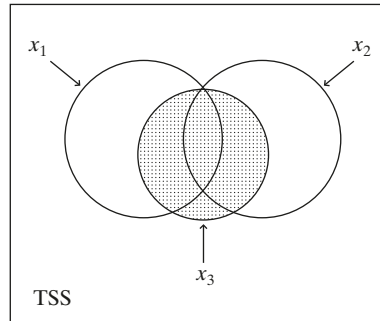
For predictive purposes, we gain little by adding explanatory variables to a model that are strongly correlated with ones already in the model, since R^2 will not increase much. Ideally, we should use explanatory variables having weak correlations with each other but strong correlations with y. In practice, this is not always possible, especially when we include certain variables in the model for theoretical reasons.

The sample size you need to do a multiple regression well gets larger when you want to use more explanatory variables. Technical difficulties caused by multi-collinearity are less severe for larger sample sizes. Ideally, the sample size should be at least about 10 times the number of explanatory variables (e.g., at least about 40 for 4 explanatory variables).

11.3 Inferences for Multiple Regression Coefficients

To make inferences about the parameters in the multiple regression function

$$E(y) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p,$$

we formulate the entire *multiple regression model*. This consists of this equation together with a set of assumptions:

- The population distribution of y is normal, for each combination of values of $x_1, \ldots, x_p$.

- The standard deviation, σ, of the conditional distribution of responses on y is the same at each combination of values of $x_1, \ldots, x_p$.

- The sample is randomly selected.

Under these assumptions, the true sampling distributions exactly equal those quoted in this section. In practice, the assumptions are never satisfied perfectly. Two-sided inferences are robust to the normality and common σ assumptions. More important are the assumptions of randomization and that the regression function describes well how the mean of y depends on the explanatory variables. We'll see ways to check the latter assumption in Sections 11.4 and 14.2.

Multiple regression analyses use two types of significance tests. The first is a global test of independence. It checks whether *any* of the explanatory variables are

statistically related to y. The second studies the partial regression coefficients individually, to assess which explanatory variables have significant partial effects on y.

TESTING THE COLLECTIVE INFLUENCE OF THE EXPLANATORY VARIABLES

Do the explanatory variables collectively have a statistically significant effect on the response variable? We check this by testing

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

that the mean of y does not depend on the values of $x_1, \ldots, x_p$. Under the inference assumptions, this states that y is statistically independent of all p explanatory variables.

The alternative hypothesis is

$$H_a: \text{At least one } \beta_i \neq 0.$$

This states that *at least one* explanatory variable is associated with y, controlling for the others. The test judges whether using $x_1, \ldots, x_p$ together to predict y, with the prediction equation $\hat{y} = a + b_1 x_1 + \cdots + b_p x_p$, is significantly better than using $\bar{y}$.

These hypotheses about $\{\beta_i\}$ are equivalent to

H_0: Population multiple correlation $= 0$ and H_a: Population multiple correlation > 0.

The equivalence occurs because the multiple correlation equals 0 only in those situations in which all the partial regression coefficients equal 0. Also, H_0 is equivalent to H_0: population R-squared $= 0$.

For these hypotheses about the p predictors, the test statistic is

$$F = \frac{R^2/p}{(1 - R^2)/[n - (p + 1)]}.$$

The sampling distribution of this statistic is called the **F distribution**.

THE *F* DISTRIBUTION

The symbol for the F test statistic and its distribution honors the most eminent statistician in history, R. A. Fisher, who discovered the F distribution in 1922. Like the chi-squared distribution, the F distribution can take only nonnegative values and it is somewhat skewed to the right. Figure 11.9 illustrates this.

FIGURE 11.9: The *F* Distribution and the *P*-Value for *F* Tests. Larger *F*-values give stronger evidence against H_0.
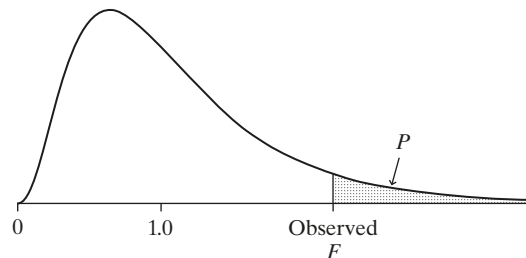
The shape of the F distribution is determined by two degrees of freedom terms, denoted by df_1 and df_2:

$$df_1 = p, \text{ the number of explanatory variables in the model.}$$

$$df_2 = n - (p + 1) = n - \text{ number of parameters in regression equation.}$$

The first of these, $df_1 = p$, is the divisor of the numerator term (R^2) in the F test statistic. The second, $df_2 = n - (p + 1)$, is the divisor of the denominator term ($1 - R^2$). The number of parameters in the multiple regression model is $p + 1$, representing the p beta terms and the y-intercept (α) term.

The mean of the F distribution is approximately[2] equal to 1. The larger the R^2-value, the larger the ratio $R^2/(1 - R^2)$, and the larger the F test statistic becomes. Thus, larger values of the F test statistic provide stronger evidence against H_0. Under the presumption that H_0 is true, the P-value is the probability the F test statistic is larger than the observed F-value. This is the right-tail probability under the F distribution beyond the observed F-value, as Figure 11.9 shows. Software for regression and Internet applets[3] report the P-value.

Example 11.4	**F Test for Mental Impairment Data** For Table 11.1 (page 325), we used multiple regression for $n = 40$ observations on $y =$ mental impairment, with $p = 2$ explanatory variables, life events and SES. The null hypothesis that mental impairment is statistically independent of life events and SES is $H_0: \beta_1 = \beta_2 = 0$.

In Example 11.3 (page 330), we found that this model has $R^2 = 0.339$. The F test statistic value is

$$F = \frac{R^2/p}{(1 - R^2)/[n - (p + 1)]} = \frac{0.339/2}{0.661/[40 - (2 + 1)]} = 9.5.$$

The two degrees of freedom terms for the F distribution are $df_1 = p = 2$ and $df_2 = n - (p + 1) = 40 - 3 = 37$, the two divisors in this statistic.

Part of the SPSS software output in Table 11.7 showed the ANOVA table

	Sum of Squares	df	Mean Square	F	Sig.
Regression	394.238	2	197.119	9.495	.000
Residual	768.162	37	20.761		

which contains the F statistic. The P-value, which rounded to three decimal places is $P = 0.000$, appears under the heading *Sig* in the table. (R reports it as *p-value*, Stata reports it as *Prob > F*, and SAS reports it as *Pr > F*.) This extremely small P-value provides strong evidence against H_0. We infer that at least one of the explanatory variables is associated with mental impairment. Equivalently, we can conclude that the population multiple correlation and R-squared are positive. ■

Normally, unless n is small and the associations are weak, this F test has a small P-value. If we choose variables wisely for a study, at least one of them should have *some* explanatory power.

INFERENCES FOR INDIVIDUAL REGRESSION COEFFICIENTS

When the P-value is small for the F test, this does not imply that *every* explanatory variable has an effect on y (controlling for the other explanatory variables in the model), but merely that *at least one* of them has an effect. More narrowly focused analyses judge *which* partial effects are nonzero and estimate the sizes of those effects. These inferences make the same assumptions as the F test.

For a particular explanatory variable x_i in the model, the test for its partial effect on y has $H_0: \beta_i = 0$. If $\beta_i = 0$, the mean of y is identical for all values of x_i, controlling

---

[2]The mean equals $df_2/(df_2 - 2)$, which is usually close to 1 unless n is quite small.
[3]For example, the *F distribution* applet at www.pearsonglobaleditions.com/Agresti.

for the other explanatory variables in the model. The alternative can be two-sided, $H_a: \beta_i \neq 0$, or one-sided, $H_a: \beta_i > 0$ or $H_a: \beta_i < 0$, to predict the direction of the partial effect.

The test statistic for $H_0: \beta_i = 0$, using sample estimate b_i of β_i, is

$$t = \frac{b_i}{se},$$

where se is the standard error of b_i. As usual, the t test statistic takes the best estimate (b_i) of the parameter (β_i), subtracts the H_0 value of the parameter (0), and divides by the standard error. The formula for se is complex, but software provides its value. If H_0 is true and the model assumptions hold, the t statistic has the t distribution with $df = n - (p + 1)$, which is the same as df_2 in the F test.

It is more informative to estimate the size of a partial effect than to test whether it is zero. Recall that β_i represents the change in the mean of y for a one-unit increase in x_i, controlling for the other variables. A confidence interval for β_i is

$$b_i \pm t(se).$$

The t-score comes from the t table, with $df = n - (p + 1)$. For example, a 95% confidence interval for the partial effect of x_1 is $b_1 \pm t_{.025}(se)$.

Example 11.5

Inferences for Individual Predictors of Mental Impairment For the multiple regression model for y = mental impairment, x_1 = life events, and x_2 = SES,

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2,$$

let's analyze the effect of life events. The hypothesis that mental impairment is statistically independent of life events, controlling for SES, is $H_0: \beta_1 = 0$. If H_0 is true, the multiple regression equation reduces to $E(y) = \alpha + \beta_2 x_2$. If H_0 is false, then $\beta_1 \neq 0$ and the full model provides a better fit than the bivariate model.

Table 11.5 contained the results,

	B	Std. Error	t	Sig.
(Constant)	28.230	2.174	12.984	.000
LIFE	.103	.032	3.177	.003
SES	-.097	.029	-3.351	.002

The point estimate of β_1 is $b_1 = 0.103$, which has standard error $se = 0.032$. The test statistic is

$$t = \frac{b_1}{se} = \frac{0.103}{0.032} = 3.177.$$

This appears under the heading t in the table in the row for the variable LIFE. The statistic has $df = n - (p + 1) = 40 - 3 = 37$. The P-value is 0.003, the probability that the t statistic exceeds 3.177 in absolute value. The evidence is strong that mental impairment is associated with life events, controlling for SES.

A 95% confidence interval for β_1 uses $t_{0.025} = 2.026$, the t-value for $df = 37$ having a probability of $0.05/2 = 0.025$ in each tail. This interval is

$$b_1 \pm t_{0.025}(se) = 0.103 \pm 2.026(0.032), \quad \text{which is} \quad (0.04, 0.17).$$

Controlling for SES, we are 95% confident that the change in mean mental impairment per one-unit increase in life events falls between 0.04 and 0.17. An increase of 100 units in life events corresponds to anywhere from a $100(0.04) = 4$-unit to a $100(0.17) = 17$-unit increase in mean mental impairment. The interval is relatively wide because of the relatively small sample size. The interval does not contain 0. This is in agreement with rejecting $H_0: \beta_1 = 0$ in favor of $H_a: \beta_1 \neq 0$ at the $\alpha = 0.05$

level. Since the confidence interval contains only positive numbers, the association between mental impairment and life events is positive, controlling for SES. ∎

How is the t test for a partial regression coefficient different from the t test of H_0: $\beta = 0$ for the bivariate model, $E(y) = \alpha + \beta x$, presented in Section 9.5? That t test evaluates whether y and x are associated, *ignoring* other variables, because it applies to the bivariate model. By contrast, the test just presented evaluates whether variables are associated, *controlling* for other variables.

A note of caution: Suppose multicollinearity occurs, that is, much overlap among the explanatory variables in the sense that any one is well predicted by the others. Then, possibly none of the individual partial effects has a small P-value, even if R^2 is large and a large F statistic occurs in the global test for the βs. Any particular variable may uniquely explain little variation in y, even though together the variables explain much variation.

VARIABILITY AND MEAN SQUARES IN THE ANOVA TABLE*

The precision of the least squares estimates relates to the size of the conditional standard deviation σ that measures variability of y at fixed values of the predictors. The smaller the variability of y-values about the regression equation, the smaller the standard errors become. The estimate of σ is

$$s = \sqrt{\frac{\sum (y - \hat{y})^2}{n - (p + 1)}} = \sqrt{\frac{\text{SSE}}{df}}.$$

The degrees of freedom value is also df for t inferences for regression coefficients, and it is df_2 for the F test about the collective effect of the explanatory variables. (When a model has only $p = 1$ predictor, df simplifies to $n - 2$, the term in the s formula in Section 9.3 on page 268.)

From the ANOVA table in Table 11.7 (page 330) that contains the sums of squares for the multiple regression model with the mental impairment data, SSE = 768.2. Since $n = 40$ for $p = 2$ predictors, we have $df = n - (p + 1) = 40 - 3 = 37$ and

$$s = \sqrt{\frac{\text{SSE}}{df}} = \sqrt{\frac{768.2}{37}} = \sqrt{20.76} = 4.56.$$

If the conditional distributions are approximately bell shaped, nearly all mental impairment scores fall within about 14 units (3 standard deviations) of the mean specified by the regression function.

SPSS reports the conditional standard deviation under the heading *Std. Error of the Estimate* in the Model Summary table that also shows R and R^2 (see Table 11.7). This is a poor choice of label by SPSS, because s refers to the variability in y-values, not the variability of a sampling distribution of an estimator.

The square of s, which estimates the conditional variance, is often called the **error mean square**, often abbreviated by MSE, or the **residual mean square**. Software shows it in the ANOVA table in the *Mean Square* column, in the row labeled *Residual* (or *Error* in some software). For example, MSE = 20.76 in Table 11.7. Some software (such as Stata and SAS) better labels the conditional standard deviation estimate s as *Root MSE*, because it is the square root of the error mean square. R reports it as *Residual standard error*.

THE F STATISTIC IS A RATIO OF MEAN SQUARES*

An alternative formula for the F test statistic for testing $H_0: \beta_1 = \cdots = \beta_p = 0$ uses the two mean squares in the ANOVA table. Specifically, for our example,

$$F = \frac{\text{Regression mean square}}{\text{Residual mean square (MSE)}} = \frac{197.1}{20.8} = 9.5.$$

This gives the same value as the F test statistic formula (page 333) based on R^2.

The regression mean square equals the regression sum of squares divided by its degrees of freedom. The df equals p, the number of explanatory variables in the model, which is df_1 for the F test. In the ANOVA table in Table 11.7, the regression mean square equals

$$\frac{\text{Regression SS}}{df_1} = \frac{394.2}{2} = 197.1.$$

RELATIONSHIP BETWEEN F AND t STATISTICS*

We've used the F distribution to test that all partial regression coefficients equal 0. Some regression software also lists F test statistics instead of t test statistics for the tests about the individual regression coefficients. The two statistics are related and have the same P-values. The square of the t statistic for testing that a partial regression coefficient equals 0 is an F test statistic having the F distribution with $df_1 = 1$ and $df_2 = n - (p + 1)$.

To illustrate, in Example 11.5 for $H_0: \beta_1 = 0$ and $H_a: \beta_1 \neq 0$, the test statistic $t = 3.18$ with $df = 37$. Alternatively, we could use $F = t^2 = 3.18^2 = 10.1$, which has the F distribution with $df_1 = 1$ and $df_2 = 37$. The P-value for this F-value is 0.002, the same as Table 11.5 reports for the two-sided t test.

In general, if a statistic has the t distribution with d degrees of freedom, then the square of that statistic has the F distribution with $df_1 = 1$ and $df_2 = d$. A disadvantage of the F approach is that it lacks information about the direction of the association. It cannot be used for one-sided alternative hypotheses.

11.4 Modeling Interaction Effects

The multiple regression equation

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

assumes that the slope β_i of the partial relationship between y and each x_i is identical for all values of the other explanatory variables. This implies a parallelism of lines relating the two variables, at various values of the other variables, as Figure 11.3 (page 321) illustrated.

This model is sometimes too simple to be adequate. Often, the relationship between two variables changes according to the value of a third variable. There is *interaction*, a concept introduced in Section 10.3 (page 306).

Interaction	For quantitative variables, *interaction* exists between two explanatory variables in their effects on y when the effect of one variable changes as the level of the other variable changes.

For example, for $y =$ annual income (thousands of dollars), $x_1 =$ years of working experience, and $x_2 =$ number of years of education, suppose $E(y) = 18 + 0.25x_1$

when $x_2 = 10$, $E(y) = 25 + 0.50x_1$ when $x_2 = 12$, and $E(y) = 39 + 1.00x_1$ when $x_2 = 16$. The slope for the partial effect of x_1 changes markedly as the value for x_2 changes. Interaction occurs between x_1 and x_2 in their effects on y.

CROSS-PRODUCT TERMS

The most common approach for modeling interaction effects introduces ***cross-product terms*** of the explanatory variables into the multiple regression model. With two explanatory variables, the model is

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

This is a special case of the multiple regression model with three explanatory variables, in which x_3 is an artificial variable created as the cross product $x_3 = x_1 x_2$ of the two primary explanatory variables.

Let's see why this model permits interaction. Consider how y is related to x_1, controlling for x_2. We rewrite the equation in terms of x_1 as

$$E(y) = (\alpha + \beta_2 x_2) + (\beta_1 + \beta_3 x_2)x_1 = \alpha' + \beta' x_1,$$

where

$$\alpha' = \alpha + \beta_2 x_2 \qquad \text{and} \qquad \beta' = \beta_1 + \beta_3 x_2.$$

So, for fixed x_2, the mean of y changes linearly as a function of x_1. But the slope of the relationship, $\beta' = (\beta_1 + \beta_3 x_2)$, depends on the value of x_2. As x_2 changes, the slope for the effect of x_1 changes. In summary, the mean of y is a linear function of x_1, but the slope of the line changes as the value of x_2 changes.

For the model containing the cross-product term, β_1 is the effect of x_1 only when $x_2 = 0$. Unless $x_2 = 0$ is a particular value of interest for x_2, it is not particularly useful to form confidence intervals or perform significance tests about β_1 (or β_2) in this model.

Similarly, the mean of y is a linear function of x_2, but the slope varies according to the value of x_1. The coefficient β_2 of x_2 refers to the effect of x_2 only at $x_1 = 0$.

Example 11.6

Allowing Interaction in Modeling Mental Impairment For the data set on $y =$ mental impairment, $x_1 =$ life events, and $x_2 =$ SES, we create a third explanatory variable x_3 that gives the cross product of x_1 and x_2 for the 40 individuals. For the first subject, for example, $x_1 = 46$ and $x_2 = 84$, so $x_3 = 46(84) = 3864$. Software makes it easy to create this variable without doing the calculations yourself. Table 11.8 shows some software output for the model that permits interaction. The prediction equation is

$$\hat{y} = 26.0 + 0.156x_1 - 0.060x_2 - 0.00087x_1 x_2.$$

Figure 11.10 portrays the relationship between predicted mental impairment and life events for a few distinct SES values. For an SES score of $x_2 = 0$, the relationship between $\hat{y}$ and x_1 is

$$\hat{y} = 26.0 + 0.156x_1 - 0.060(0) - 0.00087x_1(0) = 26.0 + 0.156x_1.$$

When $x_2 = 50$, the prediction equation is

$$\hat{y} = 26.0 + 0.156x_1 - 0.060(50) - 0.00087(50)x_1 = 23.0 + 0.113x_1.$$

When $x_2 = 100$, the prediction equation is

$$y = 20.0 + 0.069x_1.$$

TABLE 11.8: Output for Model Allowing Interaction, for y = Mental Impairment, x_1 = Life Events, and x_2 = SES from `Mental` Data File

```
              Sum of                Mean
              Squares    DF        Square        F           Sig
Regression    403.631     3        134.544      6.383       0.0014
Residual      758.769    36         21.077
Total        1162.400    39


              R: .589              R Square:    .347


                B           Std. Error        t          Sig
(Constant)    26.036649     3.948826        6.594       0.0001
LIFE           0.155865     0.085338        1.826       0.0761
SES           -0.060493     0.062675       -0.965       0.3409
LIFE*SES      -0.000866     0.001297       -0.668       0.5087
```

The higher the value of SES, the smaller the slope between predicted mental impairment and life events, and so the weaker is the effect of life events. Perhaps subjects who possess greater resources, in the form of higher SES, are better able to withstand the mental stress of potentially traumatic life events. ■

FIGURE 11.10: Portrayal of Interaction between x_1 and x_2 in Their Effects on y
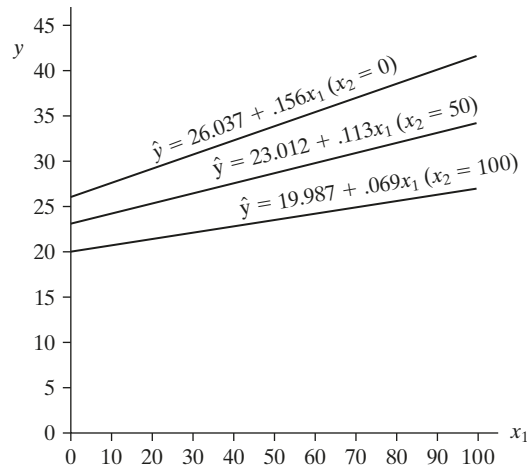
TESTING SIGNIFICANCE OF AN INTERACTION TERM

For two explanatory variables, the model allowing interaction is

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

The simpler model assuming no interaction is the special case $\beta_3 = 0$. The hypothesis of no interaction is $H_0: \beta_3 = 0$. As usual, the t test statistic divides the estimate of the parameter (β_3) by its standard error.

From Table 11.8, $t = -0.00087/0.0013 = -0.67$. The P-value for $H_a: \beta_3 \neq 0$ is $P = 0.51$. Little evidence exists of interaction. The variation in the slope of the relationship between mental impairment and life events for various SES levels could be due to sampling variability. The sample size here is small, however, which makes it difficult to estimate effects precisely.

When the evidence of interaction is weak, as it is here with a *P*-value of 0.51, it is best to drop the interaction term from the model before testing hypotheses about partial effects such as H_0: $\beta_1 = 0$ or H_0: $\beta_2 = 0$. On the other hand, *if the evidence of interaction is strong, it no longer makes sense to test these other hypotheses*. If there is interaction, then the effect of each variable exists and differs according to the level of the other variable.

CENTERING THE EXPLANATORY VARIABLES*

For the mental impairment data, x_1 and x_2 are highly significant in the model with only those predictors (Table 11.5) but lose their significance after entering the interaction term, even though the interaction is not significant (Table 11.8). But we've noted that the coefficients of x_1 and x_2 in an interaction model are not usually meaningful, because they refer to the effect of a predictor only when the other predictor equals 0.

An alternative way to parameterize the interaction model gives estimates and significance for the effect of x_1 and x_2 similar to those for the no-interaction model. The method *centers* the scores for each explanatory variable around 0, by subtracting the mean. Let $x_1^C = x_1 - \mu_{x_1}$ and $x_2^C = x_2 - \mu_{x_2}$, so that each new explanatory variable has a mean of 0. Then, we express the interaction model as

$$
\begin{aligned}
E(y) &= \alpha + \beta_1 x_1^C + \beta_2 x_2^C + \beta_3 x_1^C x_2^C \\
&= \alpha + \beta_1(x_1 - \mu_{x_1}) + \beta_2(x_2 - \mu_{x_2}) + \beta_3(x_1 - \mu_{x_1})(x_2 - \mu_{x_2}).
\end{aligned}
$$

Now, β_1 refers to the effect of x_1 at the mean of x_2, and β_2 refers to the effect of x_2 at the mean of x_1. Their estimates are usually similar to the estimated effects for the no-interaction model.

When we rerun the interaction model for the mental health data after centering the predictors about their sample means, that is, with

$$\text{LIFE\_CEN} = \text{LIFE} - 44.425 \text{ and } \text{SES\_CEN} = \text{SES} - 56.60,$$

we get software output shown in Table 11.9. The estimate for the interaction term is the same as for the model with uncentered predictors. Now, though, the estimates (and standard errors) for the effects of x_1 and x_2 alone are similar to the values for the no-interaction model. This happens because the coefficient for a variable represents its effect at the mean of the other variable, which is typically similar to the effect for the no-interaction model. Also, the statistical significance of x_1 and x_2 is similar as in the no-interaction model.

TABLE 11.9: Output for Model Allowing Interaction, Using Centered Explanatory Variables

	B	Std. Error	t	Sig
(Constant)	27.359555	0.731366	37.409	0.0001
LIFE_CEN	0.106850	0.033185	3.220	0.0027
SES_CEN	-0.098965	0.029390	-3.367	0.0018
LIFE_CEN*SES_CEN	-0.000866	0.001297	-0.668	0.5087

In summary, centering the explanatory variables before using them in a model allowing interaction has two benefits. First, the estimates of the effects of x_1 and x_2 are more meaningful, being effects at the mean rather than at 0. Second, the estimates and their standard errors are similar as in the no-interaction model.

GENERALIZATIONS AND LIMITATIONS*

When the number of explanatory variables exceeds two, a model allowing interaction can have cross products for each pair of explanatory variables. For example, with three explanatory variables, an interaction model is

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3.$$

This is a special case of multiple regression with six explanatory variables, identifying $x_4 = x_1 x_2$, $x_5 = x_1 x_3$, and $x_6 = x_2 x_3$. Significance tests can judge which, if any, of the cross-product terms are needed in the model.

When interaction exists and the model contains cross-product terms, it is more difficult to summarize simply the relationships. One approach is to sketch a collection of lines such as those in Figure 11.10 to describe graphically how the relationship between two variables changes according to the values of other variables. Another possibility is to divide the data into groups according to the value on a control variable (e.g., high on x_2, medium on x_2, low on x_2) and report the slope between y and x_1 within each subset as a means of describing the interaction.

11.5 Comparing Regression Models

When the number of explanatory variables increases, the multiple regression model becomes more difficult to interpret and some variables may become redundant. This is especially true when some explanatory variables are cross products of others, to allow for interaction. Not all the variables may be needed in the model. We next present a significance test of whether a model fits significantly better than a simpler model containing only some of the explanatory variables.

COMPLETE AND REDUCED MODELS

We refer to the full model with all the explanatory variables as the ***complete model***. The model containing only some of these variables is called the ***reduced model***. The reduced model is said to be *nested* within the complete model, being a special case of it.

The complete and reduced models are identical if the partial regression coefficients for the extra variables in the complete model all equal 0. In that case, none of the extra explanatory variables increases the explained variability in y, in the population of interest. Testing whether the complete model is identical to the reduced model is equivalent to testing whether the extra parameters in the complete model equal 0. The alternative hypothesis is that at least one of these extra parameters is not 0, in which case the complete model fits better than the reduced model.

For instance, a complete model with three explanatory variables and all two-variable interaction terms is

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3.$$

The reduced model without the interaction terms is

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

The test comparing the complete model to the reduced model has $H_0: \beta_4 = \beta_5 = \beta_6 = 0$.

COMPARING MODELS BY COMPARING SSE VALUES OR R^2-VALUES

The test statistic for comparing two regression models compares the residual sums of squares for the two models. Denote $SSE = \sum(y - \hat{y})^2$ for the reduced model by SSE_r and for the complete model by SSE_c. Now, $SSE_r \geq SSE_c$, because the reduced model has fewer explanatory variables and makes poorer overall predictions. Even if H_0 were true, we would not expect the estimates of the extra parameters and the difference $(SSE_r - SSE_c)$ to equal 0. Some reduction in error occurs from fitting the extra terms because of sampling variability.

The test statistic uses the reduction in error, $SSE_r - SSE_c$, that results from adding the extra variables. An equivalent statistic uses the R^2-values, R_c^2 for the complete model and R_r^2 for the reduced model. The two expressions for the test statistic are

$$F = \frac{(SSE_r - SSE_c)/df_1}{SSE_c/df_2} = \frac{(R_c^2 - R_r^2)/df_1}{(1 - R_c^2)/df_2}.$$

Here, df_1 is the number of extra terms in the complete model (e.g., 3 when we add three interaction terms to get the complete model) and df_2 is the residual df for the complete model. A relatively large reduction in error (or relatively large increase in R^2) yields a large F test statistic and a small P-value. As usual for F statistics, the P-value is the right-tail probability.

Example 11.7

Comparing Models for Mental Impairment   For the mental impairment data, a comparison of the complete model

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

to the reduced model

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

analyzes whether interaction exists. The complete model has only one additional term, and the null hypothesis is $H_0: \beta_3 = 0$.

The sum of squared errors for the complete model is $SSE_c = 758.8$ (Table 11.8 on page 339), while for the reduced model it is $SSE_r = 768.2$ (Table 11.7 on page 330). The difference

$$SSE_r - SSE_c = 768.2 - 758.8 = 9.4$$

has $df_1 = 1$ since the complete model has one more parameter. From Table 11.8, $df_2 = n - (p+1) = 40 - (3+1) = 36$, the residual df in that table. The F test statistic equals

$$F = \frac{(SSE_r - SSE_c)/df_1}{SSE_c/df_2} = \frac{9.4/1}{758.8/36} = 0.45.$$

Equivalently, the R^2-values for the two models are $R_r^2 = 0.339$ and $R_c^2 = 0.347$, so

$$F = \frac{(R_c^2 - R_r^2)/df_1}{(1 - R_c^2)/df_2} = \frac{(0.347 - 0.339)/1}{(1 - 0.347)/36} = 0.45.$$

From software, the P-value from the F distribution with $df_1 = 1$ and $df_2 = 36$ is $P = 0.51$. There is little evidence that the complete model is better. The null hypothesis seems plausible, so the reduced model is adequate.

When H_0 contains a single parameter, the t test is available. In fact, from the previous section (and Table 11.8), the t statistic is

$$t = \frac{b_3}{se} = \frac{-0.00087}{0.0013} = -0.67.$$

It also has a *P*-value of 0.51 for H_a: $\beta_3 \neq 0$. We get the same result with the *t* test as with the *F* test for complete and reduced models. In fact, the *F* test statistic equals the square of the *t* statistic. (Refer to page 337.) ■

The *t* test method is limited to testing one parameter at a time. The *F* test can test *several* regression parameters together to analyze whether at least one of them is nonzero, such as in the global *F* test of H_0: $\beta_1 = \cdots = \beta_p = 0$ or the test comparing a complete model to a reduced model. *F* tests are equivalent to *t* tests only when H_0 contains a single parameter.

11.6 Partial Correlation*

Multiple regression models describe the effect of an explanatory variable on the response variable while controlling for other variables of interest. Related measures describe the strength of the association. For example, to describe the association between mental impairment and life events, controlling for SES, we could ask, "Controlling for SES, what proportion of the variation in mental impairment does life events explain?"

These measures describe the partial association between *y* and a particular explanatory variable, whereas the multiple correlation and R^2 describe the association between *y* and the entire set of explanatory variables in the model. The *partial correlation* is based on the ordinary correlations between each pair of variables. For a single control variable, it is calculated as follows:

Partial Correlation

The sample ***partial correlation*** between *y* and x_1, controlling for x_2, is

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1 x_2}^2\right)}}.$$

In the symbol $r_{yx_1 \cdot x_2}$, the variable to the right of the dot represents the controlled variable. The analogous formula for $r_{yx_2 \cdot x_1}$ (i.e., controlling x_1) is

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right)\left(1 - r_{x_1 x_2}^2\right)}}.$$

Since one variable is controlled, the partial correlations $r_{yx_1 \cdot x_2}$ and $r_{yx_2 \cdot x_1}$ are called ***first-order partial correlations***.

Example 11.8

Partial Correlation between Education and Crime Rate Example 11.1 (page 320) discussed a data set for counties in Florida, with y = crime rate, x_1 = education, and x_2 = urbanization. The pairwise correlations are $r_{yx_1} = 0.468, r_{yx_2} = 0.678$, and $r_{x_1 x_2} = 0.791$. It was surprising to observe a positive correlation between crime rate and education. Can it be explained by their joint dependence on urbanization? This is plausible if the association disappears when we control for urbanization.

The partial correlation between crime rate and education, controlling for urbanization, is

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1 x_2}^2)}} = \frac{0.468 - 0.678(0.791)}{\sqrt{(1 - 0.678^2)(1 - 0.791^2)}} = -0.152.$$

Not surprisingly, $r_{yx_1 \cdot x_2}$ is much smaller than r_{yx_1}. It even has a different direction, illustrating Simpson's paradox. The relationship between crime rate and education may well be spurious, reflecting their joint dependence on urbanization. ■

INTERPRETING PARTIAL CORRELATIONS

The partial correlation has properties similar to those for the ordinary correlation between two variables. We list the properties below for $r_{yx_1 \cdot x_2}$, but analogous properties apply to $r_{yx_2 \cdot x_1}$.

- $r_{yx_1 \cdot x_2}$ falls between -1 and $+1$.

- The larger the absolute value of $r_{yx_1 \cdot x_2}$, the stronger the association between y and x_1, controlling for x_2.

- The value of a partial correlation does not depend on the units of measurement of the variables.

- $r_{yx_1 \cdot x_2}$ has the same sign as the partial slope (b_1) for the effect of x_1 in the prediction equation $\hat{y} = a + b_1 x_1 + b_2 x_2$, because the same variable (x_2) is controlled in the model as in the correlation.

- Under the assumptions for conducting inference for multiple regression (see the beginning of Section 11.3), $r_{yx_1 \cdot x_2}$ estimates the correlation between y and x_1 at every *fixed* value of x_2. If we could control x_2 by considering a subpopulation of subjects all having the same value on x_2, then $r_{yx_1 \cdot x_2}$ estimates the correlation between y and x_1 for that subpopulation.

- The sample partial correlation is identical to the ordinary bivariate correlation computed for the points in the *partial regression plot* (page 326).

INTERPRETING SQUARED PARTIAL CORRELATIONS

Like r^2 and R^2, the square of a partial correlation has a proportional reduction in error (PRE) interpretation. For example, $r^2_{yx_2 \cdot x_1}$ is the proportion of variation in y explained by x_2, controlling for x_1. This squared measure describes the effect of removing from consideration the portion of the total sum of squares (TSS) in y that is explained by x_1, and then finding the proportion of the remaining unexplained variation in y that is explained by x_2.

Squared Partial Correlation

> The square of the partial correlation $r_{yx_2 \cdot x_1}$ represents the proportion of the variation in y that is explained by x_2, out of that left unexplained by x_1. It equals
>
> $$r^2_{yx_2 \cdot x_1} = \frac{R^2 - r^2_{yx_1}}{1 - r^2_{yx_1}} = \frac{\text{Partial proportion explained uniquely by } x_2}{\text{Proportion unexplained by } x_1}.$$

From Section 9.4 (page 271), $r^2_{yx_1}$ represents the proportion of the variation in y explained by x_1. The remaining proportion $(1 - r^2_{yx_1})$ represents the variation left unexplained. When x_2 is added to the model, it accounts for some additional variation. The total proportion of the variation in y accounted for by x_1 and x_2 jointly is R^2 for the model with both x_1 and x_2 as explanatory variables. So, $R^2 - r^2_{yx_1}$ is the additional proportion of the variability in y explained by x_2, after the effects of x_1

have been removed or controlled. The maximum this difference could be is $1 - r_{yx_1}^2$, the proportion of variation yet to be explained after accounting for the influence of x_1. The additional explained variation $R^2 - r_{yx_1}^2$ divided by this maximum possible difference is a measure that has a maximum possible value of 1. In fact, as the above formula suggests, this ratio equals the squared partial correlation between y and x_2, controlling for x_1. Figure 11.11 illustrates this property of the squared partial correlation.

FIGURE 11.11:
Representation of $r_{yx_2 \cdot x_1}^2$ as the Proportion of Variability That Can Be Explained by x_2, of That Left Unexplained by x_1
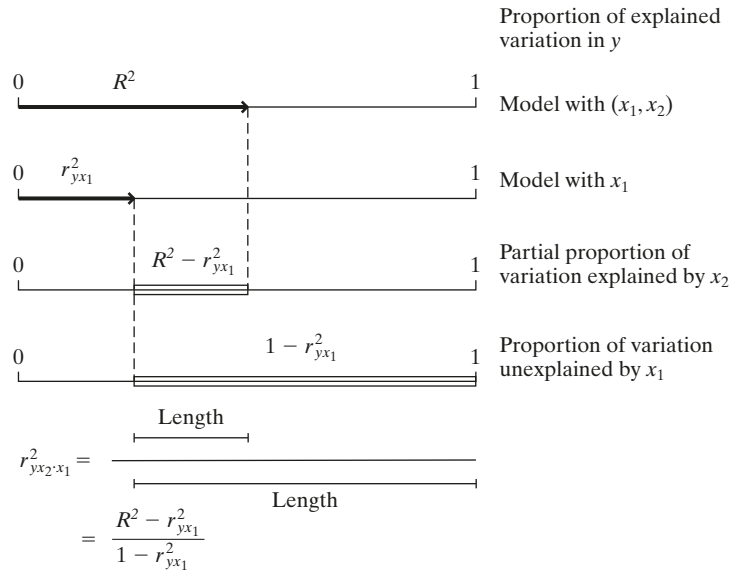
Example 11.9

Partial Correlation of Life Events with Mental Impairment We return to the example with y = mental impairment, x_1 = life events, and x_2 = SES. Software reports the correlation matrix

	IMPAIR	LIFE	SES
IMPAIR	1.000	.372	-.399
LIFE	.372	1.000	.123
SES	-.399	.123	1.000

So, $r_{yx_1} = 0.372$, $r_{yx_2} = -0.399$, and $r_{x_1x_2} = 0.123$. The partial correlation between mental impairment and life events, controlling for SES, is

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1x_2}^2\right)}} = \frac{0.372 - (-0.399)(0.123)}{\sqrt{[1 - (-0.399)^2](1 - 0.123^2)}} = 0.463.$$

The partial correlation, like the correlation of 0.37 between mental impairment and life events, is moderately positive.

Since $r_{yx_1 \cdot x_2}^2 = (0.463)^2 = 0.21$, controlling for SES, 21% of the variation in mental impairment is explained by life events. Alternatively, since $R^2 = 0.339$ (Table 11.7),

$$r_{yx_1 \cdot x_2}^2 = \frac{R^2 - r_{yx_2}^2}{1 - r_{yx_2}^2} = \frac{0.339 - (-0.399)^2}{1 - (-0.399)^2} = 0.21.$$

■

HIGHER-ORDER PARTIAL CORRELATIONS

The connection between squared partial correlation values and R-squared also applies when the number of control variables exceeds one. For example, with three explanatory variables, let $R^2_{y(x_1,x_2,x_3)}$ denote the value of R^2. The square of the partial correlation between y and x_3, controlling for x_1 and x_2, relates to how much larger this is than the R^2-value for the model with only x_1 and x_2 as explanatory variables, which we denote by $R^2_{y(x_1,x_2)}$. The squared partial correlation is

$$r^2_{yx_3 \cdot x_1,x_2} = \frac{R^2_{y(x_1,x_2,x_3)} - R^2_{y(x_1,x_2)}}{1 - R^2_{y(x_1,x_2)}}.$$

In this expression, $R^2_{y(x_1,x_2,x_3)} - R^2_{y(x_1,x_2)}$ is the increase in the proportion of explained variance from adding x_3 to the model. The denominator $1 - R^2_{y(x_1,x_2)}$ is the proportion of the variation left unexplained when x_1 and x_2 are the only explanatory variables in the model.

The partial correlation $r_{yx_3 \cdot x_1,x_2}$ is called a ***second-order partial correlation***, since it controls two variables. It has the same sign as b_3 in the prediction equation $\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3$, which also controls x_1 and x_2 in describing the effect of x_1.

11.7 Standardized Regression Coefficients*

As in bivariate regression, the sizes of regression coefficients in multiple regression models depend on the units of measurement for the variables. To compare the relative effects of two explanatory variables, it is appropriate to compare their coefficients only if the variables have the same units. Otherwise, *standardized* versions of the regression coefficients provide more meaningful comparisons.

Standardized Regression Coefficient

> The ***standardized regression coefficient*** for an explanatory variable represents the change in the mean of y, in y standard deviations, for a one standard deviation increase in that variable, controlling for the other explanatory variables in the model. We denote them by $\beta_1^*, \beta_2^*, \ldots$.

If $|\beta_2^*| > |\beta_1^*|$, for example, then a standard deviation increase in x_2 has a greater partial effect on y than does a standard deviation increase in x_1.

THE STANDARDIZATION MECHANISM

The standardized regression coefficients represent the values the regression coefficients take when the units are such that y and the explanatory variables all have equal standard deviations, such as when we use standardized variables. We can obtain the standardized regression coefficients from the unstandardized coefficients. Let s_y denote the sample standard deviation of y, and let $s_{x_1}, s_{x_2}, \ldots, s_{x_p}$ denote the sample standard deviations of the explanatory variables.

> The estimates of the standardized regression coefficients relate to the estimates of the unstandardized coefficients by
>
> $$b_1^* = b_1 \left(\frac{s_{x_1}}{s_y}\right), \quad b_2^* = b_2 \left(\frac{s_{x_2}}{s_y}\right), \ldots.$$

Example 11.10

Standardized Coefficients for Mental Impairment   The prediction equation relating mental impairment to life events and SES is

$$\hat{y} = 28.23 + 0.103x_1 - 0.097x_2.$$

Table 11.2 reported the sample standard deviations $s_y = 5.5$, $s_{x_1} = 22.6$, and $s_{x_2} = 25.3$. The unstandardized coefficient of x_1 is $b_1 = 0.103$, so the estimated standardized coefficient is

$$b_1^* = b_1 \left(\frac{s_{x_1}}{s_y} \right) = 0.103 \left(\frac{22.6}{5.5} \right) = 0.43.$$

Since $b_2 = -0.097$, the standardized value is

$$b_2^* = b_2 \left(\frac{s_{x_2}}{s_y} \right) = -0.097 \left(\frac{25.3}{5.5} \right) = -0.45.$$

The estimated change in the mean of y for a standard deviation increase in x_1, controlling for x_2, has similar magnitude as the estimated change for a standard deviation increase in x_2, controlling for x_1. However the partial effect of x_1 is positive, whereas the partial effect of x_2 is negative.

Table 11.10, which repeats Table 11.5, shows how SPSS reports the estimated standardized regression coefficients. It uses the heading BETA (as does Stata), reflecting the alternative name ***beta weights*** for these coefficients. ■

TABLE 11.10: SPSS Output for Fit of Multiple Regression Model to Mental Impairment Data from `Mental` Data File, with Standardized Coefficients

	Unstandardized coefficients		Standardized coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	28.230	2.174		12.984	.000
LIFE	.103	.032	.428	3.177	.003
SES	-.097	.029	-.451	-3.351	.002

PROPERTIES OF STANDARDIZED REGRESSION COEFFICIENTS

For bivariate regression, standardizing the regression coefficient yields the correlation. For the multiple regression model, the standardized partial regression coefficient relates to the partial correlation (Exercise 11.65), and it usually takes a similar value.

Unlike the partial correlation, however, b_i^* need not fall between -1 and $+1$. A value $|b_i^*| > 1$ occasionally occurs when x_i is highly correlated with the set of other explanatory variables in the model. In such cases, the standard errors are usually large and the estimates are unreliable.

Since a standardized regression coefficient is a multiple of the unstandardized coefficient, one equals 0 when the other does. The test of H_0: $\beta_i^* = 0$ is equivalent to the t test of H_0: $\beta_i = 0$. It is unnecessary to have separate tests for these coefficients. In the sample, the magnitudes of the $\{b_i^*\}$ have the same relative sizes as the t statistics from those tests. For example, the explanatory variable with the greatest standardized partial effect is the one that has the largest t statistic, in absolute value.

STANDARDIZED FORM OF PREDICTION EQUATION

Regression equations have an expression using the standardized regression coefficients. In this equation, the variables appear in standardized form.

Let $z_y, z_{x_1}, \ldots, z_{x_p}$ denote the standardized versions of the variables $y, x_1, \ldots, x_p$. For instance, $z_y = (y - \bar{y})/s_y$ represents the number of standard deviations that an observation on y falls from its mean. Each subject's scores on $y, x_1, \ldots, x_p$ have corresponding z-scores for $z_y, z_{x_1}, \ldots, z_{x_p}$. If a subject's score on x_1 is such that $z_{x_1} = (x_1 - \bar{x}_1)/s_{x_1} = 2.0$, for instance, then that subject falls two standard deviations above the mean $\bar{x}_1$ on x_1.

Let $\hat{z}_y = (\hat{y} - \bar{y})/s_y$ denote the predicted z-score for the response variable. For the standardized variables and the estimated standardized regression coefficients, the prediction equation is

$$\hat{z}_y = b_1^* z_{x_1} + b_2^* z_{x_2} + \cdots + b_p^* z_{x_p}.$$

This equation predicts how far an observation on y falls from its mean, in standard deviation units, based on how far the explanatory variables fall from their means, in standard deviation units. The standardized coefficients are the weights attached to the standardized explanatory variables in contributing to the predicted standardized response variable.

Example 11.11

Standardized Prediction Equation for Mental Impairment   Example 11.10 found that the estimated standardized regression coefficients for the life events and SES predictors of mental impairment are $b_1^* = 0.43$ and $b_2^* = -0.45$. The prediction equation relating the standardized variables is therefore

$$\hat{z}_y = 0.43 z_{x_1} - 0.45 z_{x_2}.$$

A subject who is two standard deviations above the mean on life events but two standard deviations below the mean on SES has a predicted standardized mental impairment of

$$\hat{z}_y = 0.43(2) - 0.45(-2) = 1.8.$$

The predicted mental impairment for that subject is 1.8 standard deviations above the mean. If the distribution of mental impairment is approximately normal, this subject might well have mental health problems, since only about 4% of the scores in a normal distribution fall at least 1.8 standard deviations above their mean.   ∎

In the prediction equation with standardized variables, no intercept term appears. Why is this? When the standardized explanatory variables all equal 0, those variables all fall at their means. Then, $\hat{y} = \bar{y}$, so

$$\hat{z}_y = \frac{\hat{y} - \bar{y}}{s_y} = 0.$$

So, this merely tells us that a subject who is at the mean on each explanatory variable is predicted to be at the mean on the response variable.

CAUTIONS IN COMPARING STANDARDIZED REGRESSION COEFFICIENTS

To assess which explanatory variable in a multiple regression model has the greatest impact on the response variable, it is tempting to compare their standardized

regression coefficients. Make such comparisons with caution. In some cases, the observed differences in the b_i^* may simply reflect sampling error. In particular, when multicollinearity exists, the standard errors are high and the estimated standardized coefficients may be unstable.

For a standardized regression coefficient to make sense, the variation in the explanatory variable must be representative of the variation in the population of interest. It is inappropriate to compare the standardized effect of an explanatory variable to others if the study purposely sampled values of that variable in a narrow range. This comment relates to a warning in Section 9.6 (page 284) about the correlation: Its value depends strongly on the range of explanatory variable values sampled.

Keep in mind also that the effects are partial ones, depending on which other variables are in the model. An explanatory variable that seems important in one system of variables may seem unimportant when other variables are controlled. For example, it is possible that $|b_2^*| > |b_1^*|$ in a model with two explanatory variables, yet when a third explanatory variable is added to the model, $|b_2^*| < |b_1^*|$.

It is unnecessary to standardize to compare the effect of the same variable for two groups, such as in comparing the results of separate regressions for females and males, since the units of measurement are the same in each group. In fact, it is usually unwise to standardize in this case, because the standardized coefficients are more susceptible than the unstandardized coefficients to differences in the standard deviations of the explanatory variables. Two groups that have the same value for an estimated regression coefficient have different standardized coefficients if the standard deviation of the explanatory variable differs for the two groups.

11.8 Chapter Summary

This chapter generalized the bivariate regression model to include additional explanatory variables. The ***multiple regression equation*** relating a response variable y to a set of p explanatory variables is

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

- The $\{\beta_i\}$ are ***partial regression coefficients***. The value β_i is the change in the mean of y for a one-unit change in x_i, controlling for the other variables in the model.

- The ***multiple correlation*** R describes the association between y and the collective set of explanatory variables. It equals the correlation between the observed and predicted y-values. It falls between 0 and 1.

- $R^2 = (\text{TSS} - \text{SSE})/\text{TSS}$ represents the *proportional reduction in error* from predicting y using the prediction equation $\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$ instead of $\bar{y}$. It equals the square of the multiple correlation.

- A ***partial correlation***, such as $r_{yx_1 \cdot x_2}$, describes the association between two variables, controlling for others. It falls between -1 and $+1$. The squared partial correlation between y and x_i represents the proportion of the variation in y that can be explained by x_i, out of that part left unexplained by a set of control variables.

- An F ***statistic*** tests H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, that the response variable is independent of all the explanatory variables. F-values are nonnegative and have two df values. A large F test statistic and small P-value suggest that the response variable is correlated with at least one of the explanatory variables.

- Individual t tests and confidence intervals for $\{\beta_i\}$ analyze partial effects of each explanatory variable, controlling for the other variables in the model.

- **_Interaction_** between x_1 and x_2 in their effects on y means that the effect of either explanatory variable changes as the value of the other one changes. We can allow this by adding cross products of explanatory variables to the model, such as the term $\beta_3(x_1 x_2)$.

- To **_compare regression models_**, a _complete_ model and a simpler _reduced_ model, an F test compares the SSE values or the R^2-values.

- **_Standardized regression coefficients_** do not depend on the units of measurement. The estimated standardized coefficient b_i^* describes the change in y, in y standard deviation units, for a one standard deviation increase in x_i, controlling for the other explanatory variables.

To illustrate, with $p = 3$ explanatory variables, the prediction equation is

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3.$$

Fixing x_2 and x_3, a straight line describes the relation between y and x_1. Its slope b_1 is the change in $\hat{y}$ for a one-unit increase in x_1, controlling for x_2 and x_3. The multiple correlation R is at least as large as the absolute values of the correlations r_{yx_1}, r_{yx_2}, and r_{yx_3}. The squared partial correlation $r^2_{yx_3 \cdot x_1, x_2}$ is the proportion of the variation of y that is explained by x_3, out of that part of the variation left unexplained by x_1 and x_2. The estimated standardized regression coefficient $b_1^* = b_1(s_{x_1}/s_y)$ describes the effect of a standard deviation change in x_1, controlling for x_2 and x_3.

Table 11.11 summarizes the basic properties and inference methods for these measures and those introduced in Chapter 9 for bivariate regression.

TABLE 11.11: Summary of Bivariate and Multiple Regression

	Bivariate Regression	Multiple Regression	
Model	$E(y) = \alpha + \beta x$	$E(y) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$	
Prediction equation	$\hat{y} = a + bx$	$\hat{y} = a + b_1 x_1 + \cdots + b_p x_p$	
	Overall effect of x	Simultaneous effect of $x_1, \ldots, x_p$	Partial effect of one x_i
Measures	$b = $ Slope		$b_i = $ Partial slope
	$r = $ Correlation, standardized slope, $-1 \leq r \leq 1$, r has the same sign as b	$R = $ Multiple correlation, $0 \leq R \leq 1$	$b_i^* = $ Standardized regression coefficient
	$r^2 = $ PRE measure, $0 \leq r^2 \leq 1$	$R^2 = $ PRE measure, $0 \leq R^2 \leq 1$	Partial correlation, $-1 \leq r_{yx_1 \cdot x_2} \leq 1$, same sign as b_i and b_i^*, $r^2_{yx_1 \cdot x_2}$ is PRE measure
Tests of no association	$H_0: \beta = 0$ or $H_0: \rho = 0$, y not associated with x	$H_0: \beta_1 = \cdots = \beta_p = 0$, y not associated with $x_1, \ldots, x_p$	$H_0: \beta_i = 0$, y not associated with x_i, controlling for other x variables
Test statistic	$t = \dfrac{b}{se} = \dfrac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ $df = n - 2$	$F = \dfrac{\text{Regression MS}}{\text{Residual MS}}$ $= \dfrac{R^2/p}{(1-R^2)/[n-(p+1)]}$, $df_1 = p, df_2 = n - (p+1)$	$t = \dfrac{b_i}{se}$ $df = n - (p+1)$

The model studied in this chapter is somewhat restrictive, requiring that all the explanatory variables be quantitative. The next chapter shows how a model can contain categorical explanatory variables.

Exercises

Practicing the Basics

11.1. For students at Walden University, the relationship between y = college GPA (with range 0–4.0) and x_1 = high school GPA (range 0–4.0) and x_2 = verbal college board score (range 200–800) satisfies $E(y) = 0.20 + 0.50x_1 + 0.002x_2$.

(a) Find the mean college GPA for students having (i) high school GPA = 4.0 and college board score = 800, (ii) $x_1 = 3.0$ and $x_2 = 300$.

(b) Show that the relationship between y and x_1 for those students with $x_2 = 500$ is $E(y) = 1.2 + 0.5x_1$.

(c) Show that when $x_2 = 600$, $E(y) = 1.4 + 0.5x_1$. Thus, increasing x_2 by 100 shifts the line relating y to x_1 upward by $100\beta_2 = 0.2$ units.

(d) Show that setting x_1 at a variety of values yields a collection of parallel lines, each having slope 0.002, relating the mean of y to x_2.

11.2. For recent data in Jacksonville, Florida, on y = selling price of home (in dollars), x_1 = size of home (in square feet), and x_2 = lot size (in square feet), the prediction equation is $\hat{y} = -10,536 + 53.8x_1 + 2.84x_2$.

(a) A particular home of 1240 square feet on a lot of 18,000 square feet sold for \$145,000. Find the predicted selling price and the residual, and interpret.

(b) For fixed lot size, how much is the house selling price predicted to increase for each square-foot increase in home size? Why?

(c) According to this prediction equation, for fixed home size, how much would lot size need to increase to have the same impact as a one-square-foot increase in home size?

(d) Suppose house selling prices are changed from dollars to thousands of dollars. Explain why the prediction equation changes to $\hat{y} = -10.536 + 0.0538x_1 + 0.00284x_2$.

11.3. The Social Progress Index (see www.socialprogressimperative.org) is a measure of national progress in delivering social and environmental value. It is an average of three component measures: BHN = basic human needs, incorporating basic medical care and personal safety; FW = foundations of well-being, incorporating access to basic knowledge and ecosystem sustainability; and Opp = opportunity, incorporating personal rights and access to advanced education. The SocialProgress data file at the text website shows the values for SP = social progress index and its three components, for nations in Europe and North America.

(a) Construct a scatterplot matrix for the four variables. Interpret.

(b) Construct a correlation matrix for the four variables. Interpret.

(c) Regress SP on BHN and FW. Report the prediction equation and interpret the effects.

(d) Show how R^2 changes when you regress SP on (i) BHN, (ii) BHN and FW, (iii) BHN, FW, and Opp. Interpret.

11.4. Use software with the Crime2 data file at the text website, with murder rate (number of murders per 100,000 people) as the response variable and with percentage of high school graduates and the poverty rate as explanatory variables.

(a) Construct the partial regression plots. Interpret. Do you see any unusual observations?

(b) Report the prediction equation. Explain how to interpret the estimated coefficients.

(c) Redo the analyses after deleting the D.C. observation. Describe the influence of this observation on the predicted effect of poverty rate. What does this tell you about how influential outliers can be?

11.5. A regression analysis with recent UN data from several nations on y = percentage of people who use the Internet, x_1 = per capita gross domestic product (in thousands of dollars), and x_2 = percentage of people using cell phones has results shown in Table 11.12.

(a) Write the prediction equation.

(b) Find the predicted Internet use for a country with per capita GDP of \$10,000 and 50% of the people using cell phones.

(c) Find the prediction equations when cell phone use is (i) 0%, (ii) 100%, and use them to interpret the effect of GDP.

(d) Use the equations in (c) to explain the *no interaction* property of the model.

TABLE 11.12

	B	Std. Error	t	Sig
(Constant)	-3.601	2.506	-1.44	0.159
GDP	1.2799	0.2703	4.74	0.000
CELLULAR	0.1021	0.0900	1.13	0.264

R Square .796

ANOVA

	Sum of Squares	DF
Regression	10316.8	2
Residual Error	2642.5	36
Total	12959.3	38

11.6. Refer to the previous exercise.

(a) Show how to obtain R-squared from the sums of squares in the ANOVA table. Interpret it.

(b) $r^2 = 0.78$ when GDP is the sole predictor. Why do you think R^2 does not increase much when cell phone use is added to the model, even though it is itself highly associated with y (with $r = 0.67$)? (*Hint*: Would you expect x_1 and x_2 to be highly correlated? If so, what is the effect?)

11.7. The Florida data file, shown partly on page 295, has data from the 67 Florida counties on y = crime rate (number per 1000 residents), x_1 = median income (thousands of dollars), and x_2 = percentage in urban environment.

(a) Figure 11.12 shows a scatterplot relating y to x_1. Predict the sign that the estimated effect of x_1 has in the prediction equation $\hat{y} = a + bx_1$. Explain.

(b) Figure 11.13 shows partial regression plots relating y to each explanatory variable, controlling for the other. Explain how these relate to the signs of b_1 and b_2 in the prediction equation $\hat{y} = a + b_1x_1 + b_2x_2$.

(c) Table 11.13 shows some software output for the bivariate and multiple regression models. Report the prediction equation relating y to x_1, and interpret the slope.

(d) Report the prediction equation relating y to both x_1 and x_2. Interpret the coefficient of x_1, and compare to (c).

(e) The correlations are $r_{yx_1} = 0.43, r_{yx_2} = 0.68$, and $r_{x_1x_2} = 0.73$. Use these to explain why the x_1 effect seems so different in (c) and (d).

(f) Report the prediction equations relating crime rate to income at urbanization levels of (i) 0, (ii) 50, (iii) 100. Interpret.

11.8. Refer to the previous exercise. Using software with the Florida data file at the text website,

(a) Construct box plots for each variable and scatterplots and partial regression plots between y and each of x_1 and x_2. Interpret these plots.

(b) Find the prediction equations for the (i) bivariate effects of x_1 and of x_2, (ii) multiple regression model. Interpret the estimated regression coefficients.
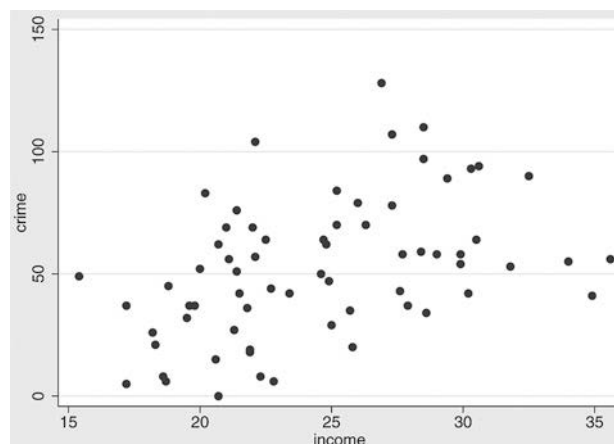
FIGURE 11.12
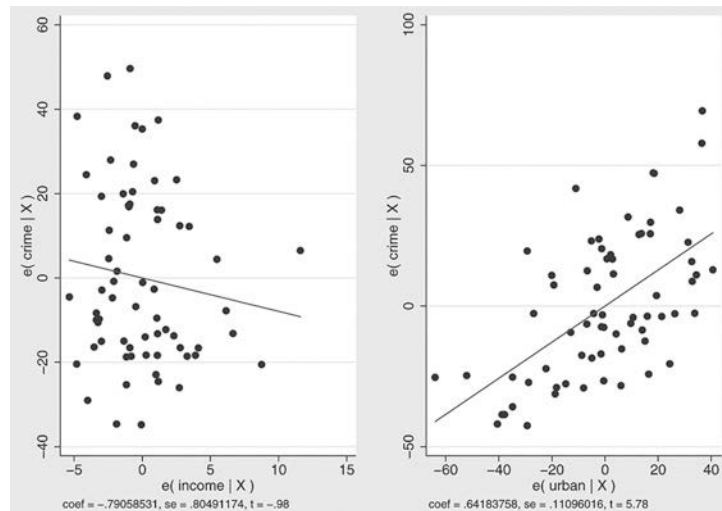
FIGURE 11.13

TABLE 11.13

BIVARIATE	Coef.	Std. Err.	t	P>|t|
income	2.609	0.675	3.866	0.0003
_cons	-11.526	16.834	-0.685	0.4960
MULTIPLE	Coef.	Std. Err.	t	P>|t|
income	-0.809	0.805	-1.005	0.3189
urban	0.646	0.111	5.811	0.0001
_cons	40.261	16.365	2.460	0.0166

(c) Find R^2 for the multiple regression model, and show that it is not much larger than r^2 for the model using urbanization alone as the predictor. Interpret.

11.9. Recent UN data from several nations on $y =$ crude birth rate (number of births per 1000 population size), $x_1 =$ women's economic activity (female labor force as percentage of male), and $x_2 =$ GNP (per capita, in thousands of dollars) has prediction equation $\hat{y} = 34.53 - 0.13x_1 - 0.64x_2$. The bivariate prediction equation with x_1 is $\hat{y} = 37.65 - 0.31x_1$. The correlations are $r_{yx_1} = -0.58$, $r_{yx_2} = -0.72$, and $r_{x_1x_2} = 0.58$. Explain why the coefficient of x_1 in the bivariate equation is quite different from that in the multiple predictor equation.

11.10. For recent UN data for several nations, a regression of carbon dioxide use (CO_2, a measure of air pollution) on gross domestic product (GDP) has a correlation of 0.786. With life expectancy as a second explanatory variable, the multiple correlation is 0.787.

(a) Explain how to interpret the multiple correlation.

(b) For predicting CO_2, did it help much to add life expectancy to the model? Does this mean that life expectancy is very weakly correlated with CO_2? Explain.

11.11. Table 11.14 shows Stata output from fitting the multiple regression model to recent statewide data, excluding D.C., on $y =$ violent crime rate (per 100,000 people), $x_1 =$ poverty rate (percentage with income below the poverty level), and $x_2 =$ percentage living in urban areas.

(a) Report the prediction equation.

(b) Massachusetts had $y = 805$, $x_1 = 10.7$, and $x_2 = 96.2$. Find its predicted violent crime rate. Find the residual, and interpret.

(c) Interpret the fit by showing the prediction equation relating $\hat{y}$ and x_1 for states with (i) $x_2 = 0$, (ii) $x_2 = 100$. Interpret.

(d) Interpret the correlation matrix.

(e) Report R^2 and the multiple correlation, and interpret.

11.12. Refer to the previous exercise.

(a) Report the F statistic for testing $H_0: \beta_1 = \beta_2 = 0$, report its df values and P-value, and interpret.

(b) Show how to construct the t statistic for testing $H_0: \beta_1 = 0$, report its df and P-value for $H_a: \beta_1 \neq 0$, and interpret.

TABLE 11.14

```
. regress violent poverty urban


      Source |       SS       df       MS        Number of obs =       50
    ---------+------------------------------     F(2, 47)      =   31.249
       Model | 2448368.07    2   1224184.        Prob > F      =   0.0001
    Residual | 1841257.15   47   39175.68        R-squared     =   0.5708
    ---------+------------------------------     Adj R-squared =   0.5525
       Total | 4289625.22   49   87543.37        Root MSE      = 197.928


     violent |    Coef.     Std. Err.     t       P>|t|
     poverty |   32.622       6.677     4.885     0.0001
       urban |    9.112       1.321     6.900     0.0001
       _cons | -498.683     140.988    -3.537     0.0009


. corr violent poverty urban
                      violent        poverty         urban
     violent         1.0000
     poverty          .3688         1.0000
       urban          .5940         -.1556         1.0000
```

(c) When we add x_3 = percentage of single-parent families to the model, we get the results in Table 11.15. Why do you think the effect of poverty rate is much lower after x_3 is added to the model?

TABLE 11.15

Variable	Coefficient	Std. Error
Intercept	−1197.538	
Poverty	18.283	6.136
Urban	7.712	1.109
Single parent	89.401	17.836
R^2	0.722	

11.13. For 2014 GSS data on y = highest year of school completed, x_1 = mother's highest year of school completed, and x_2 = father's highest year of school completed, we obtain $\hat{y} = 9.86 + 0.345x_1$ ($r^2 = 0.195$), $\hat{y} = 10.15 + 0.330x_2$ ($r^2 = 0.204$), and $\hat{y} = 9.30 + 0.194x_1 + 0.212x_2$ ($R^2 = 0.243$). In a single paragraph, summarize what you learn from these results.

11.14. Table 11.16 comes from a regression analysis[4] of y = number of children in family, x_1 = mother's educational level in years (MEDUC), and x_2 = father's socioeconomic status (FSES), for a random sample of 49 college students at Texas A&M University.

(a) Write the prediction equation. Interpret parameter estimates.

(b) Find R^2, and interpret it.

TABLE 11.16

	Sum of Squares
Regression	31.8
Residual	199.3
	b
(Constant)	5.25
MEDUC	−0.24
FSES	0.02

11.15. The General Social Survey has asked subjects to rate various groups using the "feeling thermometer." The rating is between 0 and 100, more favorable as the score gets closer to 100 and less favorable as the score gets closer to 0. For a small data set from the GSS, Table 11.17 shows results of fitting the multiple regression model with feelings toward liberals as the response, using explanatory variables political ideology (scored from 1 = extremely liberal to 7 = extremely conservative) and religious attendance, using scores (1 = never, 2 = less than once a year, 3 = once or twice a year, 4 = several times a year, 5 = about once a month, 6 = 2–3 times a month, 7 = nearly every week, 8 = every week, 9 = several times a week).

(a) Report the prediction equation and interpret the ideology partial effect.

(b) Report, and explain how to interpret, R^2.

(c) Tables of this form often put * by an effect having $P < 0.05$, ** by an effect having $P < 0.01$, and *** by an effect having $P < 0.001$. Show how this was determined for the ideology effect. Explain the disadvantage of summarizing in this manner.

---

[4]Thanks to Barbara Finlay for these results.

(d) Explain how the F-value can be obtained from the R^2-value reported. Report its df values, and explain how to interpret its result.

(e) The estimated standardized regression coefficients are -0.79 for ideology and -0.23 for religion. Interpret.

TABLE 11.17

Variable	Coefficient	Std. Error
Intercept	135.31	
Ideology	-14.07	3.16^{**}
Religion	-2.95	2.26
F	13.93^{**}	
R^2	0.799	
Adj. R^2	0.742	
n	10	

11.16. Refer to Table 11.5 on page 328. Test $H_0: \beta_2 = 0$ that mental impairment is independent of SES, controlling for life events. Report the test statistic, and report and interpret the P-value for **(a)** $H_a: \beta_2 \neq 0$, **(b)** $H_a: \beta_2 < 0$.

11.17. For a random sample of 66 state precincts, data are available on $y = $ percentage of adult residents who are registered to vote, $x_1 = $ percentage of adult residents owning homes, $x_2 = $ percentage of adult residents who are nonwhite, $x_3 = $ median family income (thousands of dollars), $x_4 = $ median age of residents, $x_5 = $ percentage of residents who have lived in the precinct for at least 10 years. Table 11.18 shows some output used to analyze the data.

(a) Fill in all the missing values.

(b) Do you think it is necessary to include all five explanatory variables in the model? Explain.

(c) To what test does "F value" refer? Interpret the result of that test.

(d) To what test does the t-value opposite x1 refer? Interpret the result of that test.

11.18. Refer to the previous exercise. Find a 95% confidence interval for the change in the mean of y for a **(a)** 1-unit increase, **(b)** 50-unit increase in the percentage of adults owning homes, controlling for the other variables. Interpret.

11.19. Use software with the `Houses` data file at the text website to conduct a multiple regression analysis of $y = $ selling price of home (dollars), $x_1 = $ size of home (square feet), $x_2 = $ number of bedrooms, $x_3 = $ number of bathrooms.

(a) Use scatterplots to display the effects of the explanatory variables on y. Explain how the highly discrete nature of x_2 and x_3 affects the plots.

(b) Report the prediction equation and interpret the estimated partial effect of size of home.

(c) Inspect the correlation matrix, and report the variable having the (i) strongest association with y, (ii) weakest association with y.

(d) Report R^2 for this model and r^2 for the simpler model using x_1 alone as the explanatory variable. Interpret.

11.20. Refer to the previous exercise.

(a) Test the partial effect of number of bathrooms, and interpret.

(b) Find the partial correlation between selling price and number of bathrooms, controlling for number of bedrooms. Compare it to the correlation, and interpret.

(c) Find the estimated standardized regression coefficients for the model, and interpret.

(d) Write the prediction equation using standardized variables. Interpret.

11.21. Exercise 11.11 showed a regression analysis for statewide data on $y = $ violent crime rate, $x_1 = $ poverty rate, and $x_2 = $ percentage living in urban areas. When we add an interaction term, we get $\hat{y} = 158.9 - 14.72x_1 - 1.29x_2 + 0.76x_1x_2$.

TABLE 11.18

	Sum of Squares	DF	Mean Square	F	Sig	R-Square ----
Regression	----	---	----	----	----	
Residual	2940.0	---	----			Root MSE
Total	3753.3	---				----

Variable	Parameter Estimate	Standard Error	t	Sig
Intercept	70.0000			
x1	0.1000	0.0450	----	----
x2	-0.1500	0.0750	----	----
x3	0.1000	0.2000	----	----
x4	-0.0400	0.0500	----	----
x5	0.1200	0.0500	----	----

(a) As the percentage living in urban areas increases, does the effect of poverty rate tend to increase or decrease? Explain.

(b) Show how to interpret the prediction equation, by finding how it simplifies when $x_2 = 0$, 50, and 100.

11.22. A study analyzes relationships among $y =$ percentage vote for Democratic candidate, $x_1 =$ percentage of registered voters who are Democrats, and $x_2 =$ percentage of registered voters who vote in the election, for several congressional elections in 2016. The researchers expect interaction, since they expect a higher slope between y and x_1 at larger values of x_2 than at smaller values. They obtain the prediction equation $\hat{y} = 20 + 0.30x_1 + 0.05x_2 + 0.005x_1x_2$. Does this equation support the direction of their prediction? Explain.

11.23. Use software with the Houses data file to allow interaction between number of bedrooms and number of bathrooms in their effects on selling price.

(a) Interpret the fit by showing the prediction equation relating $\hat{y}$ and number of bedrooms for homes with (i) two bathrooms, (ii) three bathrooms.

(b) Test the significance of the interaction term. Interpret.

11.24. A multiple regression analysis investigates the relationship between $y =$ college GPA and several explanatory variables, using a random sample of 195 students at Slippery Rock University. First, high school GPA and total SAT score are entered into the model. The sum of squared errors is SSE $= 20$. Next, parents' education and parents' income are added, to determine if they have an effect, controlling for high school GPA and SAT. For this expanded model, SSE $= 19$. Test whether this complete model is significantly better than the one containing only high school GPA and SAT. Report and interpret the P-value.

11.25. Table 11.19 shows results of regressing $y =$ birth rate (number of births per 1000 population) on $x_1 =$ women's economic activity and $x_2 =$ literacy rate, using UN data for 23 nations.

(a) Report the value of each of the following:
(i) r_{yx_1}, (ii) r_{yx_2}, (iii) R^2, (iv) TSS, (v) SSE, (vi) mean square error, (vii) s, (viii) s_y, (ix) se for b_1, (x) t for H_0: $\beta_1 = 0$, (xi) P for H_0: $\beta_1 = 0$ against H_a: $\beta_1 \neq 0$, (xii) P for H_0: $\beta_1 = 0$ against H_a: $\beta_1 < 0$, (xiii) F for H_0: $\beta_1 = \beta_2 = 0$, (xiv) P for H_0: $\beta_1 = \beta_2 = 0$.

(b) Report the prediction equation, and interpret the signs of the estimated regression coefficients.

(c) Interpret the correlations r_{yx_1} and r_{yx_2}.

(d) Report R^2, and interpret its value.

TABLE 11.19

	Mean	Std Deviation	N
BIRTHS	22.117	10.469	23
ECON	47.826	19.872	23
LITERACY	77.696	17.665	23

Correlations

		BIRTHS	ECON	LITER
Correlation	BIRTHS	1.00000	-0.61181	-0.81872
	ECON	-0.61181	1.00000	0.42056
	LITERACY	-0.81872	0.42056	1.00000

	Sum of Squares	DF	Mean Square	F	Sig
Regression	1825.969	2	912.985	31.191	0.0001
Residual	585.424	20	29.271		
Total	2411.393	22			

Root MSE (Std. Error of the Estimate) 5.410   R Square 0.7572

	Unstandardized Coeff.		Standardized		
	B	Std. Error	Coeff. (Beta)	t	Sig
(Constant)	61.713	5.2453		11.765	0.0001
ECON	-0.171	0.0640	-0.325	-2.676	0.0145
LITERACY	-0.404	0.0720	-0.682	-5.616	0.0001

(e) Report the multiple correlation, and interpret.

(f) Though inference may not be relevant for these data, report the F statistic for H_0: $\beta_1 = \beta_2 = 0$, report its P-value, and interpret.

(g) Show how to construct the t statistic for H_0: $\beta_1 = 0$, report its df and P-value for H_a: $\beta_1 \neq 0$, and interpret.

11.26. Refer to the previous exercise.

(a) Find the partial correlation between y and x_1, controlling for x_2. Interpret the partial correlation and its square.

(b) Find the estimate of the conditional standard deviation, and interpret.

(c) Show how to find the estimated standardized regression coefficient for x_1 using the unstandardized estimate and the standard deviations, and interpret its value.

(d) Write the prediction equation using standardized variables. Interpret.

(e) Find the predicted z-score for a country that is one standard deviation above the mean on both explanatory variables. Interpret.

11.27. Refer to Examples 11.1 (page 320) and 11.8 (page 343). Explain why the partial correlation between crime rate and high school graduation rate is so different (including its sign) from the bivariate correlation.

11.28. For the 2014 GSS, Table 11.20 shows estimates (with *se* values in parentheses) for four regression models for y = political party identification in the United States, scored from 1 = strong Democrat to 7 = strong Republican. The explanatory variables are number of years of education in model 1, also annual income last year (12 ordered categories, scored 1 to 12) in model 2, also religion (1 = fundamentalist, 2 = moderate, 3 = liberal) in model 3, and also political views (scored from 1 = extremely liberal to 7 = extremely conservative) in model 4.

(a) Summarize what you learn from these four model fits.

(b) The effect of religion weakens considerably (and even changes direction) after adding political views to the

model. Is it plausible that the relationship between y and religion is spurious? Explain.

11.29. A multiple regression model describes the relationship among a collection of cities between y = murder rate (number of murders per 100,000 residents) and x_1 = number of police officers (per 100,000 residents), x_2 = median length of prison sentence given to convicted murderers (in years), x_3 = median income of residents of city (in thousands of dollars), and x_4 = unemployment rate in city. These variables are observed for a random sample of 30 cities with population size exceeding 35,000. For the model with these explanatory variables, software reports the estimated standardized regression coefficients of -0.075 for x_1, -0.125 for x_2, -0.30 for x_3, and 0.20 for x_4.

(a) Write the prediction equation using standardized variables.

(b) Which explanatory variable has the greatest partial effect on y? Explain.

(c) Find the predicted z-score on murder rate for a city that is one standard deviation above the mean on x_1, x_2, and x_3, and one standard deviation below the mean on x_4. Interpret.

11.30. A recent study[5] analyzed the effect of x_1 = work hours per day and x_2 = commuting time to work on y = political participation. For the cluster sample of 1001 adult Americans, $\bar{x}_1 = 8.4$ hours ($s = 2.4$) and $\bar{x}_2 = 19.8$ minutes ($s = 13.6$). Political participation, which was a composite variable based on responses to several questions and coded to range from 0 to 1, has $\bar{y} = 0.20$ ($s = 0.24$). The multiple regression model, which also included control and mediating variables, had estimated effects 0.017 ($se = 0.070$) for work hours and -0.120 ($se = 0.040$) for commuting time.

(a) Find and interpret the estimated standardized regression coefficients.

(b) Explain how the estimates with their *se* values and the standardized coefficients support their conclusion that "While work itself exerted no effect on participation, the amount of time spent getting to and from work does."

TABLE 11.20

Variable	Model 1 Coef. (*se*)	Model 2 Coef. (*se*)	Model 3 Coef. (*se*)	Model 4 Coef. (*se*)
Education	0.004 (0.013)	−0.015 (0.017)	−0.003 (0.018)	0.010 (0.016)
Income		0.036 (0.018)	0.040 (0.018)	0.016 (0.016)
Religion			−0.255 (0.069)	0.036 (0.061)
Political views				0.752 (0.032)
Constant	3.597	3.496	3.833	0.255
Multiple *R*	0.007	0.054	0.115	0.550

[5] B. Newman, J. Johnson, and P. Lown, *American Politics Research*, vol. 42 (2014), pp. 141–170.

Concepts and Applications

11.31. Refer to the `Students` data file. Using software, conduct a regression analysis using either (**a**) y = political ideology with explanatory variables number of times per week of newspaper reading and religiosity, or (**b**) y = college GPA with explanatory variables high school GPA and number of weekly hours of physical exercise. Prepare a report, posing a research question and summarizing your graphical analyses, bivariate models and interpretations, multiple regression models and interpretations, inferences, checks of effects of outliers, and overall summary of the relationships.

11.32. Refer to the student data file created in Exercise 1.12. For variables chosen by your instructor, fit a multiple regression model and conduct descriptive and inferential statistical analyses. Interpret and summarize your findings.

11.33. Using industry-level data, a recent study[6] analyzed labor's share of income, measured as total compensation divided by total compensation plus the gross operating surplus. The authors predicted this would decrease as the degree of financialization of the company increased. Financialization was measured as the ratio of financial receipts (e.g., interest, dividends, capital gains) to business receipts from selling goods and services. The authors used regression analyses. The prediction equation reported for data from 1999 to 2008 for all nonfinance industries sampled was

$$\hat{y} = \hat{\alpha} - 0.882f - 0.906u - 0.727ci - 0.367c + 0.880wm$$
$$+ 0.052ic + 8.697es + 0.850cc - 0.207ecr,$$

where f = financialization, u = percentage of workforce in unions, ci = a measure of computer investment, c = proportion of workers who are college graduates, wm = proportion of workers who were non-Hispanic white men, ic = industrial concentration, es = employment size, cc = capital consumption, and ecr = error correction rate. The se for financialization was 0.070. The authors stated, "the model estimates support our hypothesis." Explain how the results stated here support this conclusion.

11.34. For the `OECD` data file at the text website, shown in Table 3.13 (page 70), pose a research question about how at least two of the variables shown in that table relate to carbon dioxide emissions. Conduct appropriate analyses to address that question, and prepare a one-page report summarizing your analyses and conclusions.

11.35. Using software with the `Crime` data file at the text website, conduct a regression analysis of violent crime rate with the explanatory variables poverty rate, the percentage living in urban areas, and the percentage of high school graduates. Prepare a report in which you state a re-search question you could answer with these data, conduct descriptive and inferential analyses, and provide interpretations and summarize your conclusions.

11.36. For the previous exercise, repeat the analysis, excluding the observation for D.C. Describe the effect of this observation on the various analyses.

11.37. For the `UN` data file at the text website (Table 3.9 on page 65), construct a multiple regression model containing two explanatory variables that provide good predictions for the fertility rate. How did you select this model? (*Hint*: One way uses the correlation matrix.)

11.38. In about 200 words, explain to someone who has never studied statistics what multiple regression does and how it can be useful.

11.39. Analyze the `Houses` data file at the text website (and introduced in Example 9.10 on page 280), using selling price of home, size of home, number of bedrooms, and taxes. Prepare a one-page report summarizing your analyses and conclusions.

11.40. For Example 11.2 on mental impairment, Table 11.21 shows the result of adding religious attendance as an explanatory variable, measured as the approximate number of times the subject attends a religious service over the course of a year. Write a report of about 200 words interpreting the table.

TABLE 11.21

Variable	Coefficient	Std. Error
Intercept	27.422	
Life events	0.0935	0.0313
SES	−0.0958	0.0256
Religious attendance	−0.0370	0.0219
R^2	0.3582	

11.41. A study[7] of mortality rates found in the United States that states with higher income inequality tended to have higher mortality rates. The effect of income inequality disappeared after controlling for the percentage of a state's residents who had at least a high school education. Explain how these results relate to analyses conducted using bivariate regression and multiple regression.

11.42. A study[8] relating the percentage of a child's life spent in poverty to the number of years of education completed by the mother and the percentage of a child's life spent in a single-parent home reported the results shown in Table 11.22. Prepare a one-page report explaining how to interpret the results in this table.

[6]K.-H. Lin and D. Tomaskovic-Dewey, *American Journal of Sociology*, vol. 118 (2013), pp. 1970–2008.
[7]A. Muller, *BMJ*, vol. 324 (2002), pp. 23–25.
[8]http://www.heritage.org/Research/Family/cda02-05.cfm.

TABLE 11.22

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	56.401	2.121		12.662	.000
% single parent	0.323	.014	.295	11.362	.000
mother school	-3.330	.152	-.290	-11.294	.000

F 611.6 (df = 2, 4731)   Sig .000
R 0.453   R Square 0.205

11.43. *The Economist* magazine[9] developed a quality-of-life index for nations as the predicted value obtained by regressing an average of life-satisfaction scores from several surveys on gross domestic product (GDP, per capita, in dollars), life expectancy (in years), an index of political freedom (from 1 = completely free to 7 = unfree), the percentage unemployed, the divorce rate (on a scale of 1 for lowest rates to 5 for highest), latitude (to distinguish between warmer and cold climes), a political stability measure, gender equality defined as the ratio of average male and female earnings, and community life (1 if country has high rate of church attendance or trade union membership, 0 otherwise). Table 11.23 shows results of the model fit for 74 countries, for which the multiple correlation is 0.92. The study used the prediction equation to predict the quality of life for 111 nations. The top 10 ranks were for Ireland, Switzerland, Norway, Luxembourg, Sweden, Australia, Iceland, Italy, Denmark, and Spain. Other ranks included 13 for the United States, 14 for Canada, 15 for New Zealand, 16 for the Netherlands, and 29 for the United Kingdom.

(a) Which variables would you expect to have negative effects on quality of life? Is this supported by the results?

(b) The study states that by itself "GDP explains more than 50% of the variation in life satisfaction." How does this relate to a summary measure of association?

(c) The study reported that "Using so-called Beta coefficients from the regression to derive the weights of the various factors, life expectancy and GDP were the most important." Explain what was meant by this.

(d) Although GDP seems to be an important predictor, in a bivariate sense and a partial sense, Table 11.23 reports a very small coefficient, 0.00003. Why do you think this is?

(e) The study mentioned other explanatory variables that were not included because they provided no further predictive power. For example, the study stated that education seemed to have an effect mainly through its effects on other variables in the model, such as GDP, life expectancy, and political freedom. Does this mean there is no association between education and quality of life? Explain.

TABLE 11.23

	Coefficient	Standard Error	t Statistic
Constant	2.796	0.789	3.54
GDP per person	0.00003	0.00001	3.52
Life expectancy	0.045	0.011	4.23
Political freedom	-0.105	0.056	-1.87
Unemployment	-0.022	0.010	-2.21
Divorce rate	-0.188	0.064	-2.93
Latitude	-1.353	0.469	-2.89
Political stability	0.152	0.052	2.92
Gender equality	0.742	0.543	1.37
Community life	0.386	0.124	3.13

11.44. An article[10] used multiple regression to predict a measure of tolerance toward homosexuality.

(a) The researchers found that the effect of number of years of education varied from essentially no effect for political conservatives to a considerably positive effect for political liberals. Explain how this is an example of statistical interaction. Explain how it would be handled by a multiple regression model.

(b) The best predictor of tolerance toward homosexuality was educational level, with an estimated standardized regression coefficient of 0.21. Interpret. In comparing this effect with the effects of other predictors, explain the purpose of using standardized coefficients.

11.45. An article[11] that analyzed the effects of the levels of the participant's generosity and of the spouse's generosity on a measure of marital quality reported that low levels of both were associated with low marital quality and high levels of both were associated with high marital quality. However, when the participant reported low levels of generosity toward the spouse, but the spouse reported high levels of generosity toward the participant, marital quality also tended to be low. For a regression model predicting marital quality, do you think it is adequate to use main effects alone, or do you probably also need an interaction term? Explain.

11.46. In Exercise 11.1 on y = college GPA, x_1 = high school GPA, and x_2 = college board score, $E(y)$ =

[9]www.economist.com/media/pdf/QUALITY_OF_LIFE.pdf.
[10]T. Shackelford and A. Besser, *Individual Differences Research*, vol. 5 (2007), pp. 106–114.
[11]J. Dew and W. B. Wilcox, *Journal of Marriage and Family*, vol. 75 (2013), pp. 1218–1228.

$0.20 + 0.50x_1 + 0.002x_2$. True or false: Since $\beta_1 = 0.50$ is larger than $\beta_2 = 0.002$, this implies that x_1 has the greater partial effect on y. Explain.

11.47. Table 11.24 shows results of fitting various regression models to data on $y = $ college GPA, $x_1 = $ high school GPA, $x_2 = $ mathematics entrance exam score, and $x_3 = $ verbal entrance exam score. Indicate which of the following statements are false. Give a reason for your answer.

TABLE 11.24

Estimates	Model $E(y) = \alpha + \beta x_1$	$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$	$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
Coefficient of x_1	0.450	0.400	0.340
Coefficient of x_2		0.003	0.002
Coefficient of x_3			0.002
R^2	0.25	0.34	0.38

(a) The correlation between y and x_1 is positive.

(b) A one-unit increase in x_1 corresponds to a change of 0.45 in the estimated mean of y, controlling for x_2 and x_3.

(c) In the third model, x_1 has the strongest partial effect on y.

(d) The value of $r^2_{yx_3}$ is 0.40.

(e) The partial correlation $r_{yx_1 \cdot x_2}$ is positive.

(f) Controlling for x_1, a 100-unit increase in x_2 corresponds to a predicted increase of 0.3 in college GPA.

(g) For the first model, the estimated standardized regression coefficient equals 0.50.

11.48. In regression analysis, which of the following statements must be false? Why?

(a) $r_{yx_1} = 0.01$, $r_{yx_2} = -0.75$, $R = 0.2$

(b) The value of the residual sum of squares, SSE, can increase as we add additional variables to the model.

(c) For the model $E(y) = \alpha + \beta_1 x_1$, y is significantly related to x_1 at the 0.05 level, but when x_2 is added to the model, y is not significantly related to x_1 at the 0.05 level.

(d) The estimated coefficient of x_1 is positive in the bivariate model but negative in the multiple regression model.

(e) When the model is refitted after y is multiplied by 10, R^2, r_{yx_1}, $r_{yx_1 \cdot x_2}$, and the F statistics and t statistics do not change.

(f) The F statistic for testing that all the regression coefficients equal 0 has $P < 0.05$, but none of the individual t tests have $P < 0.05$.

(g) If you compute the standardized regression coefficient for a bivariate model, you always get the correlation.

(h) $r^2_{yx_1} = r^2_{yx_2} = 0.6$ and $R^2 = 1.2$.

(i) The correlation between y and $\hat{y}$ equals -0.10.

(j) For every F test, there is an equivalent test using the t distribution.

(k) When $|b_1| > |b_2|$ in a multiple regression prediction equation, we can conclude that x_1 has a stronger effect than x_2 on y.

(l) The estimated standardized regression coefficient for an explanatory variable in a multiple regression model can be interpreted as the value the ordinary slope would equal for the linear prediction equation if that explanatory variable and y were scaled so that they both had the same standard deviation value.

(m) If $\hat{y} = 31.3 + 0.15x_1 - 0.05x_2 - 0.002x_1x_2$, then the estimated effect x_1 on y decreases as x_2 increases.

(n) Suppose $\hat{y} = 31.3 + 0.15x_1 - 0.05x_2 - 0.002x_1x_2$, with x_1 and x_2 taking values between 0 and 100. Then, since the coefficient of x_1x_2 is so small compared to the coefficients of x_1 and of x_2, we can conclude that the amount of interaction is negligible.

For Exercises 11.49–11.52, select the correct answer(s) and indicate why the other responses are inappropriate. (More than one response may be correct.)

11.49. If $\hat{y} = 2 + 3x_1 + 5x_2 - 8x_3$, then controlling for x_2 and x_3, the predicted mean change in y when x_1 is increased from 10 to 20 equals
(a) 3, **(b)** 30, **(c)** 0.3, **(d)** cannot be given—depends on specific values of x_2 and x_3.

11.50. If $\hat{y} = 2 + 3x_1 + 5x_2 - 8x_3$,

(a) The strongest correlation is between y and x_3.

(b) The variable with the strongest partial influence on y is x_2.

(c) The variable with the strongest partial influence on y is x_3, but one cannot tell from this equation which pair has the strongest correlation.

(d) None of the above.

11.51. If $\hat{y} = 2 + 3x_1 + 5x_2 - 8x_3$,

(a) $r_{yx_3} < 0$.

(b) $r_{yx_3 \cdot x_1} < 0$.

(c) $r_{yx_3 \cdot x_1, x_2} < 0$.

(d) Insufficient information to answer.

(e) Answers (a), (b), and (c) are all correct.

11.52. The F test for comparing a complete model to a reduced model can be used to test

(a) The significance of a single regression parameter in a multiple regression model.

(b) H_0: $\beta_1 = \cdots = \beta_p = 0$ in a multiple regression equation.

(c) H_0: no interaction, in the model
$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3$.

(d) Whether the model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ gives a significantly better fit than the model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_3$.

11.53. Explain the difference in the purposes of the correlation, the multiple correlation, and the partial correlation.

11.54. Let y = height, x_1 = length of right leg, and x_2 = length of left leg. Describe what you expect for the relative sizes of $r_{x_1 x_2}$, $r_{y x_2}$, R, and $r_{y x_2 \cdot x_1}$.

11.55. Give an example of three variables for which you expect $\beta \neq 0$ in the model $E(y) = \alpha + \beta x_1$ but for which it is plausible that $\beta_1 = 0$ in the model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$.

11.56. For the models $E(y) = \alpha + \beta x$ and $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$, express null hypotheses in terms of correlations that are equivalent to the following:

(a) H_0: $\beta = 0$.
(b) H_0: $\beta_1 = \beta_2 = 0$.

11.57.* Whenever x_1 and x_2 are uncorrelated, then R^2 for the model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ satisfies $R^2 = r_{y x_1}^2 + r_{y x_2}^2$. In this case, draw a figure that portrays the variability in y, the part of that variability explained by each of x_1 and x_2, and the total variability explained by both of them together.

11.58.* Which of the following sets of correlations would you expect to yield the highest R^2-value? Why?

(a) $r_{y x_1} = 0.4$, $r_{y x_2} = 0.4$, $r_{x_1 x_2} = 0.0$.
(b) $r_{y x_1} = 0.4$, $r_{y x_2} = 0.4$, $r_{x_1 x_2} = 0.5$.
(c) $r_{y x_1} = 0.4$, $r_{y x_2} = 0.4$, $r_{x_1 x_2} = 1.0$.

11.59.* Suppose the correlation between y and x_1 equals the multiple correlation between y and x_1 and x_2. What does this imply about the partial correlation $r_{y x_2 \cdot x_1}$? Interpret.

11.60.* Software reports four types of sums of squares in multiple regression models. The ***Type I sum of squares***, sometimes called *sequential SS*, represents the variability explained by a variable, controlling for variables previously entered into the model. The ***Type III sum of squares***, sometimes called *partial SS*, represents the variability explained by that variable, controlling for all other variables in the model.

(a) For any multiple regression model, explain why the Type I sum of squares for x_1 is the regression sum of squares for the bivariate model with x_1 as the explanatory variable, whereas the Type I sum of squares for x_2 equals the amount by which SSE decreases when x_2 is added to the model.

(b) Explain why the Type I sum of squares for the last variable entered into a model is the same as the Type III sum of squares for that variable.

11.61.* *Adjusted R^2* is defined as

$$R_{adj}^2 = 1 - \frac{s^2}{s_y^2},$$

where s^2 is the estimated conditional variance and s_y^2 is the sample variance of y, both of which are unbiased. This relates to ordinary R^2 by

$$R_{adj}^2 = R^2 - \left[\frac{p}{n - (p+1)} \right] (1 - R^2).$$

(a) Suppose $R^2 = 0.339$ for a model with $p = 2$ explanatory variables, as in Table 11.5. Find R_{adj}^2 when $n = 10$, 40 (as in the text example), and 1000. Show that R_{adj}^2 approaches R^2 in value as n increases.

(b) Show that $R_{adj}^2 < 0$ when $R^2 < p/(n-1)$. This is undesirable, and R_{adj}^2 is equated to 0 in such cases. (Also, unlike R^2, R_{adj}^2 can decrease when we add an explanatory variable to a model.)

11.62.* Let $R_{y(x_1, \ldots x_p)}^2$ denote R^2 for the multiple regression model with p explanatory variables. Explain why

$$r_{y x_p \cdot x_1, \ldots, x_{p-1}}^2 = \frac{R_{y(x_1, \ldots, x_p)}^2 - R_{y(x_1, \ldots, x_{p-1})}^2}{1 - R_{y(x_1, \ldots, x_{p-1})}^2}.$$

11.63.* The numerator $R^2 - r_{y x_1}^2$ of the squared partial correlation $r_{y x_2 \cdot x_1}^2$ gives the increase in the proportion of explained variation from adding x_2 to the model. This increment, denoted by $r_{y(x_2 \cdot x_1)}^2$, is called the squared ***semipartial*** correlation. One can use squared semipartial correlations to partition the variation in the response variable. For instance, for three explanatory variables,

$$R_{y(x_1, x_2, x_3)}^2 = r_{y x_1}^2 + (R_{y(x_1, x_2)}^2 - r_{y x_1}^2) + (R_{y(x_1, x_2, x_3)}^2 - R_{y(x_1, x_2)}^2)$$

$$= r_{y x_1}^2 + r_{y(x_2 \cdot x_1)}^2 + r_{y(x_3 \cdot x_1, x_2)}^2.$$

The total variation in y explained by x_1, x_2, and x_3 together partitions into (i) the proportion explained by x_1 (i.e., $r_{y x_1}^2$), (ii) the proportion explained by x_2 beyond that explained by x_1 (i.e., $r_{y(x_2 \cdot x_1)}^2$), and (iii) the proportion explained by x_3 beyond that explained by x_1 and x_2 (i.e, $r_{y(x_3 \cdot x_1, x_2)}^2$). For a particular model, the semi-partial correlations have the same ordering as the t statistics for testing the partial effects, and some researchers use them as indices of importance of the explanatory variables.

(a) In Example 11.2 on mental impairment, show that $r_{y(x_2 \cdot x_1)}^2 = 0.20$ and $r_{y(x_1 \cdot x_2)}^2 = 0.18$. Interpret.

(b) Explain why the squared semipartial correlation $r_{y(x_2 \cdot x_1)}^2$ cannot be larger than the squared partial correlation $r_{y x_2 \cdot x_1}^2$.

11.64.* The least squares prediction equation provides predicted values $\hat{y}$ with the strongest possible correlation with y, out of all possible prediction equations of that form. Based on this property, explain why the multiple correlation cannot decrease when you add a variable to a multiple regression model. (*Hint*: The prediction equation for the simpler model is a special case of a prediction equation for the full model that has coefficient 0 for the added variable.)

11.65.* Let $\bar{b}_i^*$ denote the estimated standardized regression coefficient when x_i is treated as the *response* variable and y as an *explanatory* variable, controlling for the same set of other variables. Then, $\bar{b}_i^*$ need not equal b_i^*. The squared partial correlation between y and x_i, which *is* symmetric, equals $b_i^* \bar{b}_i^*$.

(a) Explain why the partial correlation must fall between b_i^* and $\bar{b}_i^*$. (*Note*: When $a = \sqrt{bc}$, a is said to be the *geometric average* of b and c.)

(b) Even though b_i^* does not necessarily fall between -1 and $+1$, explain why $b_i^* \bar{b}_i^*$ cannot exceed 1.

11.66.* Chapters 12 and 13 show how to incorporate categorical explanatory variables in regression models. This exercise provides a preview. Table 11.25 shows some output for a model for the Houses2 data set at the text website, with y = selling price of home, x_1 = size of home, and x_2 = whether the house is new (1 = yes, 0 = no).

(a) Report the prediction equation. By setting $x_2 = 0$ and then 1, construct the two separate lines for older and for new homes. Note that the model implies that the slope effect of size on selling price is the same for each.

(b) Since x_2 takes only the values 0 and 1, explain why the coefficient of x_2 estimates the difference of mean selling prices between new and older homes, controlling for house size.

11.67.* Refer to the previous exercise. When we add an interaction term, we get $\hat{y} = -16.6 + 66.6x_1 - 31.8x_2 + 29.4(x_1 x_2)$.

(a) Interpret the fit by reporting the prediction equation between selling price and size of house separately for new homes ($x_2 = 1$) and for old homes ($x_2 = 0$). Interpret. (This fit is equivalent to fitting lines separately to the data for new homes and for old homes.)

(b) A plot of the data shows an outlier, a new home with a very high selling price. When that observation is removed from the data set and the model is refitted, $\hat{y} = -16.6 + 66.6x_1 + 9.0x_2 + 5.0(x_1 x_2)$. Redo (a), and explain how an outlier can have a large impact on a regression analysis.

TABLE 11.25

	B	Std. Error	t	Sig
(Constant)	-26.089	5.977	-4.365	0.0001
SIZE	72.575	3.508	20.690	0.0001
NEW	19.587	3.995	4.903	0.0001

REGRESSION WITH CATEGORICAL PREDICTORS: ANALYSIS OF VARIANCE METHODS

Chapter

12

The regression models presented so far have quantitative explanatory variables. This chapter shows how a regression model can contain categorical explanatory variables.

Chapter 7 presented methods for comparing the means of two groups. Those methods extend for comparing means of *several* groups. The methods relate to the association between a *quantitative* response variable and a *categorical* explanatory variable. The mean of the quantitative response variable is compared among groups that are categories of the explanatory variable. For example, for a comparison of mean annual income among blacks, whites, and Hispanics, the quantitative response variable is annual income and the categorical explanatory variable is racial–ethnic status. We can use the regression methods of this chapter to do this.

Artificial variables called *dummy variables* can represent the categories of a categorical explanatory variable in a regression model. The inferential method for testing equality of several means is often called the **analysis of variance**, abbreviated as **ANOVA**. We'll see that the name refers to the way the significance test focuses on two types of variability in the data.

Categorical explanatory variables in ANOVA are called **factors**. ANOVA methods extend to incorporate multiple factors, for example, to compare mean income across categories of both racial–ethnic status and gender. We first present analyses for *independent samples*. When each sample has the same subjects or the samples are matched, the samples are *dependent* and different methods apply. We also present such methods, referred to as **repeated-measures ANOVA**.

12.1 Regression Modeling with Dummy Variables for Categories

We can use a regression model for the relationship between a quantitative response variable and a *categorical* explanatory variable. We shall use the following example to illustrate methods.

Example 12.1

Political Ideology by Political Party ID   Table 12.1 summarizes observations on political ideology for three groups, based on data from subjects of ages 18–27 in the 2014 General Social Survey. The three groups are the (Democrat, Independent, Republican) categories of the explanatory variable, political party identification (ID). Political ideology, the response variable, is measured on a seven-point scale, ranging from extremely liberal (1) to extremely conservative (7). For each party ID, Table 12.1 shows the number of subjects who made each response. For instance, of 83 Democrats, 5 responded extremely liberal, 18 responded liberal, ..., 2 responded extremely conservative.

363

TABLE 12.1: Political Ideology by Political Party Identification (ID), for Respondents of Age 18–27

Party ID	Political Ideology							Sample Size	Mean	Standard Deviation
	1	2	3	4	5	6	7			
Democrat	5	18	19	25	7	7	2	83	3.48	1.43
Independent	4	19	27	79	13	9	6	157	3.82	1.23
Republican	1	3	1	11	10	11	1	38	4.66	1.36

Note: For political ideology, 1 = extremely liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = extremely conservative.

Since Table 12.1 displays the data as counts in a contingency table, we could use methods for categorical data (Chapter 8). The chi-squared test treats both variables as nominal, however, whereas political ideology is ordinal. That test is not directed toward detecting whether responses have a higher or lower mean in some groups than others. The ordinal measure of association gamma is inappropriate, because it requires both variables to be ordinal. Here, the groups, which are the categories of political party ID, are nominal.

When an ordinal response has many categories, one approach assigns scores to its levels and treats it as a quantitative variable. This is a reasonable strategy when we want to focus on a measure of center (such as the mean) rather than on the proportions in particular categories, and when the observations do not mainly fall at one of the boundary categories. For Table 12.1, for instance, interest might focus on how liberal or conservative the responses tend to be for each group, in some average sense. We analyze these data by assigning the scores (1, 2, 3, 4, 5, 6, 7) to the levels of political ideology and then comparing means. The higher the mean score, the more conservative the group's responses tended to be. For these scores, Table 12.1 shows the mean and standard deviation for each group. The overall sample mean is $\bar{y} = 3.83$, not far from the score of 4.0 corresponding to moderate ideology. ∎

REGRESSION WITH DUMMY (INDICATOR) VARIABLES

How can we enter a categorical explanatory variable such as party ID in a regression model? We set up an indicator variable to equal 1 if an observation comes from a particular category and 0 otherwise. With three categories (e.g., the party IDs), we use two indicator variables. The first, denoted by z_1, equals 1 for observations from the first category and equals 0 otherwise. The second, denoted by z_2, equals 1 for observations from the second category and equals 0 otherwise. That is,

$z_1 = 1$ and $z_2 = 0$: observations from category 1.
$z_1 = 0$ and $z_2 = 1$: observations from category 2.
$z_1 = 0$ and $z_2 = 0$: observations from category 3.

It is unnecessary and redundant to create a variable for the last (third) category, because values of 0 for z_1 and z_2 identify observations from it.

The indicator variables z_1 and z_2 are called ***dummy variables***. They indicate the category for an observation. That is, they give a classification, not a magnitude, for the factor. Table 12.2 summarizes the dummy variables for three categories.

For three groups, denote the population means on y by μ_1, μ_2, and μ_3. For the dummy variables just defined, consider the multiple regression equation

$$E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2.$$

TABLE 12.2:	The Two Dummy Variables for a Categorical Explanatory Variable with Three Categories	
Category	z_1	z_2
1	1	0
2	0	1
3	0	0

For observations from category 3, $z_1 = z_2 = 0$. The equation then simplifies to

$$E(y) = \alpha + \beta_1(0) + \beta_2(0) = \alpha.$$

So, α represents the population mean μ_3 of y for the last category. For observations from category 1, $z_1 = 1$ and $z_2 = 0$, so

$$E(y) = \alpha + \beta_1(1) + \beta_2(0) = \alpha + \beta_1$$

equals the population mean μ_1 for that category. Similarly, $\alpha + \beta_2$ equals the population mean μ_2 for category 2 (let $z_1 = 0$ and $z_2 = 1$).

Since $\alpha + \beta_1 = \mu_1$ and $\alpha = \mu_3$, the parameter $\beta_1 = \mu_1 - \mu_3$. Similarly, $\beta_2 = \mu_2 - \mu_3$. Table 12.3 summarizes the parameters of the regression model and their correspondence with the population means. The β coefficient of a dummy variable represents the difference between the mean for the category that dummy variable represents and the mean of the category not having its own dummy variable.

TABLE 12.3:	Interpretation of Coefficients of Dummy Variables in Model $E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2$ Having Explanatory Variable with Three Categories			
Category	z_1	z_2	Mean of y	Interpretation of β
1	1	0	$\mu_1 = \alpha + \beta_1$	$\beta_1 = \mu_1 - \mu_3$
2	0	1	$\mu_2 = \alpha + \beta_2$	$\beta_2 = \mu_2 - \mu_3$
3	0	0	$\mu_3 = \alpha$	

Dummy variable coding works because it allows the population means to take arbitrary values, with no assumed distances between categories. Using a single variable with coding such as $z = 1$ for category 1, $z = 2$ for category 2, and $z = 3$ for category 3 would not work. The model $E(y) = \alpha + \beta z$ would then assume an ordering as well as equal distances between categories. It treats the factor as if it were quantitative, which is improper. Whereas we need only one term in a regression model to represent the linear effect of a quantitative explanatory variable, for a categorical explanatory variable we need one fewer term than the number of categories.

Example 12.2

Regression Model for Political Ideology and Party ID For Table 12.1, the categorical explanatory variable (political party ID) has three categories. The regression model for $y =$ political ideology is

$$E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2,$$

with $z_1 = 1$ only for Democrats, $z_2 = 1$ only for Independents, and $z_1 = z_2 = 0$ for Republicans. Table 12.4 shows some software output for fitting this regression model. No dummy variable estimate appears in the table for party 3 (Republicans), because it is redundant to include a dummy variable for the last category.

IDEOLOGY		Coef.	Std. Err.	t	P>|t|
(Constant)		4.658	0.2126	21.91	<0.0001
PARTY	1	-1.176	0.2567	-4.58	<0.0001
	2	-0.836	0.2369	-3.53	0.0005
	3	0.000			

TABLE 12.4: Software Output for Fitting Regression Model $E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2$ to Data on y = Political Ideology with Dummy Variables z_1 and z_2 for Political Party ID

The prediction equation is $\hat{y} = 4.66 - 1.18z_1 - 0.84z_2$. The coefficients in the prediction equation relate to the sample means in the same manner that the regression parameters relate to the population means. Just as $\alpha = \mu_3$, so does its estimate $4.66 = \bar{y}_3$, the sample mean for Republicans. Similarly, the coefficient of z_1 is $-1.18 = \bar{y}_1 - \bar{y}_3$ and the coefficient of z_2 is $-0.84 = \bar{y}_2 - \bar{y}_3$.

Some software codes factors so that the first category is the one lacking its own dummy variable. The reported model parameter estimates then differ, but they yield the same estimates for differences between category means. For example, R software sets up dummy variables for categories 2 and 3 and yields estimates

	Estimate
(Intercept)	3.48
party2	0.34
party3	1.18

The estimate for party1 (Democrats) is 0, so the estimated difference between the means for Democrats and Republicans is still $0 - 1.18 = -1.18$. ∎

USING REGRESSION FOR A SIGNIFICANCE TEST COMPARING MEANS

For the three groups that are categories of a categorical explanatory variable with three categories, consider H_0: $\mu_1 = \mu_2 = \mu_3$. If H_0 is true, then $\mu_1 - \mu_3 = 0$ and $\mu_2 - \mu_3 = 0$. Recall that $\mu_1 - \mu_3 = \beta_1$ and $\mu_2 - \mu_3 = \beta_2$ in the regression model $E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2$ with dummy variables for categories 1 and 2. So, the hypothesis is equivalent to H_0: $\beta_1 = \beta_2 = 0$ in that model. If all β-values in the model equal 0, then the mean of the response variable equals α for each category.

As usual, we assume randomization. This could be either a single random sample, with subjects then classified by group, or *independent random samples* from the groups. The assumption for inferences in regression modeling that the conditional distributions of y about the regression equation are normal with constant standard deviation corresponds here to the population distributions for the groups being normal, with identical standard deviations.

We can perform the test using the F test of H_0: $\beta_1 = \beta_2 = 0$ for the regression model. As shown in Section 11.3 (page 332), the P-value is the right-tail probability that the F test statistic exceeds the observed F-value. The larger the F test statistic, the smaller the P-value. Table 12.5 shows the ANOVA table for the regression model on political ideology and party ID. The F test statistic equals 10.51, with $df_1 = 2$ and $df_2 = 275$, for testing H_0: $\beta_1 = \beta_2 = 0$, which is equivalently H_0: $\mu_1 = \mu_2 = \mu_3$ for the three party IDs. The P-value is <0.0001, strong evidence against H_0. We conclude that a difference exists among the population mean political ideology values for the three political party IDs.

TABLE 12.5: Software Output of ANOVA Table for Regression Model $E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2$ for y = Political Ideology and Political Party ID. The "regression sum of squares" is called the "model sum of squares" by Stata and SAS.

	Sum of Squares	df	Mean Square	F Value	Prob>F
Regression	36.11	2	18.05	10.51	<0.0001
Residual	472.28	275	1.72		
Total	508.39	277			

ROBUSTNESS AND EFFECTS OF VIOLATIONS OF ASSUMPTIONS

In addition to randomization, each method presented in this chapter assumes that the groups have population distributions that are normal with identical standard deviations. These are stringent assumptions that are never exactly satisfied in practice.

Moderate departures from normality of the population distributions can be tolerated. The F distribution still provides a good approximation to the actual sampling distribution of the F test statistic. This is particularly true for larger sample sizes, since the sampling distributions then have weaker dependence on the shape of the population distribution. Moderate departures from equal standard deviations can also be tolerated. When the sample sizes are identical for the groups, the F test is very robust to violations of this assumption.

Constructing histograms for each sample data distribution helps to check for extreme deviations from these assumptions. Misleading results may occur in the F tests if the population distributions are highly skewed and the sample size is small, or if there are relatively large differences among the population standard deviations (say, the largest sample standard deviation is several times as large as the smallest one) and the sample sizes are unequal. When the distributions are very highly skewed, the mean may not even be an appropriate summary measure.

As in other inferences, the quality of the sample is most crucial. Conclusions may be invalid if the observations in the separate groups compared are not independent random samples.

12.2 Multiple Comparisons of Means

When the P-value is small for comparing several means for groups corresponding to categories of the explanatory variable, this does not indicate which means are different or how different they are. In practice, it is more informative to estimate differences between the population means than merely to test whether they are all equal. Confidence intervals do this. Even if the P-value is not small, it still is informative to determine the plausible sizes of the differences among the population means.

CONFIDENCE INTERVALS COMPARING PAIRS OF MEANS

We can construct a confidence interval for each mean or for each difference between a pair of means. For a categorical variable with g categories corresponding to g groups, denote the sample means by $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_g$ and the corresponding populations by $\mu_1, \mu_2, \ldots, \mu_g$. Let $N = n_1 + n_2 + \cdots + n_g$ denote the total sample size.

A confidence interval for $\mu_i - \mu_j$ is

$$(\bar{y}_i - \bar{y}_j) \pm ts\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

In this formula, s^2 is the residual mean square in the regression model for g groups. The t-value for the chosen confidence level has $df = N - g$.

The t-value is based on df for the variance estimate s^2, which is $df = N - g$ since the model has g parameters. Evidence exists of a difference between μ_i and μ_j when the interval[1] does not contain 0.

Confidence intervals, like tests, do not depend strongly on the normality assumption. When the standard deviations are quite different, with the ratio of the largest to smallest exceeding about 2, it is preferable to use intervals based on separate standard deviations for the groups rather than a single pooled value. For instance, the confidence interval method presented in Section 7.3 for two groups does not assume equal standard deviations.

Example 12.3

Comparing Mean Ideology of Democrats and Republicans For Table 12.1, let's compare population mean ideology of Democrats (group 1) and Republicans (group 3). From Table 12.1 (page 364), $\bar{y}_1 = 3.48$ for $n_1 = 83$ Democrats and $\bar{y}_3 = 4.66$ for $n_3 = 38$ Republicans. From the regression results in Table 12.5 (page 367), the estimate of the population standard deviation is $s = \sqrt{1.72} = 1.31$, with $df = 275$. For a 95% confidence interval with $df = 275$, the t-score is $t_{.025} = 1.97$. The confidence interval for $\mu_3 - \mu_1$ is

$$(\bar{y}_3 - \bar{y}_1) \pm t_{.025}s\sqrt{\frac{1}{n_1} + \frac{1}{n_3}} = (4.66 - 3.48) \pm 1.97(1.31)\sqrt{\frac{1}{83} + \frac{1}{38}}$$

$$= 1.18 \pm 0.51, \quad \text{or} \quad (0.67, 1.68).$$

We infer that population mean ideology was between 0.67 and 1.68 units higher for Republicans than for Democrats. Since the interval contains only positive numbers, we conclude that $\mu_3 - \mu_1 > 0$; that is, μ_3 exceeds μ_1. On the average, Republicans were more conservative than Democrats, with difference between the means 0.67 to 1.68 categories on the seven-category scale. ∎

ERROR RATES WITH LARGE NUMBERS OF CONFIDENCE INTERVALS

With g groups, we can compare $g(g-1)/2$ pairs of groups. When g is relatively large, the number of comparisons can be very large. Confidence intervals for some pairs of means may suggest they are different *even if all of the population means are equal*.

When $g = 10$, for example, there are $g(g-1)/2 = 45$ pairs of means. Suppose we form a 95% confidence interval for the difference between each pair. The error probability of 0.05 applies for each comparison. For the 45 comparisons, the expected number of intervals that would not contain the true differences of means is $45(0.05) = 2.25$.

For 95% confidence intervals, the error probability of 0.05 is the probability that any particular confidence interval will not contain the true difference in population

---

[1] For $g = 2$ groups, $df = N - g = n_1 + n_2 - 2$; this interval then simplifies to the one in Section 7.5 (page 205) introduced for $\mu_2 - \mu_1$ assuming a common standard deviation.

means. When we form a large number of confidence intervals, the probability that *at least* one confidence interval will be in error is much larger than the error probability for any particular interval. The larger the number of groups to compare, the greater is the chance of at least one incorrect inference.

BONFERRONI MULTIPLE COMPARISONS OF MEANS

When we plan many comparisons, methods are available that control the probability that *all* intervals will contain the true differences. Such methods are called ***multiple comparison*** methods. They fix the probability that *all* intervals contain the true differences of population means *simultaneously*, rather than individually.

For example, with a multiple comparison method applied with $g = 10$ means and 95% confidence, the probability equals 0.95 that *all* 45 of the intervals will contain the pairwise differences $\mu_i - \mu_j$. Equivalently, the probability that *at least one* interval is in error equals 0.05. This probability is called the ***multiple comparison error rate***.

The ***Bonferroni multiple comparison*** method is simple and applies to a wide variety of situations. This method uses the same formula for a confidence interval introduced at the beginning of this section. However, it uses a more stringent confidence level for each interval, to ensure that the overall confidence level is sufficiently high.

To illustrate, suppose we would like a multiple comparison error rate of 0.10, that is, a probability of 0.90 that all confidence intervals are simultaneously correct. If we plan four comparisons of means, then the Bonferroni method uses error probability $0.10/4 = 0.025$ for each one. That is, it uses a 97.5% confidence level for each interval. This approach is somewhat conservative: It ensures that the actual overall error rate is *at most* 0.10 and that the overall confidence level is *at least* 0.90. The method is based on a probability inequality employed by the Italian probabilist Carlo Bonferroni in 1935. It states that the probability that at least one of a set of events occurs can be no greater than the sum of the separate probabilities of the events. For instance, if the probability of an error equals 0.025 for each of four confidence intervals, then the probability that at least one of the four intervals will be in error is no greater than $(0.025 + 0.025 + 0.025 + 0.025) = 0.10$.

Example 12.4

Bonferroni Intervals for Political Ideology Comparisons   For the $g = 3$ political party IDs in Table 12.1, let's compare the mean political ideologies: μ_1 with μ_2, μ_1 with μ_3, and μ_2 with μ_3. We construct confidence intervals having overall confidence level at least 0.95. For a multiple comparison error rate of 0.05 with three comparisons, the Bonferroni method uses error probability $0.05/3 = 0.0167$ for each interval. These use the *t*-score with two-tail probability 0.0167, or single-tail probability 0.0083. For $df = 275$, $t_{0.0083} = 2.41$.

The interval for $\mu_3 - \mu_1$, the difference between the population mean ideology of Republicans and Democrats, is

$$(\bar{y}_2 - \bar{y}_1) \pm ts\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (4.66 - 3.48) \pm 2.41(1.31)\sqrt{\frac{1}{83} + \frac{1}{38}}$$

$$= 1.18 \pm 0.62, \quad \text{or} \quad (0.56, 1.79).$$

We construct the intervals for the other two pairs of means in a similar way. Table 12.6 displays them. The interval comparing Democrats and Independents contains 0. They are not significantly different. The intervals comparing Republicans to Democrats and to Independents do not contain 0. They show significant evidence of a difference between the population means for Republicans and the other two groups.   ■

	Difference of Means $\mu_i - \mu_j$	Estimated Difference $\bar{y}_i - \bar{y}_j$	Bonferroni 95% CI	Tukey 95% CI
TABLE 12.6: Bonferroni and Tukey 95% Multiple Comparisons of Mean Political Ideology for Three Political Party ID Groups. The 95% confidence applies to the entire set of three intervals, rather than each individual interval.				
Groups				
(Independent, Democrat)	$\mu_2 - \mu_1$	0.34	(−0.09, 0.77)	(−0.08, 0.76)
(Republican, Democrat)	$\mu_3 - \mu_1$	1.18	(0.56, 1.79)*	(0.57, 1.78)*
(Republican, Independent)	$\mu_3 - \mu_2$	0.84	(0.27, 1.41)*	(0.28, 1.39)*

Note: An asterisk * indicates a significant difference.

The Bonferroni 95% multiple comparison confidence intervals are wider than separate 95% confidence intervals. For instance, the ordinary 95% confidence interval comparing Republicans and Democrats is (0.67, 1.68), whereas the Bonferroni interval is (0.56, 1.79). This is because the multiple comparison method uses a higher confidence level for each separate interval to ensure achieving the desired overall confidence level for the entire set of comparisons.

TUKEY MULTIPLE COMPARISONS OF MEANS

Of the other methods available for multiple comparisons, we recommend **Tukey's method**. Proposed by the great statistician John Tukey, who also developed exploratory data analysis methods such as box plots and stem-and-leaf plots as well as terminology such as *software*, this method has intervals that are slightly narrower than the Bonferroni intervals. This is because they are designed to *approximate* the nominal confidence level rather than to have *at least* that level. The Tukey method uses a probability distribution (the *Studentized range*) that refers to the difference between the largest and smallest sample means. We do not present this distribution in this text, so we rely on software rather than a formula for the Tukey intervals.

Table 12.6 shows Tukey intervals for the political ideology data. For practical purposes, they provide the same conclusions as the Bonferroni intervals.

12.3 Comparing Several Means: Analysis of Variance

The F test for comparing several population means can also be presented without reference to any regression models. The method for doing this is called **analysis of variance** (ANOVA). This is a *test of independence* between the quantitative response variable and the categorical explanatory variable that defines the groups.

For g groups, the analysis of variance is an F test for

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_g.$$
$$H_a: \text{at least two of the population means are unequal.}$$

The assumptions for the ANOVA test are as follows:

- For each group, the population distribution of the response variable y is normal.

- The standard deviation of the population distribution is the same for each group. Denote the common value by σ.

- The samples from the populations are *independent* random samples.

These correspond precisely to the assumptions for the corresponding regression model.

VARIABILITY BETWEEN AND WITHIN GROUPS

Why is a method for comparing population *means* called an analysis of *variance*? The reason is that the F test statistic compares the means by using two estimates of the variance, σ^2, for each group. One estimate uses the variability *between* each sample mean $\bar{y}_i$ and the overall mean $\bar{y}$. The other estimate uses the variability *within* each group of the sample observations about their separate means—the observations from the first group about $\bar{y}_1$, the observations from the second group about $\bar{y}_2$, and so forth.

To illustrate, suppose randomly sampled observations from three groups are as shown in Figure 12.1a. It seems clear that the means of the populations these samples represent are unequal. The basis for this conclusion is that the variability *between* sample means is large and the variability of the observations *within* each sample is small.

By contrast, look at Figure 12.1b. It has the same sample means as in Figure 12.1a, so the variability *between* sample means is the same. But, in Figure 12.1b the variability *within* the groups is much larger than in Figure 12.1a. Now it is not clear whether the population means differ. Generally, the greater the variability between sample means and the smaller the variability within each group, the stronger the evidence against H_0: equal population means.

FIGURE 12.1: Two Cases. The means are the same in each case, so variability *between* groups is the same. Variability *within* groups is less in the first case, which gives stronger evidence against H_0: $\mu_1 = \mu_2 = \mu_3$.
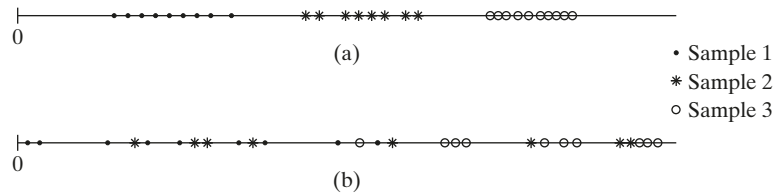
THE F TEST STATISTIC IS A RATIO OF TWO VARIANCE ESTIMATES

For testing H_0: $\mu_1 = \mu_2 = \cdots = \mu_g$, the test statistic is the ratio of the two estimates of the population variance. The estimate that uses the variability *between* each sample mean $\bar{y}_i$, and the overall sample mean $\bar{y}$ is called the **between-groups estimate**. The estimate that uses the variability *within* each sample is called the **within-groups estimate**. The F test statistic has the form

$$F = \frac{\text{Between-groups estimate of variance}}{\text{Within-groups estimate of variance}}.$$

We'll defer the computational details to later in the section. The great British statistician R. A. Fisher developed the analysis of variance method in the 1920s, deriving this test statistic and the F distribution for its sampling distribution.

The within-groups estimate is an unbiased estimate of σ^2 regardless of whether H_0 is true. By contrast, the between-groups estimate is unbiased only if H_0 is true. It then takes about the same value as the within-groups estimate. We then expect values of F near 1.0, apart from sampling error. When H_0 is true, the F test statistic has an F sampling distribution. If H_0 is false, however, the between-groups estimate tends to overestimate σ^2. It then tends to be larger than the within-groups estimate, and F tends to be larger than 1.0, more so with larger samples.

Example 12.5

F Test Comparing Party IDs on Mean Political Ideology   Software displays the results of ANOVA F tests in an **ANOVA table**. Table 12.7 shows an ANOVA table for the F test of $H_0: \mu_1 = \mu_2 = \mu_3$ comparing population mean political ideology for three political party IDs, using the data from Table 12.1.

In an ANOVA table,

- The two "mean squares" (abbreviated as MS by some software) are the *between-groups* and *within-groups* estimates of the population variance σ^2.

- The F test statistic is the ratio of the two mean squares.

From the Mean Square column of Table 12.7, the between-groups estimate of the variance is 18.05 and the within-groups estimate is 1.72. The F test statistic is $F = 18.05/1.72 = 10.5$. In other words, the between-groups estimate is more than 10 times the within-groups estimate. The P-value is < 0.0001, very strong evidence against $H_0: \mu_1 = \mu_2 = \mu_3$. ∎

TABLE 12.7: ANOVA Table for the *F* Test of Equal Means for Table 12.1. The *F* test statistic is the ratio of the mean squares.

Source	Sum of Squares	df	Mean Square	F	Prob > F
Between-groups (party ID)	36.11	2	18.05	10.5	<0.0001
Within-groups	472.28	275	1.72		
Total	508.39	277			

For the ANOVA F test,

$$df_1 = g - 1 = \text{Number of groups} - 1;$$

$$df_2 = N - g = \text{Total sample size} - \text{Number of groups}.$$

These are reported in the df column of the ANOVA table. For these data, $df_1 = g - 1 = 3 - 1 = 2$ and $df_2 = N - g = 278 - 3 = 275$.

In the "Between-groups" row of the ANOVA table, the between-groups sum of squares (SS) divided by df_1 gives a mean square, $36.11/2 = 18.05$. In the "Within-groups" row, the within-groups SS divided by df_2 gives the other mean square, $472.28/275 = 1.72$. The two df terms for the test are the denominators of the two estimates of the variance.

Table 12.7 resembles Table 12.5, the ANOVA table from the regression analysis. The *between-groups SS* in ANOVA is the *regression SS* in the regression analysis. The *within-groups SS* in ANOVA is the *residual SS* (denoted SSE) in regression. This is the variability within the groups unexplained by including parameters in the model to account for the differences between the means. The residual SS divided by its degrees of freedom is the residual mean square, which is the within-groups estimate $s^2 = 1.72$ of the variance of observations for each group. The regression mean square is the between-groups estimate.

WITHIN-GROUPS ESTIMATE OF VARIANCE*

We'll now see how to construct the variance estimates that form the F statistic. The within-groups estimate of σ^2 pools together the sums of squares of the observations about their means. For the n_1 observations from the first group, $\sum(y - \bar{y}_1)^2$ is the sum of squares of the observations about their mean. This sum of squares has $n_1 - 1$

degrees of freedom, the denominator for the sample variance s_1^2 for group 1. For the n_2 observations from the second group, $\sum(y - \bar{y}_2)^2$ is the sum of squares of the observations about their sample mean, with $n_2 - 1$ degrees of freedom. The sum of these SS terms for all the samples is called the ***within-groups sum of squares***, since the sums of squares are calculated *within* each sample.

The within-groups sum of squares has degrees of freedom equal to the sum of the *df* values of the component parts:

$$df = (n_1 - 1) + (n_2 - 1) + \cdots + (n_g - 1) = (n_1 + n_2 + \cdots + n_g) - g$$

$$= N - g = \text{Total sample size} - \text{Number of groups},$$

where N denotes the total sample size. The ratio

$$s^2 = \frac{\text{Within-groups sum of squares}}{df} = \frac{\text{Within-groups SS}}{N - g}$$

is the within-groups estimate of the population variance σ^2 for the g groups.

This estimate summarizes information about variability from the separate samples. The estimate of σ^2 using only the first group is

$$s_1^2 = \frac{\sum(y - \bar{y}_1)^2}{n_1 - 1}.$$

In Table 12.1, for example, this is the square of the reported standard deviation, $s_1 = 1.43$. Similarly, the sample variance for the second group is $s_2^2 = \sum(y - \bar{y}_2)^2/(n_2 - 1)$, and so forth for the remaining groups. Under the assumption that the population variances are identical, these terms all estimate the same parameter, σ^2. The numerator and denominator of s^2 pool the information from these estimates by adding their numerators and adding their denominators. The resulting estimate relates to the separate sample variances by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_g - 1)s_g^2}{N - g}.$$

This estimate is a weighted average of the separate sample variances, with greater weight given to larger samples. With equal sample sizes, s^2 is the mean of the g sample variances.

BETWEEN-GROUPS ESTIMATE OF VARIANCE*

The estimate of σ^2 based on variability between each sample mean and the overall sample mean equals

$$\frac{\sum_i n_i(\bar{y}_i - \bar{y})^2}{g - 1} = \frac{n_1(\bar{y}_1 - \bar{y})^2 + \cdots + n_g(\bar{y}_g - \bar{y})^2}{g - 1}.$$

Exercise 12.44 motivates this formula. Since this estimate describes variability among g means, its

$$df = g - 1 = \text{Number of groups} - 1,$$

which is the denominator of the estimate.

The numerator of this estimate is called the ***between-groups sum of squares***. When the population means are unequal, the $\bar{y}_i$-values tend to be more variable than if the population means are equal. The farther the population means fall from the H_0: $\mu_1 = \cdots = \mu_g$ case, the larger the between-groups SS, the between-groups variance estimate, and the F test statistic all tend to be.

SUMS OF SQUARES IN ANOVA TABLES*

The sum of the within-groups SS and between-groups SS in an ANOVA table is called the ***total sum of squares***. In fact, this is

$$\text{TSS} = \sum (y - \bar{y})^2 = \text{Between-groups SS} + \text{Within-groups SS},$$

the sum of squares of the combined sample of N observations about the overall mean, $\bar{y}$.

The ANOVA partitions the total variability about the overall mean, TSS, into two independent parts. One part, the between-groups SS, is the portion of the total explained by the difference between each group mean and the overall mean. This is also called the *group sum of squares*, and most software replaces the "Between-groups" label in Table 12.7 by the name of the factor that is the group variable (e.g., PARTY ID). The other part, the within-groups SS, is the portion of the total variability that cannot be explained by the differences among the groups.

FOR TWO GROUPS, THE *F* TEST IS EQUIVALENT TO A *t* TEST

With two groups, Section 7.5 (page 205) presented a *t* test that compares the means under the assumption of equal population standard deviations. In fact, if we apply the ANOVA *F* test to data from $g = 2$ groups, the *F* test statistic equals the square of that *t* test statistic. The *P*-value for the *F* test is exactly the same as the two-sided *P*-value for the *t* test. We can use either test to conduct the analysis.

THE KRUSKAL–WALLIS TEST: A NONPARAMETRIC APPROACH*

The ***Kruskal–Wallis test*** is an alternative to ANOVA for comparing several groups. It is a nonparametric method, not requiring the normality assumption. The test statistic uses only the ordinal information in the data. It ranks the observations and compares mean ranks for the various groups. The test statistic is larger when the differences among the mean ranks are larger. It has an approximate chi-squared distribution with $df = g - 1$.

This test is especially useful for samples with which the effects of severe departures from normality may be influential. It is valid for comparing the group distributions, even when the mean is not a relevant parameter. We shall not present the test statistic here. Its result is similar to that of a chi-squared test for the effect of a qualitative factor in a model for an ordinal response presented in Section 15.4.

Nonparametric tests also exist for more complex analyses. In practice, it is more informative to use a modeling approach, because the model parameter estimates give us information about the sizes of effects, which are more important than significance testing. In addition, the modeling strategy presented in Section 12.1 adapts better to multivariate analyses.

12.4 Two-Way ANOVA and Regression Modeling

We've learned how to compare means for groups that are categories of a factor. Sometimes the groups result from cross-classifying two or more factors. For example, the groups (white men, white women, black men, black women) result from cross-classifying race and sex. The ANOVA method for comparing the mean of a quantitative response variable across categories of each of two explanatory factors is called

two-way ANOVA. The ANOVA presented in Section 12.3 for a single explanatory factor is called *one-way ANOVA*.

MAIN EFFECT HYPOTHESES IN TWO-WAY ANOVA

Two-way ANOVA compares population means across categories of two factors. Each null hypothesis states that the population means are identical across categories of one factor, controlling for the other one.

To illustrate, we analyze mean political ideology using political party ID and sex as factors. Six means result from the $2 \times 3 = 6$ combinations of their categories, as Table 12.8 shows. Table 12.8a displays a set of population means satisfying the null hypothesis that mean political ideology is identical for the three party IDs, controlling for sex. Table 12.8b displays a set of population means satisfying the null hypothesis that mean political ideology is identical for females and males, controlling for party ID. The effects of individual factors tested in these two null hypotheses are called *main effects*.

TABLE 12.8: Population Mean Political Ideology Satisfying Main Effect Null Hypotheses: (a) No Effect of Political Party ID, (b) No Effect of Sex

| | | Political Party Identification | | |
Table	Sex	Democrat	Independent	Republican
(a)	Female	3.0	3.0	3.0
	Male	4.6	4.6	4.6
(b)	Female	3.5	4.0	5.0
	Male	3.5	4.0	5.0

F TESTS ABOUT MAIN EFFECTS

The F tests for two-way ANOVA have the same assumptions as the F test for one-way ANOVA: randomization, a normal population distribution for each group, with the same standard deviation for each group. The test statistics have complex formulas except when the sample sizes in all cells are equal. We'll rely on software and corresponding regression modeling. As in one-way ANOVA, the test for a factor effect uses two estimates of the variance for each group. These estimates appear in the mean square (MS) column of the ANOVA table. For testing the main effect for a particular factor, the test statistic is the ratio of mean squares,

$$F = \frac{\text{MS for the factor}}{\text{Residual MS}}.$$

The MS for the factor is a variance estimate based on between-groups variation for that factor. That estimate tends to be inflated when H_0 is not true. The residual MS is a within-groups variance estimate that is always unbiased and is also used in confidence intervals.

The MS variance estimates divide a sum of squares by its df value. For the F test statistics, $df_1 = df$ for the numerator estimate, and $df_2 = df$ for the residual MS. The value of df_1 is always one less than the number of groups being compared.

**Example
12.6** **Two-Way ANOVA for Political Ideology by Political Party ID and Sex**   Table 12.9 shows GSS data from 2014 for political ideology by political party ID and sex. The table also shows the sample means and standard deviations of political ideology, for the scores (1, 2, 3, 4, 5, 6, 7).

TABLE 12.9: GSS Data on Political Ideology by Political Party ID and Sex

Party ID	Sex	1	2	3	4	5	6	7	n	Mean	Std. Dev.
Democrat	Female	33	88	77	208	52	21	6	485	3.51	1.28
	Male	27	79	49	122	21	19	10	327	3.39	1.44
Independent	Female	13	61	60	287	67	51	12	551	3.97	1.21
	Male	12	56	51	224	68	48	13	472	4.01	1.27
Republican	Female	0	7	8	70	60	101	25	271	5.16	1.15
	Male	2	5	10	54	52	105	34	262	5.29	1.22

Let's consider the validity of the assumptions for two-way ANOVA. The sample standard deviations are similar for the six groups (between 1.15 and 1.44). Also, the sample sizes are large (262 and up), so the normality assumption is not crucial, which is important because the observations are quite discrete (a seven-point scale). The full GSS sample was randomly obtained, so we can regard the six samples as independent random samples. ANOVA is suitable for these data.

Stata software reports Table 12.10 for summarizing the analyses. The residual mean square, which estimates the population variance σ^2 within each cell, is

$$s^2 = \text{Residual MS} = \frac{\text{Residual SS}}{df} = \frac{3784.47}{2364} = 1.60.$$

TABLE 12.10: ANOVA Table (Edited from Stata) for Two-Way Analysis of Main Effects of Political Party ID and Sex on Mean Political Ideology

```
Source     | Partial SS        df         MS         F       Prob>F
Model      |    1020.30         3     340.100     212.43     0.0000
           |
    party  |    1020.27         2     510.134     318.66     0.0000
      sex  |       0.03         1       0.032       0.02     0.8870
           |
Residual   |    3784.47      2364       1.601
-----------+-----------------------------------------------------
    Total  |    4804.77      2367       2.030
```

For the null hypothesis of identical mean political ideology for the three political party IDs, controlling for sex, Table 12.10 shows that the F test statistic is

$$F = \frac{\text{Party ID mean square}}{\text{Residual mean square}} = \frac{510.13}{1.60} = 318.66,$$

with $df_1 = 2$ and $df_2 = 2364$. The P-value is 0.0000. Extremely strong evidence exists of a difference among the means. Negligible evidence occurs that mean political ideology varies by sex, within each political party ID (P-value $= 0.89$).   ∎

INTERACTION IN TWO-WAY ANOVA

In practice, before conducting the main effects tests just described, we would first test H_0: no interaction. An absence of interaction between two explanatory variables means that the effect of either variable on y (in the population) is identical at each level of the other.

If no interaction exists between sex and political party ID in their effects on political ideology, then the difference between females and males in population mean political ideology is the same for each political party ID. Table 12.11a shows population means satisfying this. The difference between females and males in mean political ideology is -1.0 for each party. Similarly, the difference between each pair of political parties in mean political ideology is the same for each sex. Figure 12.2 plots the means for the political party ID categories, within each sex. The ordering of categories on the horizontal axis is unimportant, since political party ID is nominal. The absence of interaction is indicated by the parallel sequences of points.

TABLE 12.11: Population Means for a Two-Way Classification, Displaying (a) No Interaction, (b) Interaction

		Political Party Identification		
Table	Sex	Democrat	Independent	Republican
(a)	Female	3.0	3.5	5.0
	Male	4.0	4.5	6.0
(b)	Female	3.0	4.0	5.0
	Male	5.0	4.0	3.0

FIGURE 12.2: Mean Political Ideology, by Political Party ID and Sex, Displaying No Interaction
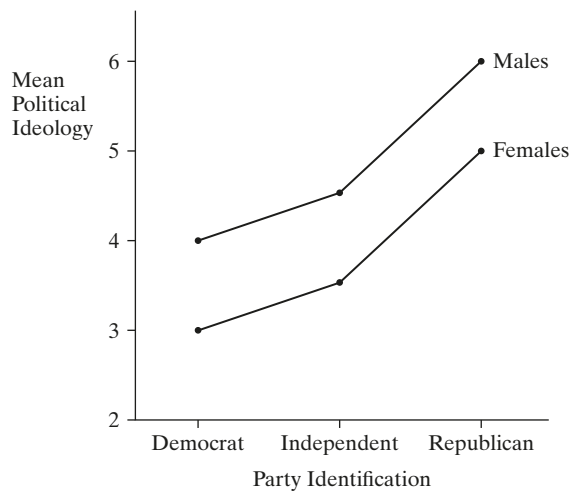
By contrast, Table 12.11b and Figure 12.3 show population means displaying interaction. The difference between females and males in mean political ideology is -2 for Democrats, 0 for Independents, and $+2$ for Republicans. So, the difference depends on the political party ID. Similarly, the political party ID effect on ideology differs for females and males. For females, Republicans are the most conservative, whereas for males, Democrats are the most conservative.

In Table 12.11b, suppose the numbers of males and females are equal, for each political party ID. Then the overall mean political ideology, ignoring sex, is 4.0 for each political party. In a one-way comparison of mean political ideology by party ID, party ID has no effect. However, in a two-way comparison, the interaction implies differing party ID effects for males and females.

The F test statistic for H_0: no interaction is the ratio of a mean square based on the sample degree of interaction divided by the residual mean square. A small P-value suggests that each factor has an effect on the response, but the size of effect varies according to the category of the other factor. In that case, it is not meaningful to test the main effects hypotheses. We conclude that each factor has an effect, but the nature of that effect changes according to the category of the other factor. It is then better to compare the means for one factor separately within categories of the other. On the other hand, if the evidence of interaction is not strong, we then remove the interaction terms from the model and test the two main effect hypotheses.

Table 12.9 showed the sample mean political ideology for the six combinations of political party ID and sex. These means show no obvious evidence of interaction. We'll see that the test of H_0: no interaction has $F = 1.57$ and a P-value of $P = 0.21$. So, a lack of interaction is plausible, and the main effect tests are valid.

Next, it is natural to use confidence intervals to find out more about the political party ID effect, controlling for sex. To do this, it is helpful to study two-way ANOVA in the context of regression modeling.

TWO-WAY ANOVA AS A MULTIPLE REGRESSION ANALYSIS

To conduct the F tests as special cases of tests about parameters of a multiple regression model, we set up dummy variables for each factor. We use the symbol p for dummy variables for political party ID and s as a dummy variable for sex (whether a subject is female). That is,

$$p_1 = \begin{cases} 1 \text{ if subject is Democrat,} \\ 0 \text{ otherwise.} \end{cases}$$

$$p_2 = \begin{cases} 1 \text{ if subject is Independent,} \\ 0 \text{ otherwise.} \end{cases}$$

Both p_1 and p_2 equal 0 when the subject is Republican. Also,

$$s = \begin{cases} 1 \text{ if subject is female,} \\ 0 \text{ if subject is male.} \end{cases}$$

It is redundant to include dummy variables for the final categories.

The multiple regression model assuming a lack of interaction is

$$E(y) = \alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s.$$

To find the correspondence between the population means and the regression parameters, we substitute the possible values for the dummy variables. To illustrate, for Republicans ($p_1 = p_2 = 0$), the mean political ideology is

$$\text{Males } (s = 0): \mu = \alpha + \beta_1(0) + \beta_2(0) + \beta_3(0) = \alpha.$$
$$\text{Females } (s = 1): \mu = \alpha + \beta_1(0) + \beta_2(0) + \beta_3(1) = \alpha + \beta_3.$$

For the six combinations of political party ID and sex, Table 12.12 shows the population means in terms of the regression parameters. For each political party ID, the difference in means between females and males equals β_3. That is, the coefficient β_3 of the dummy variable s for sex equals the difference between females and males in mean political ideology, controlling for political party ID. Also, β_1 is the difference between the means for Democrats and Republicans, and β_2 is the difference between the means for Independents and Republicans, controlling for sex. The null hypothesis of no differences among the parties in mean political ideology, controlling for sex, is $H_0: \beta_1 = \beta_2 = 0$.

TABLE 12.12: Population Mean Political Ideology for the Two-Way Classification of Political Party ID and Sex, Assuming No Interaction

Sex	Political Party ID	Dummy Variables			Population Mean of y
		p_1	p_2	s	$\alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s$
Female	Democrat	1	0	1	$\alpha + \beta_1 + \beta_3$
	Independent	0	1	1	$\alpha + \beta_2 + \beta_3$
	Republican	0	0	1	$\alpha + \beta_3$
Male	Democrat	1	0	0	$\alpha + \beta_1$
	Independent	0	1	0	$\alpha + \beta_2$
	Republican	0	0	0	α

Table 12.13 shows some output for fitting the complete regression model. The prediction equation is

$$\hat{y} = 5.23 - 1.77 p_1 - 1.24 p_2 - 0.01 s.$$

The coefficient -1.77 of p_1 is the estimated difference between Democrats and Republicans in mean political ideology, for each sex. The coefficient -1.24 of p_2 is the estimated difference between Independents and Republicans, for each sex. The estimated difference between Democrats and Independents is $(-1.77) - (-1.24) = -0.53$, for each sex.

Substituting dummy variable values into the prediction equation yields estimated means that satisfy the no interaction model. For instance, for female Republicans, $p_1 = p_2 = 0$ and $s = 1$, so $\hat{y} = 5.23 - 1.77(0) - 1.23(0) - 0.01(1) = 5.22$. These estimated means differ from the sample means in cells of the two-way classification, because they impose the restriction of no interaction.

TABLE 12.13: Fit of Regression Model for Two-Way Analysis of Mean Political Ideology by Party Identification and Sex, Assuming No Interaction. The estimate is 0 at the last level of each predictor, because a dummy variable for that level would be redundant.

Parameter		Estimate	Std. Error	t	Pr(>|t|)
Intercept		5.2289	0.0609	85.80	0.000
party	1	-1.7651	0.0707	-24.97	0.000
	2	-1.2366	0.0676	-18.29	0.000
	3	0	.	.	.
sex	1	-0.0074	0.0524	-0.14	0.887
	2	0	.	.	.

As usual, we use confidence intervals to estimate the sizes of effects. For instance, the estimated difference $\hat{\beta}_1 = -1.765$ between Democrats and Republicans in mean political ideology, controlling for sex, has standard error (reported in Table 12.13) of 0.0707. A 95% confidence interval is $-1.765 \pm 1.96(0.0707)$, or $(-1.90, -1.63)$. For each sex, Democrats are less conservative, on the average, by nearly two categories, quite substantial in practical terms.

REGRESSION MODEL FOR TWO FACTORS WITH INTERACTION

Section 11.4 showed that cross-product terms in a multiple regression model can represent interaction. Here, we take cross products of dummy variables to obtain a regression model that allows interaction effects. The interaction model for the two-way classification of party ID and sex is

$$E(y) = \alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s + \beta_4 (p_1 \times s) + \beta_5 (p_2 \times s).$$

We do not use cross products of dummy variables from categories of the same factor, such as $p_1 p_2$. This is because no more than one dummy variable for a given predictor can be nonzero for any observation, since an observation cannot fall in more than one category. All such cross products would equal 0.

Table 12.14 shows an ANOVA table (using Stata) for the model that allows interaction. The sum of squares for interaction, shown in the row with the product label *party#sex*, is the amount of variability explained by the two interaction terms. It equals the difference between the residual SS without and with these terms in the model. We test H_0: no interaction, that is, H_0: $\beta_4 = \beta_5 = 0$, using

$$F = \frac{\text{Interaction mean square}}{\text{Residual mean square}} = \frac{2.51}{1.60} = 1.57.$$

The test has P-value 0.21, not much evidence of interaction. The simpler model without interaction terms is adequate, and the main effect tests presented in Table 12.10 for party ID and sex are valid. When the evidence of interaction is not significant, it is better to use the model without interaction terms in testing the main effects of the predictors and in constructing confidence intervals for the effects.

PARTIAL SUMS OF SQUARES IN ANOVA TABLES

The sums of squares for party, sex, and their interaction in Tables 12.10 and 12.14 are called ***partial sums of squares***.[2] Some software labels them as ***Type III sums of***

---

[2] See also Exercise 11.60 on page 361.

TABLE 12.14: ANOVA Table (from Stata) for Two-Way Analysis of Mean Political Ideology by Political Party ID and Sex, Allowing Interaction

Source	Partial SS	df	MS	F	Prob>F
Model	1025.32	5	205.06	128.16	0.000
party	1020.27	2	510.13	318.81	0.000
sex	0.03	1	0.03	0.02	0.887
party#sex	5.02	2	2.51	1.57	0.208
Residual	3779.45	2362	1.60		
Total	4804.77	2367	2.03		

squares. They represent the variability in *y* explained by those terms, once the other terms are already in the model. This equals the difference between the residual SS values for the model without those terms and the model with them.

For example, consider the partial SS value of 5.02 for the interaction term in Table 12.14. This equals the difference between the residual SS of 3784.47 for the model without that interaction term (see Table 12.10 on page 376) and the residual SS of 3779.45 with the term in the model (Table 12.14).

Unless the explanatory variables are independent, partial sums of squares for different terms in a model overlap somewhat and do not add up *exactly* to the regression model sum of squares. Likewise, the partial sum of squares explained by a factor depends on which other factors or interaction terms are in the model.

FACTORIAL ANOVA

The methods of two-way ANOVA for two factors extend to several factors. A multifactor ANOVA with observations from all the combinations of the factors is called **factorial ANOVA**. For instance, for three factors, the regression model has a set of dummy variables for each factor and may contain cross products of pairs of dummy variables for the two-factor interactions.

When you have two or more factors, why not instead perform separate one-way ANOVAs? For instance, you could compare the mean political ideology for females and males using a one-way ANOVA, ignoring the information about political party ID, and you could perform a separate one-way ANOVA to compare the means for the three political party IDs, ignoring the information about sex. The main reason is that with factorial ANOVA we can learn whether there is interaction. When there is, it is more informative to compare levels of one factor separately at each level of the other factor. This enables us to investigate how the effect depends on that other factor.

12.5 Repeated-Measures Analysis of Variance*

The methods presented so far assume that the samples in the groups are *independent*, each group having a separate sample of subjects. In many studies, however, each group has the same subjects. Most commonly, this happens when there is *repeated measurement* of the subjects over time or on several related response variables. The samples are then *dependent*, and the analysis must take this into account.

Influences of Entertainment on Children A recent General Social Survey asked subjects to respond about the influence on the daily lives of their children of movies, programs on network television, and rock music. The possible responses for influence were (very negative, negative, neutral, positive, very positive). Table 12.15 shows responses for 12 of the sampled subjects, using scores $(-2, -1, 0, 1, 2)$ for the categories. This is part of a much larger data file for more than 1000 respondents. We analyze only this small sample here, which is the `Influences` data file at the text website.   ∎

TABLE 12.15: Influences on Children of Three Entertainment Types. The influence scores represent $-2 =$ very negative, $-1 =$ negative, $0 =$ neutral, $1 =$ positive, $2 =$ very positive.			
Subject	Movies	TV	Rock
1	−1	0	−1
2	1	0	0
3	0	1	−2
4	2	0	1
5	0	−1	−1
6	−2	−2	−2
7	−1	−1	0
8	0	1	−1
9	−1	−1	−1
10	1	0	1
11	1	1	−1
12	−1	−1	−2
Mean	−0.08	−0.25	−0.75

ONE-WAY ANOVA WITH REPEATED MEASUREMENT

For Table 12.15, H_0 is the same as in ordinary one-way ANOVA: equal population means for the groups. Is there much evidence that the population mean influence differs among movies, TV, and rock music? Ordinary ANOVA is inappropriate because the three samples are not independent. Each sample has the same subjects.

Suppose we regard the rows of Table 12.15, like the columns, as a factor. Then, the data layout resembles a two-way ANOVA. Each cell cross-classifies a subject (a row) with entertainment type (a column). The test comparing population means is then the main effect test for the column variable in the two-way ANOVA. In fact, this approach provides the correct test statistic for this setting. This approach does not extend to more complex analyses, but statistical software has specialized programs for repeated-measures ANOVA[3] that account for the dependence.

Table 12.16 shows the ANOVA table for a repeated-measures analysis of Table 12.15. For

H_0: Equal population mean influence for the three entertainment types,

---

[3] For example, the R or Stata commands shown in Appendix A, or in SPSS using the *Repeated Measures* option after selecting *General Linear Model* in the *Analyze* menu.

the test statistic is $F = 2.55$, with $df_1 = 2$ and $df_2 = 22$. The P-value equals 0.10. The evidence against H_0 is not strong. But, with only 12 subjects, if H_0 is false, the power is probably low.

	TABLE 12.16: Output for Repeated-Measures ANOVA of Influence by Entertainment Type (Movies, TV, Rock Music), for Data in Table 12.15 from the Influences Data File

	Test of Within-Subjects Effects (Sphericity assumed)				
Source	Partial (Type III) SS	df	Mean Square	F	Sig.
Entertainment	2.889	2	1.444	2.55	.101
Residual (Error)	12.444	22	.566		

CONFIDENCE INTERVALS COMPARING DEPENDENT SAMPLES

To compare the three pairs of means simultaneously with confidence intervals, we can use multiple comparison methods. Since $n = 12$ is small, we weaken the multiple comparison confidence level a bit so that the intervals are not overly wide. The 90% Bonferroni confidence intervals use error probability $0.10/3 = 0.0333$ for each interval. The error $df = 22$, and the t-score with probability $0.0333/2 = 0.0167$ in each tail is 2.27. The square root of the residual mean square is $s = \sqrt{0.566} = 0.75$. Each group has 12 observations, so the margin of error for each confidence interval is

$$ ts\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = 2.27(0.75)\sqrt{\frac{1}{12} + \frac{1}{12}} = 0.70. $$

For instance, the confidence interval for the difference between the mean on movies and the mean on rock music is $(-0.08) - (-0.75) \pm 0.70$, or $(-0.03, 1.37)$. It is plausible that the means are equal, but also plausible that the mean for movies is much more in the positive direction than the mean for rock music. Table 12.17 shows all three Bonferroni comparisons. Confidence intervals can convey useful information even if the overall test statistic is not significant.

	TABLE 12.17: Bonferroni Multiple Comparison 90% Confidence Intervals for Comparing Mean Influence for Three Entertainment Types	
Entertainment Types	Difference of Means	Confidence Interval
Movies, TV	0.17	(−0.53, 0.87)
Movies, Rock	0.67	(−0.03, 1.37)
TV, Rock	0.50	(−0.20, 1.20)

THE SPHERICITY ASSUMPTION AND COMPOUND SYMMETRY

The standard repeated-measures ANOVA assumes *sphericity*. Roughly speaking, this means the following. For each pair of groups in the one-way ANOVA, consider the difference between two paired observations, one from each group. This difference is a variable, and the sphericity condition states that the standard deviation of the

distribution of this difference is identical for each pair of groups. It is easier to get a feel for a special case of sphericity, called **compound symmetry**. This condition holds when the different groups have the same standard deviations and when each pair of responses has the same correlation.

When the sphericity assumption is badly violated, the *P*-value tends to be too small. Most software provides a formal significance test (*Mauchly's test*) of the sphericity assumption. When the data strongly contradict that assumption, an approximate test adjusts the degrees of freedom downward for the usual *F* test statistic, using the *Greenhouse–Geisser adjustment*. The technical details for these tests and adjustments are beyond the scope of this text, but standard software reports them.

Using a repeated measurement design can improve precision of estimation. For a one-way ANOVA with repeated measures, having the same subjects in each group helps to eliminate extraneous sources of error. For instance, other variables that affect the response have the same values for each group, so differences between group means cannot reflect differences between groups on those variables. Controlling for possibly confounding factors by keeping them constant for each observation is referred to as **blocking**.

With only two groups and the same subjects in each, Section 7.4 showed that inference uses the *t* distribution with difference scores. For testing equality of means, the *F* statistic from repeated-measures ANOVA then simplifies to the square of the *t* statistic from that matched-pairs *t* test.

FIXED EFFECTS AND RANDOM EFFECTS

Regarding the data file like one for a two-way ANOVA, we can express a regression model for the previous analysis as

$$E(y) = \alpha + \beta_1 m + \beta_2 t + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \gamma_{11} z_{11},$$

where y is the influence response, m is a dummy variable for movies ($m = 1$ for a response on movies, 0 otherwise), t is a dummy variable for TV ($t = 1$ for a response on TV, 0 otherwise), and $m = t = 0$ for a response on rock music. Similarly, z_1 is a dummy variable for subject 1, equaling 1 for that subject's three responses and 0 otherwise, and likewise for 10 other subject dummy variables. We use γ (gamma) instead of β for the coefficients of these terms for convenience, so the index of the parameter agrees with the index of the dummy variable. As usual, each factor has one fewer dummy variable than its number of categories.

A short-hand way to express this regression model is

$$E(y) = \alpha + \beta_j + \gamma_i,$$

where β_j denotes the effect for entertainment type j and γ_i is the effect for subject i. This equation expresses the expected response in the cell in row i and column j in terms of a row main effect and a column main effect. Testing equality of the population mean of y for the three types corresponds to testing H_0: $\beta_1 = \beta_2 = \beta_3$. As usual, to avoid redundancy, we can set $\beta_3 = 0$.

In this model, the $\{\gamma_i\}$ depend on which subjects are in the sample. The subject effect is called a **random effect**, because the categories of the subject factor represent a random sample of all the possible ones. By contrast, the factor that defines the groups, entertainment type, is called a **fixed effect**. The analyses use *all* the categories of interest of a fixed effect, rather than a random sample of them. Models studied in earlier sections of this chapter contained only fixed effects.

We treat random effects in statistical models differently from fixed effects. We'll learn more about this in Section 13.5, which presents regression models of more general form containing both random and fixed effects.

12.6 Two-Way ANOVA with Repeated Measures on a Factor*

Repeated measurement data sets often have more than one fixed effect. The repeated measures usually occur across categories of one factor, but the categories of the other factor(s) have independent samples. The following example illustrates this:

Example 12.8

Comparing Three Treatments for Anorexia For 72 young girls suffering from anorexia, Table 12.18 (the `Anorexia` data file at the text website) shows their weights before and after an experimental period.[4] The girls were randomly assigned to receive one of three therapies during this period. One group, a control group, received the standard therapy. The study analyzed whether one treatment is better than the others, with the girls tending to gain more weight under that treatment.

TABLE 12.18: Weights of Anorexic Girls, in Pounds, before and after Receiving One of Three Treatments

Cognitive Behavioral		Family Therapy		Control	
Weight before	Weight after	Weight before	Weight after	Weight before	Weight after
80.5	82.2	83.8	95.2	80.7	80.2
84.9	85.6	83.3	94.3	89.4	80.1
81.5	81.4	86.0	91.5	91.8	86.4
82.6	81.9	82.5	91.9	74.0	86.3
79.9	76.4	86.7	100.3	78.1	76.1
88.7	103.6	79.6	76.7	88.3	78.1
94.9	98.4	76.9	76.8	87.3	75.1
76.3	93.4	94.2	101.6	75.1	86.7
81.0	73.4	73.4	94.9	80.6	73.5
80.5	82.1	80.5	75.2	78.4	84.6
85.0	96.7	81.6	77.8	77.6	77.4
89.2	95.3	82.1	95.5	88.7	79.5
81.3	82.4	77.6	90.7	81.3	89.6
76.5	72.5	83.5	92.5	78.1	81.4
70.0	90.9	89.9	93.8	70.5	81.8
80.4	71.3	86.0	91.7	77.3	77.3
83.3	85.4	87.3	98.0	85.2	84.2
83.0	81.6			86.0	75.4
87.7	89.1			84.1	79.5
84.2	83.9			79.7	73.0
86.4	82.7			85.5	88.3
76.5	75.7			84.4	84.7
80.2	82.6			79.6	81.4
87.8	100.4			77.5	81.2
83.3	85.2			72.3	88.2
79.7	83.6			89.0	78.8
84.5	84.6				
80.8	96.2				
87.4	86.7				

Source: Thanks to Prof. Brian Everitt, Kings College, London, for these data, which are in the `Anorexia` data file at the text website.

---

[4] We previously analyzed parts of this data set on pages 129, 160, and 206.

Figure 12.4 shows box plots, graphically describing the response distributions before and after the experimental period for each treatment. Table 12.19 shows the summary sample means. The three treatments have similar distributions originally. This is not surprising, because subjects were randomly allocated to the three groups at that time. There is some evidence of a greater mean weight gain for the family therapy group, though there are a few low outlying weight values. ■

FIGURE 12.4: Box Plots for Weights of Anorexic Girls, by Treatment and Time of Measurement
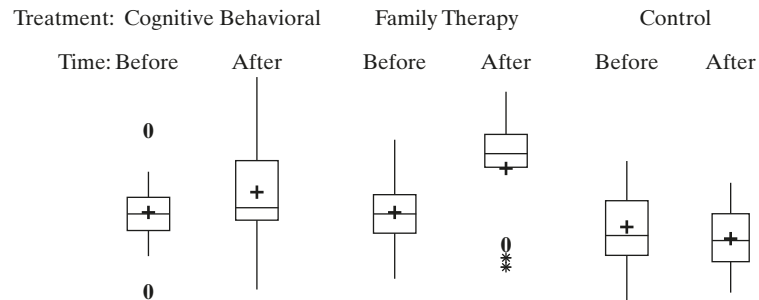
TABLE 12.19: Sample Mean Weight, by Treatment and Time of Measurement, in Anorexia Study

| | Time | |
Treatment	Before	After
Cognitive behavioral (CB)	82.7	85.7
Family therapy (FT)	83.2	90.5
Control (C)	81.6	81.1

REPEATED MEASURES ON ONE OF TWO FIXED EFFECTS

This experiment had two fixed effects. One of them, "Treatment," has categories (CB = cognitive behavioral, FT = family therapy, C = control). It defines three groups of girls, represented by three independent samples. The second, "Time," consists of the two times for observations (before, after). Each time has the same subjects, so the samples at its levels are dependent. Time is called a ***within-subjects factor***, because comparisons of its categories use repeated measurements on subjects. Treatment is called a ***between-subjects factor***, because comparisons of its categories use different subjects.

Although the two factors (treatment and time) are fixed effects, the analysis differs from ordinary two-way ANOVA. This is because the repeated measurements on the within-subjects factor (time) create a third effect, a random effect for subjects. Each subject is measured at every category of time. Subjects are said to be ***crossed*** with the time factor. Each subject occurs at only one category of the between-subjects factor (treatment). Subjects are said to be ***nested*** within the treatment factor.

As in ordinary two-way ANOVA, we can test each main effect as well as their interaction. However, tests about the within-subjects factor (both its main effect and its interaction with the other fixed effect) use a different error term than the test about the between-subjects main effect. The ordinary sum of squared errors term is partitioned into two parts. One part uses the variability between mean scores of

subjects. It forms an error term for testing the between-subjects factor. The other part is based on how the pattern of within-subject scores varies among subjects. It forms an error term for any test involving the within-subjects factor. Figure 12.5 shows the partitioning of the total sum of squares for a two-way ANOVA with repeated measures on one factor.

FIGURE 12.5: Partitioning of Variability in Two-Way ANOVA with Treatment and Time Factors and Repeated Measures on Time. Tests involving the within-subjects factor (time) use a separate error term.
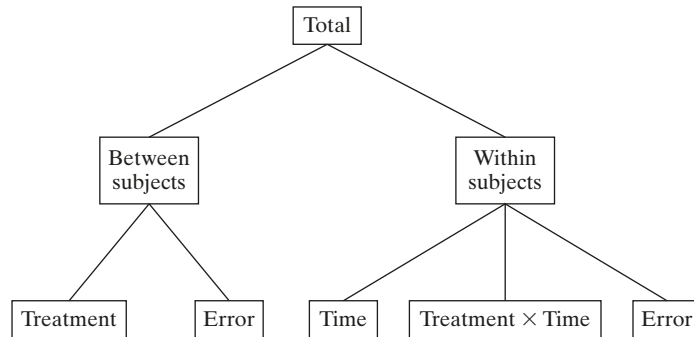
ANOVA *F* Tests Comparing Anorexia Treatments   Table 12.20 shows software output for the analysis of the anorexia study data. Since treatment has three categories, it has two dummy variables in the regression model, and its sum of squares has $df = 2$. Since time has two levels, it has one dummy variable, and its sum of squares has $df = 1$. The interaction between these effects has two terms in the model, based on the cross product of the two dummy variables for treatment with the dummy variable for time, so its $df = 2$.

TABLE 12.20: Software Output for Two-Way ANOVA of the `Anorexia` Data File with Treatment and Time Fixed Effects and Repeated Measures on Time

```
                 Tests of Within-Subject Effects
                              Partial            Mean
Source                     (Type III) SS  df    Square    F      Sig
TIME                           366.04      1    366.04  12.92   0.001
TIME*TREATMENT                 307.32      2    153.66   5.42   0.006
Residual (subject|treatment)  1955.37     69     28.34


                Tests of Between-Subjects Effects
                    Partial
Source           (Type III) SS   df   Mean Square   F     Sig
TREATMENT            644.23       2      322.12     6.20  0.003
Residual (Error)    3584.03      69       51.94
```

The residual term for the between-subjects part of the table has $df = 69$, based on 28 dummy variables for the 29 subjects receiving CB therapy, 16 dummy variables for the 17 subjects receiving FT, and 25 dummy variables for the 26 subjects in group C ($28 + 16 + 25 = 69$). The remaining variability, not accounted for by this residual term or by the main effects and interaction terms, is the residual sum of squares for the within-subjects effects.

The TIME*TREATMENT row of the ANOVA table indicates that the interaction is highly significant (P-value $= 0.006$). The difference between population means for the two times differs according to the treatment, and the difference between population means for a pair of treatments varies according to the time. Because of the

significant interaction, we do not test the main effects. We instead use confidence intervals to describe the interaction. ∎

FOLLOW-UP CONFIDENCE INTERVALS

From the sample means in Table 12.19, the evidence of interaction is clear. The sample means for the three treatments are similar at the initial time. At the second time, by contrast, the mean for the control group is similar to its mean at the initial time, but the mean is larger for the other two treatments than their initial means, especially for the FT group.

To construct confidence intervals comparing means at the two times, for each treatment, the appropriate common standard deviation estimate is the square root of the residual mean square from the within-subjects analysis. From Table 12.20, this equals $\sqrt{28.34} = 5.3$, with $df = 69$. We illustrate by constructing a 95% confidence interval comparing the two means for family therapy (FT), which 17 girls received at each time. The t-score for 95% confidence when $df = 69$ is 1.99. The confidence interval is

$$(90.5 - 83.2) \pm 1.99(5.3)\sqrt{\frac{1}{17} + \frac{1}{17}}, \quad \text{or} \quad 7.3 \pm 3.6, \quad \text{or} \quad (3.6, 10.9).$$

We conclude that the population mean weight is between 3.6 and 10.9 pounds higher following the treatment period. Similarly, a 95% confidence interval comparing the two means equals (0.2, 5.8) for the CB therapy and (−3.4, 2.5) for the control group. There is evidence of an increase, albeit a small one, for the CB therapy, but no evidence of change for the control group.

To make between-subjects comparisons of treatments, for each time, we cannot use the residual mean square from the between-subjects analysis. The reason is that these separate comparisons involve both the treatment main effect and the interaction, and these two sources of variation have different error terms in the repeated-measures ANOVA. At a particular time, however, the subjects in the three treatments are independent samples. Thus, we can compare three means at a given time using a one-way ANOVA F test or using confidence intervals for those data alone.

For instance, for the 72 observations at time = after, the F test statistic for the one-way ANOVA comparing the three means has P-value = 0.0004, very strong evidence of a difference among the treatment means. For this one-way ANOVA, the square root of the residual mean square is $s = 7.3$. The 95% confidence interval for the difference between the FT and the CB treatment means, based on the $17 + 29$ observations for the two groups, is

$$(90.5 - 85.7) \pm 1.99(7.3)\sqrt{\frac{1}{17} + \frac{1}{29}}, \quad \text{or} \quad 4.8 \pm 4.4, \quad \text{or} \quad (0.4, 9.2).$$

The true means may be essentially equal at the follow-up time, but if they differ, the advantage could be quite noticeable for the family therapy. Table 12.21 shows the confidence intervals for each pair of treatments.

TABLE 12.21: 95% Confidence Intervals Comparing Treatment Means after the Treatment Period			
Treatments Compared	Difference of Sample Means	Confidence Interval	Bonferroni Interval
FT − CB	4.8	(0.4, 9.2)	(−0.7, 10.3)
FT − C	9.4	(4.9, 13.9)	(3.8, 15.0)
CB − C	4.6	(0.7, 8.5)	(−0.2, 9.4)

In summary, evidence suggests that the mean weight increases during the experimental period for both noncontrol treatments. Marginal evidence suggests that the mean is higher after the experiment for the FT treatment than for the CB treatment.

At this stage, further interest may relate to whether the change in means, between time = after and time = before, differed for the two noncontrol treatments. That is, do the difference scores for the FT treatment have a significantly higher mean than the difference scores for the CB treatment? The difference scores have a mean of $(90.5 - 83.2) = 7.3$ for the FT treatment and $(85.7 - 82.7) = 3.0$ for the CB treatment, and we used these as the basis of separate confidence intervals for the mean change, above. Since the two groups are independent samples, the variance of the difference of these means is the sum of the variances. A 95% confidence interval for the difference between the mean changes in weight is

$$(7.3 - 3.0) \pm 1.99(5.3)\sqrt{\frac{1}{17} + \frac{1}{17} + \frac{1}{29} + \frac{1}{29}}, \quad \text{or} \quad 4.3 \pm 4.6, \quad \text{or} \quad (-0.3, 8.8).$$

Although the mean change could be considerably larger for the FT treatment, it is also plausible that the mean changes could be identical.

BONFERRONI MULTIPLE COMPARISONS OF TREATMENTS

To control the overall error rate for several comparisons, we can use the Bonferroni multiple comparison method. Suppose we use three confidence intervals to compare treatments at time = after and three confidence intervals to compare times within the treatments. To ensure at least 90% confidence for the entire set, since $0.10/6 = 0.0167$, we use a 98.33% confidence interval for each individual comparison. If we focus on the three intervals comparing treatments at time = after, these are Bonferroni 95% confidence intervals.

Such intervals are wider than the ones just reported, since they use a *t*-score of 2.45 instead of 1.99. Table 12.21 shows them for the pairwise comparisons of treatments at time = after. With this more conservative approach, only the difference between the FT and C treatments is significant, with the interval not containing 0.

OTHER REPEATED-MEASURES ANALYSES

In the anorexia study, the repeated measurements occurred at two times. When observations occur at several times, the repeated-measures ANOVA is more complex. In particular, the results depend on assumptions about the correlation structure of the repeated measurements. The standard test for the within-subject effect assumes *sphericity*, as in one-way repeated-measures ANOVA. Tests of the between-subjects effects are not affected by violation of the sphericity assumption, so no adjustment is needed for that *F* test.

In some studies with two fixed effects, repeated measures occur on both factors. For instance, we may observe the same subjects for each treatment at each of several times. Then, subjects (a random effect) are crossed with both factors (fixed effects), and an observation occurs for every subject at every combination of factor levels. As in ordinary two-way ANOVA, the effects of interest refer to the fixed effects—their main effects and interaction.

Two-way ANOVA with repeated measures on one or both factors extends to multiple factors. For example, a study may have three factors, A, B, and C, with repeated measures on C. Then subjects are crossed with C but nested within combinations of levels of A and B. The complicating factor of these cases is that each test may require a separate residual mean square. However, software can easily conduct the

analyses. For details on the linkage of ANOVA methods with experimental designs, see Howell (2012), Kirk (2012), and Winer et al. (1991).

MANOVA: MULTIVARIATE ANALYSIS OF VARIANCE

A more general ANOVA can test hypotheses comparing the means without any assumptions about the correlation structure. It treats the set of repeated measures on a subject as a multivariate vector of responses that has a multivariate version of the normal distribution. By doing this, methods are available that are designed for multivariate response data. The tests are called MANOVA, short for ***multivariate analysis of variance***. The particular MANOVA test referred to by most software as *Wilks' lambda* is a *likelihood-ratio test*, the idea of which we will introduce in Section 15.3.

The MANOVA approach has its own disadvantages, however. The main one is that it loses power as the sacrifice for having weaker assumptions. If the assumptions for traditional repeated-measures ANOVA are not badly violated, that method has higher power for testing effects. MANOVA has lower power because it requires estimating a larger number of parameters.

REPEATED MEASURES USING MIXED MODELS AND CORRELATION STRUCTURE

Repeated-measures ANOVA methods have limitations:

- The index for the repeated measurement enters in the model the same way for each subject. In longitudinal studies, this means that we need to observe all subjects at the same time points.

- The method cannot deal with missing data. Subjects with *any* missing observations get dropped from the analysis. This can lead to significant bias and inefficiency when the missing data are considerable.

- The correlations among observations on the same subject are assumed to satisfy a *sphericity* structure, which is implied by common variability at all times and the same correlation between each pair of observations.

Other types of methods are available for repeated measurement data. They vary in the assumptions they make and in how they model the correlation structure of the repeated observations. We just mentioned the multivariate ANOVA approach. It also has the first two disadvantages, and has potentially low power. Section 13.5 introduces a more general linear modeling approach with random effects that models the correlation structure for the repeated responses, permits observations at different time points, and accommodates subjects in the analysis when some of their observations are missing. The models also include fixed effects for the explanatory variables (such as treatment) for which the analyses use all the categories of interest, and can have both categorical and quantitative explanatory variables. Because its effects are a mixture of random and fixed effects, the model is called a ***mixed model***.

12.7 Chapter Summary

Chapters 9 and 11 presented regression models for a quantitative response variable when the explanatory variables are also *quantitative*. This chapter has modeled a quantitative response variable as a function of *categorical* explanatory variables, called ***factors***. Models of the next chapter include both quantitative and categorical explanatory variables.

This chapter also presented **analysis of variance** (ANOVA) methods for comparing several groups according to their means on a quantitative response variable. The groups are categories of categorical explanatory variables.

- Analysis of variance methods are a special case of multiple regression analyses. **Dummy variables** in the regression model are indicators that represent the groups. Each dummy variable equals 1 for a particular group and 0 otherwise.

- **Multiple comparison** methods provide confidence intervals for differences between pairs of means, while controlling the overall error probability. The **Bonferroni** method does this using an error probability for each comparison that equals the desired overall error probability divided by the number of comparisons.

- **One-way ANOVA** methods compare means for categories of a single factor. **Two-way ANOVA** methods compare means across categories of each of two factors. Assuming no interaction, the main effects describe the effect of each factor while controlling for the other one. Ordinary ANOVA methods compare groups with *independent* random samples from the groups.

- For longitudinal and repeated-measures studies, different samples have the same subjects, and are *dependent*. Methods for **repeated-measures ANOVA** result from regression models with **random effects** that represent the effects of the random sample of observed subjects. Such methods treat **within-subjects** effects (for repeated measurements on subjects) differently from **between-subjects** effects (for independent samples of subjects).

Exercises

Practicing the Basics

12.1. For GSS data comparing the reported number of good friends for those who are (married, widowed, divorced, separated, never married), an ANOVA table reports $F = 0.80$.

(a) Specify the null and alternative hypotheses for the test.

(b) Software reports a P-value of 0.53. Explain how to interpret it.

(c) State the hypotheses tested in terms of parameters of a regression model. Define variables in that model.

12.2. A General Social Survey asked subjects how many good friends they have. Is this associated with the respondent's astrological sign (the 12 symbols of the zodiac)? The ANOVA table for the GSS data reports $F = 0.61$ based on $df_1 = 11$, $df_2 = 813$.

(a) Specify the null and alternative hypotheses for the analysis.

(b) Software reports a P-value of 0.82. Explain how to interpret it.

(c) State a regression model and corresponding null hypothesis that can yield these results. Define variables in that model.

12.3. A recent General Social Survey asked, "What is the ideal number of kids for a family?" Show how to define dummy variables, and formulate a model for this response with explanatory variable religious affiliation (Christian, Muslim, Jewish, Other or none).

12.4. Refer to the previous exercise. Table 12.22 shows an ANOVA table for the model.

(a) Specify the hypotheses tested in this table.

(b) Report the F test statistic value and the P-value. Interpret the P-value.

(c) Based on (b), can you conclude that *every* pair of religious affiliations has different population means for ideal family size? Explain.

TABLE 12.22

Source	SS	df	Mean Square	F	Sig (Prob>F)
Religion	11.72	3	3.91	5.48	0.001
Residual (Error)	922.82	1295	0.71		
Total	934.54	1298			

12.5. A recent GSS asked, "How often do you go to a bar or tavern?" Table 12.23 shows descriptive statistics and an ANOVA table for comparing the mean reported number of good friends at three levels of this variable.

(a) State the (i) hypotheses, (ii) test statistic value, (iii) *P*-value, (iv) decision for an $\alpha = 0.05$-level test.

(b) Does any aspect of the summary here suggest that an assumption for the *F* test may be badly violated? Explain.

(c) Set up dummy variables, and show the prediction equation you would obtain, based on the results shown in the table.

TABLE 12.23

	How often go to bar or tavern?		
	Very often	Occasional	Never
Mean no. good friends	12.1	6.4	6.2
Standard deviation	21.3	10.2	14.0
Sample size	41	166	215

Source	Sum of Squares	df	Mean Square	F	Prob>F
Group	1116.8	2	558.4	3.03	0.049
Residual (Error)	77171.8	419	184.2		
Total	78288.5	421			

12.6. Table 12.24 shows scores on the first quiz (maximum score 10 points) in a beginning French course. Students in the course are grouped as follows:

Group A: Never studied foreign language before, but have good English skills.
Group B: Never studied foreign language before; have poor English skills.
Group C: Studied other foreign language.

Using software for regression or ANOVA, conduct a test comparing the means. Report the assumptions, hypotheses, test statistic, and *P*-value. Interpret the *P*-value.

TABLE 12.24

Group A	Group B	Group C
4	1	9
6	5	10
8		5

12.7. Refer to the previous exercise.

(a) Suppose that the first observation in the second group was actually 9, not 1. Then, the standard deviations are the same, but the sample means are 6, 7, and 8 rather than 6, 3, and 8. Do you think the *F* test statistic would be larger, the same, or smaller? Explain your reasoning, without doing any calculations.

(b) Suppose you had the same means as these data, but the sample standard deviations were 1.0, 1.8, and 1.6, instead of the actual 2.0, 2.8, and 2.6. Do you think the *F* test statistic would be larger, the same, or smaller? Explain your reasoning.

(c) Suppose you had the same means and standard deviations as these data, but the sample sizes were 30, 20, and 30, instead of 3, 2, and 3. Do you think the *F* test statistic would be larger, the same, or smaller? Explain your reasoning.

(d) In (a), (b), and (c), would the *P*-value be larger, the same, or smaller? Why?

12.8. In a study to compare customer satisfaction at service centers for PC technical support in San Jose (California), Toronto (Canada), and Bangalore (India), each center randomly sampled 100 people who called during a two-week period. Callers rated their satisfaction on a scale of 0 to 10, with higher scores representing greater satisfaction. The sample means were 7.6 for San Jose, 7.8 for Toronto, and 7.1 for Bangalore. Table 12.25 shows an ANOVA table.

(a) Explain how to obtain the *F* test statistic value reported in the table from the mean square values shown. Report the df_1 and df_2 values for the *F* distribution, and report and interpret the *P*-value.

TABLE 12.25

Source	SS	df	MS	F	Prob>F
Group	26.00	2	13.00	27.6	0.000
Residual (Error)	140.00	297	0.47		
Total	60.00	299			

(b) Explain why the margin of error for separate 95% confidence intervals is the same (0.19) for comparing the population means for each pair of cities. Construct and interpret the three intervals.

(c) The margin of error for Bonferroni or for Tukey 95% multiple comparison confidence intervals is 0.23. Why is it different from that in (b), and what is an advantage of this approach?

(d) With dummy variables to represent the service centers, the prediction equation is $\hat{y} = 7.1 + 0.5z_1 + 0.7z_2$. Show how the terms in this equation relate to the sample means of 7.6, 7.8, and 7.1.

12.9. For g groups with $n = 100$ each, we plan to compare all pairs of population means. We want the probability to equal at least 0.80 that the entire set of confidence intervals contains the true differences. For the Bonferroni method, which t-score multiple of the standard error should we use for each interval if **(a)** $g = 10$, **(b)** $g = 5$? Describe how the t-score depends on g, and explain the implication regarding width of the intervals.

12.10. A recent GSS asked, "Would you say that you are very happy, pretty happy, or not too happy?" and "About how many good friends do you have?" Table 12.26 summarizes results, with number of friends as the response variable.

(a) State a research question you could answer with these data.

(b) Interpret the result of the F test, but indicate one assumption of the test that is clearly violated.

(c) Software reports Tukey 95% confidence intervals of (0.3, 5.7) comparing very happy and pretty happy, $(-2.3, 6.5)$ comparing very happy and not too happy, and $(-5.1, 3.3)$ comparing pretty happy and not too happy. Interpret.

12.11. When we use the GSS to evaluate how the mean number of hours a day watching TV depends on sex and race, for subjects of age 18–25, we get the results shown in Table 12.27. The sample means were 2.66 for white females, 2.62 for white males, 3.48 for black females, and 3.14 for black males. Explain how these results seem to be compatible with the results of the tests shown.

12.12. A recent GSS asked, "What is the ideal number of kids for a family?" Table 12.28 shows results of evaluating the effects of gender and race.

(a) Explain how to interpret the results of the F tests.

(b) Let $s = 1$ for females and 0 for males, and let $r = 1$ for blacks and 0 for whites. The no interaction model has $\hat{y} = 2.42 + 0.04s + 0.37r$. Find the estimated mean for each combination of gender and race. Explain how these means satisfy "no interaction."

TABLE 12.26

	Very happy	Pretty happy	Not too happy
Mean	10.4	7.4	8.3
Standard deviation	17.8	13.6	15.6
Sample size	276	468	87

Source	Sum of Squares	df	MS	F	Prob>F
Group	1626.8	2	813.4	3.47	0.032
Residual (Error)	193900.9	828	234.2		
Total	195527.7	830			

TABLE 12.27

Source	SS	df	MS	F	Prob>F
Sex	2.22	1	2.22	0.35	0.555
Race	489.65	1	489.65	76.62	0.000
Residual (Error)	11094.16	1737	6.39		
Total	11583.81	1739			

TABLE 12.28

Source	SS	df	MS	F	P-value
Gender	0.25	1	0.25	0.36	0.550
Race	16.98	1	16.98	24.36	0.000
Residual	868.67	1246	0.70		
Total	886.12	1248			

12.13. Table 12.13 on page 380 gave the prediction equation $\hat{y} = 5.23 - 1.77p_1 - 1.24p_2 - 0.01s$ relating political ideology to political party ID and to sex. Find the estimated means for the six cells, and show that they satisfy a lack of interaction.

12.14. Using software with the Houses data set at the text website, conduct an ANOVA for y = house selling price with factors whether the house is new and whether number of bathrooms exceeds two.

TABLE 12.29

Religion	Sex	Political Ideology Mean	Political Ideology Std Dev.	Sex	Political Ideology Mean	Political Ideology Std. Dev.
Protestant	Female	4.18	1.39	Male	4.28	1.51
Catholic	Female	3.97	1.26	Male	4.05	1.46
Jewish	Female	3.22	1.63	Male	3.00	1.52
None	Female	3.86	1.60	Male	3.85	1.61

(a) Using $\alpha = 0.05$, test the hypothesis of no interaction between the factors in their effects on y.

(b) Assuming no interaction, conduct the test of the hypothesis that the population mean of y is the same for new and older homes, controlling for the bathrooms indicator. Interpret.

12.15. For the 2014 GSS, when we regress y = number of hours per day watching TV on s = sex (1 = male, 0 = female) and religious affiliation ($r_1 = 1$ for Protestant, $r_2 = 1$ for Catholic, $r_3 = 1$ for Jewish, $r_1 = r_2 = r_3 = 0$ for none or other), we get $\hat{y} = 2.7 + 0.1s + 0.4r_1 + 0.2r_2 - 0.2r_3$.

(a) Interpret the coefficient of r_1.

(b) State a corresponding model for the population, and indicate which parameters must equal zero for y to be independent of religious affiliation, for each sex.

12.16. In the United States, the Bureau of Labor Statistics recently reported that for males the current population mean hourly wage is $22 for white-collar jobs, $11 for service jobs, and $14 for blue-collar jobs. For females, the means are $15 for white-collar jobs, $8 for service jobs, and $10 for blue-collar jobs.

(a) Identify the response variable and the two factors.

(b) Show these means in a two-way classification of the two factors.

(c) Compare the differences between males and females for (i) white-collar jobs, (ii) blue-collar jobs. Explain why there is interaction, and describe it.

12.17. In 2013, the U.S. Census Bureau reported that the population median income was $29,127 for white females, $26,006 for black females, $41,086 for white males, and $30,394 for black males.

(a) Identify the response variable and the two factors, and show these medians in a two-way classification of the factors.

(b) Explain why there is interaction in terms of the median.

(c) Show four population median incomes that would satisfy H_0: no interaction.

12.18. Table 12.29 summarizes responses on political ideology in the 2014 General Social Survey by religion and sex. The P-value is <0.01 for testing H_0: no interaction. Explain what this means in the context of this example, and indicate one place in the table that may be responsible for the small P-value.

12.19. Table 12.30 shows results of an ANOVA on y = depression index by gender and marital status (married, never married, divorced). State the sample size and fill in the blanks in the ANOVA table. Interpret results.

TABLE 12.30

Source	Sum of Squares	df	Mean Square	F	Sig
Gender	100	—	—	—	—
Marital status	200	—	—	—	—
Interaction	100	—	—	—	—
Residual (error)		—	—		
Total	4000	205			

12.20. Twenty-five students in a statistics class were surveyed about their attitudes toward divorce. Each received a response score according to how many from a list of seven possible reasons were regarded as legitimate for a woman to seek a divorce. The students were also asked whether they were fundamentalist or nonfundamentalist in their religious beliefs and whether their religious attendance was frequent (more than once a month) or infrequent. Table 12.31 below shows the data.

(a) Using regression methods in software, fit the model that assumes no interaction. Interpret parameter estimates.

TABLE 12.31

Visit Religious Places	Religion Fundamentalists	Religion Nonfundamentalists
Frequent	0, 1, 0, 0, 2, 3, 0, 0, 2, 2	1, 4, 0, 2, 2
Infrequent	3, 2, 2	5, 6, 3, 4, 5, 5, 4

(b) Test the main effects for a two-way ANOVA, assuming no interaction. Interpret.

(c) Analyze whether an analysis permitting interaction is more appropriate.

12.21. The prediction equation $\hat{y} = 16 + 2s + 3r + 8(s \times r)$ relates $y =$ annual income (thousands of dollars), $s =$ sex ($s = 1$ for men, $s = 0$ for women), and $r =$ race ($r = 1$ for whites, $r = 0$ for blacks). By finding the four predicted means for this equation, show that the coefficient 8 of the interaction term is the amount by which the mean for one of the four groups must increase or decrease for the interaction to disappear.

12.22. For the 2014 GSS, Table 12.32 shows sample means of political ideology (higher values being more conservative), classified by gender and by race, for those over 50 in age. For H_0: no interaction, software reports $F = 21.7$, $df_1 = 1$ and $df_2 = 1081$, and P-value <0.001.

(a) Suppose that instead of the two-way ANOVA, you performed separate one-way ANOVAs for gender and for race. Suppose the ANOVA for gender does not show a significant effect. Explain how this could happen, even though the two-way ANOVA implies that the gender effect varies by race. (*Hint*: Will the overall sample means for females and males be more similar than they are for each race?)

(b) Summarize what you would learn about the gender effect from a two-way ANOVA that you would fail to learn from a one-way ANOVA.

TABLE 12.32

Gender	Race Black	Race White
Female	3.75 (n = 95)	4.23 (n = 484)
Male	3.46 (n = 52)	4.36 (n = 454)

12.23. Refer to Table 12.15 (page 382) about the influence of three entertainment types on children.

(a) Using software, conduct the repeated-measures analyses of Section 12.5.

(b) Suppose you scored the influence categories $(-3, -2, 0, 2, 3)$. What would this assume about the response categories? Repeat the analyses using these scores. Are the conclusions sensitive to the choice of scores?

12.24. Recently the General Social Survey asked respondents, "Compared with 10 years ago, would you say that American children today are (1) much better off, (2) better off, (3) about the same, (4) worse off, or (5) much worse off." Table 12.33 shows opinion responses for 10 of the subjects on three issues: quality of their education, safety of the neighborhoods they live in, and getting health care when they need it.

(a) For each of the following, indicate whether it is a fixed effect, random effect, or response variable: (i) opinion, (ii) issue, (iii) subject.

(b) Test the hypothesis that the population means are equal. Report the P-value, and interpret.

(c) The first five respondents were female, and the last five were male. Analyze these data using both gender and issue as factors.

TABLE 12.33

	Issue		
Subject	Education	Neighborhood	Healthy Care
1	4	4	3
2	2	4	2
3	3	3	4
4	1	2	1
5	3	4	3
6	2	5	4
7	1	4	2
8	3	3	3
9	4	5	3
10	2	4	2

12.25. The General Social Survey asks respondents to rate various groups using the "feeling thermometer" on a scale of 0 (most unfavorable) to 100 (most favorable). We plan to study how the mean compares for rating liberals and rating conservatives, for ratings in 2016 and ratings in 1986. Explain why a two-way ANOVA using time (1986, 2016) and group rated (Liberal, Conservative) as factors would require methods for repeated measures. Identify the within-subjects and between-subjects factors.

12.26. Using software, conduct the repeated-measures ANOVA of the anorexia data in Table 12.18 (page 385), available at the text website. Interpret results.

Concepts and Applications

12.27. Refer to the Students data file (Exercise 1.11 on page 21), with response variable the number of weekly hours engaged in sports and other physical exercise. Using software, conduct an analysis of variance and follow-up estimation, and prepare a report summarizing your analyses and interpretations using **(a)** gender as the sole factor; **(b)** gender and whether a vegetarian as factors.

12.28. For $y =$ number of times used public transportation in previous week and $x =$ number of cars in family (which takes value 0, 1, or 2 for the given sample), explain the difference between conducting a test of independence of the variables using the ANOVA F test for comparing three means and using a regression t test for the coefficient of the number of cars in an ordinary regression model with a linear effect for number of cars. Give an example of three population means for which the regression test would be less appropriate than the ANOVA test. (*Hint*: What does

the regression linear model assume that the ANOVA F test does not?)

12.29. Go to the GSS website `sda.berkeley.edu/GSS`.

(a) Analyze the change over time (GSS variable YEAR) in the mean of political ideology (POLVIEWS) by political party identification (PARTYID). Compare strong Republicans to strong Democrats in 1974 and in the latest survey, and summarize the rather dramatic change.

(b) For the latest survey, report the sample mean political ideology for the 2×2 cross-classification of RACE (using black and white) and SEX. For the model $E(y) = \alpha + \beta_1 r + \beta_2 s$, where $r = 1$ for white and $r = 0$ for black and $s = 1$ for male and $s = 0$ for female, give approximate values (to one decimal place) that you would obtain for $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$.

12.30. A study[5] described an experiment that randomly assigned participants to receive \$3 to spend on themselves (self-interest), or to receive \$3 to donate to a nonprofit charity (imposed charity), or to receive \$3 that they could either spend on themselves or donate to charity (choice). After receiving or donating the money, the participants rated how happy they were with this experience, on a seven-point scale (with $1 = $ not at all, $7 = $ an extreme amount). The authors reported a significant effect ($F = 9.08, P < 0.001$), with follow-up t-tests confirming their hypothesis that imposing self-interest increases outcome happiness. For their analysis, identify the response variable, the explanatory factor, and the hypothesis tested to yield the reported F statistic and P-value.

12.31. A study[6] compared verbal memory of men and women for abstract words and for concrete words. It found a gender main effect in favor of women. It also reported, "There was no sex $\times$ word-type interaction ($F = 0.408, P = 0.525$), indicating that women were equally advantaged on the two kinds of words." How would you explain what this sentence means to someone who has never studied statistics?

12.32. (a) Explain carefully the difference between a probability of Type I error of 0.05 for a single comparison of two means and a multiple comparison error rate of 0.05 for comparing all pairs of means.

(b) In multiple comparisons following a one-way ANOVA with equal sample sizes, the margin of error with a 95% confidence interval for comparing each pair of means equals 10. Give three sample means illustrating that it is possible that group A is not significantly different from group B and group B is not significantly different from group C, yet group A is significantly different from group C.

12.33. For a two-way classification of means by factors A and B, at each level of B the means are equal for the levels of A. Does this imply that the overall means are equal at the various levels of A, ignoring B? Explain the implications, in terms of how results may differ between two-way ANOVA and one-way ANOVA.

12.34. Table 7.29 (page 224) summarized a study that reported the mean number of dates in the past three months. For men, the mean was 9.7 for the more attractive and 9.9 for the less attractive. For women, the mean was 17.8 for the more attractive and 10.6 for the less attractive. Identify the response variable and the factors, and indicate whether these data appear to show interaction. Explain what you learn from a two-way ANOVA that you cannot learn from a one-way ANOVA.

12.35. Construct a numerical example of means for a two-way classification under the following conditions:

(a) Main effects are present only for the row variable.

(b) Main effects are present for each variable, with no interaction.

(c) Interaction effects are present.

(d) No effects of any type are present.

12.36. The 25 women faculty in the humanities division of a college have a mean salary of \$76,000, and the five women in the science division have a mean salary of \$90,000. The 20 men in the humanities division have a mean salary of \$75,000, and the 30 men in the science division have a mean salary of \$89,000.

(a) Construct a table of sample mean incomes for the 2×2 cross-classification of gender and division of college. Find the overall means for men and women. Interpret.

(b) Discuss how the results of a one-way comparison of mean incomes by gender would differ from the results of a two-way comparison of mean incomes by gender, controlling for division of college. (*Note*: This reversal of which gender has the higher mean salary, according to whether one controls division of college, illustrates *Simpson's paradox*. See Exercise 10.14 in Chapter 10.)

12.37. Refer to Exercise 12.20. The students were also asked about their attitudes toward abortion. Each received a score according to how many from a list of eight possible reasons for abortion she would accept as a legitimate reason for a woman to seek abortion. Table 12.34 below displays the scores, classified by religion and

TABLE 12.34

Visit Religious	Religion	
Places	Fundamentalists	Nonfundamentalists
Frequent	0, 1, 3, 0, 0, 3, 4, 2, 2	2, 3, 1, 5, 2, 1
Infrequent	3, 4, 3	6, 4, 3, 6, 4, 6, 7, 8

---

[5] By J. Berman and D. Small, *Psychological Science*, vol. 23 (2012), pp. 1193–1199.

frequency of visit to religious places. Using software, analyze the data and report your findings in a short report.

12.38. True or false? Suppose that for subjects aged under 55, there is little difference in mean annual medical expenses for smokers and nonsmokers, but for subjects over 55 there is a large difference. Then, there is no interaction between smoking status and age in their effects on annual medical expenses.

Select the correct response(s) in Exercises 12.39–12.42. (More than one response may be correct.)

12.39. Analysis of variance and regression are similar in the sense that

(a) They both assume a quantitative response variable.

(b) They both have F tests for testing that the response variable is statistically independent of the explanatory variable(s).

(c) For inferential purposes, they both assume that the response variable y is normally distributed with the same standard deviation at all combinations of levels of the explanatory variable(s).

(d) They both provide ways of partitioning the variation in y into explained and unexplained components.

12.40. One-way ANOVA provides relatively more evidence that $H_0: \mu_1 = \cdots = \mu_g$ is false

(a) The smaller the between-groups variation and the larger the within-groups variation.

(b) The smaller the between-groups variation and the smaller the within-groups variation.

(c) The larger the between-groups variation and the smaller the within-groups variation.

(d) The larger the between-groups variation and the larger the within-groups variation.

12.41. For four means, a multiple comparison method provides 95% confidence intervals for the differences between the six pairs. Then

(a) For each confidence interval, there is a 0.95 chance that it contains the population difference.

(b) P(all six confidence intervals are correct) $= 0.70$.

(c) P(all six confidence intervals are correct) $= 0.95$.

(d) P(all six confidence intervals are correct) $= (0.95)^6$.

(e) P(at least one confidence interval does not contain the true difference) $= 0.05$.

(f) The confidence intervals are wider than separate 95% confidence intervals for each difference.

12.42. Interaction terms are needed in a two-way ANOVA model when

(a) Each pair of variables is associated.

(b) Both explanatory variables have significant effects in the model without interaction terms.

(c) The difference in means between two categories of one explanatory variable varies greatly among the categories of the other explanatory variable.

(d) The mean square for interaction is huge compared to the error mean square.

12.43. Use the *ANOVA* applet at `www.pearsonglobal editions.com/Agresti` to illustrate how between-groups and within-groups variability affect the result of the ANOVA F test. Print results of two scenarios that result in relatively large and relatively small P-values.

12.44.* This exercise motivates the formula for the between-groups variance estimate in one-way ANOVA. Suppose the sample sizes all equal n and the population means all equal μ. The sampling distribution of each $\bar{y}_i$ then has mean μ and variance σ^2/n. The sample mean of the $\bar{y}_i$ values is $\bar{y}$.

(a) Treating $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_g$ as g observations having sample mean $\bar{y}$, explain why $\sum(\bar{y}_i - \bar{y})^2/(g-1)$ estimates the variance σ^2/n of the sampling distribution of the $\bar{y}_i$-values.

(b) Using (a), explain why $\sum n(\bar{y}_i - \bar{y})^2/(g-1)$ estimates σ^2. For the unequal sample size case, replacing n by n_i yields the between-groups estimate.

12.45.* You form a 95% confidence interval in five different situations, with independent samples.

(a) Find the probability that (i) *all* five intervals contain the parameters they are designed to estimate, (ii) at least one interval is in error. (*Hint*: Use the binomial distribution.)

(b) If you use confidence level 0.9898 for each interval, the probability that all five intervals contain the parameters equals exactly 0.95. Explain why. (*Hint*: What is $(0.9898)^5$?) Compare 0.9898 to the confidence coefficient for each interval in the Bonferroni method.

---

[6] By D. Kimura and P. Clarke, *Psychological Reports*, vol. 91 (2002), pp. 1137–1142.

This page intentionally left blank

MULTIPLE REGRESSION WITH QUANTITATIVE AND CATEGORICAL PREDICTORS

Chapter 11 introduced the multiple regression model to analyze the relationship between a quantitative response variable and *quantitative* explanatory variables. Chapter 12 showed that multiple regression models can also handle *categorical* explanatory variables, by constructing dummy variables. In this chapter, we see that multiple regression can simultaneously handle quantitative and categorical explanatory variables.

In the last chapter, we learned that models with a single categorical explanatory variable focus on comparing the mean of y for several groups. The analysis of variance (ANOVA) F test relates to that model. In many applications, it is useful to compare means while controlling for other variables, some of which may be quantitative. For example, in comparing mean income for men and women in some profession, we might control for possibly differing levels of job experience between men and women. The quantitative control variable measuring job experience is called a ***covariate***. The use of regression for this type of comparison is often called ***analysis of covariance***. It is one of the many statistical contributions of R. A. Fisher, the brilliant British statistician.

Because effects may change after controlling for a variable, the results of analysis of covariance may differ from the results of analysis of variance. For instance, job experience is usually positively correlated with income. If men tend to have higher levels of experience than women in the profession studied, the results of a comparison of mean income for men and women will depend on whether we control for experience.

In this chapter, we first show graphic representations of using both categorical and quantitative explanatory variables. In regression models, we again use dummy variables for qualitative explanatory variables. The models enable us to analyze effects of variables while controlling for both quantitative and categorical explanatory variables. For example, we can adjust sample means of y for different groups to reflect their predicted values after controlling for covariates.

The final section of the chapter introduces a more general model, called the *linear mixed model*, which can have both quantitative and categorical explanatory variables but also includes *random effects*. Whereas the ordinary regression model assumes that all observations are independent, the linear mixed model handles situations in which some observations are correlated. This type of model is useful for repeated-measures experiments, longitudinal studies, and for applications with clusters of observations such as families, since the observations are not all independent in such studies.

13.1 Models with Quantitative and Categorical Explanatory Variables

We introduce concepts using a single quantitative explanatory variable, denoted by x, and a single categorical factor, denoted by z. When the categorical variable has two categories, z is a dummy variable; when it has several categories, we use a set

of dummy variables. The analysis of the effect of x refers to the regression of y on x within each category of the categorical variable, treating z as a control variable. The analysis of the effect of the categorical variable z refers to comparing the means of y for the groups defined by z, treating x as the control variable.

COMPARING REGRESSION LINES

Table 9.5 (page 280) introduced a data file on $y = $ selling price of homes. One quantitative explanatory variable is $x = $ size of home. One categorical variable is $z = $ whether a house is new ($1 = $ yes, $0 = $ no). Studying the effect of x on y while controlling for z is equivalent to analyzing the regression of y on x separately for new and older homes. We could find the best-fitting straight line for each set of points, one line for new homes and a separate line for older homes. We could then compare characteristics of the lines, for instance, whether they climb with similar or different slopes.

In this context, *no interaction* means that the true slope of the line relating expected selling price to the size of home is the same for new and older homes. Equality of slopes implies that the regression lines are parallel. See Figure 13.1a. When the y-intercepts are also equal, the regression lines coincide. See Figure 13.1b. If the rate of increase in selling price as a function of size of home differed for new and existing homes, then the two regression lines would not be parallel. There is then interaction. See Figure 13.1c.

FIGURE 13.1:
Regression Lines between Quantitative Response and Quantitative Explanatory Variable, within Categories of a Categorical Variable with Two Categories

(a) No interaction

(b) No interaction, with identical y-intercepts

(c) Interaction
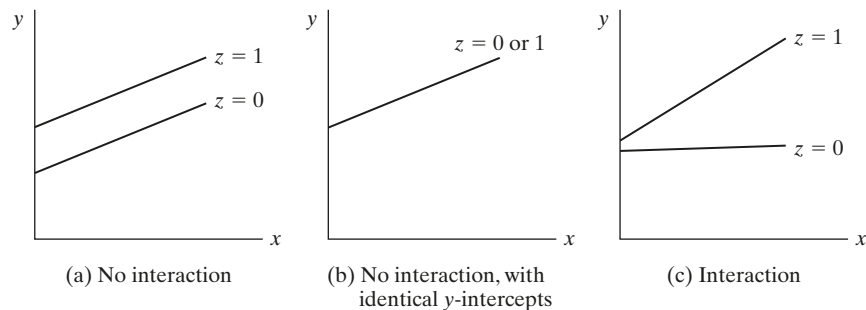
The effect of x while controlling for z may differ in substantial ways from the bivariate relationship. For instance, the effect could disappear when we control for z. Figure 13.2 displays a set of points having an overall positive relationship when z is

FIGURE 13.2: An Association between Two Quantitative Variables that Disappears after Controlling for a Categorical Variable
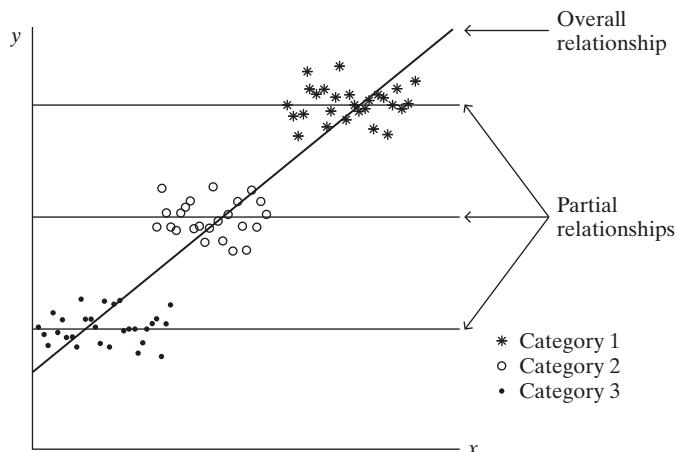
ignored. Within each category of z, however, the regression line relating y to x is horizontal. The overall positive trend is due to the tendency for the categories with high (low) scores on y to have high (low) scores on x also. Example 10.1 in Chapter 10 (page 303) presented an example of this type, with y = math achievement test score and x = height. The categorical variable was grade of school, with students coming from grades 2, 5, and 8.

COMPARING MEANS ON y, CONTROLLING FOR x

Likewise, the effect of the categorical variable z may change substantially when we control for x. For example, consider the relationship between y = annual income and z = gender for managerial employees of a chain of fast-food restaurants. From a two-sample comparison of men and women, mean annual income is higher for men than for women. In this company, annual income of managers tends to increase with x = number of years of experience. In addition, only recently have women received many managerial appointments, so on the average they have less experience than the men. In summary, men tend to have greater experience, and greater experience tends to correlate with higher income. Perhaps this is why the overall mean annual income is higher for men. A chain relationship may exist, with gender affecting experience, which itself affects income. The difference between the mean incomes of men and women could disappear when we control for experience.

To study whether the difference in mean incomes can be explained by differing experience levels of men and women, we compare mean incomes for men and women having equal levels of experience. If there is no interaction, then the regression line between income and experience for the male employees is parallel to the one for the female employees. In that case, the difference between mean incomes for men and women is identical for all fixed values of x = number of years of experience. Figure 13.3a illustrates this. If the same regression line applies to each gender, as in Figure 13.3b, the mean income for each gender is identical at each level of experience. In that case, no difference occurs between male and female incomes, controlling for experience.

FIGURE 13.3: Three Scenarios for the Regression of y = Income on x = Number of Years of Experience and z = Gender

(a) No interaction   (b) No interaction, with identical regression lines   (c) Interaction
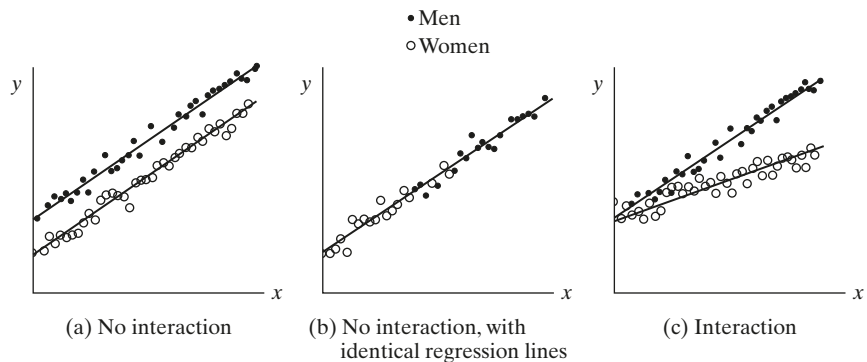
The results of this analysis may differ considerably from a comparison of mean incomes while ignoring rather than controlling for experience. For example, Figure 13.3b depicts a situation in which the sample mean income for men is much greater than that for women. However, the reason for the difference is that men have more experience. In fact, the same regression line fits the relationship between income and experience for both genders. It appears that the mean incomes are equal, controlling for experience.

If interaction exists, then the regression lines are not parallel. In that case, the difference between the mean incomes varies by level of experience. In Figure 13.3c, for example, the mean income for men is higher than the mean income for women at all experience levels, and the difference increases as experience increases. Example 13.6 in this chapter shows an example of this type.

Example 13.1

Regression of Income on Education and Racial–Ethnic Group   For a sample of adult Americans aged over 25, Table 13.1 shows $y =$ annual income (thousands of dollars), $x =$ number of years of education (where $12 =$ high school graduate, $16 =$ college graduate), and $z =$ racial–ethnic group (black, Hispanic, white). The data exhibit patterns of a much larger sample taken by the U.S. Bureau of the Census. The sample contains $n_1 = 16$ blacks, $n_2 = 14$ Hispanics, and $n_3 = 50$ whites, for a total sample size of $N = 80$.

TABLE 13.1: Observations on $y =$ Annual Income (in Thousands of Dollars) and $x =$ Number of Years of Education, for Three Racial–Ethnic Groups

Black		Hispanic		White		White		White	
y	x	y	x	y	x	y	x	y	x
16	10	32	16	30	14	62	16	50	16
18	7	16	11	48	14	24	10	50	14
26	9	20	10	40	7	50	13	22	11
16	11	58	16	84	18	32	10	26	12
34	14	30	12	50	10	34	16	46	16
22	12	26	10	38	12	52	18	22	9
42	16	20	8	30	12	24	12	24	9
42	16	40	12	76	16	22	14	64	14
16	9	32	10	48	16	20	13	28	12
20	10	22	11	36	11	30	14	32	12
66	16	20	10	40	11	24	13	38	14
26	12	56	14	44	12	120	18	44	12
20	10	32	12	30	10	22	10	22	12
30	15	30	11	60	15	82	16	18	10
20	10			24	9	18	12	24	12
30	19			88	17	26	12	56	20
				46	16	104	14		

Note: The data are in the `Income` data file at the text website.

Table 13.2 reports the mean income and education for these subjects. Although the mean incomes differ among the three groups, these differences could result from the differing educational levels. For instance, although white subjects had higher mean incomes than blacks or Hispanics, they also had higher mean education. Perhaps the differences would disappear if we could control for education, making comparisons among the racial–ethnic groups at fixed levels of education.

TABLE 13.2: Mean Income and Education, by Racial–Ethnic Group

	Black	Hispanic	White	Overall
Mean income	$\bar{y}_1 = 27.8$	$\bar{y}_2 = 31.0$	$\bar{y}_3 = 42.5$	$\bar{y} = 37.6$
Mean education	$\bar{x}_1 = 12.2$	$\bar{x}_2 = 11.6$	$\bar{x}_3 = 13.1$	$\bar{x} = 12.7$
Sample size	$n_1 = 16$	$n_2 = 14$	$n_3 = 50$	$N = 80$

As in Section 12.1, we represent a categorical factor in a regression model using dummy variables, one fewer than the number of categories. With three categories, the regression model is

$$E(y) = \alpha + \beta x + \beta_1 z_1 + \beta_2 z_2.$$

Here, β (without a subscript) describes the effect of $x =$ education on the mean of y for each racial–ethnic group. For racial–ethnic status, one way to set up the dummy variables is

$z_1 = $ 1 if subject is black, $z_1 = 0$ otherwise;
$z_2 = $ 1 if subject is Hispanic, $z_2 = 0$ otherwise;
$z_1 = $ $z_2 = 0$ if subject is white.

Table 13.3 shows some output from using software to fit the regression model. The [race = b] and [race = h] parameters refer to the coefficients of the dummy variables z_1 for blacks and z_2 for Hispanics. The prediction equation is

$$\hat{y} = -15.7 + 4.4x - 10.9z_1 - 4.9z_2.$$

For blacks, $z_1 = 1$ and $z_2 = 0$, so the prediction equation is

$$\hat{y} = -15.7 + 4.4x - 10.9(1) - 4.9(0) = -26.6 + 4.4x.$$

The prediction equations for the other two racial–ethnic groups are

$$\hat{y} = -20.6 + 4.4x \quad \text{(Hispanics)};$$

$$\hat{y} = -15.7 + 4.4x \quad \text{(whites)}.$$

TABLE 13.3: Output for Fitting Model to Table 13.1 from the `Income` Data File on $y =$ Income and Explanatory Variables Education and Racial–Ethnic Status, with Dummy Variables for Black and Hispanic Categories

Parameter	Coef.	Std. Error	t	Sig	95% Conf. Int. Lower	Upper
Intercept	-15.663	8.412	-1.862	.066	-32.4	1.09
education	4.432	.619	7.158	.000	3.2	5.7
[race = b]	-10.874	4.473	-2.431	.017	-19.8	-2.0
[race = h]	-4.934	4.763	-1.036	.304	-14.4	4.6
[race = w]	0	.	.	.		

race=w parameter is set to zero because it is redundant
R-Squared = .462

Figure 13.4 is a scatterplot showing the prediction equations for the three groups. The lines are parallel, since they each have the same slope, 4.43. In each prediction equation, 4.43 is the coefficient of x, reflecting the increase for each group in the mean of y per one-year increase in education. The parallelism reflects the lack of interaction terms for this model. Since z_1 is a dummy variable for blacks, the coefficient -10.9 of z_1 represents the difference ($-\$10,900$) between the estimated annual mean income for blacks and for whites, controlling for education. The estimated mean income is \$10,900 lower for blacks than for whites, at each fixed level of education. Since z_2 is a dummy variable for Hispanics, the coefficient -4.9 of z_2 represents the difference ($-\$4900$) between the estimated mean income for Hispanics and whites, controlling for education. ∎
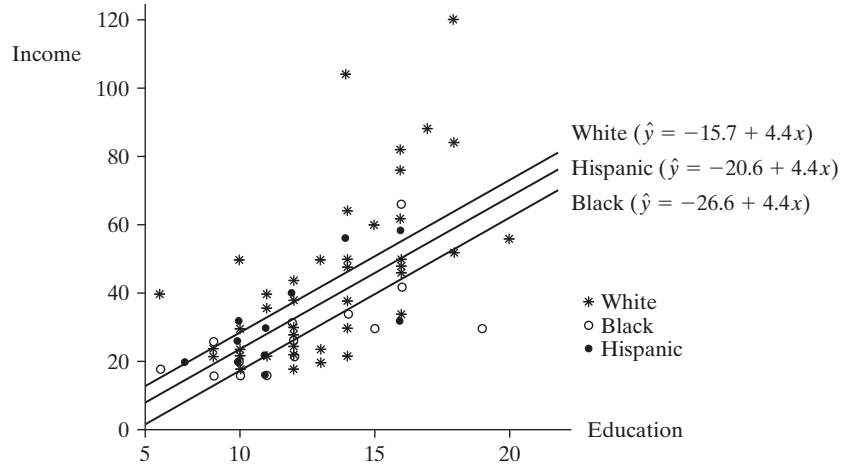
In summary, the coefficients of the dummy variables estimate differences in means between each category and the final category, which does not have its own dummy variable. Some software (such as R and Stata) uses the first category instead of the final category as the baseline that does not have its own dummy variable. The coefficients of the dummy variables then estimate differences in means between each category and the first category, controlling for the other variables in the model.

Example 13.2

Regression of Income on Education and Racial–Ethnic Group, Permitting Interaction
A model that allows interaction between a quantitative explanatory variable x and a categorical factor z allows a different slope for the effect of x in each category of z. To allow interaction, as usual we take cross products of the explanatory variables. For Table 13.1, we take cross products $x \times z_1$ and $x \times z_2$ of the dummy variables z_1 and z_2 for blacks and Hispanics with the education explanatory variable.

Software provides the results shown in Table 13.4. The overall prediction equation is

$$\hat{y} = -25.9 + 5.2x + 19.3z_1 + 9.3z_2 - 2.4(x \times z_1) - 1.1(x \times z_2).$$

TABLE 13.4: Output for Fitting Interaction Model to Table 13.1 from the `Income` Data File on Income, Education, and Racial–Ethnic Status

Parameter	Coef.	Std. Error	t	Sig
Intercept	-25.869	10.498	-2.464	.016
education	5.210	.783	6.655	.000
[race=b]	19.333	18.293	1.057	.294
[race=h]	9.264	24.282	.382	.704
[race=w]	0	.	.	.
[race=b]*education	-2.411	1.418	-1.700	.093
[race=h]*education	-1.121	2.006	-.559	.578
[race=w]*education	0	.	.	.

race=w parameters are set to zero because they are redundant
R-Squared   0.482

The prediction equation with both dummy variables equal to zero ($z_1 = z_2 = 0$) refers to the third racial–ethnic category, namely, whites. For that group,

$$\hat{y} = -25.9 + 5.2x + 19.3(0) + 9.3(0) - 2.4x(0) - 1.1x(0) = -25.9 + 5.2x.$$

For the first category (blacks), $z_1 = 1$, $z_2 = 0$, and

$$\hat{y} = -6.6 + 2.8x.$$

For the second category (Hispanics), $z_1 = 0$, $z_2 = 1$, and

$$\hat{y} = -16.6 + 4.1x.$$

The coefficient 19.3 of z_1 describes the difference between the y-intercepts for blacks and whites. However, this is the difference *only* at $x = 0$, since the equations have different slopes. Since the 5.2 coefficient of x represents the slope for whites, the coefficient of $(x \times z_1)$ (i.e., -2.4) represents the *difference in slopes* between blacks and whites. The two lines are parallel only when that coefficient equals 0. Similarly, for the second category, the coefficient of z_2 is the difference between the y-intercepts for Hispanics and whites, and the coefficient of $(x \times z_2)$ is the difference between their slopes. Table 13.5 summarizes the interpretations of the estimated parameters in the model.

TABLE 13.5: Summary of Prediction Equation
$\hat{y} = -25.9 + 5.2x + 19.3z_1 + 9.3z_2 - 2.4(x \times z_1) - 1.1(x \times z_2)$
Allowing Interaction, with $z_1 = 1$ for Blacks and $z_2 = 1$ for Hispanics

Group	y-Intercept	Slope	Prediction Equation	Difference from White of y-Intercept	Slope
Black	$-25.9 + 19.3$	$5.2 - 2.4$	$(-25.9 + 19.3) + (5.2 - 2.4)x$	19.3	-2.4
Hispanic	$-25.9 + 9.3$	$5.2 - 1.1$	$(-25.9 + 9.3) + (5.2 - 1.1)x$	9.3	-1.1
White	-25.9	5.2	$-25.9 + 5.2x$	0	0

Figure 13.5 plots the three prediction equations. The sample slopes are all positive. Over nearly the entire range of education values observed, whites have the highest estimated mean income, and blacks have the lowest.

FIGURE 13.5: Plot of Prediction Equations for Model with Interaction Terms
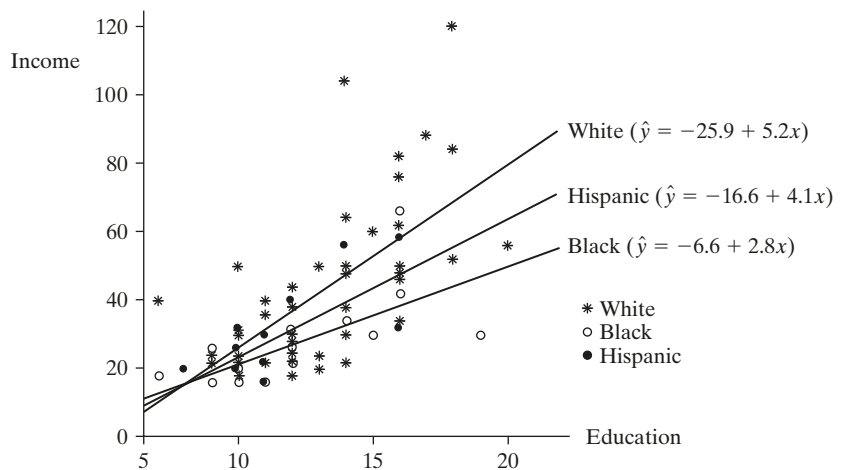
When interaction exists, the difference between means of y for two groups varies as a function of x. For example, the difference between the estimated mean of y for

whites and Hispanics at a particular x-value is

$$(-25.9 + 5.2x) - (-16.6 + 4.1x) = -9.3 + 1.1x.$$

This depends on the value of x. As the education level x increases, the difference between the estimated mean incomes is larger. Figure 13.5 shows that the difference between the mean incomes of whites and blacks also gets larger at higher education levels. When a variable occurs in an interaction term, it is inappropriate to use the main effect term to summarize its effect, because that variable's effect changes as the value changes of a variable with which it interacts.

To summarize how much better the model permitting interaction fits, we can check the increase in R^2 or in the multiple correlation R. From the output for the no-interaction model (Table 13.3 on page 403), $R^2 = 0.462$. From the output for the interaction model (Table 13.4), $R^2 = 0.482$. The corresponding multiple correlation values are $\sqrt{0.462} = 0.680$ and $\sqrt{0.482} = 0.695$. Little is gained by fitting the more complex model, as R^2 and R do not increase much. ∎

REGRESSION WITH MULTIPLE CATEGORICAL AND QUANTITATIVE PREDICTORS

The models generalize to add explanatory variables of either type. To introduce additional quantitative variables, add a βx term for each one. To introduce another categorical variable, add a set of dummy variables for its categories. To permit interaction, introduce cross-product terms.

With several explanatory variables, the number of potential models is quite large when we consider the possible main effect and interaction terms. Also, some variables may overlap considerably in the variation they explain in the response variable, so it may be possible to simplify the model by dropping some terms. Using inference, as described in the next section, helps us select a model.

13.2 Inference for Regression with Quantitative and Categorical Predictors

This section presents inference methods for models that contain both quantitative and categorical explanatory variables. As in other multivariable models, we first test whether the model needs interaction terms. We test hypotheses about model parameters using the F test comparing complete and reduced regression models, introduced in Section 11.5. For instance, the test of H_0: no interaction between two explanatory variables compares the complete model containing their cross-product interaction terms to the reduced model deleting them. This test has a small P-value if the addition of the interaction terms provides a significant improvement in the fit.

Example 13.3

Testing Interaction of Education and Racial–Ethnic Group in Their Effects on Income
For Table 13.1, we now test H_0: no interaction between education and racial–ethnic group, in their effects on income. The complete model,

$$E(y) = \alpha + \beta x + \beta_1 z_1 + \beta_2 z_2 + \beta_3 (x \times z_1) + \beta_4 (x \times z_2),$$

contains two interaction terms. The null hypothesis is H_0: $\beta_3 = \beta_4 = 0$. The model under H_0 has a common slope β for all three lines relating $E(y)$ to x. Figure 13.6 depicts the hypotheses for this test.

FIGURE 13.6: Graphical Representation of Null and Alternative Hypotheses in a Test of No Interaction, for a Categorical Factor with Three Categories and a Quantitative Explanatory Variable x
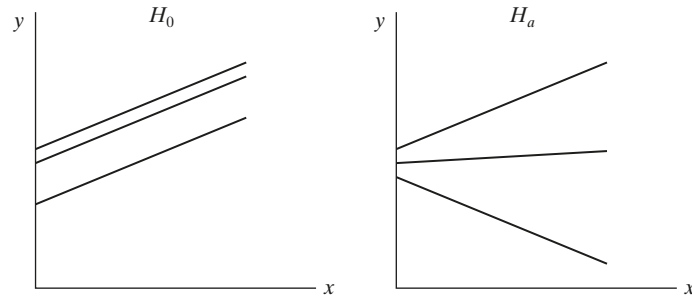
Table 13.6 shows how software summarizes sums of squares explained by various sets of terms in the model with interaction terms. The variability explained by the interaction terms, 691.8, equals the difference between the SSE values without and with those terms in the model. These sums of squares are *partial sums of squares* (see page 380). They represent the variability explained after the other terms are already in the model.

TABLE 13.6: Software Output of Partial Sums of Squares Explained by Education, Racial–Ethnic Group, and Their Interaction, in the Model Permitting Interaction

```
                   Partial Sum            Mean
Source             of Squares   df       Square      F      Sig
Race                  267.319    2       133.659    .566    .570
Education            6373.507    1      6373.507  26.993    .000
Race*Education        691.837    2       345.918   1.465    .238
Residual (Error)    17472.412   74       236.114
Total               33761.950   79
```

For H_0: no interaction, the F test statistic is the ratio of the interaction mean square to the residual mean square. Table 13.6 shows that the test statistic is $F = 345.9/236.1 = 1.46$, with a P-value of 0.24. There is not much evidence of interaction. We are justified in using the simpler model without cross-product terms. ∎

TESTS FOR INDIVIDUAL PARTIAL EFFECTS

Possibly the model can be simplified further, if either of the main effects is not significant. For the test of the main effect for the categorical factor, racial–ethnic group, the null hypothesis states that each racial–ethnic group has the same regression line between x and y. Equivalently, each group has the same mean on y, controlling for x. This test compares the complete model

$$E(y) = \alpha + \beta x + \beta_1 z_1 + \beta_2 z_2$$

to the reduced model

$$E(y) = \alpha + \beta x$$

lacking effects of racial–ethnic group. The null hypothesis is

$$H_0: \beta_1 = \beta_2 = 0 \quad \text{(coefficients of dummy variables = 0).}$$

The complete model represents three different but parallel regression lines between income and education, one for each racial–ethnic group. The reduced model states

that the same regression line applies for all three groups. Figure 13.7 depicts this test. The P-value is small if the complete model with separate parallel lines provides a significantly better fit to the data than the reduced model of a common line.

FIGURE 13.7: Graphical Representation of Null and Alternative Hypotheses in a Test of Equivalence of Regression Lines, when the Categorical Factor Represents Three Groups (Test Assumes No Interaction)
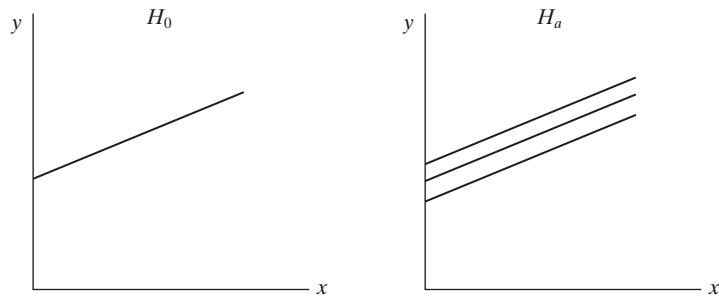
We can also test for the effect of the quantitative variable ($x =$ education), by testing $H_0: \beta = 0$ in the model. The hypothesis states that the straight line relating x to the mean of y has slope 0 for each racial–ethnic group. Since H_0 specifies a value for a single parameter, we can perform the test using the t test.

Example 13.4

Testing Partial Effects of Racial–Ethnic Group and Education   Table 13.7 shows how software reports the results of tests for the no-interaction model. The F statistic for the test of no effect of racial–ethnic group is $730.29/239.00 = 3.06$. Its P-value equals 0.053. There is some evidence, but not strong, that the regressions of y on x are different for at least two of the racial–ethnic groups. The sample sizes for two of the three groups are very small, so this test does not have much power.

TABLE 13.7: Software Output of Inferences about Education and Racial–Ethnic Group, in the Model without Interaction for the `Income` Data File

Source	Partial Sum of Squares	df	Mean Square	F	Sig
Race	1460.58	2	730.29	3.06	.053
Education	12245.23	1	12245.23	51.23	.000
Residual (Error)	18164.25	76	239.00		
Total	33761.95	79			

From Table 13.3 (page 403), the estimated slope for the effect of education on income of 4.432 has a standard error of 0.619. The test statistic is $t = 4.432/0.619 = 7.2$, which has a P-value of 0.000. The evidence is very strong that the true slope is positive. Equivalently, the square of this t statistic equals the F statistic of 51.2 reported for the effect of education in Table 13.7.   ∎

Table 13.8 summarizes the hypotheses and R^2-values for the models. In bivariate models, education is a good predictor of income ($R^2 = 0.42$), considerably better than racial–ethnic group ($R^2 = 0.10$). Some further reduction in error results from using both explanatory variables, assuming no interaction, to predict income ($R^2 = 0.46$). A small and insignificant reduction in error occurs by allowing interaction ($R^2 = 0.48$).

TABLE 13.8: Summary of Comparisons of Four Models for Predicting Income (y) Using Education (x) and Racial–Ethnic Status (z)

Model:	$E(y) = \alpha + \beta x$ $+\beta_1 z_1 + \beta_2 z_2$ $+\beta_3(xz_1) + \beta_4(xz_2)$	$E(y) = \alpha + \beta x$ $+\beta_1 z_1 + \beta_2 z_2$	$E(y) = \alpha + \beta x$	$E(y) = \alpha$ $+\beta_1 z_1 + \beta_2 z_2$
R^2	0.48	0.46	0.42	0.10
H_0: no interaction $F = 1.5, P = 0.24$	Complete model	Reduced model	—	—
H_0: $\beta_1 = \beta_2 = 0$ (equal means, control for x) $F = 3.1, P = 0.053$	—	Complete model	Reduced model	—
H_0: $\beta = 0$ (zero slopes) $F = 51.2, P = 0.000$	—	Complete model	—	Reduced model

13.3 Case Studies: Using Multiple Regression in Research

Multiple regression analysis is a common statistical tool in social research. Many studies start with a simple model containing an explanatory variable of primary focus, with the goal of studying its effect on the response and how that effect changes as other explanatory variables enter the model. Each new model adds potential confounding variables to try to help account for the bivariate effect of the primary explanatory variable on the response. The study often also adds potential mediating variables that could be responsible for the original association. Social scientists typically attempt to evaluate some causal dynamics by using a sequence of models, with primary interest in mediation processes and elimination of the possibility of spuriousness due to confounding variables.

The explanatory variables entered in the model often include both categorical variables and quantitative variables. Now that you've learned how regression can use both these types of explanatory variables, you have sufficient background to understand most regression analyses in social research. This section summarizes three such research studies[1] for which the conclusions are based on results of regression analyses.

Example 13.5

Regression for Modeling Adolescent Sexual Behavior A research study[2] about adolescent sexual behavior by Brian Soller and Dana Haynie used multiple regression with a response variable that is a composite measure of sexual risk-taking. This index incorporates information about inconsistent condom use, sexual intercourse without first discussing contraception or sexually transmitted infections, and sexual intercourse with more than one partner. The higher the composite measure, the greater

---

[1] Thanks to Prof. Alfred DeMaris for suggesting this section and two of these articles.
[2] Published in *Sociological Inquiry*, vol. 83 (2013), pp. 537–569; you may be able to access a PDF file of this article through your university's library at `http://onlinelibrary. wiley.com/doi/10.1111/ soin.12019/abstract`.

the sexual risk-taking. The sample of 6255 adolescents, taken from a random sample of high schools in the United States, ranged from 7th to 12th graders. The explanatory variable of main interest was peer anticipation of college completion. If most of one's peers believed they would attend and complete college, does this tend to reduce a respondent's sexual risk-taking, controlling for relevant confounding variables?

Because of the very large sample size, multiple regression models can use many explanatory variables. The study measured each respondent's anticipation of college completion as a binary variable (1 = pretty likely or more, 0 = less than pretty likely). Peer anticipation of college completion was measured as the mean of the same binary variable for up to five male and five female friends. Other explanatory variables measured the consequences of pregnancy or romantic relationships. The study included control variables thought to be potential confounding variables, such as measures intended to capture individual and peer investment in scholastic achievement and measures of peer delinquency, impulsivity, family attachment, and religiosity. Other quantitative control variables were age, parental SES, and prior sexual risk-taking. Qualitative control variables included race, family structure, and whether the respondent had taken an abstinence pledge (i.e., to remain a virgin until married).

The authors fitted four regression models in which peer anticipation of college completion was an explanatory variable for y = sexual risk-taking. Model 1 analyzed its effect, adjusting for the control variables and the respondent's own anticipation of college completion. The estimated effect of peer anticipation was $\hat{\beta} = -0.13$ ($SE = 0.05$). So, the estimated mean of sexual risk-taking was 0.13 lower for those whose peers all felt pretty likely or more to attend college (and so had a variable value of 1) than for those whose peers all felt less than pretty likely to attend college (and so had a variable value of 0), controlling for other variables.

The authors then investigated whether other variables mediated that association. Model 2 added the consequences of pregnancy variables. The estimated effect of peer anticipation then weakened a bit ($\hat{\beta} = -0.11$, $SE = 0.05$), suggesting a slight mediating effect. Model 3 removed the pregnancy variables and added relationship measures. The estimated effect of peer anticipation was then similar ($\hat{\beta} = -0.12$, $SE = 0.05$). Model 4 (the full model) added both the pregnancy variables and the relationship variables. The effect again weakened only slightly ($\hat{\beta} = -0.10$, $SE = 0.05$).

Table 13.9 shows some of the explanatory variables and their means and standard deviations and estimated effects in Model 4. Here are some results worth noting:

- The respondents' own anticipation of college completion was not significantly associated with sexual risk-taking, controlling for the other variables in the model.

- The authors concluded, "Results from our study underscore the importance of peers in shaping adolescent sexual behavior." For all the models, however, the effect of peer anticipation of college completion on sexual risk-taking is only about -0.1. This effect seems quite weak, because its magnitude is a small fraction of the standard deviation of y = sexual risk-taking, which was reported to equal 0.81. For example, Table 13.9 reports peer anticipation of college completion to have a standard deviation of 0.25, so the estimated standardized regression coefficient for this variable for the full model is only $(-0.10)(0.25)/0.81 = -0.03$. Although the effect was statistically significant at the 0.05 level for all these models and confirmed the authors' theoretical prediction that it would be negative, could this be a case of statistical significance but not practical significance, reflecting the very large n? In practice, often social scientists investigate whether a theoretical effect is truly there, even if it is quite small, to confirm a research hypothesis. With human behavior and imperfect measurement of constructs, large observed effect sizes are not common.

• The article does not mention anything about checking for interactions. Could the effect of peer anticipation of college completion depend on the respondent's level of anticipation of college completion, or on sex or some other variable? ∎

TABLE 13.9: Explanatory Variables and Effects in Multiple Regression Model Predicting Adolescent Sexual Risk-Taking		
Variable	Mean (Std. Dev.)	$\hat{\beta}$ (*SE*)
Control variables		
Age	16.63 (0.99)	0.05 (0.01)
Sex (male=1, female=0)	0.54	−0.10 (0.03)
SES	0.00 (0.79)	0.01 (0.02)
Religiosity	−0.02 (0.82)	−0.04 (0.02)
Abstinence pledge	0.12	−0.06 (0.03)
GPA	2.75 (0.79)	−0.04 (0.02)
Anticipation of college completion	0.76 (0.43)	0.02 (0.04)
Peer variables		
Peer anticipation of college completion	0.76 (0.25)	−0.10 (0.05)
Peer delinquency	0.93 (0.13)	0.25 (0.08)
Peer GPA	2.77 (0.50)	−0.05 (0.03)

Note: Standard deviations were not reported for sex and abstinence pledge.

Example 13.6

Regression for Modeling the Earnings Gender Gap   For many professions, men and women have similar mean salaries when first employed after college graduation, but over time men's salaries tend to grow more quickly than women's salaries. The difference between the mean salaries increases with time. What is responsible for this? A research study[3] by Marianne Bertrand, Claudia Goldin, and Lawrence Katz addressed this using multiple regression models for a sample of 1856 men and 629 women MBA graduates from the University of Chicago.

Their use of multiple regression started with a dummy variable for gender, to enable comparing means, and then successively added various explanatory variables that could potentially explain differences in mean incomes. Some, such as undergraduate GPA and verbal and quantitative GMAT scores, were quantitative. Some, such as race, reasons for choosing the job, and whether the undergraduate institution was a "top 10" institution, were categorical. Some were potentially quantitative but were measured with ordered categories and represented by dummy variables, such as weekly hours worked (<20, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99, ≥100) and number of years since receiving the MBA (0, 1, 3, 6, 9, ≥10). Fitting such models enabled the authors to study how the difference in mean income between men and women changed over time, controlling for relevant variables.

Upon starting a job after receiving the MBA, the mean salary was $130,000 for men and $115,000 for women. After nine years on the job, the mean was about $400,000 for men and about $250,000 for women. For the overall pooled sample, the mean was 36% higher for men. The initial regression model had a dummy variable for gender, dummy variables for five of the six categories for number of years since receiving the MBA, and five interaction terms to allow the difference between men and women to vary by time. This regression model had $R^2 = 0.15$.

---

[3] Published in *American Economic Journal: Applied Economics*, vol. 2 (2010), pp. 228–255; you may be able to access a pdf file of this article through your university's library at `https://www.aeaweb.org/articles .php?doi=10.1257/app.2.3.228`.

When the model adds the number of weekly hours worked as an explanatory variable, R^2 increases to 0.26. Controlling for this explanatory variable, the mean salary for men is now 19% higher than the mean salary for women. When the model next includes pre-MBA characteristics such as race and undergraduate GPA and GMAT scores, MBA GPA, and the fraction of MBA classes that were in finance, $R^2 = 0.40$, and controlling for all the explanatory variables, the mean salary for men is 10% higher than the mean salary for women. Finally, when the model adds a dummy variable for the presence of any post-MBA career interruption (such as caring for a baby) and variables dealing with reasons for choosing the job, the job function, and the employer type, $R^2 = 0.54$. At this stage, controlling for all explanatory variables in the model, the mean salary for men is only 4% higher than the mean salary for women, and the difference is not statistically significant.

- The authors concluded that three factors account for most of the gender gap in earnings: a modest male advantage in training prior to the MBA; greater weekly hours working for men, the difference increasing with years since MBA; greater career interruptions for women combined with large earnings losses associated with any career interruption. They noted that the greater career interruptions and shorter work hours for women than men were largely associated with motherhood.

- The authors used the logarithm of salary as the response variable but failed to clearly explain how to interpret the estimated regression coefficients. Data analysts sometimes use the log transform for variables such as income that have distributions very highly skewed to the right, as it "pulls in" values that are far out in the right tail and makes the distribution less skewed. Section 14.4 in the next chapter (page 447) presents an alternative model for such skewed response data, and Section 14.6 shows another setting in which the logarithm transform is effective. ∎

Example 13.7

Modeling the Consequences of Stigma for Self-Esteem of the Mentally Ill   A research study[4] by Bruce Link, Elmer Struening, Sheree Neese-Todd, Sara Asmussen, and Jo Phelan analyzed whether stigma affects the self-esteem of people who have serious mental illnesses, using a sample of 70 members of a clubhouse program for people with mental illness. To measure self-esteem, the study asked participants whether they strongly agreed, agreed, disagreed, or strongly disagreed with 10 statements such as "At times, you think you are no good at all." Each item had scores (1, 2, 3, 4) with a high score reflecting high self-esteem. The overall self-esteem measure was the mean of these 10 scores. The study measured self-esteem initially, after six months of an intervention designed to facilitate coping with stigma, and after 24 months. The initial measure had a mean of 2.7 and standard deviation of 0.5.

The primary explanatory variables were two quantitative stigma measures: One (perceived devaluation discrimination) measured the extent to which a person believes that other people devalue someone who has a mental illness. The other (stigma withdrawal) quantified the extent to which participants endorse withdrawal as a way to avoid rejection. Each of these was also scaled from 1 to 4, and had means of 2.76 and 2.82 and standard deviations of 0.50 and 0.42. Control variables included sex (male = 1, female = 0), diagnosis (schizophrenia and other nonaffective psychotic disorders = 1, other diagnoses = 0), and a quantitative assessment of depression (which ranged from 0 to 42).

Table 13.10 shows results of four regression models fitted to the self-esteem response after six months. The first model uses the control variables as explanatory variables and the initial self-esteem as a covariate. Models 2 and 3 add each stigma variable separately, and each shows a significant effect. Model 4 adds them together to determine their combined effect, which increased R^2 from 0.43 to 0.55. The stigma effects were negative in Models 2–4, although the effect of stigma withdrawal was weaker and not statistically significant in Model 4.

TABLE 13.10: Regression Analyses for Self-Esteem at Six Months				
	Model 1	Model 2	Model 3	Model 4
Variable	Coeff. (*se*)	Coeff. (*se*)	Coeff. (*se*)	Coeff. (*se*)
Sex	−0.133 (0.088)	−0.165 (0.081)	−0.121 (0.084)	−0.152 (0.080)
Diagnosis	−0.098 (0.084)	−0.144 (0.077)	−0.081 (0.080)	−0.101 (0.076)
Self-esteem initial	0.352 (0.113)	0.343 (0.103)	0.338 (0.108)	0.337 (0.102)
Depression	−0.018 (0.007)	−0.010 (0.007)	−0.013 (0.007)	−0.008 (0.006)
Stigma devaluation		−0.321 (0.085)		−0.261 (0.091)
Stigma withdrawal			−0.302 (0.104)	−0.182 (0.107)
R^2	0.43	0.53	0.49	0.55

- The authors stated, "We also tested for interactions between the stigma variables and age, sex, diagnosis, and depressive symptoms. Only one interaction was significant, and, given that we tested 16, this one may have occurred by chance." So, they did not report results of models with interaction terms.

- The authors noted that an unmeasured confounding variable could potentially account for the association between stigma and self-esteem. However, they argued that the stigma measures strongly predicted self-esteem, and thus any unmeasured confounder would need to have very strong associations with both the stigma measures and self-esteem in order to eliminate the associations.

- How would you assess whether the stigma measures truly did strongly predict self-esteem, controlling for the other explanatory variables? ■

13.4 Adjusted Means*

We have seen that categorical explanatory variables often refer to groups to be compared. This section shows how to estimate means on *y* for the groups, while controlling for the other variables in the model.

ADJUSTING RESPONSE MEANS, CONTROLLING FOR OTHER VARIABLES

To estimate the means of *y* for the groups while taking into account the groups' differing means on the other explanatory variables, we can report the values expected for the means if the groups all had the same means on those other variables. These values, which adjust for the groups' differing distributions on the other variables, are *adjusted means* (also called *least squares means*).

Adjusted Mean

> The ***adjusted mean*** of *y* for a particular group is the regression function for that group evaluated at the overall means of the explanatory variable values for all the groups. It represents the expected value for *y* at the means of the explanatory variables for the combined population.

Adjusted means are mainly relevant for models without interaction terms among the explanatory variables, so differences among them are the same at all values of those variables.

Figure 13.8 illustrates the population adjusted means for three groups when the model has a single covariate x. Let μ_x denote the mean of x for the combined population. The adjusted mean of y for a particular group equals that group's regression function evaluated at μ_x. The *sample adjusted mean* of y for a group is the prediction equation for that group evaluated at $\bar{x}$, the overall sample mean of x. This estimates the value for the group's mean of y if the mean of x for the group had equaled the overall mean of x. We denote the sample adjusted mean for group i by $\bar{y}'_i$.

FIGURE 13.8: Population Adjusted Means with a Covariate x, when a Categorical Explanatory Factor Has Three Categories
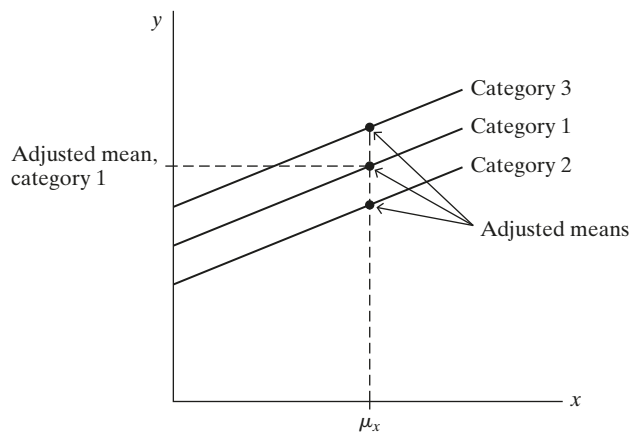
Example 13.8

Adjusted Mean Incomes, Controlling for Education   From Table 13.3 (page 403) for the example regressing income on education and racial–ethnic status, the prediction equation for the model is

$$\hat{y} = -15.7 + 4.4x - 10.9z_1 - 4.9z_2.$$

Table 13.11 lists the equations predicting income using education, for the three racial–ethnic groups. The table also shows the unadjusted mean incomes and the adjusted mean incomes, controlling for education.

TABLE 13.11: Prediction Equations, Sample Unadjusted Mean Incomes, and Adjusted Means (Controlling for $x =$ Education)

Group	Prediction Equation	Mean of x	Mean of y	Adjusted Mean of y
Blacks	$\hat{y} = -26.54 + 4.43x$	12.2	27.8	29.7
Hispanics	$\hat{y} = -20.60 + 4.43x$	11.6	31.0	35.6
Whites	$\hat{y} = -15.66 + 4.43x$	13.1	42.5	40.6

From Table 13.2 (page 402), the mean education for the combined sample of 80 observations is $\bar{x} = 12.7$. Using the three prediction equations, the sample adjusted means for blacks, Hispanics, and whites are

$$\bar{y}'_1 = -26.54 + 4.43\bar{x} = -26.54 + 4.43(12.7) = 29.7,$$

$$\bar{y}'_2 = -20.60 + 4.43(12.7) = 35.6,$$

$$\bar{y}'_3 = -15.66 + 4.43(12.7) = 40.6.$$
■

COMPARING ADJUSTED MEANS, AND GRAPHICAL INTERPRETATION

The coefficients of the dummy variables in the model refer to differences between adjusted means. To illustrate, the estimated difference between adjusted mean incomes of blacks and whites is $\bar{y}'_1 - \bar{y}'_3 = 29.7 - 40.6 = -10.9$ (i.e., $-\$10,900$). This is precisely the coefficient of the dummy variable z_1 for blacks in the above prediction equation. Similarly, the estimated difference between the adjusted means of Hispanics and whites is $\bar{y}'_2 - \bar{y}'_3 = -4.9$, which is the coefficient of z_2. Figure 13.9 depicts the sample adjusted means. The vertical distances between the lines represent the differences between these adjusted means.

FIGURE 13.9: Sample Adjusted Means on Income, Controlling for Education, for Three Racial–Ethnic Groups
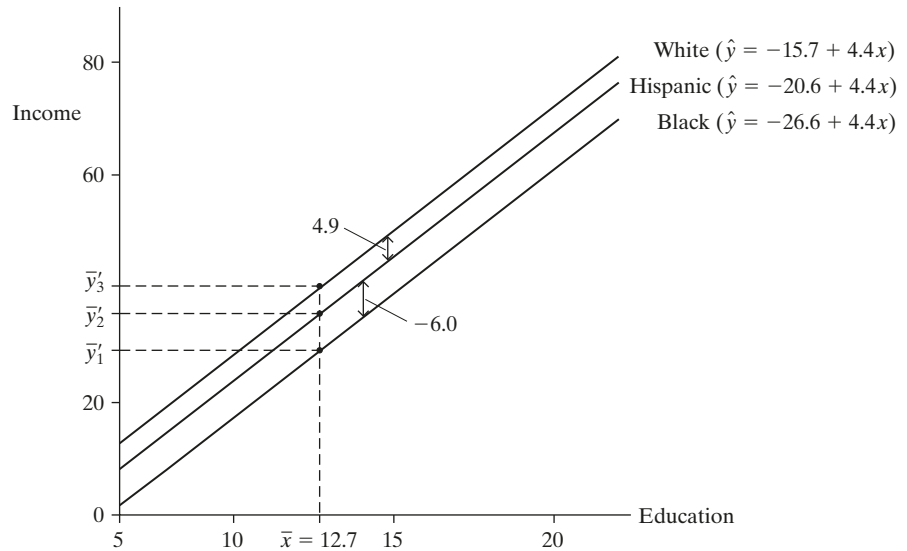
Figure 13.10 depicts the relationship between the adjusted and unadjusted means. The prediction equation predicts a value of $\bar{y}_i$ at the x-value of $x = \bar{x}_i$ for group i. In particular, the prediction line for the group i passes through the point with coordinates $(\bar{x}_i, \bar{y}_i)$. In other words, the *unadjusted mean* $\bar{y}_i$ is the value of the prediction equation for that group evaluated at the x-value of $\bar{x}_i$, the mean of the x-values *for that group alone* [see points such as $(\bar{x}_1, \bar{y}_1)$ in this figure].

FIGURE 13.10: Adjustment Process for Income by Racial–Ethnic Group, Controlling for Education. The ordinary mean for a group is the predicted value at the mean of x for that group alone, whereas the adjusted mean is the predicted value at the mean $\bar{x}$ for all the data.
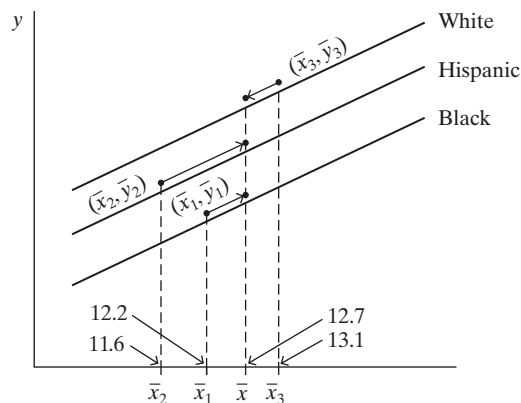
The *adjusted mean* $\bar{y}'_i$ for group i is the value of that prediction equation evaluated at the *overall mean* $\bar{x}$ for the combined sample. Hence, the prediction line for

that category also passes through the point $(\bar{x}, \bar{y}'_i)$, as Figure 13.10 shows for each of the three groups.

The adjustment process moves an ordinary sample mean upward or downward according to whether mean education for the group is below or above average. For whites, for instance, the adjusted mean income of 40.6 is smaller than the unadjusted mean of 42.5. The reason is that the mean education for whites ($\bar{x}_3 = 13.1$) is larger than the mean education for the combined sample ($\bar{x} = 12.7$). Since a positive relationship exists between income and education, the model predicts that whites would have a lower mean income if their mean education were lower (equal to $\bar{x} = 12.7$).

The difference between a group's adjusted and unadjusted means depends directly on the difference between $\bar{x}$ for the combined sample and $\bar{x}_i$ for that group. The adjusted means are similar to the unadjusted means if the $\bar{x}_i$-values are close to the overall $\bar{x}$, or if the slope of the prediction equations is small.

MULTIPLE COMPARISONS OF ADJUSTED MEANS

Following an analysis of variance, the Bonferroni method compares all pairs of means simultaneously with a fixed overall confidence level. This method extends directly to multiple comparison of *adjusted means*. We can form t confidence intervals using these estimates and their standard errors.

Example 13.9

Confidence Intervals for Comparing Adjusted Mean Incomes   Let's construct 95% confidence intervals for differences between the three pairs of adjusted mean incomes, using the Bonferroni multiple comparison approach. The error probability for each interval is $0.05/3 = 0.0167$. The t-score with single-tail probability $0.0167/2 = 0.0083$ and $df = 76$ (which is the residual df for the model) is 2.45.

Table 13.3 (page 403) showed the racial–ethnic effects from the model fit,

```
Parameter     Coef.   Std. Error    t      Sig
[race = b]   -10.874    4.473    -2.431    .017
[race = h]    -4.934    4.763    -1.036    .304
```

The estimated difference between adjusted mean incomes of Hispanics and whites is the coefficient -4.934 of the dummy variable z_2 for Hispanics in the prediction equation. This coefficient has a standard error of 4.763, so the Bonferroni confidence interval equals

$$-4.934 \pm 2.45(4.763), \quad \text{or} \quad (-16.6, 6.7).$$

Controlling for education, the difference in mean incomes for Hispanics and whites is estimated to fall between $-\$16,600$ and $\$6700$. Since the interval contains 0, it is plausible that the true adjusted mean incomes are equal. The sample contained only 14 Hispanics, so the interval is wide. The confidence interval comparing blacks and whites is $-10.874 \pm 2.45(4.473)$, or $(-21.8, 0.1)$. To get the standard error for the estimate $b_1 - b_2 = (-10.87 - (-4.93)) = -5.94$ comparing blacks and Hispanics, we could fit the model with one of these categories as the baseline category lacking a dummy variable. Or, we could use the general expression to get se from the values se_1 for b_1 and se_2 for b_2 as

$$se = \sqrt{(se_1)^2 + (se_2)^2 - 2\text{Cov}(b_1, b_2)},$$

where $\text{Cov}(b_1, b_2)$ is taken from the *covariance matrix* of the parameter estimates, which software can provide. For these data, the standard error for $b_1 - b_2$ equals 5.67, and the confidence interval is $(-19.8, 8.0)$.

Table 13.12 summarizes the comparisons. We can be 95% confident that all three of these intervals contain the differences in population adjusted means. None of the intervals show a significant difference, which is not surprising because the F test of the group effect has a P-value of 0.053. Nonetheless, the intervals show that the adjusted means could be quite a bit smaller for blacks or Hispanics than for whites. More precise estimation requires a larger sample. ∎

TABLE 13.12: Bonferroni Multiple Comparisons of Differences in Adjusted Mean Income by Racial–Ethnic Group, Controlling for Education		
Racial–Ethnic Groups	Estimated Difference in Adjusted Means	95% Bonferroni Confidence Intervals
Blacks, whites	$\bar{y}_1' - \bar{y}_3' = -10.9$	$(-21.8, 0.1)$
Hispanics, whites	$\bar{y}_2' - \bar{y}_3' = -4.9$	$(-16.6, 6.7)$
Blacks, Hispanics	$\bar{y}_1' - \bar{y}_2' = -5.9$	$(-19.8, 8.0)$

A CAUTION ABOUT HYPOTHETICAL ADJUSTMENT

Adjusted means can be useful for comparing several groups by adjusting for differences in the means of a covariate x. *Use them with caution, however, when the means on x are greatly different.* The control process is a hypothetical one that infers what would happen *if* all groups had the same mean for x. If large differences exist among the groups in their means on x, the results of this control may be purely speculative. We must assume (1) that it makes sense to conceive of adjusting the groups on this covariate and (2) that the relationship between y and x would continue to have the same linear form within each category as the x mean shifts for each category.

To illustrate, recall the relationship between annual income and experience and gender shown in Figure 13.3b (page 401). The same line fits the relationship between income and experience for each gender, so it is plausible that the adjusted means are equal. However, nearly all the women have less experience than the men. The conclusion that the mean incomes are equal, controlling for experience, assumes that the regression line shown also applies to women with more experience than those in the sample and to men with less experience. If it does not, then the conclusion is incorrect.

Figure 13.11 portrays a situation in which the conclusion would be misleading. The dotted lines show the relationship for each group over the x-region not observed.

FIGURE 13.11: A Situation in Which Adjusted Means Are Misleading, Comparing Mean Incomes for Men and Women while Controlling for Experience
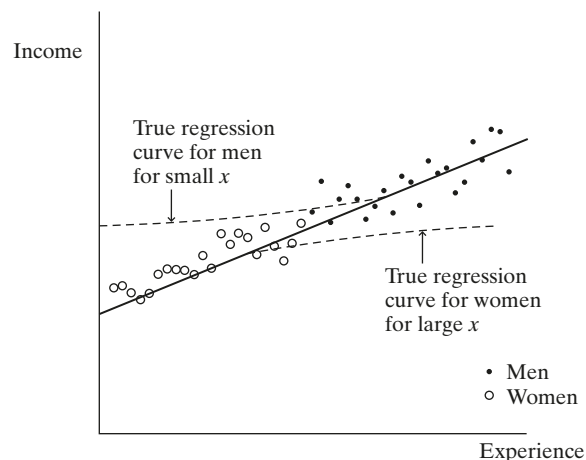
At each fixed x-value, a difference persists between the means of y. In practice, in observational studies we cannot manipulate x-values to force groups to have the same means on covariates, so inferences about what would happen if this could be done are merely hypothetical.

Whenever we use adjusted means, we should check the degree to which the distributions differ on the mean of x. Excessively large differences may mean that the conclusions need strong qualification.

13.5 The Linear Mixed Model*

The analyses so far in this chapter apply with independent observations. Some studies have sets of observations that are correlated. For example, longitudinal studies observe each person at multiple times, and different observations of the same person are typically positively correlated. We next introduce a regression model for quantitative and categorical explanatory variables that permits correlated observations.

MIXED EFFECTS MODELS: RANDOM EFFECTS AND FIXED EFFECTS

As explained on page 384, with repeated measures on the subjects, models can include a dummy variable for each subject. The coefficient of a dummy variable represents a *random effect* for a particular subject. For example, a positive random effect means that each observation for that subject tends to be higher than the average for all the subjects who share the same values of the explanatory variables as that subject. The subject factor consists of all these subject-specific random effects. The term *random effects* reflects that we regard the subjects observed as a random sample of all the possible subjects who could have been sampled.

Regression models also include *fixed effects* for ordinary explanatory variables. These are ordinary parameters. By contrast, we treat the random effects as unobserved random variables rather than as parameters. That is, the terms in the model for the subjects are assumed to come from a particular probability distribution, usually the normal distribution. Regression models for which the effects of explanatory variables are a mixture of random and fixed effects are called **linear mixed models**. The *linear* adjective refers to the effects in the right-hand side of the regression model equations having an additive rather than multiplicative structure.

MODELING CORRELATION STRUCTURE WITH LINEAR MIXED MODELS

For the ordinary regression model, with one observation per subject, let y_i denote the value of the response variable y for subject i, let x_{i1} denote the value of the explanatory variable x_1 for subject i, and so forth. To identify variables at the subject level in the model, we express the model as

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i.$$

As usual, when some explanatory variables are categorical, we use dummy variables for them. In this equation, the *error term* ϵ_i reflects variability in responses for subjects at particular values of explanatory variables. Its sample value is the residual for subject i. The error term has expected value 0, so we delete it when we write the corresponding model formula for $E(y_i)$. The ordinary regression model assumes independent $\{\epsilon_i\}$, and hence independent $\{y_i\}$.

We now generalize this ordinary regression model to the linear mixed model, which permits multiple observations per subject that are correlated. Let y_{ij} denote observation j for subject i. For example, in a longitudinal study, the index j may refer to the jth time of observation. For three times of observations, we have response outcomes (y_{i1}, y_{i2}, y_{i3}) for subject i. Likewise, let x_{ij1} denote observation j for subject i on explanatory variable x_1, and similarly for the other explanatory variables. Some of these explanatory variables take the same value for each observation, whereas some may vary. For instance, in a study of health over time, demographic characteristics such as race and sex are constant, but health characteristics such as weight and blood pressure level may vary. Let s_i denote a random effect for subject i. For particular fixed values of explanatory variables, a subject with a large positive s_i tends to make a relatively high response on y for each j, whereas a subject with a large negative s_i tends to make a relatively low response on y for each j. The random effects are usually assumed to have a normal distribution with mean 0 and unknown variance σ_s^2.

The **linear mixed model** has the form

$$ y_{ij} = \alpha + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijk} + s_i + \epsilon_{ij}. $$

The fixed effects are the effect parameters $(\beta_1, \ldots, \beta_p)$ for the explanatory variables. The random effects are the subject-specific terms $(s_1, \ldots, s_n)$. These subject terms do not involve the explanatory variables, so we can add them to the ordinary intercept term. That is, $(\alpha + s_i)$ is a *random intercept* for subject i.

The model has a separate error term ϵ_{ij} for each observation on a subject. The model is completed by making an assumption about the correlation structure of $(\epsilon_{i1}, \epsilon_{i2}, \ldots)$ and hence the cluster of observations $(y_{ij1}, y_{ij2}, \ldots)$. A popular choice is the **compound symmetry** structure of equal pairwise correlations for the different observations of a subject. With this choice, the random effects term is redundant, because it also implies compound symmetry. In practice, with software we can obtain the compound symmetry structure either (i) deleting the random intercept and assuming equally correlated error terms or (ii) inserting the random intercept but taking the error terms to be independent. In longitudinal studies, often observations closer together in time tend to be more highly correlated than observations farther apart. The **autoregressive** structure is a way to permit this. Another possible structure is an **unstructured** one that makes no assumption about the correlation pattern. When there are many observations per subject, this has the disadvantage of a very large number of correlation parameters.

The model-fitting process for linear mixed models accounts for the assumption about the correlated observations by yielding estimates of $\{\beta_1, \ldots, \beta_p\}$ that have standard errors based on that assumption and that therefore recognize the within-subject correlation. As usual, statistical inference assumes normality for y, and this assumption becomes less important with larger sample sizes. The correlations for the assumed correlation structure among the error terms are themselves parameters that are estimated. This structure also implies correlations among the repeated responses. For example, with the compound symmetry structure based on having a random intercept but independent error terms, the model implies theoretical correlations equal to

$$ \sigma_s^2/(\sigma_s^2 + \sigma^2) $$

for each pair of observations by a subject. This is called an **intraclass correlation**. Greater variability σ_s^2 among random effects implies that the repeated responses are more strongly positively correlated.

LINEAR MIXED MODELS FOR CLUSTERED DATA

In linear mixed models, the random effects are not restricted to individual subjects. They can represent *clusters* of subjects that are similar in some way.

For example, some studies sample families and observe variables for every person in each family. We can regard all the people in a particular family as a cluster. For given values of explanatory variables, two people in the same family tend to be more alike than two people in different families. Identifying the families in the model using a random effect for each family accounts for the correlation among observations within a family.

Example 13.10

Regression Modeling of Family-Clustered Data Here, we use a small example[5] to illustrate the use of linear mixed models with clustered data. Table 13.13 shows responses for people from eight families on $y = $ evaluation of President Obama's performance, $x_1 = $ political party orientation, and $x_2 = $ sex (1 = female, 0 = male). The quantitative variables y and x_1 are measured on an integer scale of 0 to 10, with higher values representing better performance for y and stronger orientation toward Democrat for x_1.

TABLE 13.13: Data for People within Families on $y = $ Evaluation of President's Performance, $x_1 = $ Political Party Orientation, and $x_2 = $ Sex. Data are in the `Family` data file at the text website.

Family	y	x_1	x_2	Family	y	x_1	x_2	Family	y	x_1	x_2
1	8	8	1	4	9	9	1	6	8	9	0
1	7	9	0	4	8	10	0	7	1	3	1
1	7	7	1	5	2	1	1	7	1	1	0
2	4	6	0	5	1	1	0	7	2	3	1
2	3	3	1	5	3	4	0	7	1	3	0
3	1	1	1	6	9	9	1	8	8	6	1
3	1	2	0	6	9	10	0	8	9	5	1

For observation j in family i, we use the model

$$y_{ij} = \alpha + \beta_1 x_{ij1} + \beta_2 x_{ij2} + s_i + \epsilon_{ij},$$

where s_i is a random effect for family i. We assume that $\{s_i\}$ are independent from a normal distribution with mean 0 and standard deviation σ_s. For each family, we treat $\{\epsilon_{i1}, \epsilon_{i2}, \ldots\}$ as independent from a normal distribution with mean 0 and standard deviation σ. Variation among the random effects induces a common positive correlation (compound symmetry) among responses within a family. A family with a large positive s_i has a high probability that everyone in the family gives the President a high rating, whereas a family with a large negative s_i has a high probability that everyone in the family gives a low rating.

Table 13.14 shows linear mixed model parameter estimates and standard errors. It also shows naive estimates and standard errors that we would get if we treated the observations within a family as independent, instead of allowing them to be correlated. Estimates and standard errors can be quite far from the more sensible ones that recognize observations within families as being correlated.

---

[5] The example is unrealistically small, to make it simple to show the entire data file. To read about an actual study that used a linear mixed model, see Exercise 13.37.

	TABLE 13.14: Estimates and Standard Errors for Clustered Family Data of Table 13.13			
	Linear Mixed Model		Naive Linear Model	
Effect	$\hat{\beta}$	Std. Error	$\hat{\beta}$	Std. Error
Intercept	1.449	0.974	−0.921	0.628
Party	0.598	0.125	0.949	0.087
Sex	0.806	0.300	1.538	0.562

Software reports that the estimated variance of the family random effects is $\hat{\sigma}_s^2 = 3.05$ and the estimated variance of the error term is $s^2 = 0.36$. The ratio $\hat{\sigma}_s^2/(\hat{\sigma}_s^2 + s^2) = 0.89$ estimates the within-family (intraclass) correlation implied by this linear mixed model. Such estimates are extremely imprecise with such a small number of families. ∎

LINEAR MIXED MODELS FOR REPEATED-MEASURES ANALYSES

Sections 12.5 and 12.6 introduced ANOVA for repeated measurement of subjects on one or two categorical factors. This ANOVA method has limitations. For example, the method cannot deal with missing data, so subjects with any missing observations get dropped from the analysis. Also, the correlations among observations on the same subject are assumed to satisfy a *sphericity* structure, implied by common variability at all times and the same correlation between each pair of observations. Linear mixed models do not have these limitations. In addition, linear mixed models can accommodate both quantitative and categorical explanatory variables.

Suppose a study has repeated measures of subjects across categories of a single factor. That is, the data file has one within-subjects factor but no between-subjects factor, the scenario for repeated-measures ANOVA in Section 12.5. For simplicity, suppose each subject has T observations, such as at T times. When $T = 2$, for example, the purpose of the study may be to compare means before and after receiving some treatment. The linear mixed model is

$$y_{ij} = \alpha + \beta_1 t_1 + \beta_2 t_2 + \cdots + \beta_{T-1} t_{T-1} + s_i + \epsilon_{ij},$$

where t_1 is a dummy variable for time 1 (i.e., $t_1 = 1$ when $j = 1$), t_2 is a dummy variable for time 2, and so forth. The parameters of main interest to be estimated are the fixed-effects parameters $\{\beta_t\}$.

We can generalize this model by adding additional factors, such as between-subjects effects in addition to the within-subjects effects. We can also add quantitative explanatory variables, by adding terms of the form βx. For instance, if we expect a linear trend in the means over time in the above model, we could replace the $T - 1$ fixed effects terms for the T times by βt, thus using one slope parameter instead of $T - 1$ separate time parameters. A more general version of this linear mixed model has a second type of random-effect term to allow the slopes to vary by subject around the mean of β. The linear mixed model then has the form

$$y_{ij} = \alpha + (\beta + b_i)t + s_i + \epsilon_{ij},$$

in which s_i are subject-specific random effects that permit variability in the intercept and b_i are subject-specific random effects that permit variability in the slopes. Then, $(\beta + b_i)$ is a *random slope* and $(\alpha + s_i)$ is a random intercept. Moreover, such a linear mixed model can permit b_i and s_i to be correlated.

MISSING DATA: ASSUMPTION OF "MISSING AT RANDOM"

The linear mixed model, unlike ordinary repeated-measures ANOVA, permits some observations to be missing. For any particular subject, it uses the available data.

Depending on what causes observations to be missing, resulting estimates of fixed effects may or may not be unbiased. They are unbiased when the missing data are *missing at random*. This means that the probability an observation is missing does not depend on the value of the unobserved response.

***

Example 13.11

Quality of Life with Treatments for Alcohol Dependence Gueorguieva and Krystal (2004) analyzed data from a clinical trial for the effect of using a particular drug (naltrexone) in addition to psychosocial therapy in treating 627 veterans suffering from chronic, severe alcohol dependence. The response variable was a satisfaction score that averaged four items on a quality of life scale. Each item had potential values 1 (terrible) to 7 (delighted). For each subject, this response was observed initially and then after 4, 26, 52, and 78 weeks. However, complete results for all five times were available for only 211 of the 627 subjects. The rest of the data are missing. The study had three treatments: 12 months of the drug, 3 months of the drug followed by 9 months of placebo, or 12 months of placebo. This is a between-subjects factor. Table 13.15 shows the sample mean satisfaction scores for the 3×5 combinations of treatment and time.

TABLE 13.15: Sample Mean Satisfaction for Subjects Suffering Alcohol Dependence, by Treatment and Time of Measurement. Treatment is a between-subjects factor and time is a within-subjects factor.

Treatment	Initial	4 Weeks	26 Weeks	52 Weeks	78 Weeks
Long-term drug	3.9	4.0	4.3	4.5	4.4
Short-term drug	3.7	4.0	4.1	4.3	4.3
Placebo	3.6	3.9	3.9	4.2	4.3

This data set is the sort we analyzed in Section 12.6 using ANOVA with two factors and repeated measures on one of them. Here, the repeated measures occur across the five levels of time. Let's consider the linear mixed model that we'll abbreviate by

$$y_{ij} = S_i + D + T + (D \times T) + \epsilon_{ij},$$

where S_i denotes a random intercept $\alpha + s_i$ for subject i, D denotes drug treatment, T denotes time, and $(D \times T)$ denotes a drug-by-time interaction, which allows the effect of time to vary by the drug treatment. In the model formula, we represent D by $(\beta_1 d_1 + \beta_2 d_2)$ for dummy variables $\{d_1, d_2\}$ for the two drugs. If we treat time as categorical, we represent T by $(\beta_3 t_1 + \beta_4 t_2 + \beta_5 t_3 + \beta_6 t_4)$ using dummy variables $\{t_1, t_2, t_3, t_4\}$ for the first four times, and then $(T \times D)$ denotes interaction terms for the eight cross products of T and D dummy variables. If we instead expect linear trends in the response means, with possibly different slopes for different drug treatments, then we use a simpler model that treats T as quantitative. Then we represent T by a term $\beta_3 t$ for scores for t such as (0, 4, 26, 52, 78) or (0, 1, 2, 3, 4), with $(T \times D)$ then denoting cross products of t with each of the two dummy variables for D.

Gueorguieva and Krystal (2004) noted that repeated-measures two-way ANOVA, treating time as categorical, shows no evidence of a treatment-by-time interaction (P-value = 0.69) and no evidence of a treatment effect (P-value = 0.80). However, that analysis could only use results for the 211 of the 627 subjects with no missing data, thus ignoring the data for the 416 subjects who had some missing data. With the linear mixed model approach, which also uses data for the subjects missing some observations, the power is much greater. When used with an unstructured pattern for the correlations, no evidence occurs of interaction (P-value = 0.63), but slight evidence occurs of a treatment effect (P-value = 0.07). This shows the benefit of a linear mixed model approach when many observations are missing. The standard ANOVA approach discards much information, such that evidence about the treatment effect changes from marginal ($P = 0.07$) to very weak ($P = 0.80$). ■

Statistical software can fit linear mixed models assuming various correlation structures for the repeated observations. For further details, see Fitzmaurice et al. (2011) and Hedeker and Gibbons (2006). The analysis requires some sophistication to avoid inappropriate models. For simple factorial designs having observations taken at the same times and no missing data and for which common correlations seem like a plausible assumption, it is simpler to use standard repeated-measures ANOVA.

13.6 Chapter Summary

This chapter showed that multiple regression can describe the relationship between a quantitative response variable and both quantitative and categorical explanatory variables.

- The multiple regression model has linear terms (such as βx) for quantitative explanatory variables and dummy variable terms for categorical factors.

- In this context, **no interaction** between a quantitative variable x and a categorical group factor in their effects on y means that the slope of the line relating x to $E(y)$ is the same for each group. The model then provides a set of parallel lines. We allow interaction and different slopes by entering cross products of quantitative explanatory variables with dummy variables in the model.

- **Adjusted means** summarize the means on y for the groups while controlling for other variables. They represent the model's prediction for the means of y for the groups at the overall means of the other variables. Adjusted means are meaningful only when there is no interaction.

- One can test the hypothesis of no interaction as well as the hypothesis of equal adjusted means using F tests for the relevant parameters in the model.

- When the model is used to compare the mean of y for different groups that are the categories of a categorical variable while controlling for a quantitative *covariate*, the analysis is called an **analysis of covariance**.

The **linear mixed model** is a generalization of the multiple regression model that contains random effects as well as fixed effects and permits various correlation structures among observations within clusters indexed by the random effect term. This general model can handle the correlations that occur from having multiple observations on subjects, such as in longitudinal studies and with various types of clustered data.

Exercises

Practicing the Basics

13.1. The regression equation relating y = education (number of years completed) to race ($z = 1$ for whites, $z = 0$ for nonwhites) in a certain country is $E(y) = 11 + 2z$. The regression equation relating education to race and to father's education (x) is $E(y) = 3 + 0.8x − 0.6z$.

(a) Ignoring father's education, find the mean education for whites, the mean education for nonwhites, and the difference between them.

(b) Plot the relationship between x and the mean of y for whites and for nonwhites.

(c) Controlling for father's education, find the difference between the mean education of whites and nonwhites. Illustrate by finding the mean education for each group when father's education equals 12 years.

13.2. Table 3.9 on page 65 showed data for several nations on $y = CO2$ emissions (in metric tons per capita) and $x =$ per capita GDP (in thousands of dollars). Let $z =$ whether the nation is in Europe ($1 =$ yes, $0 =$ no).

(a) The prediction equation for the effect of z is $\hat{y} = 10.61 − 2.48z$. Interpret the coefficients.

(b) The prediction equation for the effects of x and z is $\hat{y} = 2.10 + 0.22x − 3.58z$. Interpret the coefficients.

13.3. A regression analysis for the 100th Congress predicted the proportion of each representative's votes on abortion issues that took the "pro-choice" position.[6] The prediction equation was

$$\hat{y} = 0.350 + 0.011id + 0.094r + 0.005nw + 0.005inc$$
$$+ 0.063s − 0.167p,$$

where $r =$ religion (1 for non-Catholics), $s =$ sex (1 for women), $p =$ political party (1 for Democrats), $id =$ ideology is the member's ADA score (ranging from 0 at most conservative to 100 at most liberal), $nw =$ nonwhite is the percentage nonwhite of the member's district, and $inc =$ income is the median family income of the member's district.

(a) Interpret the coefficient for political party.

(b) Using standardized variables, the prediction equation is

$$\hat{z}_y = 0.83z_{id} + 0.21z_r + 0.18z_{nw} + 0.05z_{inc} + 0.03z_s − 0.18z_p.$$

Comment on the relative sizes of the partial effects. Interpret the coefficient of ideology.

13.4. For 2014 data, the GSS website yields the prediction equation $\hat{y} = 9.59 + 0.166x_1 + 0.347x_2$ for $y =$ highest year of school completed, $x_1 =$ sex ($1 =$ male, $2 =$ female), and $x_2 =$ highest year of mother's education completed.

(a) Interpret the estimated partial effects.

(b) A more usual dummy coding for sex would be $0 =$ male and $1 =$ female. Would the estimated effects of x_1 and x_2 then change? Explain.

13.5. Based on a national survey, Table 13.16 shows results of a prediction equation for $y =$ alcohol consumption, measured as the number of alcoholic drinks the subject drank during the past month.

(a) For $x =$ alcohol consumption three years ago and dummy variables f for whether father died in the past three years, s for sex, and (m_1, m_2, m_3) for the four categories of marital status, report the prediction equation.

(b) Find the predicted alcohol consumption for a divorced male whose father died in the previous three years and whose consumption three years previously was (i) 0 drinks per month, (ii) 10 drinks per month.

TABLE 13.16

Explanatory Variable	Estimate	Std. Error
Intercept	8.3	
Death of father in past three years ($0 =$ no)	9.8	2.9
Sex ($0 =$ male)	−5.3	1.6
Marital status ($0 =$ married)		
Divorced, separated	7.0	2.0
Widowed	2.0	3.6
Never married	1.2	2.4
Alcohol consumption three years ago	0.501	0.023

13.6. Consider the results in the previous exercise.

(a) Marital status has three estimates. Dividing the coefficient of the divorced dummy variable by its standard error yields a t statistic. What hypothesis does it test?

(b) What would you need to do to test the effect of marital status (all categories at once), controlling for the other variables?

13.7. For the Houses data file at the text website, Table 13.17 shows results of modeling $y =$ selling price (in dollars) in terms of size of home (in square feet) and whether the home is new ($1 =$ yes; $0 =$ no).

TABLE 13.17

Parameter	Coef.	Std. Error	t	Sig
Intercept	-40230.867	14696.140	-2.738	.007
size	116.132	8.795	13.204	.000
new	57736.283	18653.041	3.095	.003

(a) Report and interpret the prediction equation, and form separate equations relating selling price to size for new and for not new homes.

(b) Find the predicted selling price for a home of 3000 square feet that is (i) new, (ii) not new.

13.8. For the previous exercise, Table 13.18 shows results of fitting the model allowing interaction.

(a) Report the lines relating the predicted selling price to the size for homes that are (i) new, (ii) not new.

(b) Find the predicted selling price for a home of 3000 square feet that is (i) new, (ii) not new.

(c) Find the predicted selling price for a home of 1500 square feet that is (i) new, (ii) not new. Comparing to (b), explain how the difference in predicted selling prices changes as size of home increases.

TABLE 13.18

Parameter	Coef.	Std. Error	t	Sig
Intercept	-22227.808	15521.110	-1.432	.155
size	104.438	9.424	11.082	.000
new	-78527.502	51007.642	-1.540	.127
new#size	61.916	21.686	2.855	.005

13.9. Using software, replicate all the analyses shown in Sections 13.1 and 13.2 using the `Income` data file at the text website.

13.10. The software outputs in Table 13.19 show results of fitting two models to data from a study of the relationship between y = percentage of adults voting, percentage of adults registered to vote, and racial–ethnic representation, for a random sample of 40 precincts in the state of Texas for a gubernatorial election. Racial–ethnic representation of a precinct is the group (Anglo, black, or Mexican American) having the strongest representation in a precinct.

(a) State a research question that could be addressed using these data.

(b) Report the prediction equation for the model assuming no interaction. Interpret the parameter estimates.

(c) Report the prediction equation for the model allowing interaction. Interpret the parameter estimates and describe the nature of the estimated interaction.

(d) Test whether the regression lines for the three categories have the same slope. Report the test statistic and P-value, and interpret.

(e) For the model assuming no interaction, test whether the mean voting percentages are equal for the three categories of racial–ethnic representation, controlling for percentage registered. Report the test statistic and P-value, and interpret.

(f) Report the test statistic and P-value for testing the null hypothesis that percentage voting and percentage registered are independent, controlling for racial–ethnic representation. Interpret.

TABLE 13.19

NO INTERACTION MODEL

Source	Partial SS	df	Mean Square	F	Sig
race	40.08	2	20.04	1.07	.354
register	2317.43	1	2317.43	123.93	.000

Source	Sum of Squares	df	Mean Square	Parameter	Estimate
Model	7936.734	3	2645.578	INTERCEPT	-2.7786
Residual	673.166	36	18.699	REGISTER	0.7400
Total	8609.900	39		RACE a	-1.3106
				b	-2.8522
				ma	0.0000

INTERACTION MODEL

Source	Partial SS	df	Mean Square	F	Sig
race*register	53.79	2	27.89	1.47	.243

Source	Sum of Squares	df	Mean Square	Parameter	Estimate
Model	7990.523	5	1598.105	INTERCEPT	-8.245
Residual	619.377	34	18.217	REGISTER	0.878
Total	8609.900	39		RACE a	6.974
				b	9.804
				ma	0.000
				REGISTER*RACE a	-0.175
				b	-0.283
				ma	0.000

(f) Summarize what you have learned about the question posed in (a) from your analyses.

13.11. Refer to the previous exercise. The means of percentage registered for the three categories are $\bar{x}_1 = 76.2$, $\bar{x}_2 = 49.5$, and $\bar{x}_3 = 39.7$. The overall mean $\bar{x} = 60.4$.

(a) Find the adjusted mean of the percentage voting for Anglos. Compare it to the unadjusted mean of 52.3, and interpret.

(b) Sketch a plot of the no-interaction model for these data, and identify on it the unadjusted and adjusted means for Anglos.

13.12. Table 13.1 did not report the observations for 10 Asian Americans. Their (x, y) values were

Subject	1	2	3	4	5	6	7	8	9	10
Education	16	14	12	18	13	12	16	16	14	10
Income	70	42	24	56	32	38	58	82	36	20

(a) Conduct the analyses for the no-interaction model shown in Sections 13.2 and 13.4, after adding these data to the `Income` data file at the text website. Summarize your analyses, and interpret.

(b) Conduct the analyses for the interaction model and for comparing that model to the no-interaction model, as shown in Sections 13.1 and 13.2, after adding these data.

13.13. Exercise 13.1 reported the regression equation relating y = education to race ($z = 1$ for whites) and to father's education (x) of $E(y) = 3 + 0.8x - 0.6z$. The means $\bar{y} = 11$ for nonwhites, $\bar{y} = 13$ for whites, and overall $\bar{y} = 12$.

(a) Find the adjusted mean educational levels for whites and nonwhites, controlling for father's education.

(b) Explain why the adjusted means differ as they do from the unadjusted means.

13.14. Refer to the regression modeling of the family-clustered data in Table 13.13. Add to the `Family` data file the data for family 9, who had (y, x_1, x_2) values (0, 2, 0) and (1, 2, 1). Fit the linear mixed model to all the data, and interpret results.

Concepts and Applications

13.15. Refer to the `Students` data file (Exercise 1.11). Using software, prepare a report presenting graphical, descriptive, and inferential analyses with

(a) y = political ideology and the predictors religiosity and whether a vegetarian.

(b) y = college GPA with predictors high school GPA, gender, and religiosity.

13.16. Refer to the data file your class created in Exercise 1.12. For variables chosen by your instructor, use

regression analysis as the basis of descriptive and inferential statistical analyses. Summarize your findings in a report in which you state the research question posed and describe and interpret the fitted models and the related analyses.

13.17. Refer to the `OECD` data file at the text website, shown in Table 3.13 (page 70). Pose a research question about how the human development index and whether a nation is in Europe relate to carbon dioxide emissions. Conduct appropriate analyses to address that question, and prepare a report summarizing your analyses and conclusions.

13.18. An article[7] on predicting attitudes toward homosexuality modeled a response variable with a four-point scale in which homosexual relations were scaled from 1 = always wrong to 4 = never wrong, with x_1 = education (in years), x_2 = age, x_3 = political conservative (1 = yes, 0 = no), x_4 = religious fundamentalist (1 = yes, 0 = no), and x_5 = whether live in same city as when age 16 (1 = yes, 0 = no). The prediction equation allowing interaction between x_1 and x_3 is

$$\hat{y} = 0.94 + 0.13x_1 - 0.01x_2$$
$$+ 1.10x_3 - 0.38x_4 - 0.15x_5 - 0.12(x_1 \times x_3).$$

Report the prediction equations for political conservatives and nonconservatives. Explain how these suggest that greater education corresponds to less negative views about homosexuality for nonconservatives but may have no effect for conservatives.

13.19. For the 2014 GSS, Table 13.20 shows estimates (with *se* values in parentheses) for four regression models for y = political party identification in the United States, scored from 1 = strong Democrat to 7 = strong Republican. The explanatory variables are sex (0 = male, 1 = female), race (0 = white, 1 = black), religion (scored 1 = fundamentalist, 2 = moderate, 3 = liberal), and political views (scored from 1 = extremely liberal to 7 = extremely conservative). Summarize your main conclusions from these model fits in a report of about 200 words.

TABLE 13.20

	Model 1	Model 2	Model 3	Model 4
Variable	Coef. (*se*)	Coef. (*se*)	Coef. (*se*)	Coef. (*se*)
Sex	−0.332 (0.081)	−0.228 (0.083)	−0.255 (0.083)	−0.165 (0.075)
Race		−1.777 (0.110)	−1.947 (0.113)	−1.630 (0.102)
Religion			−0.410 (0.055)	−0.098 (0.051)
Political views				0.674 (0.027)
Constant	3.95	4.23	5.13	1.60
Multiple R	0.082	0.333	0.365	0.581

---

[7] T. Shackelford and A. Besser, *Individual Differences Research*, vol. 5 (2007), pp. 106–114.

13.20. Table 13.21 shows output for GSS data with $y =$ index of attitudes toward premarital, extramarital, and homosexual sex, for which higher scores represent more permissive attitudes. The categorical explanatory variables are race (0 for whites, 1 for blacks), gender (0 for males, 1 for females), region (0 for South, 1 for non-South), and religion ($r_1 = 1$ for liberal Protestant sect, $r_2 = 1$ for conservative Protestant, $r_3 = 1$ for fundamentalist Protestant sect, $r_4 = 1$ for Catholic, $r_5 = 1$ for Jewish; no religious affiliation when $r_1 = \cdots = r_5 = 0$). The quantitative explanatory variables are age, education (number of years), attendance at church (higher values represent more frequent attendance), and a variable for which higher values represent greater intolerance of freedom of speech for atheists and communists.

(a) Based on the parameter estimates, give a profile of a person you would expect to be (i) least permissive, (ii) most permissive, with respect to sexual attitudes.

(b) Summarize your main conclusions from studying the output.

cation might differ among the three groups, you control that variable. Table 13.22 shows some of the data for a random sample of married women above age 45. Analyze the complete data, which are the `Fertility` data file at the text website. In your report, provide graphical presentations as well as interpretations for all your analyses, and summarize the main results.

TABLE 13.22

Urban Natives		Urban Migrants		Rural Migrants	
Education	Fertility	Education	Fertility	Education	Fertility
0	7	0	7	0	4
0	5	0	6	0	6
1	5	0	7	0	10
1	4	1	5	0	8

13.22. Analyze the `Houses2` data file at the text website by modeling selling price in terms of size of house and whether it is new.

TABLE 13.21

```
Analysis of Variance
              Sum of Squares   df   Mean Square      F
Regression      2583.326       12     215.277     54.098
Residual        4345.534     1092       3.979

R Square             0.373

Variable    Coef.   Std. Error   Beta       t       Sig
(Constant)  9.373
RACE        0.993    0.2040      0.125     4.869    .000
AGE        -0.029    0.0042     -0.189    -6.957    .000
SEX        -0.289    0.1230     -0.058    -2.353    .019
EDUC        0.073    0.0223      0.092     3.281    .001
REGION      0.617    0.1401      0.115     4.403    .000
ATTEND     -0.286    0.0255     -0.304   -11.217    .000
R1         -0.296    0.2826     -0.049    -1.048    .295
R2         -0.605    0.2782     -0.113    -2.174    .030
R3         -1.187    0.3438     -0.128    -3.454    .001
R4         -0.127    0.2856      0.023     0.446    .656
R5          0.521    0.4417      0.034     1.179    .238
FREESPCH   -0.465    0.0581     -0.227    -8.011    .000
```

13.21. You plan a study of factors associated with fertility (a woman's number of children) in a Latin American city. Of particular interest is whether migrants from other cities or migrants from rural areas differ from natives of the city in their family sizes. The groups to be compared are urban natives, urban migrants, and rural migrants. Since fertility is negatively related to educational level, and since edu-

(a) Fit the model allowing interaction, and test whether the interaction term is needed in the model.

(b) Construct a scatterplot, identifying the points by whether the home is new or not. The observation with the highest selling price is a new home that is somewhat removed from the general trend of points. Fit the interaction model after removing this single observation. Again,

test whether the interaction term is needed in the model. Note what a large impact one observation can have on the conclusions.

13.23. For the `Crime2` data file at the text website, let z be a dummy variable for whether a state is in the South, with $z = 1$ for AL, AR, FL, GA, KY, LA, MD, MS, NC, OK, SC, TN, TX, VA, WV.

(a) Not including the observation for D.C., analyze the relationship between y = violent crime rate and z, both ignoring and controlling for x = poverty rate. Summarize results.

(b) Repeat the analysis with D.C. in the data set, setting $z = 1$ for it. Is this observation influential? Summarize results.

13.24. You have two groups, and you want to compare their regressions of y on x, to test the hypothesis that the true slopes are identical for the two groups. Explain how to do this using regression modeling.

13.25. In analyzing GSS data relating y = frequency of having sex in the past year to frequency of going to bars, DeMaris (2004, p. 62) noted that the slope for unmarried subjects is more than double the slope for married subjects. Introducing notation, state a model that you think would be appropriate.

13.26. Let y = death rate and x = mean age of residents, measured for each county in Louisiana and in Florida. Sketch a hypothetical scatterplot, identifying points for each state, when the mean death rate is higher in Florida than in Louisiana when mean age is ignored but lower when it is controlled.

13.27. Draw a scatterplot with sets of points representing two groups such that H_0: equal means would be rejected in a one-way ANOVA but not in an analysis of covariance.

13.28. For a regression model fitted to annual income (thousands of dollars) using predictors age and marital status, Table 13.23 shows the sample mean incomes and the adjusted means. How could the adjusted means be so different from the unadjusted means? Draw a sketch to help explain.

TABLE 13.23

Group	Mean Age	Mean Income	Adjusted Mean Income
Married	44	40	30
Divorced	35	30	30
Single	26	20	30

In Exercises 13.29–13.30, select the correct response(s). (More than one response may be correct.)

13.29. In the model $E(y) = \alpha + \beta_1 x + \beta_2 z$, where $z = 1$ for females and $z = 0$ for males,

(a) The categorical factor has two categories.

(b) One line has slope β_1 and the other has slope β_2.

(c) β_2 is the difference between the mean of y for females and males.

(d) β_2 is the difference between the mean of y for females for males, controlling for x.

13.30. In the United States, the mean annual income for blacks (μ_1) is smaller than for whites (μ_2), the mean number of years of education is smaller for blacks than for whites, and annual income is positively related to number of years of education. Assuming that there is no interaction, the difference in the mean annual income between whites and blacks, controlling for education, is

(a) Less than $\mu_2 - \mu_1$.

(b) Greater than $\mu_2 - \mu_1$.

(c) Possibly equal to $\mu_2 - \mu_1$.

13.31. Summarize the differences in purpose of a one-way analysis of variance and an analysis of covariance.

13.32. * Suppose we use a centered variable for the covariate and express the interaction model when the categorical factor has two categories as

$$E(y) = \alpha + \beta_1(x - \mu_x) + \beta_2 z + \beta_3(x - \mu_x) \times z.$$

Explain how to interpret β_2, and explain how this differs from the interpretation for the model without a centered covariate.

13.33. * Using the graphical representation in Figure 13.10, explain why

$$\bar{y}_i' = \bar{y}_i + b(\bar{x} - \bar{x}_i),$$

where b is the estimated slope. So, when $b > 0$, $\bar{y}_i$ is adjusted upward if $\bar{x} > \bar{x}_i$ and adjusted downward if $\bar{x}_i < \bar{x}$.

13.34. Explain the reason for entering random effects into a regression model. Describe a study in which it would be helpful to use this approach.

13.35. Explain what is meant by the term *mixed model*, and explain the distinction between a *fixed effect* and a *random effect*.

13.36. Summarize advantages of using a linear mixed model to analyze repeated-measures data, compared to using standard repeated-measures ANOVA.

13.37. A recent study[8] examined the role of family structure in the financial support parents provide for their children's college education. Using data for 5070 children from 1519 families from the Health and Retirement Study, one aspect of the study modeled the parents' financial support of tuition costs. Access the article at www.ncbi.nlm.nih.gov/pmc/articles/

---

[8] By J. C. Henretta, D. Wolf, M. Van Voorhis, and B. Soldo, *Social Science Research*, vol. 41 (2012), pp. 876–887.

PMC3461181. Consider the results shown in Table 2 of the article, for this response variable.

(a) Identify the explanatory variables that seem especially relevant, and describe the direction of their estimated effects.

(b) Explain why this analysis used a linear mixed model with random effects. Report the estimated standard deviation of those random effects.

(c) Using the estimated standard deviations in Table 2, find the intraclass correlation. Interpret.

This page intentionally left blank

Chapter

14

MODEL BUILDING WITH MULTIPLE REGRESSION

This chapter introduces tools for building regression models and evaluating the effects on their fit of unusual observations or highly correlated predictors. It also shows ways of modeling variables that badly violate the assumptions of straight-line relationships with a normal response variable.

We first discuss criteria for *selecting a regression model* by deciding which of a possibly large collection of variables to include in the model. We then introduce methods for *checking regression assumptions* and evaluating the influence of individual observations. We also discuss effects of *multicollinearity*—such strong "overlap" among the explanatory variables that no one of them seems useful when the others are also in the model.

Section 14.4 introduces a **generalized linear model** that can handle response variables having distributions other than the normal. For example, the *gamma distribution* is useful for positive variables that exhibit skew to the right and have variability that grows with the mean. We also introduce models for nonlinear relationships, such as *exponential* increase or decrease. The final section introduces alternative regression methods with weaker assumptions, such as not assuming a functional form for the relationship or common response variability.

14.1 Model Selection Procedures

Social research studies usually have several explanatory variables. For example, for modeling mental impairment, potential predictors include income, educational attainment, an index of life events, social and environmental stress, marital status, age, self-assessment of health, number of jobs held in previous five years, number of relatives who live nearby, number of close friends, membership in social organizations, and frequency of church attendance.

Usually, the regression model for a study includes some explanatory variables for theoretical reasons, such as to analyze whether a predicted effect truly occurs under certain controls. Other explanatory variables may be included to see if they mediate the predicted effects. Others may be included for exploratory purposes, to check whether they explain other variability in the response variable. The model might also include terms to allow for interactions. In such situations, it is not simple to decide which variables to include and which to exclude from a final model.

SELECTING EXPLANATORY VARIABLES FOR A MODEL

A strategy that you might first consider is to include every potentially useful explanatory variable and then delete those terms not making statistically significant partial contributions at some preassigned α-level. Unfortunately, this usually is inadequate.

431

Because of correlations among the explanatory variables, any one variable may have little unique predictive power, especially when the number of predictors is large. It is conceivable that few, if any, explanatory variables would make significant *partial* contributions, given that all of the other explanatory variables are in the model.

Here are three general guidelines for selecting explanatory variables:

1. Include the relevant variables to make the model useful for theoretical purposes, so you can address hypotheses posed by the study, with sensible control and mediating variables.

2. Include enough variables to obtain good predictive power.

3. Keep the model simple.

Goal 3 is a counterbalance to goal 2. Having a large number of explanatory variables in a model has disadvantages. The correlations among them can result in inflated standard errors of the parameter estimates, and may make it impossible to assess the partial contributions of variables that are important theoretically. To avoid multicollinearity, it is helpful for the explanatory variables to be correlated with the response variable but not highly correlated among themselves.

Goal 2 of obtaining good predictive power might suggest "Maximize R^2" as a criterion for selecting a model. Because R^2 cannot decrease as you add variables to a model, however, this approach would lead you to the most complex model in the set being considered. Related to the goal 3 of simplicity, don't try to build a complex model if the data set is small. If you have only 25 observations, you won't be able to untangle the complexity of effects among 10 explanatory variables. With small to moderate sample sizes (say, 100 or less), it is safer to use relatively few predictors.

Keeping these thoughts in mind, no unique or optimal approach exists for selecting explanatory variables. For p potential predictors, since each can be either included or omitted (two possibilities for each variable), there are 2^p potential subsets. For $p = 2$, for example, there are $2^p = 2^2 = 4$ possible models: one with both x_1 and x_2, one with x_1 alone, one with x_2 alone, and one with neither variable. The set of potential models is too large to evaluate practically if p is even moderate; with $p = 7$, there are $2^7 = 128$ potential models.

Statistical software has automated variable selection procedures that scan the explanatory variables to construct a model. These routines sequentially enter or remove variables, one at a time according to some criterion. For any particular sample and set of variables, however, different procedures may select different subsets of variables, with no guarantee of selecting a sensible model. The most popular automated variable selection methods are *backward elimination*, *forward selection*, and *stepwise regression*.

BACKWARD ELIMINATION

Backward elimination begins by placing all of the explanatory variables under consideration in the model. It deletes one at a time until reaching a point where the remaining variables all make significant partial contributions to predicting y. The variable deleted at each stage is the one that is the least significant, having the largest P-value in the significance test for its effect.

Here is the sequence of steps for backward elimination: The initial model contains all potential explanatory variables. If all variables make significant partial contributions at some fixed α-level, according to the usual t test or F test, then that model is the final one. Otherwise, the explanatory variable having the largest P-value, controlling for the other variables in the model, is removed. Next, the model is refitted with that variable removed, and the partial contributions of the variables

remaining in the model are reassessed, controlling for the other variables still in the model. If they are all significant, that model is the final model. Otherwise, the variable having the largest *P*-value is removed. The process continues until each remaining predictor explains a significant partial amount of the variability in *y*.

Selecting Explanatory Variables for House Selling Price   Example 9.10 (page 280) introduced a data set consisting of 100 observations on house selling prices with several explanatory variables. The data are in the Houses data file at the text website. We use *y* = selling price of home, with explanatory variables size of home (denoted SIZE), annual taxes (TAXES), number of bedrooms (BEDS), number of bathrooms (BATHS), and a dummy variable for whether the home is new (NEW). We use backward elimination with these variables as potential explanatory variables but without interaction terms, requiring a variable to reach significance at the $\alpha = 0.05$ level for inclusion in the model.

Table 14.1 shows the first stage of the process, fitting the model containing all the explanatory variables. The variable making the least partial contribution to the model is BATHS. Its *P*-value ($P = 0.85$) is the largest. Although BATHS is moderately correlated with the selling price ($r = 0.56$), the other explanatory variables together explain most of the same variability in selling price. Once those variables are in the model, BATHS is essentially redundant.

TABLE 14.1: Model Fit at Initial Stage of Backward Elimination for Predicting House Selling Price

Variable	Coef.	Std. Error	t	Sig
(Constant)	4525.75	24474.05		
SIZE	68.35	13.94	4.90	.000
NEW	41711.43	16887.20	2.47	.015
TAXES	38.13	6.81	5.60	.000
BATHS	-2114.37	11465.11	-.18	.854
BEDS	-11259.10	9115.00	-1.23	.220

When we refit the model after dropping BATHS, the only nonsignificant variable is BEDS, having a *t* statistic of -1.31 and *P*-value $= 0.19$. Table 14.2 shows the third stage, refitting the model after dropping BATHS and BEDS as explanatory variables. Each variable now makes a significant contribution, controlling for the others in the model. Thus, this is the final model. Backward elimination provides the prediction equation

$$\hat{y} = -21{,}353.8 + 61.7(\text{SIZE}) + 46{,}373.7(\text{NEW}) + 37.2(\text{TAXES}).$$

Other things being equal, an extra thousand square feet of size increases the predicted selling price by about 62 thousand dollars, and having a new home increases

TABLE 14.2: Model Fit at Third Stage of Backward Elimination for Predicting House Selling Price

Variable	Coef.	Std. Error	Std. Coeff	t	Sig
(Constant)	-21353.8	13311.49			
SIZE	61.70	12.50	0.406	4.94	.000
NEW	46373.70	16459.02	0.144	2.82	.006
TAXES	37.23	6.74	0.466	5.53	.000

it by about 46 thousand dollars. With standardized variables, the equation is

$$\hat{z}_y = 0.406z_S + 0.144z_N + 0.464z_T.$$

SIZE and TAXES have similar partial effects.

If we had included interactions in the original model, we would have ended up with a different final model. However, the model given here has the advantage of simplicity, and it has good predictive power ($R^2 = 0.790$, compared to 0.793 with all the explanatory variables). ■

FORWARD SELECTION AND STEPWISE REGRESSION PROCEDURES

Whereas backward elimination begins with *all* the potential explanatory variables in the model, *forward selection* begins with *none* of them. It adds one variable at a time to the model until no remaining variable not yet in the model makes a significant partial contribution to predicting y. At each step, the variable added is the one that is most significant, having the smallest P-value. For quantitative explanatory variables, this is the variable having the largest t test statistic, or equivalently the one providing the greatest increase in R^2.

For the data on selling prices of homes, Table 14.3 depicts the process. The variable most highly correlated with selling price is TAXES, so it is added first. Once TAXES is in the model, SIZE provides the greatest boost to R^2, and it is significant ($P = 0.000$), so it is the second variable added. Once both TAXES and SIZE are in the model, NEW provides the greatest boost to R^2 and it is significant ($P = 0.006$), so it is added next. At this stage, BEDS gives the greatest boost to R^2 (from 0.790 to 0.793), but it does not make a significant contribution ($P = 0.194$), so the final model does not include it. In this case, forward selection reaches the same final model as backward elimination.

TABLE 14.3: Steps of Forward Selection for Predicting House Selling Price. The model chosen has predictors TAXES, SIZE, and NEW.

Step	Variables in Model	P-Value for New Term	R^2
0	None	—	0.000
1	TAXES	0.000	0.709
2	TAXES, SIZE	0.000	0.772
3	TAXES, SIZE, NEW	0.006	0.790
4	TAXES, SIZE, NEW, BEDS	0.194	0.793

Once forward selection provides a final model, not all the explanatory variables appearing in it are necessarily significantly related to y. The variability in y explained by a variable entered at an early stage may overlap with the variability explained by variables added later, so it may no longer be significant. Figure 14.1 illustrates this. The figure portrays the portion of the total variability in y explained by each of three explanatory variables. Variable x_1 explains a similar amount of variability, by itself, as x_2 or x_3. However, x_2 and x_3 between them explain much of the same variation that x_1 does. Once x_2 and x_3 are in the model, the unique variability explained by x_1 is minor.

Stepwise regression is a modification of forward selection that drops variables from the model if they lose their significance as other variables are added. The approach is the same as forward selection except that at each step, after entering the new variable, the procedure drops from the model any variables that no longer make

FIGURE 14.1: Variability in y Explained by x_1, x_2, and x_3. The shaded portion is the amount explained by x_1 that is also explained by x_2 and x_3.
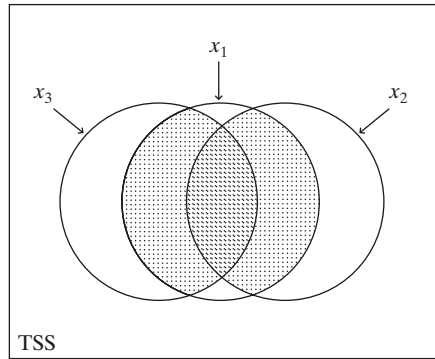
significant partial contributions. A variable entered into the model at some stage may eventually be eliminated because of its overlap with variables entered at later stages.

For the home sales data, stepwise regression behaves the same way as forward selection. At each stage, each variable in the model makes a significant contribution, so no variables are dropped. For these variables, backward elimination, forward selection, and backward elimination all agree. This need not happen.

LIMITATIONS AND ABUSES OF AUTOMATIC SELECTION PROCEDURES

It may seem appealing to select explanatory variables automatically according to established criteria. But any variable selection method should be used with caution and should not substitute for theory and careful thought. There is no guarantee that the final model chosen will be sensible.

For instance, suppose we specify all the pairwise interactions as well as the main effects as the potential explanatory variables. In this case, it is inappropriate to remove a main effect from a model that contains an interaction composed of that variable. Yet, most software does not have this safeguard. To illustrate, we used forward selection with the home sales data, including the 5 explanatory variables as well as their 10 cross-product interaction terms. The final model has $R^2 = 0.866$, using four interaction terms (SIZE×TAXES, SIZE ×NEW, TAXES×NEW, BATHS×NEW) and the TAXES main effect. It is inappropriate, however, to use these interactions as explanatory variables without the SIZE, NEW, and BATHS main effects.

Also, a variable selection procedure may exclude an important explanatory variable that really should be in the model according to other criteria. For instance, using backward elimination with the five explanatory variables of home selling price and their interactions, TAXES was removed. At a certain stage, TAXES explained an insignificant part of the variation in selling price. Nevertheless, it is the best single predictor of selling price, having $r^2 = 0.709$ by itself. (Refer to step 1 of the forward selection process in Table 14.3.) Since TAXES is such an important determinant of selling price, it seems sensible that any final model should include it as an explanatory variable.

Although P-values provide a guide for making decisions about adding or dropping variables in selection procedures, they are not the *true P*-values for the tests conducted. We add or drop a variable at each stage according to a minimum or maximum P-value, but the sampling distribution of the maximum or minimum of a set of t or F statistics differs from the sampling distribution for the statistic for an a priori chosen test. For instance, suppose we add variables in forward selection according to whether the P-value is less than 0.05. Even if none of the potential explanatory variables truly affect y, the probability is considerably larger than 0.05 that at least one

of the separate test statistics provides a *P*-value below 0.05 (Exercise 14.48). At least one variable that is not really important may look impressive merely due to chance.

Similarly, for the final model suggested by a particular selection procedure, any inferences conducted with it are highly approximate. In particular, *P*-values are likely to appear smaller than they should be and confidence intervals are likely to be too narrow, because the model was chosen that most closely reflects the data, in some sense. The inferences are more believable if performed for that model with a new set of data. (See the related discussion about *cross-validation* on page 437.)

EXPLORATORY VERSUS EXPLANATORY (THEORY-DRIVEN) RESEARCH

There is a basic difference between *explanatory* and *exploratory* modes of model selection. **Explanatory research** has a theoretical model to test using multiple regression. We might test whether a hypothesized spurious association disappears when other variables are controlled, for example. In such research, automated selection procedures are usually not appropriate, because theory dictates which variables are in the model.

Exploratory research, by contrast, has the goal not of examining theoretically specified relationships but merely finding a good set of explanatory variables. This approach searches for explanatory variables that give a large R^2, without concern about theoretical explanations. Thus, educational researchers might use a variable selection procedure to search for a set of test scores and other factors that predict well how students perform in college. They should be cautious about giving causal interpretations to the effects. For example, possibly the best predictor of students' success in college is whether their parents use the Internet for voice communication (with a program such as Skype).

In summary, automated variable selection procedures are no substitute for careful thought in formulating models. For most scientific research, they are not appropriate.

INDICES FOR SELECTING A MODEL: ADJUSTED R^2, PRESS, AND AIC

Instead of using an automated algorithm to choose a model, we could specify a set of potentially adequate models, and then use some established criterion to select among them. We next present some possible criteria.

Recall that maximizing R^2 is not a sensible criterion, because the most complicated model will have the largest R^2-value. This reflects the upward bias that R^2 has as an estimator of the population value of R^2. This bias can be considerable with small n or with many explanatory variables. In comparing predictive power of different models, it is more helpful to use *adjusted* R^2 instead of R^2. This is

$$R^2_{\text{adj}} = \frac{s^2_y - s^2}{s^2_y} = 1 - \frac{s^2}{s^2_y},$$

where $s^2 = \sum(y - \hat{y})^2/[n - (p + 1)]$ is the estimated conditional variance (i.e., the residual mean square) and $s^2_y = \sum(y - \bar{y})^2/(n - 1)$ is the sample variance of *y*. This is a less biased estimator of the population R^2. Unlike ordinary R^2, if we add a term to a model that is not especially useful, then R^2_{adj} may even *decrease*. This happens when the new model has poorer predictive power, in the sense of a larger value of s^2. A possible criterion for selecting a model is to choose the one having the greatest value of R^2_{adj}. This is, equivalently, the model with smallest residual MS.

Most other criteria for selecting a model attempt to find the model for which the predicted values tend to be closest to the true expected values. One type of method

for doing this uses ***cross-validation***. For a given model, you fit the model using some of the data and then analyze how well its prediction equation predicts the rest of the data. In one version, you use all observations except one to fit the model, and then check how well it predicts the remaining observation. Suppose we fit a model using all the data except observation 1. Using the prediction equation we get, let $\hat{y}_{(1)}$ denote the predicted selling price for observation 1. That is, we find a prediction equation using the data for observations $2, 3, \ldots, n$, and then we substitute the values of the explanatory variables for observation 1 into that prediction equation to get $\hat{y}_{(1)}$. Likewise, let $\hat{y}_{(2)}$ denote the prediction for observation 2 when we fit the model to observations $1, 3, 4, \ldots, n$, leaving out observation 2. In general, for observation i, we leave it out in fitting the model and then use the resulting prediction equation to get $\hat{y}_{(i)}$. Then, $(y_i - \hat{y}_{(i)})$ is a type of residual, measuring how far observation i falls from the value predicted for it using the prediction equation generated by the other $(n - 1)$ observations.

In summary, for a model for n observations, this version of cross-validation fits the model n times, each time leaving out one observation and using the prediction equation to predict that observation. We then get n predicted values and corresponding prediction residuals. The ***predicted residual sum of squares***, denoted by PRESS, is

$$\text{PRESS} = \sum(y_i - \hat{y}_{(i)})^2.$$

The smaller the value of PRESS, the better the predictions tend to be, in a summary sense. According to this criterion, the best-fitting model is the one with the smallest value of PRESS.

The ***AIC***, short for ***Akaike information criterion***, attempts to find a model for which the $\{\hat{y}_i\}$ tend to be closest to $\{E(y_i)\}$ in an average sense. The AIC is also scaled in such a way that the lower the value, the better the model. The best model is the one with the smallest AIC. We do not show its formula here, but it is sufficient to know that for ordinary regression models, minimizing the AIC corresponds to minimizing

$$n \log(\text{SSE}) + 2p,$$

where p is the number of model parameters. So, this criterion penalizes a model for having more parameters than are useful for getting good predictions. An advantage of the AIC is that its general definition also makes it applicable for models that assume nonnormal distributions for y, in which case a sum of squared errors is often not a useful summary.

Example 14.2

Using Indices to Select a Model for House Selling Price   Table 14.4 shows the model selection indices for five models for the house selling price data. The table shows the models in the order built by forward selection (reverse order for backward elimination).

TABLE 14.4: Model Selection Criteria for Models for House Selling Price

Variables in Model	R^2	R^2_{adj}	PRESS	AIC
TAXES	0.709	0.706	3.17	2470.5
TAXES, SIZE	0.772	0.767	2.73	2448.0
TAXES, SIZE, NEW	0.790	0.783	2.67	2442.0
TAXES, SIZE, NEW, BEDS	0.793	0.785	2.85	2442.2
TAXES, SIZE, NEW, BEDS, BATHS	0.793	0.782	2.91	2444.2

Note: Actual PRESS equals value reported times 10^{11}.

According to the criterion of minimizing adjusted R^2, the selected model has all the explanatory variables except BATHS. It has $R^2_{\text{adj}} = 0.785$. To illustrate that R^2_{adj} can decrease when variables are added, note that this model as well as the model with SIZE, NEW, and TAXES predictors have R^2_{adj} values that are higher than $R^2_{\text{adj}} = 0.782$ for the full model with all the explanatory variables.

According to the criterion of minimizing the predicted residual sum of squares, the selected model has explanatory variables TAXES, SIZE, and NEW. It has the minimum PRESS = 2.67. (The y values were in dollars, so squared residuals tended to be huge numbers, and the actual PRESS values are the numbers reported multiplied by 10^{11}.) This was also the model selected by backward elimination and by forward selection.

According to the criterion of minimizing AIC, the selected model is also the one with explanatory variables TAXES, SIZE, and NEW. It has the minimum AIC = 2442.0. The model also containing NEW fits essentially as well. ∎

14.2 Regression Diagnostics

Once we have selected the explanatory variables for a model, how do we know that model fits the data adequately? This section introduces diagnostics that indicate (1) when model assumptions are grossly violated and (2) when certain observations are highly influential in affecting the model fit or inference about model parameters.

Recall that inference about parameters in a regression model has these assumptions:

- The true regression function has the form used in the model (e.g., linear).

- The conditional distribution of y is normal.

- The conditional distribution of y has constant standard deviation throughout the range of values of the explanatory variables. This condition is called *homoscedasticity*.

- The sample is randomly selected.

In practice, the assumptions are never perfectly fulfilled, but the regression model can still be useful. It is adequate to check that no assumption is grossly violated.

EXAMINE THE RESIDUALS

Several checks of assumptions use the residuals, $y - \hat{y}$. One check concerns the normality assumption. If the observations are normally distributed about the true regression equation with constant conditional standard deviation σ, then the residuals should be approximately normally distributed. To check this, plot the residuals about their mean value 0, using a histogram. They should have approximately a bell shape about 0.

A standardized version of the residual equals the residual divided by its standard error, which describes how much residuals vary because of ordinary sampling variability. In regression, this is called[1] a *studentized residual*. Under the normality assumption, a histogram of these residuals should have the appearance of a standard normal distribution (bell shaped with mean of 0 and standard deviation of 1).

---

[1] Some software reports also a *standardized residual*, which divides $y - \hat{y}$ by s, which is slightly larger than the standard error of the residual.

If a studentized residual is larger than about 3 in absolute value, the observation is a potential outlier and should be checked. If an outlier represents a measurement error, it could cause a major bias in the prediction equation. Even if it is not an error, it should be investigated. It represents an observation that is not typical of the sample data, and it may have too much impact on the model fit. Consider whether there is some reason for the peculiarity. Sometimes the outliers differ from the other observations on some variable not included in the model, and once that variable is added, they cease to be outliers.

Example 14.3

Residuals for Modeling Home Selling Price   For the Houses data file, with $y =$ selling price, variable selection procedures in Example 14.1 (page 433) and the AIC and PRESS indices in Example 14.2 suggested the model with prediction equation

$$\hat{y} = -21,353.8 + 61.7(\text{SIZE}) + 46,373.7(\text{NEW}) + 37.2(\text{TAXES}).$$

Figure 14.2 is a histogram of the studentized residuals for this fit. No severe nonnormality seems to be indicated, since they are roughly bell shaped about 0. However, the plot indicates that two observations have relatively large residuals. On further inspection, we find that observation 6 had a selling price of $499,900, which was $168,747 higher than the predicted selling price for a new home of 3153 square feet with a tax bill of $2997. The residual of $168,747 has a studentized value of 3.88. Observation 64 had a selling price of $225,000, which was $165,501 lower than the predicted selling price for a non-new home of 4050 square feet with a tax bill of $4350. Its residual of $-$165,501 has a studentized value of -3.93.

FIGURE 14.2: Histogram of Studentized Residuals for Multiple Regression Model Fitted to House Selling Prices, with Explanatory Variables Size, Taxes, and New
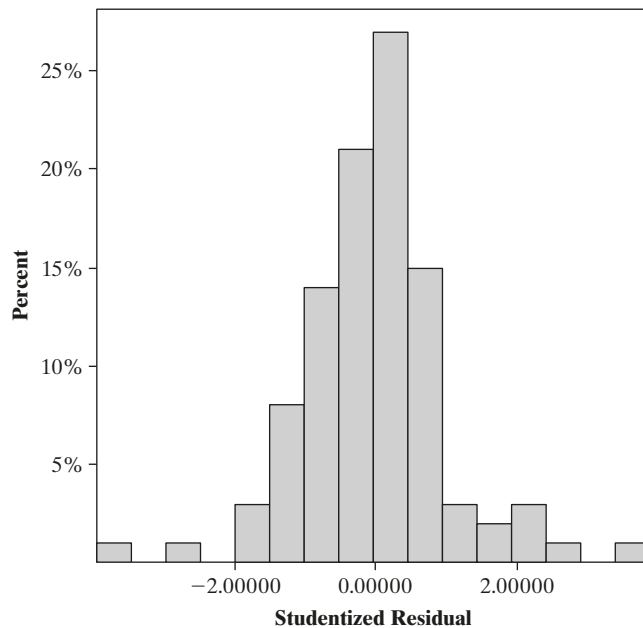
A severe outlier on y can substantially affect the fit, especially when the values of the explanatory variables are not near their means. So, we refitted the model without these two observations. The R^2-value changes from 0.79 to 0.83, and the prediction equation changes to

$$\hat{y} = -32,226 + 68.9(\text{SIZE}) + 20,436(\text{NEW}) + 38.3(\text{TAXES}).$$

The parameter estimates are similar for SIZE and TAXES, but the estimated effect of NEW drops from 46,374 to 20,436. Moreover, the effect of NEW is no longer significant, having a *P*-value of 0.17. Because the estimated effect of NEW is affected substantially by these two observations, we should be cautious in making conclusions about its effect. Of the 100 homes in the sample, only 11 were new. It is difficult to make precise estimates about the NEW effect with so few new homes, and results are highly affected by a couple of unusual observations. ∎

PLOTTING RESIDUALS AGAINST EXPLANATORY VARIABLES

The normality assumption is not as important as the assumption that the model provides a good approximation for the true relationship between the explanatory variables and the mean of *y*. If the model assumes a linear effect but the effect is actually strongly nonlinear, some conclusions may be faulty.

For bivariate models, the scatterplot provides a simple check on the form of the relationship. For multiple regression, it is also useful to construct a scatterplot of each explanatory variable against the response variable. This displays only the *bivariate* relationships, however, whereas the model refers to the *partial* effect of each explanatory variable, with the others held constant. The *partial regression plot* introduced on page 326 provides some information about this.

For multiple regression models, plots of the residuals (or studentized residuals) against the predicted values $\hat{y}$ or against each explanatory variable also help us check for potential problems. If the residuals appear to fluctuate randomly about 0 with no obvious trend or change in variation as the values of a particular x_i increase, then no violation of assumptions is indicated. The pattern should be roughly like Figure 14.3a.

FIGURE 14.3: Possible Patterns for Residuals (*e*), Plotted against an Explanatory Variable *x*

(a) Assumptions satisfied   (b) Nonconstant standard deviation   (c) Nonlinear term needed
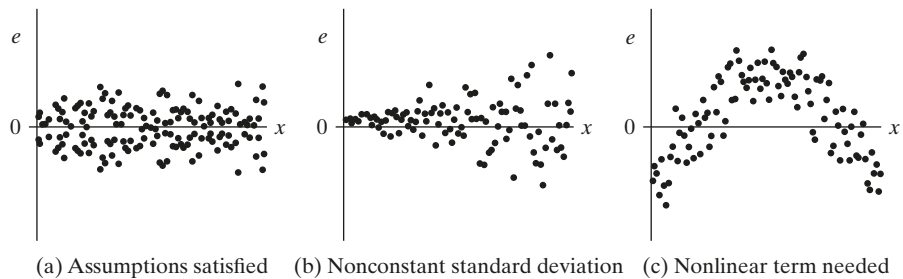
In practice, most response variables can take only nonnegative values. For such responses, a fairly common occurrence is that the variability increases as the mean increases. For example, suppose we model y = annual income (in dollars) using several explanatory variables. For those subjects having $E(Y)$ = \$10,000, the standard deviation of income is probably much less than for those subjects having $E(Y)$ = \$200,000. Plausible standard deviations might be \$4000 and \$80,000. When this happens, the conditional standard deviation of y is not constant, whereas ordinary regression assumes that it is. An indication that this is happening is when the residuals are more spread out as the y_i-values increase. If we plot the residuals against a predictor that has a positive partial association with y, such as number of years of education, the residuals are then more spread out for larger values of the predictor, as in Figure 14.3b.

Figure 14.3c shows another possible abnormality, in which y tends to be below $\hat{y}$ for very small and very large x_i-values (giving negative residuals) and above $\hat{y}$ for medium-sized x_i-values (giving positive residuals). Such a scattering of residuals

suggests that y is actually nonlinearly related to x_i. Sections 14.5 and 14.6 show how to address nonlinearity.

For the model relating selling price of home to size, taxes, and whether new for all 100 observations, Figure 14.4 plots the residuals against size. There is some suggestion of more variability at the higher size values. It does seem sensible that selling prices would vary more for very large homes than for very small homes. A similar picture occurs when we plot the residuals against taxes.

FIGURE 14.4: Scatterplot of Studentized Residuals of Home Selling Price Plotted against Size of Home, for Model with Explanatory Variables Size, Taxes, and Whether Home Is New
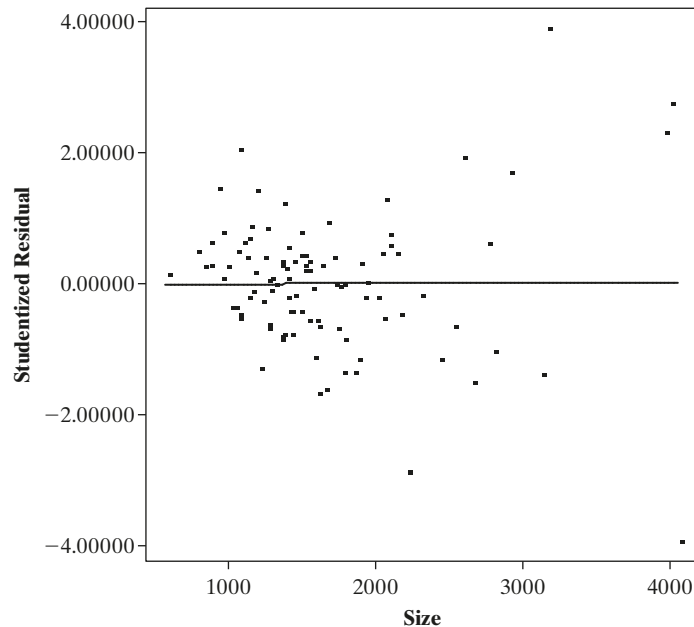
If the change in variability is severe, then a method other than ordinary least squares provides better estimates with more valid standard errors. Section 14.4 presents a generalized regression model that allows the variability to be greater when the mean is greater.

In practice, residual patterns are rarely as neat as the ones in Figure 14.3. Don't let a few outliers or ordinary sampling variability influence too strongly your interpretation of a plot. Also, the plots described here just scratch the surface of the graphical tools now available for diagnosing potential problems. Fox (2015, Section III) described a variety of modern graphical displays and diagnostic tools.

TIME SERIES DATA AND LONGITUDINAL STUDIES

Some social research studies collect observations sequentially over time. For economic variables such as a stock index or the unemployment rate, for example, the observations often occur daily or monthly. The observations are then recorded in sequence, rather than randomly sampled. Sampling subjects randomly from some population ensures that one observation is not statistically dependent on another, and this simplifies derivations of sampling distributions and their standard errors. However, neighboring observations from a time sequence are usually correlated rather than independent. For example, if the unemployment rate is relatively low in January 2018, it will probably also be relatively low in February 2018.

A plot of the residuals against the time of making the observation checks for this type of dependence. Ideally, the residuals should fluctuate in a random pattern about

0 over time, rather than showing a trend or periodic cycle. The methods presented in this text are based on independent observations and are inappropriate when time effects occur. For example, when observations next to each other tend to be positively correlated, the standard error of the sample mean is larger than the $\sigma / \sqrt{n}$ formula that applies for independent observations.

The term *time series* refers to relatively long sequences of observations over time. Books specializing in econometrics, such as Kennedy (2008), present methods for time series data. The term *longitudinal data* refers to studies, common in the social sciences and public health, that observe subjects over a relatively small number of times. For analyzing such data, see the linear mixed model in Section 13.5 and books by Fitzmaurice et al. (2011) and Hedeker and Gibbons (2006).

DETECTING INFLUENTIAL OBSERVATIONS: RESIDUAL AND LEVERAGE

Least squares estimates of parameters in regression models can be strongly influenced by an outlier, especially when *n* is small. A variety of statistics summarize the influence each observation has. These statistics refer to how much the predicted values $\hat{y}$ or the model parameter estimates change when we remove an observation from the data set. An observation's influence depends on two factors: (1) how far its *y*-value falls from the overall trend in the data and (2) how far the values of the explanatory variables fall from their means.

The first factor on influence (how far *y* falls from the overall trend) is measured by the observation's residual, $y - \hat{y}$. The larger the residual, the farther the observation falls from the overall trend. We can search for observations with large studentized residuals (say, larger than about 3 in absolute value) to find observations that may be influential.

The second factor on influence (how far the explanatory variables fall from their means) is summarized by the **leverage** of the observation. The leverage is a nonnegative statistic such that the larger its value, the greater weight that observation receives in determining the $\hat{y}$-values (hence, it also is sometimes called a *hat value*). The formula for the leverage in multiple regression is complex. For the bivariate model, the leverage for observation *i* simplifies to

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x - \bar{x})^2}.$$

So, the leverage gets larger as the *x*-value x_i for observation *i* gets farther from the mean. It gets smaller as the sample size increases. When calculated for each observation in a sample, the average leverage equals the number *p* of parameters in the model divided by *n*.

DETECTING INFLUENTIAL OBSERVATIONS: DFFIT AND DFBETA

For an observation to be influential, it must have both a relatively large residual and a relatively large leverage. Statistical software reports diagnostics that depend on the residuals and the leverages. Two popular ones are called **DFFIT** and **DFBETA**.

For an observation, DFBETA summarizes the *effect on the model parameter estimates* of removing the observation from the data set. For the effect β_j of x_j, DFBETA equals the change in the estimate $\hat{\beta}_j$ due to deleting the observation. The larger the absolute value of DFBETA, the greater the influence of the observation on that parameter estimate. Each observation has a DFBETA value for each parameter in the model.

DFFIT summarizes the *effect on the fit* of deleting the observation. For observation i, DFFIT equals the change in the predicted value due to deleting that observation (i.e., $\hat{y}_i - \hat{y}_{(i)}$). The DFFIT value has the same sign as the residual. ***Cook's distance*** is an alternative measure with the same purpose. Cook's distance and DFFIT are based on the effect that observation i has on *all* the parameter estimates. They summarize more broadly the influence of an observation, as each observation has a single DFFIT value and a single Cook's distance, whereas it has a separate DFBETA for each parameter. The larger their absolute values, the greater the influence that observation has on the fitted values.

Some software reports *standardized* versions of the DFBETA and DFFIT measures, often denoted by DFBETAS and DFFITS. The standardized DFBETA divides the change in the estimate $\hat{\beta}_j$ due to deleting the observation by the standard error of $\hat{\beta}_j$ for the adjusted data set. For observation i, the standardized DFFIT equals the change in the predicted value due to deleting that observation, divided by the standard error of $\hat{y}$ for the adjusted data set.

In practice, scan or plot these diagnostic measures to see if some observations stand out from the rest, having relatively large values. Each measure has approximate cutoff points for noteworthy observations. For example, a Cook's distance larger than about $4/n$ indicates a potentially large influence. A standardized DFBETA larger than 1 suggests a substantial influence on that parameter estimate. However, Cook's distance, DFBETA, and DFFIT tend to decrease as n increases, so normally it is a good idea to examine observations having extreme values relative to the others. Individual data points have less influence for larger sample sizes.

| Example 14.4 | **DFBETA and DFFIT for an Influential Observation**   Example 14.3 (page 439) showed that observations 6 and 64 were influential on the equation for predicting home selling price using size of home, taxes, and whether the house is new. The prediction equation for all 100 observations is |

$$\hat{y} = -21{,}354 + 61.7(\text{SIZE}) + 46{,}373.7(\text{NEW}) + 37.2(\text{TAXES}).$$

For observation 6, the DFBETA values are 12.5 for size, 16,318.5 for new, and -5.7 for taxes. This means, for example, that if this observation is deleted from the data set, the effect of NEW changes from 46,373.7 to $46{,}373.7 - 16{,}318.5 = 30{,}055.2$. Observation 6 had a predicted selling price of $\hat{y} = 331{,}152.8$. Its DFFIT value is 29,417.0. This means that if observation 6 is deleted from the data set, then $\hat{y}$ at the explanatory variable values for observation 6 changes to $331{,}152.8 - 29{,}417.0 = 301{,}735.8$. This analysis shows that this observation is quite influential. ■

| Example 14.5 | **Influence Diagnostics for Crime Data**   Table 9.1 (page 260) listed $y =$ murder rate for the 50 states and the District of Columbia (D.C.), with explanatory variables $x_1 =$ percentage of families below the poverty level and $x_2 =$ percentage of single-parent families. The data are in the `Crime2` data file at the text website. The least squares fit of the multiple regression model is |

$$\hat{y} = -40.7 + 0.32x_1 + 3.96x_2.$$

Table 14.5 shows the influence diagnostics for the model fit, including the standardized versions of DFBETA and DFFIT. The studentized residuals fall in a reasonable range except the one for the last observation (D.C.), which equals 14.2. The observed murder rate of 78.5 for D.C. falls far above the predicted value of 55.3, causing a large positive residual. This is an extreme outlier. In addition, the leverage for D.C. is 0.54, more than three times as large as any other leverage and nine times the average

TABLE 14.5: Influence Diagnostics for Model Using Poverty Rate and Single-Parent Percentage to Predict Murder Rate for 50 U.S. States and District of Columbia

Obs	Dep Var MURDER	Predict Value	Residual	Student Resid	Leverage h	Dffits	POVERTY Dfbeta	SINGLE Dfbeta
AK	9.0	18.88	−9.88	−2.04	0.162	−0.895	0.714	−0.761
AL	11.6	10.41	1.18	0.22	0.031	0.039	0.024	−0.011
AR	10.2	8.07	2.13	0.40	0.079	0.117	0.100	−0.069
AZ	8.6	12.16	−3.55	−0.65	0.022	−0.099	−0.005	−0.025
CA	13.1	14.63	−1.53	−0.28	0.034	−0.053	−0.027	−0.004
CO	5.8	10.41	−4.61	−0.87	0.060	−0.220	0.174	−0.134
CT	6.3	2.04	4.25	0.79	0.051	0.185	−0.130	0.015
DE	5.0	7.73	−2.73	−0.50	0.043	−0.107	0.079	−0.045
FL	8.9	6.97	1.92	0.35	0.048	0.080	0.059	−0.047
GA	11.4	15.12	−3.72	−0.69	0.042	−0.145	0.071	−0.105
HI	3.8	−2.07	5.87	1.11	0.059	0.279	−0.153	−0.058
IA	2.3	−1.74	4.04	0.75	0.045	0.164	−0.034	−0.081
ID	2.9	1.12	1.77	0.32	0.035	0.063	0.012	−0.040
IL	11.4	9.21	2.18	0.40	0.020	0.058	−0.013	0.011
IN	7.5	5.99	1.50	0.27	0.023	0.043	−0.014	−0.000
KS	6.4	2.71	3.68	0.68	0.029	0.117	0.013	−0.062
KY	6.6	7.79	−1.19	−0.22	0.088	−0.070	−0.061	0.043
LA	20.3	26.74	−6.44	−1.29	0.161	−0.568	−0.412	−0.055
MA	3.9	5.91	−2.01	−0.37	0.033	−0.068	0.042	−0.014
MD	12.7	9.95	2.74	0.51	0.060	0.130	−0.104	0.077
ME	1.6	4.72	−3.12	−0.57	0.031	−0.104	0.058	−0.008
MI	9.8	15.72	−5.92	−1.10	0.033	−0.204	0.035	−0.124
MN	3.4	2.23	1.16	0.21	0.029	0.037	−0.007	−0.013
MO	11.3	7.62	3.67	0.67	0.027	0.115	0.059	−0.049
MS	13.5	25.40	−11.90	−2.45	0.126	−0.933	−0.623	−0.151
MT	3.0	6.84	−3.84	−0.70	0.023	−0.108	−0.033	0.039
NC	11.3	7.87	3.42	0.62	0.020	0.090	0.009	−0.013
ND	1.7	−3.83	5.53	1.04	0.057	0.259	0.016	−0.184
NE	3.9	−0.15	4.05	0.75	0.039	0.153	−0.047	−0.056
NH	2.0	−1.07	3.07	0.57	0.044	0.123	−0.039	−0.047
NJ	5.3	0.82	4.47	0.83	0.035	0.158	−0.041	−0.058
NM	8.0	19.53	−11.53	−2.25	0.046	−0.499	−0.017	−0.308
NV	10.4	11.57	−1.17	−0.22	0.069	−0.060	0.048	−0.040
NY	13.3	14.85	−1.55	−0.28	0.028	−0.048	−0.005	−0.019
OH	6.0	8.62	−2.62	−0.48	0.022	−0.072	0.024	−0.015
OK	8.4	9.62	−1.22	−0.22	0.067	−0.061	−0.051	0.031
OR	4.6	7.84	−3.24	−0.59	0.027	−0.101	0.054	−0.029
PA	6.8	1.55	5.24	0.97	0.034	0.183	0.036	−0.115
RI	3.9	5.67	−1.77	−0.32	0.028	−0.056	0.029	−0.006
SC	10.3	13.99	−3.69	−0.68	0.038	−0.137	−0.084	0.008
SD	3.4	1.07	2.32	0.43	0.042	0.091	0.036	−0.067
TN	10.2	9.92	0.27	0.05	0.060	0.013	0.010	−0.006
TX	11.9	11.60	0.29	0.05	0.029	0.009	0.005	−0.001
UT	3.1	2.34	0.75	0.13	0.032	0.025	−0.010	−0.004
VA	8.3	3.21	5.08	0.94	0.039	0.192	−0.119	0.010
VT	3.6	6.08	−2.48	−0.46	0.040	−0.094	0.067	−0.028
WA	5.2	9.52	−4.32	−0.80	0.029	−0.139	0.078	−0.059
WI	4.4	4.53	−0.13	−0.02	0.023	−0.003	0.000	0.001
WV	6.9	3.60	3.29	0.66	0.178	0.307	0.274	−0.229
WY	3.4	6.34	−2.94	−0.54	0.021	−0.079	0.006	0.012
DC	78.5	55.28	23.22	14.20	0.536	15.271	−0.485	12.792

leverage of $p/n = 3/51 = 0.06$. Since D.C. has both a large studentized residual and a large leverage, it has considerable influence on the model fit.

Not surprisingly, DFFIT for D.C. is much larger than for the other observations. This suggests that the predicted values change considerably if we refit the model after removing this observation. The DFBETA value for the single-family explanatory variable x_2 is much larger for D.C. than for the other observations. This suggests that the effect of x_2 could change substantially with the removal of D.C. By contrast, DFBETA for poverty is not so large.

These diagnostics suggest that the D.C. observation has a large influence, particularly on the coefficient of x_2 and on the fitted values. The prediction equation for the model fitted without the D.C. observation is

$$\hat{y} = -14.6 + 0.36x_1 + 1.52x_2.$$

Not surprisingly, the estimated effect of x_1 did not change much, but the coefficient of x_2 is now less than half as large. The standard error of the coefficient of x_2 also changes dramatically, decreasing from 0.44 to 0.26. ∎

An observation with a large studentized residual does not have a major influence if its values on the explanatory variables do not fall far from their means. Recall that the leverage summarizes how far the explanatory variables fall from their means. For instance, New Mexico has a relatively large negative studentized residual (-2.25) but a relatively small leverage (0.046), so it does not have large values of DFFIT or DFBETA. Similarly, an observation far from the mean on the explanatory variables (i.e., with a large leverage) need not have a major influence if it falls close to the prediction equation and has a small studentized residual. For instance, West Virginia has a relatively large poverty rate and its leverage of 0.178 is triple the average. However, its studentized residual is small (0.66), so it has little influence on the fit.

14.3 Effects of Multicollinearity

In many social science studies using multiple regression, the explanatory variables "overlap" considerably. A variable may be nearly redundant, in the sense that it can be predicted well using the others. If we regress an explanatory variable on the others and get a large R^2-value, this suggests that it may not be needed in the model once the others are there. This condition is called *multicollinearity*, or sometimes simply *collinearity*. This section describes the effects of multicollinearity and ways to diagnose it.

VIF: MULTICOLLINEARITY CAUSES VARIANCE INFLATION

Multicollinearity causes inflated standard errors for estimates of regression parameters. The standard error of the estimator of the coefficient β_j of x_j in the multiple regression model can be expressed as

$$se = \frac{1}{\sqrt{1 - R_j^2}} \left[\frac{s}{\sqrt{n-1}s_{x_j}} \right],$$

where s is the square root of the residual mean square and s_{x_j} denotes the sample standard deviation of x_j values. Let R_j^2 denote R^2 from the regression of x_j on the other explanatory variables from the model. So, when x_j overlaps a lot with the other explanatory variables, in the sense that R_j^2 is large for predicting x_j using the other explanatory variables, this se is relatively large. Then, the confidence interval for β_j

is wide, and the test of H_0: $\beta_j = 0$ has a large P-value unless the sample size is very large or the effect is very strong.

In this se formula for the estimate of β_j, the quantity

$$\text{VIF} = 1/(1 - R_j^2)$$

is called a **_variance inflation factor_** (VIF). It represents the multiplicative increase in the variance (squared standard error) of the estimator due to x_j being correlated with the other explanatory variables. When any R_j^2-value from regressing each explanatory variable on the other explanatory variables in the model is close to 1, say above 0.90, severe multicollinearity exists.

For example, if $R_j^2 > 0.90$, then VIF > 10 for the effect of that explanatory variable. That is, the variance of the estimate of β_j inflates by a factor of more than 10. The standard error inflates by a factor of more than $\sqrt{10} = 3.2$, compared to the standard error for uncorrelated explanatory variables. When an explanatory variable is in the model primarily as a control variable, and we do not need precise estimates of its effect on the response variable, it is not crucial to worry about its VIF value.

For the model selected in Section 14.1 that predicts house selling price using taxes, size, and whether the house is new, software reports the VIF values

```
        VIF
TAXES   3.082
SIZE    3.092
NEW     1.192
```

The standard error for whether the house is new is not affected much by correlation with the other explanatory variables, but the other two standard errors multiply by a factor of roughly $\sqrt{3.1} = 1.76$.

OTHER INDICATORS OF MULTICOLLINEARITY

Even without checking VIFs, various types of behavior in a regression analysis can indicate potential problems due to multicollinearity. A warning sign occurs when the estimated coefficient for a predictor already in the model changes substantially when another variable is introduced. For example, perhaps the estimated coefficient of x_1 is 2.4 for the bivariate model, but when x_2 is added to the model, the coefficient of x_1 changes to 25.9.

Another indicator of multicollinearity is when a highly significant R^2 exists between y and the explanatory variables, but individually each partial regression coefficient is not significant. In other words, H_0: $\beta_1 = \cdots = \beta_k = 0$ has a small P-value in the overall F test, but H_0: $\beta_1 = 0$, H_0: $\beta_2 = 0$, and so forth do not have small P-values in the separate t tests. Thus, it is difficult to assess individual partial effects when severe multicollinearity exists. Other indicators of multicollinearity are surprisingly large standard errors, or standardized regression coefficients that are larger than 1 in absolute value.

When multicollinearity exists, it is rather artificial to interpret a regression coefficient as the effect of an explanatory variable when other variables are held constant. For instance, when $|r_{x_1 x_2}|$ is high, then as x_1 changes, x_2 also tends to change in a linear manner, and it is artificial to envision x_1 or x_2 as being held constant.

REMEDIAL ACTIONS WHEN MULTICOLLINEARITY EXISTS

Remedial measures can help to reduce the effects of multicollinearity. One solution is to choose a subset of the explanatory variables, removing those variables that

explain a small portion of the remaining unexplained variation in y. If x_4 and x_5 have a correlation of 0.96, it is only necessary to include one of them in the model.

When several explanatory variables are highly correlated and are indicators of a common feature, you could construct a summary index by combining responses on those variables. For example, suppose that a model for predicting $y =$ opinion about president's performance in office uses 12 explanatory variables, of which three refer to the subject's opinion about whether a woman should be able to obtain an abortion (1) when she cannot financially afford another child, (2) when she is unmarried, and (3) anytime in the first three months. Each of these items is scaled from 1 to 5, with a 5 being the most conservative response. They are likely to be highly positively correlated, contributing to multicollinearity. A possible summary measure for opinion about abortion averages (or sums) the responses to these items. Higher values on that summary index represent more conservative responses. If the items were measured on different scales, we could first standardize the scores before averaging them. Socioeconomic status is a variable of this type, summarizing the joint effects of education, income, and occupational prestige.

Often multicollinearity occurs when the explanatory variables include interaction terms. Since cross-product terms are composed of other explanatory variables in the model, it is not surprising that they tend to be highly correlated with the other terms. The effects of this are diminished if we center the explanatory variables by subtracting their sample means before entering them in the interaction model (see page 340).

Other procedures, beyond the scope of this chapter, can handle multicollinearity. For example, *factor analysis* is a method for creating artificial variables from the original ones in such a way that the new variables can be uncorrelated. In most applications, though, it is more advisable to use a subset of the variables or create some new variables directly, as just explained.

Multicollinearity does not adversely affect all aspects of regression. Although multicollinearity makes it difficult to assess *partial* effects of explanatory variables, it does not hinder the assessment of their *joint* effects. If newly added explanatory variables overlap substantially with ones already in the model, then R and R^2 will not increase much, but the fit will not be poorer. So, the presence of multicollinearity does not diminish the predictive power of the equation. For further discussion of the effects of multicollinearity and methods for dealing with it, see DeMaris (2004), Fox (2015, Chapter 13), and Kutner et al. (2008).

14.4 Generalized Linear Models

The models presented in this book are special cases of ***generalized linear models***. This broad class of models includes ordinary regression models for response variables assumed to have a normal distribution, alternative models for continuous variables that do not assume normality, and models for discrete response variables including categorical variables. This section introduces generalized linear models. We use the acronym *GLM*.

NONNORMAL DISTRIBUTIONS FOR A RESPONSE VARIABLE

As in other regression models, a GLM identifies a response variable y and a set of explanatory variables. The regression models presented in Chapters 9–14 are GLMs that assume that y has a normal distribution.

In many applications, the potential outcomes for y are binary rather than continuous. Each observation might be labeled as a *success* or *failure*, as in the methods for

proportions presented in Sections 5.2, 6.3, and 7.2. For instance, in a study of factors that influence votes in U.S. presidential elections, the response variable indicates the preferred candidate in the previous presidential election—the Democratic or the Republican candidate. In this case, models usually assume a *binomial* distribution for *y*.

In some applications, each observation is a count. For example, in a study of factors associated with family size, the response variable is the number of children in a family. GLMs for count data most often use two distributions for *y* not presented in this text, called the *Poisson* and the *negative binomial*.

Binary outcomes and counts are examples of discrete variables. Regression models that assume normal distributions are not optimal for models with discrete responses. Even when the response variable is continuous, the normal distribution may not be optimal. When each observation must take a positive value, for instance, the distribution is often skewed to the right with greater variability when the mean is greater. In that case, a GLM can assume a *gamma* distribution for *y*, as discussed later in this section.

THE LINK FUNCTION FOR A GLM

In a GLM, as in ordinary regression models, $\mu = E(y)$ varies according to values of explanatory variables, which enter linearly as predictors on the right-hand side of the model equation. However, a GLM allows a function $g(\mu)$ of the mean rather than just the mean μ itself on the left-hand side. The GLM formula states that

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

The function $g(\mu)$ is called the **link function**, because it links the mean of the response variable to the explanatory variables.

For instance, the link function $g(\mu) = \log(\mu)$ models the log of the mean. The log function applies to positive numbers, so this **log link** is appropriate when μ cannot be negative, such as with count data. GLMs that use the log link,

$$\log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

are often called **loglinear models**. The final section of this chapter shows an example.

For binary data, the most common link function is $g(\mu) = \log[\mu/(1-\mu)]$. This is called the **logit link**. It is appropriate when μ falls between 0 and 1, such as a probability, in which case $\mu/(1-\mu)$ is the *odds*. When *y* is binary, this link is used in models for the probability of a particular outcome, for instance, to model the probability that a subject votes for the Republican candidate. A GLM using the logit link, called a **logistic regression model**, is presented in the next chapter.

The simplest possible link function is $g(\mu) = \mu$. This models the mean directly and is called the **identity link**. It specifies a linear model for the mean response,

$$\mu = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

When employed with a normal assumption for *y*, this is the ordinary regression model.

A GLM generalizes ordinary regression in two ways: First, *y* can have a distribution other than the normal. Second, it can model a function of the mean. Both generalizations are important, especially for discrete responses.

GLMS VERSUS ORDINARY REGRESSION FOR TRANSFORMED DATA

Before GLMs were developed in the 1970s, the traditional way of analyzing "nonnormal" data was to transform the *y*-values. The goal of this approach is to find a function

g such that $g(y)$ has an approximately normal distribution, with constant standard deviation at all levels of the explanatory variables. Square root or log transforms are often applied to do this. If the variability is then more nearly constant, least squares works well with the transformed data. In practice, this may not work well. Simple linear models for the explanatory variables may fit poorly on that scale. If the original relationship is linear, it is no longer linear after applying the transformation. If we fit a straight line and then transform back to the original scale, the fit is no longer linear. Also, technical problems can occur, such as taking logarithms of 0. Moreover, conclusions that refer to the mean response on the scale of the transformed variable are less relevant.

With the GLM approach, it is not necessary to transform data and use normal methods. This is because the GLM fitting process utilizes the ***maximum likelihood*** estimation method (page 138), for which the choice of distribution for y is not restricted to normality. Maximum likelihood employs a generalization of least squares called ***weighted least squares***. It gives more weight to observations over regions that show less variability. In addition, in GLMs the choice of link function is separate from the choice of distribution for y. If a certain link function makes sense for a particular type of data, it is not necessary that it also stabilize variation or produce normality.

The family of GLMs unifies a wide variety of statistical methods. Ordinary regression models as well as models for discrete data (Chapter 15) are special cases of one highly general model. In fact, the same fitting method yields parameter estimates for all GLMs. Using GLM software, there is tremendous flexibility and power in the model-building process. You pick a probability distribution that is most appropriate for y. For instance, you might select the normal option for a continuous response or the binomial option for a binary response. You specify the variables that are the explanatory variables. Finally, you pick the link function, determining which function of the mean to model. Software then fits the model and provides the maximum likelihood estimates of model parameters. For further details about GLMs, see Fox (2015), Gill (2000), and King (1989).

GLMS FOR A RESPONSE ASSUMING A GAMMA DISTRIBUTION

The residual analysis in Example 14.3 for modeling selling prices of homes showed a tendency for greater variability of selling prices at higher house size values. (See Figure 14.4 on page 441.) Small homes show little variability in selling price, whereas large homes show high variability. Large homes are the ones that tend to have higher selling prices, so variability in y increases as its mean increases.

This phenomenon often happens for positive-valued response variables. When the mean response is near 0, less variation occurs than when the mean response is high. For such data, least squares is not optimal. Least squares is identical to maximum likelihood for a GLM in which y is assumed to have a normal distribution with *identical* standard deviation σ at all values of explanatory variables.

An alternative approach for data of this form assumes a distribution for y for which the standard deviation increases as the mean increases (i.e., that permits *heteroscedasticity*). The family of ***gamma distributions*** has this property. When y has a gamma distribution with mean μ, then y has

$$\text{Variance} = \phi\mu^2, \quad \text{Standard deviation} = \sqrt{\phi}\mu,$$

where ϕ is called a *scale parameter*. The standard deviation increases proportionally to the mean: When the mean doubles, the standard deviation doubles. The gamma distribution falls on the positive part of the line. It exhibits skewness to the right, like the chi-squared distribution, which is a special case of the gamma.

The scale parameter, or an equivalent *shape parameter* that is the reciprocal of the scale parameter, determines the shape of the distribution. The gamma distribution becomes more bell shaped as ϕ decreases, being quite bell shaped when $\phi < 0.1$. It becomes more skewed as ϕ increases, being so highly skewed when $\phi \geq 1$ that the mode is 0.

With GLMs, you can fit a regression model assuming a gamma distribution for y instead of a normal distribution. Even if the data are close to normal, this alternative fit is more appropriate than the least squares fit when the standard deviation increases proportionally to the mean. Just as ordinary regression models assume that the variance is constant for all values of the explanatory variables, software for gamma GLMs assumes a constant scale parameter and estimates it as part of the model-fitting process.[2]

When the relationship is closer to linear on a log scale for $E(y)$, it is preferable to apply the log as a link function with a gamma GLM. The log link is also used when a linear model for the mean would give *negative* values at some explanatory variable values, because negative values are not permitted with a gamma distribution.

Example 14.6

Gamma GLM for House Selling Price   The least squares fit of the model to the data on $y =$ selling price using explanatory variables size of home, taxes, and whether new, discussed in Example 14.1 (page 433), is

$$\hat{y} = -21{,}353.8 + 61.7(\text{SIZE}) + 46{,}373.7(\text{NEW}) + 37.2(\text{TAXES}).$$

However, Example 14.3 (page 439) showed that two outlying observations had a substantial effect on the estimated effect of NEW. Figure 14.4 showed that the variability in selling prices seems to increase as its mean does. This suggests that a model assuming a gamma distribution may be more appropriate, because the gamma permits the standard deviation to increase as the mean does.

We can use software, as explained in Appendix A, to fit GLMs that assume a gamma distribution for y. For these data, we obtain

$$\hat{y} = -940.9 + 48.7(\text{SIZE}) + 32{,}868.0(\text{NEW}) + 37.9(\text{TAXES}).$$

The estimated effect of TAXES is similar as with least squares, but the estimated effect of SIZE is weaker and the estimated effect of NEW is much weaker. Moreover, the effect of NEW is no longer significant, as the ratio of the estimate to the standard error is 1.5. This result is similar to what we obtained in Example 14.4 (page 443) after deleting observation 6, an outlier corresponding to a large new house with an unusually high selling price. The outliers are not as influential for the gamma fit, because that model expects more variability in the data when the mean is larger.

The estimate of the scale parameter is $\hat{\phi} = 0.07$. The estimated standard deviation $\hat{\sigma}$ of the conditional distribution of y relates to the estimated conditional mean $\hat{\mu}$ by

$$\hat{\sigma} = \sqrt{\hat{\phi}}\hat{\mu} = \sqrt{0.07}\hat{\mu} = 0.27\hat{\mu}.$$

For example, at explanatory variable values such that the estimated mean selling price is $\hat{\mu} = \$100{,}000$, the estimated standard deviation of selling prices is $\hat{\sigma} = 0.27(\$100{,}000) = \$27{,}000$. By contrast, at explanatory variable values such that $\hat{\mu} = \$400{,}000$, $\hat{\sigma} = 0.27(\$400{,}000) = \$108{,}000$, four times as large. ∎

---

[2] R, SPSS, and Stata estimate the scale parameter, whereas SAS estimates the shape parameter.

14.5 Nonlinear Relationships: Polynomial Regression

The ordinary multiple regression model assumes that the relationship between the mean of y and each quantitative explanatory variable is linear, controlling for other explanatory variables. Although social science relationships are not *exactly* linear, the degree of nonlinearity is often so minor that they can be reasonably well approximated with linearity. Occasionally, though, such a model is inadequate, even for approximation. A scatterplot may reveal a highly nonlinear relationship. Or, you might expect a nonlinear relationship because of the nature of the variables. For example, you might expect $y =$ medical expenses to have a curvilinear relationship with $x =$ age, being relatively high for the very young and the very old but lower for older children and young adults (Figure 14.5a). The relationship between $x =$ per capita income and $y =$ life expectancy for a sample of countries might be approximately a linearly increasing one, up to a certain point. However, beyond a certain level, additional income would probably result in little, if any, improvement in life expectancy (Figure 14.5b).

FIGURE 14.5: Two Nonlinear Relationships

(a) Age   (b) Per Capita GNP
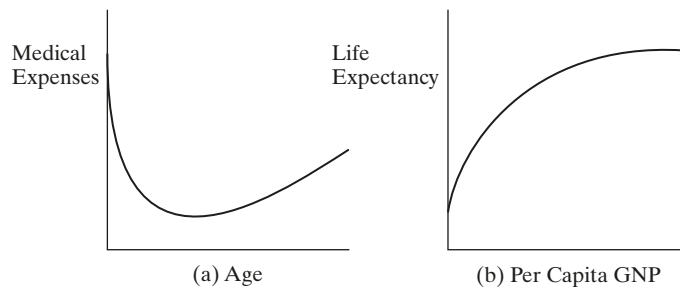
If we use straight-line regression to describe a curvilinear relationship, what can go wrong? Measures of association designed for linearity, such as the correlation, may underestimate the true association. Estimates of the mean of y at various x-values may be badly biased, since the prediction line may poorly approximate the true regression curve. Two approaches are common to deal with nonlinearity. The first approach, presented in this section, uses a *polynomial* regression function. The class of polynomial functions includes a diverse set of functional patterns, including straight lines. The second approach, presented in Section 14.6, uses a generalized linear model with a link function such as the logarithm. For example, for certain curvilinear relationships, the logarithm of the mean of the response variable is linearly related to the explanatory variables.

QUADRATIC REGRESSION MODELS

A ***polynomial regression function*** for a response variable y and single explanatory variable x has the form

$$E(y) = \alpha + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p.$$

In this model, x occurs in powers from the first $(x = x^1)$ to some integer p. For $p = 1$, this is the straight line $E(y) = \alpha + \beta_1 x$. The index p, the highest power in the polynomial equation, is called the ***degree*** of the polynomial function.

The polynomial function most commonly used for nonlinear relationships is the *second-degree polynomial*

$$E(y) = \alpha + \beta_1 x + \beta_2 x^2.$$

This is called a *quadratic regression model*. The graph of this function is parabolic, as Figure 14.6 portrays. This shape is limited in scope for applications, because it is symmetric about a vertical axis. That is, its appearance when increasing is a mirror image of its appearance when decreasing.

FIGURE 14.6: Graphs of Two Second-Degree Polynomials (Quadratic Functions) and a Third-Degree Polynomial (Cubic Function)
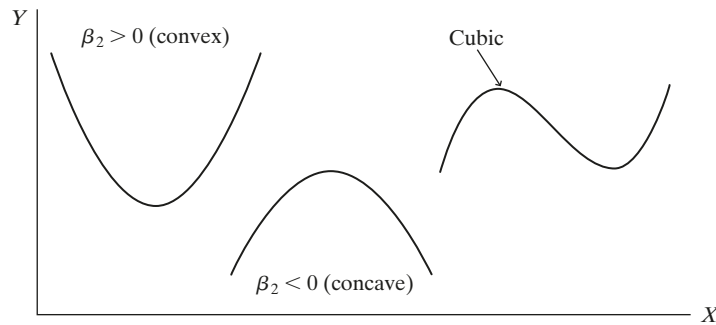
If a scatterplot reveals a pattern of points with one bend, then a second-degree polynomial usually improves upon the straight-line fit. A third-degree polynomial $E(y) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$, called a *cubic function*, is a curvilinear function having *two* bends. See Figure 14.6. But it is rarely necessary to use higher than a second-degree polynomial to describe the trend.

INTERPRETING AND FITTING QUADRATIC REGRESSION MODELS

The quadratic regression model $E(y) = \alpha + \beta_1 x + \beta_2 x^2$, plotted for the possible values of α, β_1, and β_2, describes the possible parabolic shapes. Unlike straight lines, for which the slope remains constant over all x-values, the mean change in y for a one-unit increase in x *depends on the value of x*. For example, as the value of x increases, a straight line drawn tangent to the parabola in Figure 14.7 first has positive slope, then zero slope where the parabola achieves its maximum value, and then negative slope. The rate of change of the line varies to produce a curve having a smooth bend.

FIGURE 14.7: Interpretation of Parameters of the Quadratic Model $E(y) = \alpha + \beta_1 x + \beta_2 x^2$
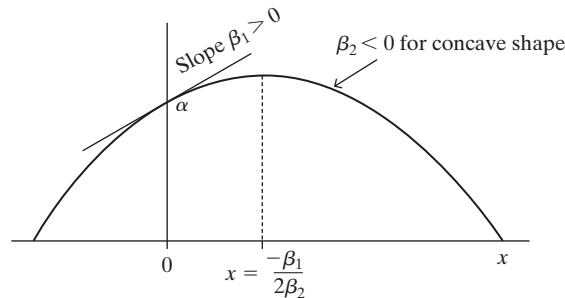
The sign of the coefficient β_2 of the x^2 term determines whether the function is bowl shaped (opens up) relative to the x-axis or mound shaped (opens down). Bowl-shaped functions, also called *convex* functions, have $\beta_2 > 0$. Mound-shaped functions, also called *concave* functions, have $\beta_2 < 0$. See Figure 14.6.

As usual, the coefficient α is the y-intercept. The coefficient β_1 of x is the slope of the line that is tangent to the parabola as it crosses the y-axis. If $\beta_1 > 0$, for example, then the parabola is sloping upward at $x = 0$ (as Figure 14.7 shows). At the point at which the slope is zero, the relationship changes direction from positive to negative

or from negative to positive. This happens at $x = -\beta_1/2\beta_2$. This is the x-value at which the mean of y takes its maximum if the parabola is mound shaped and its minimum if it is bowl shaped.

To fit quadratic regression models, we treat them as a special case of the multiple regression model

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 = \alpha + \beta_1 x + \beta_2 x^2$$

with two explanatory variables. We identify x_1 with the explanatory variable x and x_2 with its square, x^2. The data for the model fit consist of the y-values for the subjects in the sample, the x-values (called x_1), and an artificial variable (x_2) consisting of the squares of the x-values. Software can create these squared values for us. It then uses least squares to find the best-fitting function out of the class of all second-degree polynomials.

Example 14.7

Fertility Predicted Using Gross Domestic Product (GDP)   Table 14.6 shows values reported by the United Nations for several nations on y = fertility rate (the mean number of children per adult woman) and x = per capita gross domestic product (GDP, in tens of thousands of dollars). Fertility tends to decrease as GDP increases. However, a straight-line model may be inadequate, since it might predict negative fertility for sufficiently high GDP. In addition, some demographers predict that after GDP passes a certain level, fertility rate may increase, since the nation's wealth makes it easier for a parent to stay home and take care of children rather than work.

TABLE 14.6: Data on Fertility Rate and Per Capita Gross Domestic Product GDP (`FertilityGDP` Data File at the Text Website)

Nation	GDP	Fertility Rate	Nation	GDP	Fertility Rate	Nation	GDP	Fertility Rate
Algeria	0.21	2.5	Germany	2.91	1.3	Pakistan	0.06	4.3
Argentina	0.35	2.4	Greece	1.56	1.3	Philippines	0.10	3.2
Australia	2.63	1.7	India	0.06	3.1	Russia	0.30	1.3
Austria	3.13	1.4	Iran	0.21	2.1	S. Africa	0.35	2.8
Belgium	2.91	1.7	Ireland	3.85	1.9	Saudi Ar.	0.95	4.1
Brazil	0.28	2.3	Israel	1.65	2.9	Spain	2.04	1.3
Canada	2.71	1.5	Japan	3.37	1.3	Sweden	3.37	1.6
Chile	0.46	2.0	Malaysia	0.42	2.9	Switzerland	4.36	1.4
China	0.11	1.7	Mexico	0.61	2.4	Turkey	0.34	2.5
Denmark	3.93	1.8	Netherlands	3.15	1.7	UK	3.03	1.7
Egypt	0.12	3.3	New Zealand	1.98	2.0	United States	3.76	2.0
Finland	3.11	1.7	Nigeria	0.04	5.8	Vietnam	0.05	2.3
France	2.94	1.9	Norway	4.84	1.8	Yemen	0.06	6.2

Source: hdr.undp.org/en/data.

Figure 14.8, a scatterplot for the 39 observations, shows a clear decreasing trend. The linear prediction equation is $\hat{y} = 3.04 - 0.415x$, and the correlation equals -0.56. This prediction equation gives absurd predictions for very large x-values; $\hat{y}$ is negative for $x > 7.3$ (i.e., \$73,000). However, the predicted values are positive over the range of x-values for this sample. To allow for potential nonlinearity and for the possibility that fertility rate may increase for sufficiently large GDP, we could fit a quadratic regression model to these data. ∎

FIGURE 14.8: Scatterplot and Best-Fitting Straight-Line Model and Quadratic Model for Data on Fertility Rate and Per Capita GDP from the `FertilityGDP` Data File
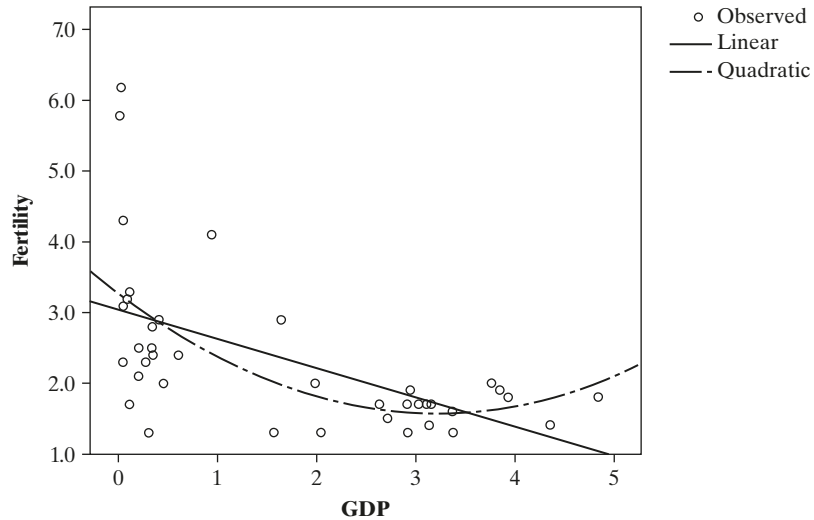
Table 14.7 shows some output for the quadratic regression of y = fertility rate on x = GDP. Here, GDP2 denotes an artificial variable constructed as the square of GDP. The prediction equation is

$$\hat{y} = 3.28 - 1.054x + 0.163x^2.$$

Figure 14.8 plots the linear and quadratic prediction equations in the scatter diagram.

TABLE 14.7: Some Output for Quadratic Regression Model for y = Fertility Rate and x = GDP from the `FertilityGDP` Data File

Variable	Coef.	Std. Error	t	Sig
INTERCEP	3.278	.257	12.750	.000
GDP	-1.054	0.366	-2.880	.007
GDP2	.163	0.090	1.810	.079
R-square	0.375			

A bowl-shaped quadratic equation takes its minimum at $x = -\beta_1/2\beta_2$. For these data, we estimate this point to be $x = 1.054/2(0.163) = 3.23$. The predicted fertility rate increases as GDP increases above this point (i.e., \$32,300).

DESCRIPTION AND INFERENCE ABOUT THE NONLINEAR EFFECT

For a polynomial model, R^2 for multiple regression describes the strength of the association. In this context, it describes the proportional reduction in error obtained from using the quadratic model, instead of $\bar{y}$, to predict y. Comparing this measure to r^2 for the straight-line model indicates how much better a fit the curvilinear model provides. Since a polynomial model has additional terms besides x, R^2 always is at least as large as r^2. The difference $R^2 - r^2$ measures the additional reduction in prediction error obtained by using the polynomial instead of the straight line.

For Table 14.6, the best-fitting straight-line prediction equation has $r^2 = 0.318$. From Table 14.7 for the quadratic model, $R^2 = 0.375$. The best quadratic equation explains about 6% more variability in y than does the best-fitting straight-line equation.

If $\beta_2 = 0$, the quadratic regression equation $E(y) = \alpha + \beta_1 x + \beta_2 x^2$ simplifies to the linear regression equation $E(y) = \alpha + \beta_1 x$. Therefore, to test the null hypothesis that the relationship is linear against the alternative that it is quadratic, we test H_0: $\beta_2 = 0$. The usual t test for a regression coefficient does this, dividing the estimate of β_2 by its standard error. The assumptions for applying inference are the same as for ordinary regression: randomization for gathering the data, a conditional distribution of y-values that is normal about the mean, with constant standard deviation σ at all x-values. The set of nations in Table 14.6 is not a random sample of nations, so inference is not relevant for those data.

CAUTIONS IN USING POLYNOMIAL MODELS

Some cautions are in order before you take the conclusions in this example too seriously. The scatterplot (Figure 14.8) suggests that the variability in fertility rates is considerably higher for nations with low GDPs than it is for nations with high GDPs. The fertility rates show much greater variability when their mean is higher. A GLM that permits nonconstant standard deviation by assuming a gamma distribution for y (see page 449) provides somewhat different results, including stronger evidence of nonlinearity (Exercise 14.14).

In fact, before we conclude that fertility rate increases above a certain value, we should realize that other models for which this does not happen are also consistent with these data. For instance, Figure 14.8 suggests that a "piecewise linear" model that has a linear decrease until GDP is about $25,000 and then a separate, nearly horizontal, line beyond that point fits quite well. A more satisfactory model for these data is one discussed in the next section of this chapter for *exponential regression*. Unless a data set is very large, several models may be consistent with the data.

In examining scatterplots, be cautious not to read too much into the data. Don't let one or two outliers suggest a curve in the trend. Good model building follows the principle of ***parsimony***: Models should have no more parameters than necessary to represent the relationship adequately. One reason is that simple models are easier to understand and interpret than complex ones. Another reason is that when a model contains unnecessary variables, the standard errors of the estimates of the regression coefficients tend to inflate, hindering efforts at making precise inferences. Estimates of the conditional mean of y also tend to be poorer than those obtained with well-fitting simple models.

When a polynomial regression model is valid, the regression coefficients do not have the partial slope interpretation usual for coefficients of multiple regression models. It does not make sense to refer to the change in the mean of y when x^2 is increased one unit and x is held constant. Similarly, it does not make sense to interpret the partial correlations $r_{yx^2 \cdot x}$ or $r_{yx \cdot x^2}$ as measures of association, controlling for x or x^2. However, the coefficient $r^2_{yx^2 \cdot x}$ does measure the proportion of the variation in y unaccounted for by the straight-line model that is explained by the quadratic model. In Example 14.7, applying the formula for $r^2_{yx_2 \cdot x_1}$ from page 344 yields

$$r^2_{yx^2 \cdot x} = \frac{R^2 - r^2_{yx}}{1 - r^2_{yx}} = \frac{0.375 - 0.318}{1 - 0.318} = 0.08.$$

Of the variation in y unexplained by the linear model, about 8% is explained by the introduction of the quadratic term.

With multiple explanatory variables, we may find that the fit improves by permitting one or more of them to have quadratic effects. For example, the model

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$$

allows nonlinearity in x_2. For fixed x_1, the mean of y is a quadratic function of x_2. For fixed x_2, the mean of y is a linear function of x_1 with slope β_1. This model is a special case of multiple regression with three explanatory variables, in which x_3 is the square of x_2. Models allowing both nonlinearity and interaction are also possible.

14.6 Exponential Regression and Log Transforms*

Although polynomials provide a diverse collection of functions for modeling nonlinearity, other mathematical functions are often more appropriate. The most important case is when the mean of the response variable is an *exponential* function of the explanatory variable.

Exponential Regression Function

> An *exponential regression* function has the form $E(y) = \alpha \beta^x$.

In this equation, the explanatory variable appears as the exponent of a parameter. Unlike a quadratic function, an exponential function can take only positive values, and it continually increases (if $\beta > 1$) or continually decreases (if $\beta < 1$). In either case, it has a convex shape, as Figure 14.9 shows. We provide interpretations for the model parameters later in this section.

FIGURE 14.9: The Exponential Regression Function $E(y) = \alpha \beta^x$
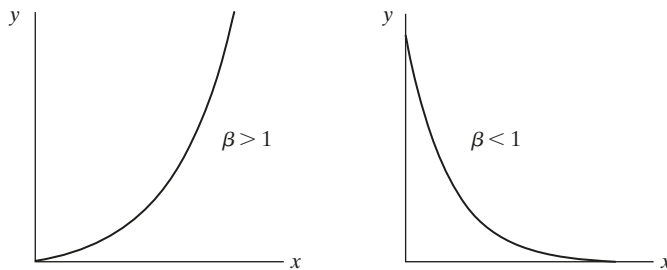
For the exponential regression function, the *logarithm* of the mean is linearly related to the explanatory variable. That is, if $\mu = E(y) = \alpha \beta^x$, then

$$\log(\mu) = \log \alpha + (\log \beta)x.$$

The right-hand side of this equation has the straight-line form $\alpha' + \beta'x$ with intercept $\alpha' = \log(\alpha)$, the log of the α parameter, and slope $\beta' = \log(\beta)$, the log of the β parameter for the exponential regression function. This model form is the special case of a generalized linear model (GLM) using the log link function. If the model holds, a plot of the log of the y-values should show approximately a linear relation with x. (Don't worry if you have forgotten your high school math about logarithms. You will not need to know this in order to understand how to fit or interpret the exponential regression model.)

You can use GLM software to estimate the parameters in the model $\log[E(y)] = \alpha' + \beta'x$. The antilogs of these estimates are the estimates for the parameters in the exponential regression model $E(y) = \alpha \beta^x$, as shown below.

Example 14.8

Exponential Population Growth   Exponential regression models well the growth of some populations over time. If the rate of growth remains constant, in percentage terms, then the size of that population grows exponentially fast. Suppose that the population size at some fixed time is α and the growth rate is 2% per year. After one year, the population is 2% larger than that at the beginning of the year. This means that the population size grows by a multiplicative factor of 1.02 each year. The population size after one year is $\alpha(1.02)$. Similarly, the population size after two years is

$$\text{(Population size at the end of one year)}(1.02) = [\alpha(1.02)]1.02 = \alpha(1.02)^2.$$

After three years, the population size is $\alpha(1.02)^3$. After x years, the population size is $\alpha(1.02)^x$. The population size after x years follows an exponential function $\alpha\beta^x$ with parameters given by the initial population size α and the rate of growth factor, $\beta = 1.02$, corresponding to 2% growth.

Table 14.8 shows the U.S. population size (in millions) at 10-year intervals beginning in 1890. Figure 14.10 plots these values over time. Table 14.8 also shows the natural logarithm of the population sizes. (This uses the base e, where $e = 2.718\dots$ is an irrational number that appears often in mathematics. The model makes sense with logs to any base, but software fits the GLM using natural logs, often denoted by $\log_e$ or LN.)

TABLE 14.8: U.S. Population Sizes and Log Population Sizes by Decade from 1890 to 2010, with Predicted Values for Exponential Regression Model

Year	No. Decades Since 1890 x	Population Size y	$\log_e(y)$	$\hat{y}$
1890	0	62.95	4.14	73.2
1900	1	75.99	4.33	82.7
1910	2	91.97	4.52	93.5
1920	3	105.71	4.66	105.6
1930	4	122.78	4.81	119.4
1940	5	131.67	4.88	134.9
1950	6	151.33	5.02	152.4
1960	7	179.32	5.19	172.3
1970	8	203.30	5.31	194.7
1980	9	226.54	5.42	220.0
1990	10	248.71	5.52	248.7
2000	11	281.42	5.64	281.0
2010	12	308.75	5.73	317.5

Source: U.S. Census Bureau; data file `Population` at the text website.

Figure 14.11 plots these log of population size values over time. The log population sizes appear to grow approximately linearly. This suggests that population growth over this time period was approximately exponential, with a constant rate of growth. We now estimate the regression curve, treating time as the explanatory variable x. For convenience, we identify the time points 1890, 1900, $\dots$, 2010 as times $0, 1, \dots, 12$; that is, x represents the number of decades since 1890.

We use software to estimate the generalized linear model $\log(\mu) = \alpha' + \beta'x$, assuming a normal distribution for y. The prediction equation is

$$\log_e(\hat{\mu}) = 4.29285 + 0.12233x.$$

FIGURE 14.10: U.S.
Population Size Since 1890.
The fitted curve is the
exponential regression,
$\hat{y} = 73.2(1.1301)^x$.
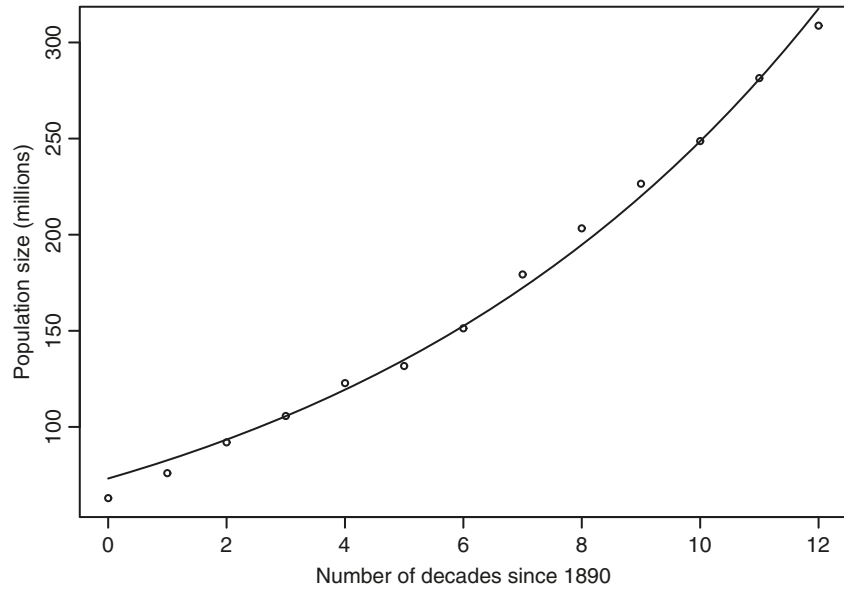
FIGURE 14.11: Log
Population Sizes Since
1890. The prediction
equation is
$\widehat{\log y} = 4.29 + 0.122x$.
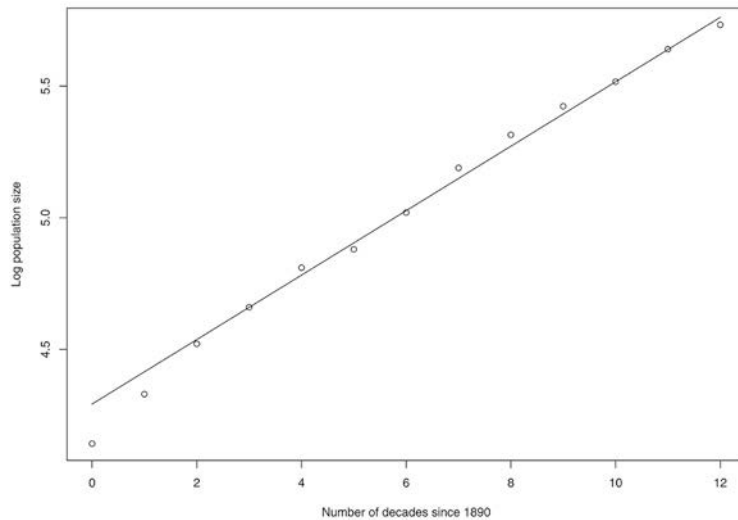
Antilogs of these estimates are the parameter estimates for the exponential regression model. For natural logs, the antilog function is the exponential function e^x. That is, antilog(4.29285) $= e^{4.29285} = 73.175$, and antilog(0.12233) $= e^{0.12233} = 1.1301$. Thus, for the exponential regression model $E(y) = \alpha\beta^x$, the estimates are $\hat{\alpha} = 73.175$ and $\hat{\beta} = 1.1301$. The prediction equation is

$$\hat{y} = \hat{\alpha}\hat{\beta}^x = 73.175(1.1301)^x.$$

The predicted initial population size (in 1890) is $\hat{\alpha} = 73.2$ million. The predicted population size x decades after 1890 is $\hat{y} = 73.175(1.1301)^x$. For 2010, for instance, $x = 12$, and the predicted population size is $\hat{y} = 73.175(1.1301)^{12} = 317.5$ million. Table 14.8 shows the predicted values for each decade. Figure 14.10 plots the exponential prediction equation.

The predictions are quite good, except for the first couple of observations. The total sum of squares of population size values about their mean is TSS = 76,791, whereas the sum of squared errors about the prediction equation is SSE = 419. The proportional reduction in error is $(76,791 - 419)/76,791 = 0.995$. The ordinary linear model $E(y) = \alpha + \beta x$ also fits quite well, having $r^2 = 0.980$. ∎

A caution: The fit of the model $\log[E(y)] = \alpha' + \beta'x$ that you get with GLM software will *not* be the same as you get by taking logarithms of all the y-values and then fitting a straight-line model using least squares. The latter approach[3] gives the fit for the model $E[\log(y)] = \alpha' + \beta'x$. For that model, taking antilogs does not take you back to $E(y)$, because $E[\log(y)]$ is not equivalent to $\log[E(y)]$. So, in software it is preferable to use a generalized linear modeling option rather than an ordinary regression option.

INTERPRETING EXPONENTIAL REGRESSION MODELS

Now let's take a closer look at how to interpret parameters in the exponential regression model, $E(y) = \alpha\beta^x$. The parameter α represents the mean of y when $x = 0$. The parameter β represents the ***multiplicative*** change in the mean of y for a one-unit increase in x. The mean of y at $x = 12$ equals β *multiplied* by the mean of y at $x = 11$. For instance, for the equation $\hat{y} = 73.175(1.1301)^x$, the predicted population size at a particular date equals 1.1301 times the predicted population size a decade earlier.

By contrast, the parameter β in the *linear* model $E(y) = \alpha + \beta x$ represents the ***additive*** change in the mean of y for a one-unit increase in x. In the linear model, the mean of y at $x = 12$ is β *plus* the mean of y at $x = 11$. The prediction equation for the linear model (i.e., identity link) fitted to Table 14.8 is $\hat{y} = 46.51 + 20.33x$. This model predicts that the population size increases by 20.33 million people every decade.

In summary, for the linear model, $E(y)$ changes by the same *quantity* for each one-unit increase in x, whereas for the exponential model, $E(y)$ changes by the same *percentage* for each one-unit increase. For the exponential regression model with Table 14.8, the multiplicative effect of 1.1301 for each decade corresponds to a predicted 13.01% growth per decade.

Suppose the growth rate is 15% per decade, to choose a rounder number. This corresponds to a multiplicative factor of 1.15. After five decades, the population grows by a factor of $(1.15)^5 = 2.0$. That is, after five decades, the population size doubles. If the rate of growth remained constant at 15% per decade, the population would double every 50 years. After 100 years, the population size would be quadruple the original size, after 150 years it would be 8 times as large, after 200 years it would be 16 times its original size, and so forth.

The exponential function with $\beta > 1$ has the property that its doubling time is a constant. As can be seen from the sequence of population sizes at 50-year intervals, this is an extremely fast increase even though the annual rate of growth (1.4% annually for a decade increase of 15%) seems small. In fact, this has been the approximate growth of the world population in the past century. (See Exercise 14.22.)

Example 14.9

Exponential Regression for Fertility Data   When $\beta < 1$ in the exponential regression model, $\beta' = \log(\beta) < 0$ in the log-transformed GLM. In this case, the mean of y *decreases* exponentially fast as x increases. The curve then looks like the second curve in Figure 14.9.

---

[3] For example, as SPSS gives by selecting *Regression* in the *Analyze* menu, followed by the choice of *Curve Estimation* with the *Exponential* option.

In Example 14.7 with Table 14.6 (page 453), we modeled $y = $ fertility rate for several countries, with $x = $ per capita GDP. If we expect $E(y)$ to continually decrease as x increases, an exponentially decreasing curve may be more appropriate. In fact, the exponential regression model provides a good fit for those data. Using the GLM with log link for $y = $ fertility rate and $x = $ per capita GDP and assuming a normal distribution for y, we get the prediction equation

$$\log_e(\hat{\mu}) = 1.148 - 0.206x.$$

Taking antilogs yields the exponential prediction equation

$$\hat{y} = \hat{\alpha}\hat{\beta}^x = e^{1.148}(e^{-0.206})^x = 3.15(0.81)^x.$$

The predicted fertility rate at GDP value $x + 1$ equals 81% of the predicted fertility rate at GDP value x; that is, it decreases by 19% for a \$10,000 increase in per capita GDP.

With this fit, the correlation between the observed and predicted fertility rates equals 0.59, nearly as high as the value of 0.61 achieved with the quadratic model, which has an extra parameter. If we expect fertility rate to decrease continuously as GDP increases, the exponential regression model is a more realistic model than the quadratic regression model, which predicted increasing fertility above a certain GDP level. Also, unlike the straight-line model, the exponential regression model cannot yield negative predicted fertility rates.

Since the scatterplot in Figure 14.8 suggests greater variability when the mean fertility rate is higher, it may be even better to assume a gamma distribution for y with this exponential regression model. The prediction equation is then

$$\log_e(\hat{\mu}) = 1.112 - 0.177x, \quad \text{for which} \quad \hat{y} = e^{1.112}(e^{-0.177})^x = 3.04(0.84)^x.$$

This gives a slightly shallower rate of decrease than the fit $3.15(0.81)^x$ for the normal response model. ∎

TRANSFORMING THE EXPLANATORY VARIABLE TO ACHIEVE LINEARITY

Other transformations of the response mean or of explanatory variables are useful in some situations. For example, suppose y tends to increase or decrease over a certain range of x-values, but once a certain x-value has been reached, further increases in x have less effect on y, as in Figure 14.5b. For this concave increasing type of trend, x behaves like an exponential function of y. Taking the logarithms of the x-values often linearizes the relationship. Another possible transform for this case uses $1/x$ as the explanatory variable.

14.7 Robust Variances and Nonparametric Regression*

Recent years have seen yet other ways developed to generalize regression to handle violations of assumptions for the ordinary linear model. Detailed explanations of such generalizations are beyond the scope of this book, but in this section we briefly introduce two popular ones.

ROBUST VARIANCE ESTIMATES

The ordinary regression model assumes a normal distribution for y with constant variability at all settings of the explanatory variables. Section 14.4 introduced the

generalized linear model, which permits alternative distributions that have nonconstant variability, such as the gamma distribution. An alternative approach uses the least squares estimates but does not assume constant variance in finding standard errors. Instead, it adjusts ordinary standard error formulas to reflect the empirical variability displayed by the sample data.

This alternative standard error estimate is sometimes called the ***sandwich estimate***, because of how its formula sandwiches the empirical variability between two terms from the ordinary formula. It is also referred to as a ***robust standard error estimate***, because it is more valid than the ordinary *se* when the true response variability is not constant. Some software[4] now makes this available. If you use it and find standard errors quite different from those given in an ordinary regression analysis, basic assumptions are likely violated and you should treat its results skeptically.

To illustrate, we found robust standard errors for the house selling price data analyzed on page 450 with least squares and with a gamma GLM. For the effects of (size, new, taxes), the robust *se* values are (22.4, 26245, 9.3), compared to (12.5, 16459, 6.7) for the ordinary *se* values. Such highly different results make us wary of the ordinary *se* values. As explained previously, the clear increase in the variability of $y =$ selling price as its mean increases made us skeptical of the ordinary regression results.

This robust variance approach also extends to handle violations of the assumption of independent observations, such as those that occur with clustered data and longitudinal studies. This approach incorporates the empirical variability and correlation within clusters to generate standard errors that are more reliable than ones that treat observations within clusters as independent. This method for clustered correlated data uses ***generalized estimating equations*** (GEEs) that resemble equations used to obtain maximum likelihood estimates, but without a parametric probability distribution incorporating correlations.

This way of handling clustered data is an alternative to the *linear mixed model* introduced in Section 13.5. The linear mixed model assumes normality for the response variable and adds *random effects* to an ordinary model. Likewise, we can add random effects to a generalized linear model to obtain a *generalized linear mixed model* to handle clustering with nonnormal responses such as the binomial. The robust variance approach has the advantage of not requiring an assumption about the distribution of *y* or the correlation structure within clusters. However, it has the disadvantage that (because of the lack of a distribution assumption) likelihood-based methods such as maximum likelihood estimates and likelihood ratio tests are not available.

NONPARAMETRIC REGRESSION

Recent advances make it possible to fit models to data without assuming particular functional forms, such as straight lines or parabolas, for the relationship. These approaches are *nonparametric*, in terms of having fewer (if any) assumptions about the functional form and the distribution of *y*. It is helpful to look at a plot of a fitted nonparametric regression model to learn about trends in the data.

One nonparametric regression method, called ***generalized additive modeling***, is a further generalization of the generalized linear model. It has the form

$$g(\mu) = f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p),$$

where $f_1, \ldots, f_p$ are unspecified and potentially highly complex functions. The GLM is the special case in which each of these functions is linear. The estimated functional

---

[4] For example, Stata with the *robust* option for its *regress* command.

form of the relationship for each explanatory variable is determined by a computer algorithm, using the sample data. As in GLMs, with this model you can select a particular link function g and also a distribution for y. This model is useful for smoothing data to reveal overall trends.

Other nonparametric smoothing methods do not even require selecting a link function or a distribution for y. Popular smoothers are **LOESS** and **kernel** methods that get the prediction at a particular point by smoothly averaging nearby values. The smoothed value is found by fitting a low-degree polynomial while giving more weight to observations near the point and less weight to observations further away. You can achieve greater smoothing by choosing a larger *bandwidth*, essentially by letting the weights die out more gradually as you move away from each given point.

Figure 14.12 shows two plots of nonparametric regression fits for the fertility rate data of Table 14.6. The first plot employs greater smoothing and has a curved, decreasing trend. It is evident that the response may not eventually increase, as a quadratic model predicts. This fit suggests that the exponential regression model is more satisfactory than the quadratic model for these data.

To learn more about robust regression and nonparametric regression, see Fox (2015, Chapters 18 and 19).

FIGURE 14.12: Fits of Nonparametric Regression Model (Using SPSS) to Smooth the Fertility Rate Data of Table 14.6. Fit (a) employs greater smoothing (bandwidth = 5) than fit (b) (bandwidth = 1).
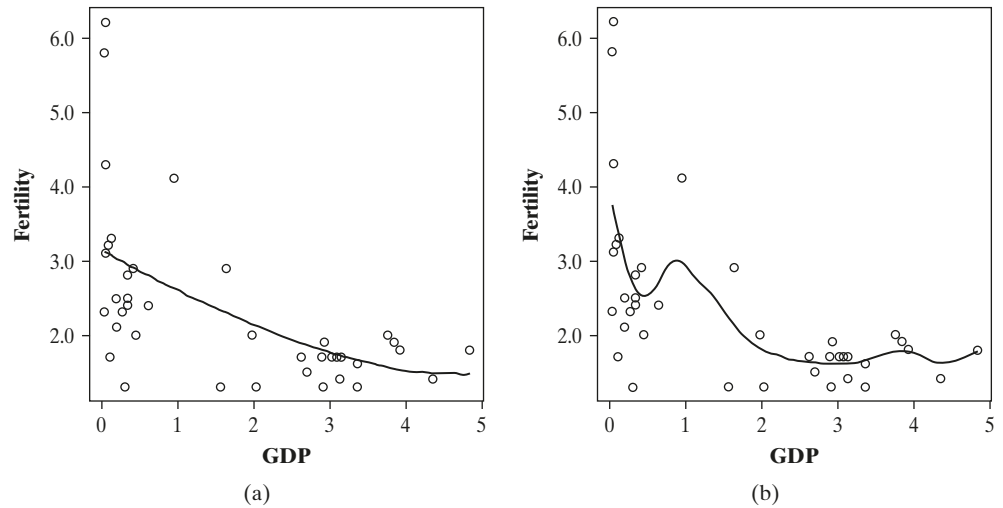
14.8 Chapter Summary

This chapter discussed issues about building regression models and showed how to check assumptions and how to ease some restrictions of the basic linear model.

- With a large number of potential explanatory variables for a model, the **backward elimination** and **forward selection** procedures use a sequential algorithm to select variables. These are exploratory in purpose and should be used with caution. Fit indices such as adjusted R^2, PRESS, and AIC also provide criteria for model selection.

- Plots of the **residuals** check whether the model is adequate and whether the assumptions for inferences are reasonable. Observations having a large leverage and large studentized residual have a strong influence on the model fit. Diagnostics such as DFBETA and DFFIT describe which observations have a strong influence on the parameter estimates and the model fit.

- *Multicollinearity*, the condition by which the set of explanatory variables contains some redundancies, causes inflation of standard errors of estimated regression coefficients and makes it difficult to evaluate partial effects.

- *Generalized linear models* allow the response variable to have a distribution other than the normal, such as the binomial for binary data and the gamma for positive responses having greater variation at greater mean values. Such models permit modeling a function of the mean, called the *link function*.

- *Nonlinear* relationships are modeled through the use of *polynomial* (particularly *quadratic*) functions and *exponential* functions. Quadratic functions have a parabolic appearance, whereas exponential functions have a convex increasing or convex decreasing appearance. The *exponential regression model* is a generalized linear model for the log of the mean.

Exercises

Practicing the Basics

14.1. For Example 11.2 (page 324) on y = mental impairment, x_1 = life events, and x_2 = SES, the multiple regression model has output

	Coef.	Std. Error	t	Sig.
(Constant)	28.230	2.174	12.984	.000
LIFE	.103	.032	3.177	.003
SES	−.097	.029	−3.351	.002

and the model allowing interaction has output

	Coef.	Std. Error	t	Sig
(Constant)	26.036649	3.948826	6.594	0.0001
LIFE	0.155865	0.085338	1.826	0.0761
SES	−0.060493	0.062675	−0.965	0.3409
LIFE*SES	−0.000866	0.001297	−0.668	0.5087

SES had a P-value of 0.011 in the bivariate model containing only that explanatory variable, and LIFE had a P-value of 0.018 in the bivariate model containing only that explanatory variable. Select explanatory variables from the set $x_1, x_2, x_3 = x_1x_2$, with $\alpha = 0.05$,

(a) Using backward elimination.

(b) Using forward selection.

14.2. Table 11.23 (page 359) showed results of a multiple regression using nine predictors of the quality of life in a country.

(a) In backward elimination with these nine predictors, can you predict which variable would be deleted (i) first? (ii) second? Explain.

(b) In forward selection with these nine predictors, can you predict which variable would be added first? Explain.

14.3. For the `Houses2` data file at the text website, Table 14.9 shows a correlation matrix and a model fit using four predictors of selling price. With these four predictors,

(a) For backward elimination, which variable would be deleted first? Why?

(b) For forward selection, which variable would be added first? Why?

(c) Why do you think that BEDS has such a large P-value in the multiple regression model, even though it has a substantial correlation with PRICE?

TABLE 14.9

	Correlation coefficients				
	price	size	beds	baths	new
price	1.000	0.899	0.590	0.714	0.357
size	0.899	1.000	0.669	0.662	0.176
beds	0.590	0.669	1.000	0.334	0.267
baths	0.714	0.662	0.334	1.000	0.182
new	0.357	0.176	0.267	0.182	1.000

Variable	Estimate	Std. Error	t	Sig
INTERCEP	−41.795	12.104	3.45	0.0009
SIZE	64.761	5.630	11.50	0.0001
BEDS	−2.766	3.960	0.70	0.4868
BATHS	19.203	5.650	3.40	0.0010
NEW	18.984	3.873	4.90	0.0001

14.4. Refer to the previous exercise. Using software with these four predictors, find the model that would be selected using the criterion. **(a)** R^2_{adj}, **(b)** PRESS, **(c)** AIC.

14.5. Use software with the `Crime2` data file at the text website, excluding the observation for D.C. Let y = murder rate. For the five explanatory variables in that data file (excluding violent crime rate), with $\alpha = 0.10$ in tests,

(a) Use backward elimination to select a model. Interpret the result.

(b) Use forward selection to select a model. Interpret the result.

(c) Use stepwise regression. Interpret the result.

FIGURE 14.13
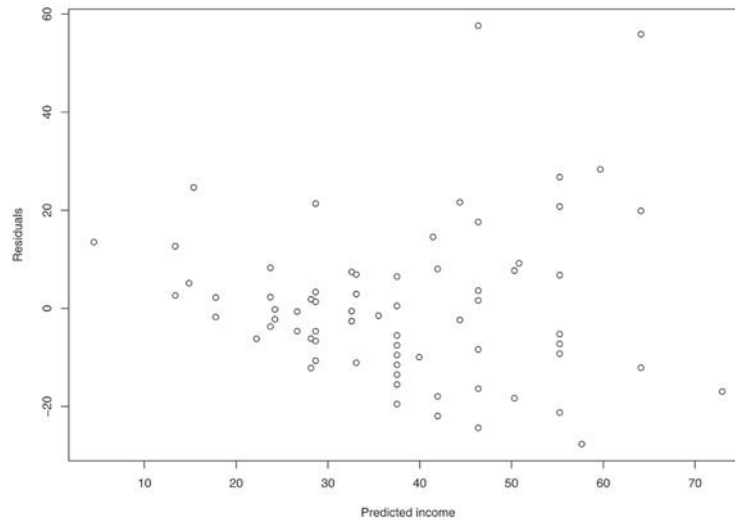
(d) Compare results of the three selection procedures. How is it possible that a variable (percentage with a high school education) can be the first variable dropped in (a) yet the second added in (b)?

(e) Now include the D.C. observation. Repeat (a) and (b), and compare to results excluding D.C. What does this suggest about the influence outliers can have on automatic selection procedures?

14.6. Figure 14.13 is a plot of the residuals versus the predicted y-values for the model discussed in Example 13.1 (page 402) relating income to education and racial–ethnic group. What does this plot suggest?

14.7. For the data for 21 nations in the UN2 data file at the text website that are not missing observations on literacy, Table 14.10 shows various diagnostics from fitting the multiple regression model relating fertility (mean number of births per woman) to literacy rate and women's economic activity.

(a) Study the studentized residuals. Are there any apparent outliers?

TABLE 14.10

Obs	Residual	Studentized Residual	Leverage h	DFFIT	DFBETA Fem_econ	Literacy
1	-1.1374	-1.3088	0.0935	-0.4204	0.1989	0.0726
2	0.2782	0.3216	0.1792	0.1503	-0.1001	0.1226
6	-0.1299	-0.1424	0.0915	-0.0452	0.0235	-0.0302
8	-0.1695	-0.1921	0.1490	-0.0804	0.0496	-0.0640
9	-0.5515	-0.6682	0.2378	-0.3732	-0.3215	0.1017
11	-0.9491	-1.1198	0.1589	-0.4868	-0.0620	0.3707
15	-1.0803	-1.2174	0.0665	-0.3249	0.0285	-0.1583
16	-0.9529	-1.1093	0.1372	-0.4424	-0.1055	0.3435
17	-1.1469	-1.3358	0.1118	-0.4738	0.3516	-0.1326
19	0.8765	0.9912	0.0982	0.3270	0.0799	0.1407
21	0.4208	0.4559	0.0596	0.1148	0.0142	0.0336
22	-0.0490	-0.0543	0.1102	-0.0191	0.0119	-0.0133
25	2.2503	3.0631	0.0867	0.9438	0.3476	-0.6337
27	-0.2954	-0.3522	0.2273	-0.1910	-0.0300	0.1562
28	1.0084	1.1396	0.0808	0.3380	-0.0232	0.1929
29	-0.4741	-0.5551	0.1901	-0.2689	-0.1750	-0.0323
30	0.7329	0.8843	0.2165	0.4648	-0.4057	0.1705
31	0.1204	0.1292	0.0512	0.0300	0.0015	0.0057
35	-0.1409	-0.1517	0.0571	-0.0373	-0.0119	-0.0014
38	0.1027	0.1294	0.3124	0.0872	0.0780	-0.0270
39	1.2868	1.7217	0.2847	1.0862	0.0098	-0.8342

(b) Which, if any, observations have relatively large leverage values?

(c) Based on the answers in (a) and (b), does it seem as if any observations may be especially influential? Explain.

(d) Study the DFFIT values. Identify an observation that may have a strong influence on the fitted values.

(e) Study the DFBETA values. Identify an observation that is influential for the literacy estimate but not for the economic activity estimate.

14.8. For the `Crime2` data file at the text website, fit the linear regression model with $y =$ violent crime rate and $x =$ percentage living in metropolitan areas, for all 51 observations.

(a) Plot the studentized residuals. Are there any clear outliers?

(b) Identify any observations with noticeable leverage.

(c) Based on (a) and (b), do any observations seem to be particularly influential? Explain.

(d) Study the DFFIT values. Which, if any, observations have a strong influence on the fitted values?

(e) Study the DFBETA values. For each term, which if any observations have a strong influence on the parameter estimate?

(f) Remove the observation that seems most influential, and refit the model. Is the prediction equation substantively different in any way?

14.9. In Exercise 14.3, backward elimination and forward selection choose the model with explanatory variables SIZE, BATHS, and NEW.

(a) Fit this model with the `Houses2` data set. Inspect the leverages and the DFFIT and DFBETA values for SIZE. Refit the model without the three highly influential observations. Compare the prediction equation, standard errors, and R^2 to the fit for the complete data set. Summarize the influence of the influential observations.

(b) For this model, report the VIF values. Interpret them, and indicate whether the degree of multicollinearity is severe.

14.10. For the `Houses2` data file, fit the model to $y =$ selling price using house size, whether the house is new, and their interaction.

(a) Show that the interaction term is highly significant.

(b) Show that observation 5 is highly influential in affecting the fit in (a).

(c) Show that the interaction effect is not significant when observation 5 is removed from the data set.

(d) Now fit the model in (a) using a GLM assuming a gamma distribution for y. Note how the estimated interaction effect differs considerably from that in (a), and note that it is not significant. (Observation 5, highly influential for ordinary least squares, is not so influential in this analysis. See also Exercise 13.22.)

14.11. Three variables have population correlations $\rho_{x_1 x_2} = 0.85$, $\rho_{yx_1} = 0.65$, and $\rho_{yx_2} = 0.65$. For these, the partial correlations are $\rho_{yx_1 \cdot x_2} = \rho_{yx_2 \cdot x_1} = 0.244$. In a sample, $r_{x_1 x_2} = 0.90$, $r_{yx_1} = 0.70$, and $r_{yx_2} = 0.60$, not far from the population values. For these, the sample partial correlations are $r_{yx_1 \cdot x_2} = 0.46$ and $r_{yx_2 \cdot x_1} = -0.10$. What does this large difference suggest about standard errors of partial correlations when multicollinearity exists? (An unwary observer might conclude that the partial effects of x_1 and x_2 have opposite signs and that the partial effect of x_1 is much stronger, when they are identical in the population.)

14.12. For a data set for 100 adults on $y =$ height, $x_1 =$ length of left leg, and $x_2 =$ length of right leg, the model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ is fitted. Neither H_0: $\beta_1 = 0$ nor H_0: $\beta_2 = 0$ has a P-value below 0.05.

(a) Does this imply that length of leg is not a good predictor of height? Why?

(b) Does this imply that H_0: $\beta_1 = \beta_2 = 0$ would not have a P-value below 0.05? Why?

(c) Suppose $r_{yx_1} = 0.901$, $r_{yx_2} = 0.902$, and $r_{x_1 x_2} = 0.999$. Using forward selection and the potential predictors x_1 and x_2 with $\alpha = 0.05$ for tests, which model would you expect to be selected? Why?

14.13. Refer to the plot of residuals in Figure 14.13 for Exercise 14.6.

(a) Explain why a more valid fit may result from assuming that income has a gamma distribution, rather than a normal distribution.

(b) Table 14.11 shows results for the normal GLM and the gamma GLM. Summarize how results differ for the two models.

TABLE 14.11

Parameter	NORMAL GLM			GAMMA GLM		
	Coef.	se	Sig	Coef.	se	Sig
Intercept	-15.663	8.412	.066	-1.927	5.169	.709
education	4.432	.619	.000	3.289	.412	.000
[race = b]	-10.874	4.473	.017	-8.905	2.842	.002
[race = h]	-4.934	4.763	.304	-5.953	3.187	.062
[race = w]	0	.	.	0	.	.
(Scale)				.117		

(c) Interpret the scale parameter estimate by estimating the standard deviation of income when its conditional mean is (i) $20,000 (ii) $50,000.

14.14. Refer to the data from Example 14.7 on fertility rates and GDP (page 453). To allow for greater variation at higher values of mean fertility, fit a quadratic GLM with a gamma distribution for fertility rate and the identity link function. Find the GDP value at which predicted fertility rate takes its minimum value. Compare estimates and their significance to those using least squares.

14.15. Table 14.12 shows the results of fitting two models to 54 observations on y = mental health score, x_1 = degree of social interaction, and x_2 = SES. The variables x_1 and x_2 are measured on scales of 0–100, and larger y-scores represent better mental health. The variable symbol x1**2 represents x_1^2, and x1*x2 represents $x_1 x_2$.

(a) When model 1 is fitted, which best describes the result over the range 0–100 of x_1-values?

(i) $\hat{y}$ is a bowl-shaped function of x_1, first decreasing and then increasing.
(ii) $\hat{y}$ is an increasing bowl-shaped function of x_1.
(iii) $\hat{y}$ is a mound-shaped function of x_1, first increasing and then decreasing.
(iv) $\hat{y}$ is an increasing mound-shaped function of x_1.

(b) When model 2 is fitted, which best describes the result over the observed ranges?

(i) $\hat{y}$ is a linear function of x_1 with positive slope that is the same for all x_2.
(ii) $\hat{y}$ is a linear function of x_1 with positive slope for some values of x_2 and negative slope for others.
(iii) $\hat{y}$ is a linear function of x_1 with positive slope, but the magnitude of that slope is smaller for larger values of x_2.
(iv) $\hat{y}$ is a quadratic function of x_1 and x_2.

TABLE 14.12

Model	Variable	Estimate	Model	Variable	Estimate
1.	Intercept	15	2.	Intercept	16
	x1	0.200		x1	0.07
	x1**2	−0.001		x2	0.04
				x1*x2	−0.0006

14.16. Sketch the following mathematical functions on the same set of axes, for values of x between 0 and 4. Use these curves to describe how the coefficients of x and x^2 affect their shape.

(a) $\hat{y} = 10 + 4x$ (b) $\hat{y} = 10 + 4x + x^2$
(c) $\hat{y} = 10 + 4x - x^2$ (d) $\hat{y} = 10 - 4x$
(e) $\hat{y} = 10 - 4x + x^2$ (f) $\hat{y} = 10 - 4x - x^2$

14.17. For the Houses data file, Table 14.13 shows results of fitting a quadratic regression model with s = size as the predictor.

(a) Interpret the coefficients of this equation. What shape does it have?

(b) Find the predicted selling price for homes with (i) $s = 1000$ square feet, (ii) $s = 2000$ square feet, (iii) $s = 3000$ square feet. Explain why the effect of a 1000-square-foot increase in s increases as s increases.

TABLE 14.13

Variable	Estimate	Std. Error	t	Sig
Intercept	5507.551	35626.650	.155	.877
size	65.156	36.289	1.795	.076
size*size	.014	.008	1.740	.085

14.18. Refer to the previous exercise.

(a) Using size as a straight-line predictor, $r^2 = 0.695$, whereas $R^2 = 0.704$ for the quadratic model. Is the degree of nonlinearity major, or minor? Is the linear association strong, or weak?

(b) Test whether the quadratic model gives a significantly better fit than the straight-line model. Interpret.

14.19. The Crime2 data file at the text website illustrates how a single observation can be highly influential in determining whether the model should allow nonlinearity.

(a) With all 51 observations, fit the quadratic model between y = murder rate and x = percentage in poverty. Test whether the quadratic term is needed. Report the P-value, and interpret.

(b) Refit the model, deleting the observation for D.C. Report the P-value for testing the quadratic term, and interpret.

(c) Compare (a) and (b), and use the scatterplot to explain how a single observation can have a large impact on whether the quadratic term seems needed. Show how you would be warned of this by influence diagnostics for the fit in (a).

14.20. For data from 2005 to 2011 from Facebook on y = number of people (in millions) worldwide using Facebook, the prediction equation $\hat{y} = 2.13(2.72)^x$ fits well, where x = number of years since January 1, 2005.

(a) Predict the number using the Internet at the beginning of (i) 2005 (take $x = 0$), (ii) 2011.

(b) Interpret the estimate 2.72.

(c) Illustrate the dangers of extrapolation, by predicting y on January 1, 2015.

(d) The straight-line model fitted to the data gives $\hat{y} = -114 + 95x$. Explain why this model is inappropriate for these data.

14.21. For data shown in the article "Wikipedia: Modelling Wikipedia's growth" at en.wikipedia.org, the number of English language articles in Wikipedia was well approximated from 2001 to 2008 by $\hat{y} = 22{,}700(2.1)^x$, where x is the time (in years) since January 1, 2001.

(a) Interpret the values 22,700 and 2.1 in this prediction equation.

(b) A plot in this article shows that the growth has been more linear than exponential since 2007. If the exponential equation had continued to hold, predict the number of English Wikipedia articles on January 1, 2016. Compare to the actual number of about 5 million, and explain the dangers of model extrapolation.

14.22. For United Nations data on y = world population size (billions) between 1900 and 2010, the exponential regression model with x = number of years since 1900 gives $\hat{y} = 1.4193(1.014)^x$.

(a) Explain why the model fit corresponds to a rate of growth of 1.4% per year.

(b) Show that the predicted population size (i) doubles after 50 years, (ii) quadruples after 100 years.

(c) The correlation equals 0.948 between y and x and 0.985 between $\log(y)$ and x. Based on this, which model seems more appropriate? Why?

14.23. Draw rough sketches of the following mathematical functions on the same set of axes, for x between 0 and 35.

(a) $\hat{y} = 6(1.02)^x$. ($\hat{y}$ = predicted world population size in billions x years after 2000, if there is a 2% rate of growth every year.)

(b) $\hat{y} = 6(0.95)^x$. What does this represent?

(c) Use these plots to explain the effect of whether $\beta > 1$ or $\beta < 1$ in the model $E(y) = \alpha\beta^x$.

14.24. Consider the formula $\hat{y} = 4(2)^x$.

(a) Plot $\hat{y}$ for integer x between 0 and 5.

(b) Plot $\log_e \hat{y}$ against x. Report the intercept and slope of this line.

14.25. For white men in the United States, Table 14.14 presents the number of deaths per thousand individuals of a fixed age within a period of a year.

TABLE 14.14

Age	Death Rate (Per Thousand)
30	3
40	6
50	14
60	27
70	60
80	125

(a) Plot x = age against y = death rate and against $\log y$. What do these plots suggest about a good model for the relationship?

(b) Find the correlation between (i) x and y, (ii) x and $\log(y)$. What do these suggest about an appropriate model?

(c) Using generalized linear models, find the prediction equation for the model $\log[E(y)] = \alpha + \beta x$.

(d) Find the prediction equation for $\hat{y}$. Interpret the parameter estimates.

14.26. Consider the fertility and GDP data in Table 14.6, from the `FertilityGDP` data file.

(a) Using GLM software, fit the exponential regression model, assuming fertility rate has a (i) normal, (ii) gamma distribution. Interpret the effect of GDP on fertility rate for the gamma fit.

(b) What advantages does the exponential regression model have over the quadratic model?

Concepts and Applications

14.27. Refer to the `Students` data file (Exercise 1.11).

(a) Conduct and interpret a regression analysis using y = political ideology, selecting predictors from the variables in that file. Prepare a report describing the research question(s) posed and analyses and diagnostic checks that you conducted, and indicate how you selected a final model. Interpret results.

(b) Repeat the analysis, using y = college GPA.

14.28. Refer to the data file the class created in Exercise 1.12. Select a response variable, pose a research question, and build a model using other variables in the data set. Interpret and summarize your findings.

14.29. Analyze the `Crime` data set at the text website, deleting the observation for D.C., with y = violent crime rate. Use methods of this chapter. Prepare a report describing the analyses and diagnostic checks that you conducted, and indicate how you selected a model. Interpret results.

14.30. For the `Mental` data file at the text website and the model predicting mental impairment using life events and SES, conduct an analysis of residuals and influence diagnostics.

TABLE 14.15

Year	Population	Year	Population
1830	34,730	1920	968,470
1840	54,477	1930	1,468,211
1850	87,445	1940	1,897,414
1860	140,424	1950	2,771,305
1870	187,748	1960	4,951,560
1880	269,493	1970	6,791,418
1890	391,422	1980	9,746,324
1900	528,542	1990	12,937,926
1910	752,619	2000	15,982,378
		2010	18,804,623

Source: U.S. Census Bureau.

14.31. Table 14.15 shows the population size of Florida, by decade from 1830 to 2010. Analyze these data, which are the data file `FloridaPop` at the text website. Explain

why a linear model is reasonable for the restricted period 1970–2010.

14.32. For the UN2 data file at the text website, using methods of this chapter,

(a) Find a good model relating x = per capita GDP to y = life expectancy. (*Hint*: What does a plot of the data suggest?)

(b) Find a good prediction equation for y = fertility. Explain how you selected variables for the model.

14.33. Give an example of a response variable and a pair of explanatory variables for which an automated variable selection procedure would probably produce a model with only one explanatory variable. Explain.

14.34. A sociologist's first reaction upon studying automated variable selection routines was that they had the danger of leading to "crass empiricism" in theory building. From a theoretical perspective, describe the dangers with such methods. What guidelines would you suggest for avoiding these problems?

14.35. Give an example of two variables that you expect to have a nonlinear relationship. Describe the pattern you expect for the relationship. Explain how to model that pattern.

14.36. You plan to model y = fertility rate (the mean number of children per adult woman) and x = per capita gross domestic product (GDP, in tens of thousands of dollars). For the ordinary bivariate model, explain what might be inappropriate about the **(a)** constant standard deviation assumption and **(b)** straight-line assumption. State a model that you think might be more valid.

14.37. Using the formula $s/s_j\sqrt{(n-1)(1-R_j^2)}$ for the standard error of the estimator of β_j in multiple regression, explain how precision of estimation is affected by

(a) Multicollinearity.

(b) The conditional variability of the response variable.

(c) The variability of the explanatory variables.

(d) The sample size.

14.38. A recent newspaper article quoted a planner in a city as saying, "This city has been growing at the rate of 3.2% per year. That is not slow growth by any means. It corresponds to 32% growth per decade." Explain what is incorrect about this statement. If, in fact, the current population size of the city is 10,000 and in each of the next 10 years the city increases in size by 3.2% relative to the previous year, then

(a) What is the population size after a decade?

(b) What percentage growth occurs for the decade?

14.39. Example 14.8 showed a predicted U.S. population size (in millions) x decades after 1890 of $\hat{y}$ = $73.175(1.130)^x$.

(a) Show this is equivalent to 1.23% predicted growth *per year*. [*Hint*: $(1.0123)^{10} = 1.130$.]

(b) Explain why the predicted U.S. population size x *years* after 1890 is $73.175(1.0123)^x$.

14.40. You invest $1000 in an account with interest compounded annually at 9%.

(a) How much money do you have after x years?

(b) How long does it take your savings to triple in size?

For multiple-choice exercises 14.41–14.44, select the correct response(s). (There may be more than one.)

14.41. In the model $E(y) = \alpha + \beta_1 x + \beta_2 x^2$, the coefficient β_2

(a) Is the mean change in y as x^2 is increased one unit with x held constant.

(b) Is a curvature coefficient that describes whether the regression equation is bowl shaped or mound shaped.

(c) Equals 0 if the relationship between y and x is linear.

(d) Equals 0 if the population value of R^2 for this model equals ρ_{yx}^2.

14.42. The log transformation of the mean response in regression is useful when

(a) $E(y)$ is approximately a logarithmic function of x.

(b) $E(y)$ is approximately an exponential function of x.

(c) $\log E(y)$ is approximately a linear function of x.

(d) Unit changes in x have a multiplicative, rather than additive, effect on the mean of y.

14.43. Forward selection and stepwise regression are similar in the sense that, if they have the same α-level for testing a term,

(a) They always select the same final regression model.

(b) They always select the same initial regression model (when they enter the first explanatory variable).

(c) Any variable not in the final model does not have a significant partial association with y, controlling for the variables in the final model.

(d) It is impossible that all the variables listed for potential inclusion are in the final model.

14.44. Evidence of multicollinearity exists in a multiple regression fit when

(a) Strong intercorrelations occur among explanatory variables.

(b) The R^2-value is very large.

(c) The F test of $H_0: \beta_1 = \cdots = \beta_k = 0$ has a small P-value, but the individual t tests of $H_0: \beta_1 = 0, \ldots, H_0: \beta_k = 0$ do not.

(d) A predictor variable has VIF = 12.

14.45. True or false?

(a) Adjusted R^2 can possibly *decrease* when an explanatory variable is added to a regression model.

(b) Possible effects of an influential observation include changing a correlation from positive to negative, a P-value from 0.01 to 0.99, and R^2 from 0.01 to 0.99.

(c) When multicollinearity exists, one can still obtain good estimates of regression parameters, but R^2 may be adversely affected.

(d) If $y =$ annual medical expenses relates to $x =$ age by $E(y) = 1400 - 22x + 0.4x^2$, then the change in the mean of y for every unit change in x equals -22.

14.46. Select the best response for each of the following terms (not every response is used):

     Heteroscedasticity   ———
     Multicollinearity   ———
     Forward selection   ———
     Interaction   ———
     Exponential model   ———
     Stepwise regression   ———
     Studentized residual   ———
     Generalized linear model ———

(a) The mean of y multiplies by β for each unit increase in x.

(b) The log of $E(y)$ is linearly related to the log of x.

(c) A residual plot indicates that the residuals are much more spread out at high x than at low x.

(d) The bivariate effect of x_1 on y differs from the partial effect of x_1 on y, controlling for x_2.

(e) There are strong intercorrelations among explanatory variables.

(f) At each stage, the variable considered for entry into the model has the smallest P-value in the test of its partial effect on y.

(g) The response variable need not be normal, and we can model a function of the mean as a linear function of the explanatory variables.

(h) At each stage after entering a new variable, all variables in the model are retested to see if they still have a significant partial effect on y.

(i) The slope between $E(y)$ and x_1 changes as the value of x_2 changes.

(j) Measures the number of standard errors that an observation falls from its predicted value.

14.47.* Show that using a cross-product term to model interaction assumes that the slope of the relationship between y and x_1 changes linearly as x_2 changes. How would you suggest modeling interaction if, instead, the slope of the linear relationship between y and x_1 first increases as x_2 changes from low to moderate values and then decreases as x_2 changes from moderate to high values?

14.48.* Forward selection is used with 10 potential explanatory variables for y. In reality, none are truly correlated with y or with each other. For a random sample, show that the probability equals 0.40 that at least one is entered into the regression model when the criterion for admission is a P-value below 0.05 for the t test. (*Hint*: Use the binomial distribution.)

This page intentionally left blank

LOGISTIC REGRESSION: MODELING CATEGORICAL RESPONSES

The regression models studied in the past six chapters assume that the response variable is quantitative. This chapter presents generalized linear models for response variables that are categorical.

The **logistic regression model** applies to **binary** response variables—variables having only two possible outcomes. For instance, logistic regression can model

- A voter's choice in a U.S. presidential election (Democrat or Republican), with explanatory variables political ideology, annual income, education level, and religious affiliation.

- Whether a person uses illegal drugs (yes or no), with explanatory variables education level, whether employed, religiosity, marital status, and annual income.

Multicategory versions of logistic regression can handle ordinal response variables and nominal response variables. **Loglinear models** describe association structure among a set of categorical response variables. We can check the goodness of fit of models when data have contingency table form. All the models of this chapter use the *odds ratio* to summarize the strength of the associations.

15.1 Logistic Regression

For a binary response variable y, denote its two categories by 1 and 0, commonly referred to as *success* and *failure*. Recall[1] that the mean of 0 and 1 outcomes equals the *proportion* of outcomes that equal 1. Regression models for binary response variables describe the population proportion, which also represents the probability $P(y = 1)$ for a randomly selected subject. This probability varies according to the values of the explanatory variables.

Models for binary data ordinarily assume a *binomial distribution* for the response variable (Section 6.7). This is natural for binary outcomes. The models are special cases of generalized linear models (Section 14.4).

LINEAR PROBABILITY MODEL

For a single explanatory variable, the simple model

$$P(y = 1) = \alpha + \beta x$$

implies that the probability of success is a linear function of x. This is called the **linear probability model**.

---

[1] From the discussion of Table 3.6 on page 52 and Example 4.6 on page 95.

This model is simple but often inappropriate. As Figure 15.1 shows, it implies that probabilities fall below 0 or above 1 for sufficiently small or large x-values, whereas probabilities must fall between 0 and 1. The model may be valid over a restricted range of x-values, but it is rarely adequate when the model has several explanatory variables.
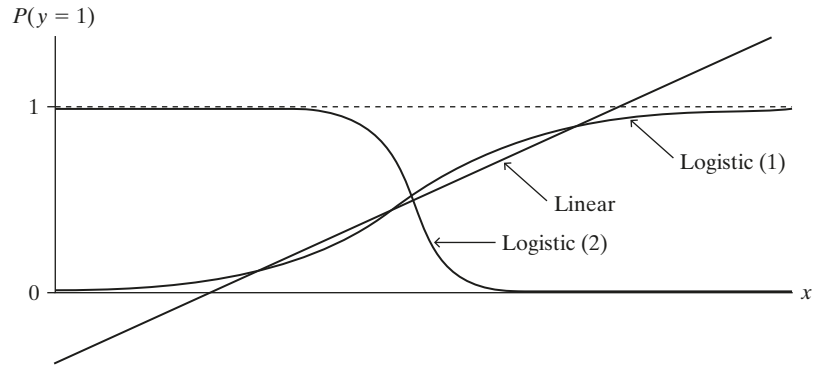
THE LOGISTIC REGRESSION MODEL FOR BINARY RESPONSES

Figure 15.1 also shows more realistic response curves, which have an S-shape. With these curves, the probability of a success falls between 0 and 1 for all possible x-values. These curvilinear relationships are described by the formula

$$\log\left[\frac{P(y = 1)}{1 - P(y = 1)}\right] = \alpha + \beta x.$$

The ratio $P(y = 1)/[1 - P(y = 1)]$ equals the **odds** (see page 242). For instance, when $P(y = 1) = 0.75$, the odds are $0.75/0.25 = 3.0$, meaning that a success is three times as likely as a failure. Software uses natural logarithms in fitting the model. However, we won't need to use (or understand) logarithms to interpret the model and conduct inference using it.

This formula uses the log of the odds, $\log[P(y = 1)/(1 - P(y = 1))]$, called the **logistic transformation**, or **logit** for short. The model, abbreviated as

$$\text{logit}[P(y = 1)] = \alpha + \beta x,$$

is called the **logistic regression model**.

When the logit follows this straight-line model, the probability $P(y = 1)$ itself follows a curve such as in Figure 15.1. When $\beta > 0$, $P(y = 1)$ increases as x increases, as in curve (1) in Figure 15.1. When $\beta < 0$, $P(y = 1)$ decreases as x increases, as in curve (2) in Figure 15.1. If $\beta = 0$, $P(y = 1)$ does not change as x changes, so the curve flattens to a horizontal straight line. The steepness of the curve increases as $|\beta|$ increases. For instance, $|\beta|$ for curve (2) is greater than β for curve (1).

When $P(y = 1) = 0.50$, the odds $P(y = 1)/[1 - P(y = 1)] = 1$, and $\log[P(y = 1)/(1 - P(y = 1))] = 0$. So, to find the value of x at which $P(y = 1) = 0.50$, we equate this log odds value of 0 to $\alpha + \beta x$ and then solve for x. We then find that $P(y = 1) = 0.50$ when $x = -\alpha/\beta$.

Software uses *maximum likelihood* (page 138) to fit the model. For binary data, this method is more appropriate than least squares.

Political Ideology and Belief in Evolution   The 2014 General Social Survey asked, "Human beings, as we know them today, developed from earlier species of animals. True or false?" Is the response to this question associated with one's political ideology? Let y = opinion about evolution (1 = true, 0 = false). Let x = political ideology (1 = extremely conservative, 2 = conservative, 3 = slightly conservative, 4 = moderate, 5 = slightly liberal, 6 = liberal, 7 = extremely liberal). Table 15.1 shows 4 of the 1064 observations.

TABLE 15.1: GSS Data on y = Opinion about Evolution (1 = True, 0 = False) and x = Political Ideology (from 1 = Extremely Conservative to 7 = Extremely Liberal)		
Subject	x	y
1	4	1
2	3	0
3	2	0
4	6	1

Source: Complete data file `Evolution` for n = 1064 is at the text website.

Political ideology is an ordinal variable. As with any model, we can treat an ordinal explanatory variable in a quantitative manner if we expect a trend upward or a trend downward in y as x increases. We treat it as categorical with dummy variables for more general effects other than trends. For these data, if we expect that the probability of belief in evolution continually increases or continually decreases as a person is more liberal, we treat x as quantitative. The model is then more parsimonious and simpler to interpret than if we use dummy variables for categories of x. In fact, the sample proportion of responses in the *true* category, shown in Figure 15.2, continually increases from 0.23 to 0.86 as political ideology moves from the most conservative to the most liberal. For a quantitative approach, it seems sensible to assign equally spaced scores for political ideology, such as the category numbers (1, 2, 3, 4, 5, 6, 7).

FIGURE 15.2: Sample Proportions Believing in Evolution for Seven Political Ideology Categories, and Logistic Regression Prediction Curve for the Probability of Believing in Evolution
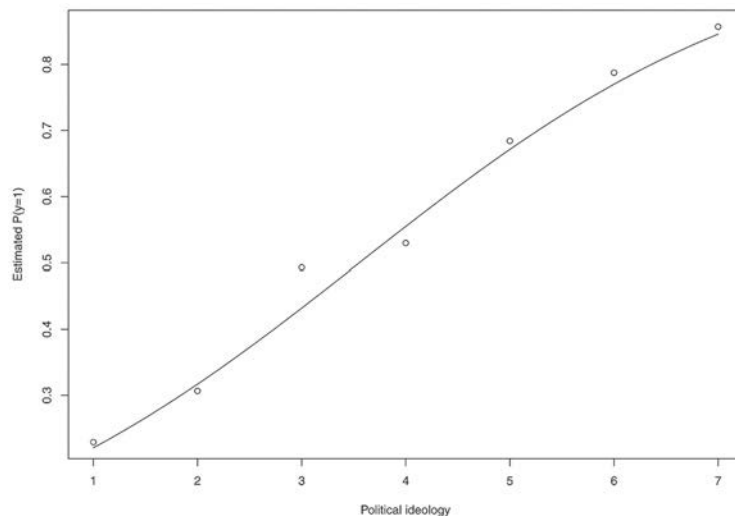
Figure 15.2 also shows the prediction curve for the fit of the logistic regression model. Table 15.2 shows some results that software provides for the model. The prediction equation is

$$\text{logit}[\hat{P}(y=1)] = -1.757 + 0.494x.$$

Since $\hat{\beta} = 0.494 > 0$, the estimated probability of believing in evolution increases as political ideology moves in the more liberal direction (i.e., higher x scores). The estimated probability equals 0.50 at $x = -\hat{\alpha}/\hat{\beta} = 1.757/0.494 = 3.55$. The estimated probability of believing in evolution is below 0.50 for the three categories for which political ideology is conservative.

TABLE 15.2: Logistic Regression Model Output (R Software) for the `Evolution` Data File on Belief in Evolution and Political Ideology

```
> summary(glm(y ~ polviews, family=binomial(link="logit"))

             Estimate  Std. Error  z value   Pr(>|z|)
(Intercept)  -1.75658     0.20500   -8.569    <2e-16
polviews      0.49422     0.05092    9.706    <2e-16
```

LOGISTIC REGRESSION EQUATION FOR PROBABILITIES

An alternative equation for logistic regression expresses the probability of success directly. It is

$$P(y=1) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}.$$

Here, e raised to a power represents the antilog of that number, using natural logs.[2] We use this formula to estimate values of $P(y=1)$ at particular predictor values.

From the estimates in Table 15.2, a person with political ideology x has estimated probability of believing in evolution

$$\hat{P}(y=1) = \frac{e^{-1.757+0.494x}}{1 + e^{-1.757+0.494x}}.$$

For subjects with ideology $x = 1$, the most conservative category, the estimated probability equals

$$\hat{P}(y=1) = \frac{e^{-1.757+0.494(1)}}{1 + e^{-1.757+0.494(1)}} = \frac{e^{-1.262}}{1 + e^{-1.262}} = \frac{0.283}{1.283} = 0.221.$$

For $x = 7$, the most liberal category, the estimated probability equals 0.846.

INTERPRETING THE LOGISTIC REGRESSION MODEL

We've seen how to estimate $P(y = 1)$, and we've seen that the sign of β tells us whether $P(y = 1)$ is increasing or decreasing as x increases. How else can we interpret β? Unlike in the linear probability model, β is not the slope for the change in $P(y = 1)$ as x changes. Since the curve for $P(y = 1)$ is S-shaped, the rate at which the curve climbs or descends changes according to the value of x.

The simplest way to use β to interpret the steepness of the curve uses a straight-line approximation to the logistic regression curve. A straight line drawn tangent to

---

[2] Most calculators have an e^x key that provides these antilogs, and software can report estimated probabilities.

the curve at a particular x-value has slope $\beta P(y = 1)[1 - P(y = 1)]$, where $P(y = 1)$ is the probability at that x. Figure 15.3 illustrates this. The slope is greatest when $P(y = 1) = 1/2$, where it is $\beta(1/2)(1/2) = \beta/4$. When $P(y = 1)$ is near 1/2, one-fourth of the β effect parameter in the logistic regression model is the approximate rate at which $P(y = 1)$ changes per one-unit increase in x.

FIGURE 15.3: A Line Drawn Tangent to a Logistic Regression Curve Has Slope $\beta P(y = 1)[1 - P(y = 1)]$, Which Is about $\beta/4$ when $P(y = 1)$ Is near 1/2
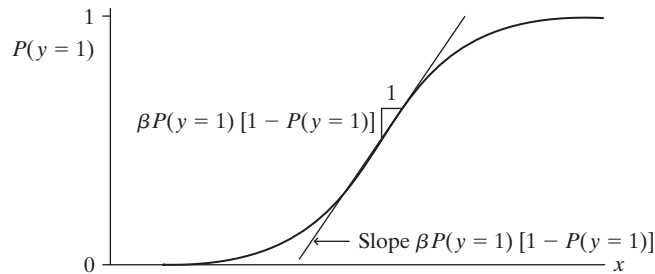
For the political ideology and belief in evolution data, $\hat{\beta} = 0.494$. At $x = 4$ (*moderate* ideology), the estimated probability of believing in evolution is $\hat{P}(y = 1) = 0.555$. A line drawn tangent to the curve at that point has slope approximately equal to $\hat{\beta}/4 = 0.494/4 = 0.12$. So, a one-category increase in political ideology (i.e., from "moderate" to "slightly liberal") has approximately a 0.12 increase in the estimated probability of belief in evolution.

Software can also fit the linear probability model, $P(y = 1) = \alpha + \beta x$. This model seems reasonable for these data, since the logistic fit in Figure 15.2 is not much different from a straight line, and a linear fit is simpler to interpret. The maximum likelihood fit is $\hat{P}(y = 1) = 0.108 + 0.110x$. This formula suggests about a 0.11 increase in $\hat{P}(y = 1)$ per category increase in political ideology.

Another way to describe the effect of x compares $\hat{P}(y = 1)$ at different values of x. We've seen that when x increases from its smallest to its largest value in the sample, $\hat{P}(y = 1)$ increases from 0.221 to 0.846. Such a very large change represents a strong effect.

INTERPRETATION USING THE ODDS AND ODDS RATIO

Another interpretation of the logistic regression parameter β uses the *odds ratio* (page 242). Applying antilogs to both sides of the logistic regression equation $\log[P(y = 1)/(1 - P(y = 1))] = \alpha + \beta x$ yields the model expressed in terms of the *odds*,

$$\frac{P(y = 1)}{1 - P(y = 1)} = e^{\alpha + \beta x} = e^{\alpha}(e^{\beta})^x.$$

The right-hand side of this equation has the *exponential regression* form studied in Section 14.6, a constant multiplied by another constant raised to the x power. This exponential relationship implies that every unit increase in x has a multiplicative effect of e^{β} on the odds.

In Example 15.1, the antilog of $\hat{\beta}$ is $e^{\hat{\beta}} = e^{0.494} = 1.64$. When political ideology increases by one category in the liberal direction, the estimated odds of belief in evolution multiply by 1.64; that is, they increase by 64%. When $x = 5$, for example, the estimated odds of belief in evolution are 1.64 times what they are when $x = 4$. When $x = 4$,

$$\text{Estimated odds} = \frac{\hat{P}(y = 1)}{1 - \hat{P}(y = 1)} = e^{-1.757 + 0.494(4)} = 1.246,$$

whereas when $x = 5$,

$$\text{Estimated odds} = \frac{\hat{P}(y = 1)}{1 - \hat{P}(y = 1)} = e^{-1.757 + 0.494(5)} = 2.043,$$

which is 1.64 times the value of 1.246 at $x = 4$. In other words, $e^{\hat{\beta}} = e^{0.494} = 1.64 = 2.043/1.246$ is an estimated *odds ratio*, equaling the estimated odds at $x = 5$ divided by the estimated odds at $x = 4$.

Odds ratios also apply to changes in x other than 1. For example, the six-unit change in x from category 1 to category 7 in political ideology corresponds to a change of 6β in the log odds, and a multiplicative effect of $e^{6\beta} = (e^{\beta})^6$ on the odds. When $x = 7$, the estimated odds of belief in evolution equal $(1.64)^6 = 19.4$ times the estimated odds when $x = 1$. This is an extremely strong effect.

LOGISTIC REGRESSION CAN USE GROUPED OR UNGROUPED DATA FILES

Table 15.1 showed the data file in the usual form of one row of data for each subject. When the explanatory variables are categorical and the data can take the form of a contingency table, an alternative data file has a row of data for each cell count. The data are then said to be *grouped* instead of *ungrouped*.

For example, for the data set on political ideology and opinion about evolution, in category 1, 11 people said evolution was true and 37 people said it was false. In Stata software, we could represent this as two rows of a data file with columns x, y, *count*, having rows

```
---------------------
1 1 11
1 0 37
---------------------
```

As Appendix A shows, the model-fitting command then includes an [*fweight = count*] part to tell Stata that a row has as many observations as the variable *count* lists. In other software, such as SPSS and R, one could include both the count of $y = 1$ values and the sample size for that setting of x in a single row, such as

```
---------------------
1 11 48
---------------------
```

which indicates that when $x = 1$, 11 observations out of 48 total had $y = 1$. See Appendix A for details. The results (estimates, standard errors, tests, confidence intervals) are the same for both forms of the data file.

PROBIT MODELS AND INTERPRETATIONS*

The logit link is the most popular for modeling binary response variables. We've seen that the linear probability model, which models the probability itself as a linear function of explanatory variables, has limited scope but is simpler to interpret.

Another alternative, using the ***probit link***, results from the following construction: Suppose there is some underlying continuous variable y^* such that we observe $y = 1$ when $y^* \geq T$ for some particular threshold T and we observe $y = 0$ when $y^* < T$. We cannot actually observe the variable y^*, which is called a ***latent variable***,

or the threshold T. But, assuming this underlying model, we can interpret a coefficient β of an explanatory variable x in the probit model as the predicted change in y^*, in standard deviation units, per one-unit increase in x.

We illustrate for Example 15.1 on $x =$ political ideology and $y =$ belief in evolution ($1 =$ yes, $0 =$ no). When we use software to fit the model

$$\text{probit}[P(y = 1)] = \alpha + \beta x,$$

we obtain $\hat{\beta} = 0.3036$ with $se = 0.0300$. We estimate that each one-unit increase in political ideology (i.e., a one-category increase in liberalism) corresponds to a 0.3 standard deviation increase in the underlying latent variable y^* that measures belief in evolution. For a six-unit increase (from extremely conservative to extremely liberal), we predict a $6(0.3036) = 1.82$ standard deviation increase in the underlying response. This is an enormous effect.

BINARY REGRESSION MODELS WITH RANDOM EFFECTS*

Section 13.5 showed how to handle repeated measurement and other forms of correlated response data by including random effects in the model, called a *linear mixed model*. That section referred to continuous responses and modeling the mean. Similar approaches have been developed for categorical responses and modeling proportions using logits or probits.

For example, we can include random effects in a logistic regression model to account for within-subject associations in studies with repeated measures on a binary variable. In Example 15.1 on political ideology and belief in evolution, suppose we sampled families and measured $y =$ belief in evolution ($1 =$ yes, $0 =$ no) and $x =$ political ideology for everyone in each family. We would expect the family-specific responses to be associated. Let (y_{ij}, x_{ij}) be the outcomes for subject j in family i. We could use the model

$$\text{logit}[P(y_{ij} = 1)] = \alpha + \beta x_{ij} + s_i,$$

where s_i is a random effect for family i.

Logistic models with random effects can be computationally difficult to fit, but software is now widely available. For details about ways of handling categorical responses with correlated observations, see Agresti (2007, Chapters 9 and 10).

15.2 Multiple Logistic Regression

Logistic regression can handle multiple predictors. The multiple logistic regression model has the form

$$\text{logit}[P(y = 1)] = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p.$$

The formula for the probability itself is

$$P(y = 1) = \frac{e^{\alpha + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\alpha + \beta_1 x_1 + \cdots + \beta_p x_p}}.$$

Exponentiating a beta parameter provides the multiplicative effect of that explanatory variable on the odds, controlling for the other variables. The farther a β_i falls from 0, the stronger the effect of the predictor x_i, in the sense that the odds ratio falls farther from 1.

As in ordinary regression, cross-product terms allow interactions between pairs of explanatory variables. To include categorical explanatory variables, you can use dummy variables, as the next example illustrates.

<div style="margin-left:2em">

Example
15.2

Death Penalty and Racial Predictors Table 15.3 is a three-dimensional contingency table from a study[3] of the effects of racial characteristics on whether individuals convicted of homicide receive the death penalty. The variables in Table 15.3 are *death penalty verdict*, the response variable, having categories (yes, no), and the explanatory variables *race of defendant* and *race of victims*, each having categories (white, black). The 674 subjects were defendants in indictments involving cases with multiple murders in Florida.

</div>

TABLE 15.3: Death Penalty Verdict by Defendant's Race and Victims' Race, for Cases with Multiple Murders in Florida

Defendant's Race	Victims' Race	Death Penalty Yes	Death Penalty No	Percentage Yes
White	White	53	414	11.3
	Black	0	16	0.0
Black	White	11	37	22.9
	Black	4	139	2.8

For each of the four combinations of defendant's race and victims' race, Table 15.3 also lists the percentage of defendants who received the death penalty. For white defendants, the death penalty was imposed 11.3% of the time when the victims were white and 0.0% of the time when the victims were black, a difference of 11.3% − 0.0% = 11.3%. For black defendants, the death penalty was imposed 22.9% − 2.8% = 20.1% more often when the victims were white than when the victims were black. Thus, controlling for defendant's race by keeping it fixed, the percentage of *yes* death penalty verdicts was considerably higher when the victims were white than when they were black.

Now, consider the association between defendant's race and the death penalty verdict, controlling for victims' race. When the victims were white, the death penalty was imposed 22.9% − 11.2% = 11.7% more often when the defendant was black than when the defendant was white. When the victims were black, the death penalty was imposed 2.8% more often when the defendant was black than when the defendant was white. In summary, controlling for victims' race, black defendants were somewhat more likely than white defendants to receive the death penalty.

For y = death penalty verdict, let $y = 1$ denote the *yes* verdict. Since defendant's race and victims' race each have two categories, a single dummy variable can represent each. Let d be a dummy variable for defendant's race and v a dummy variable for victims' race, where

$$d = 1, \text{defendant} = \text{white}, \ d = 0, \text{defendant} = \text{black},$$
$$v = 1, \text{victims} = \text{white}, \ v = 0, \text{victims} = \text{black}.$$

The logistic model with main effects for these explanatory variables is

$$\text{logit}[P(y = 1)] = \alpha + \beta_1 d + \beta_2 v.$$

Here, e^{β_1} is the odds ratio between the response variable and defendant's race, controlling for victims' race, and e^{β_2} is the odds ratio between the response and victims' race, controlling for defendant's race.

---

[3] M. Radelet and G. Pierce, *Florida Law Review*, vol. 43 (1991), pp. 1–34; see Exercise 15.5 for more recent data.

Table 15.4 shows software output for the model fit. The prediction equation is

$$\text{logit}[\hat{P}(y = 1)] = -3.596 - 0.868d + 2.404v.$$

Since $d = 1$ for white defendants, the *negative* coefficient of d means that the estimated odds of receiving the death penalty are *lower* for white defendants than for black defendants. Since $v = 1$ for white victims, the *positive* coefficient of v means that the estimated odds of receiving the death penalty are *higher* when the victims were white than when they were black.

TABLE 15.4: Parameter Estimates for Logistic Model for Death Penalty Data (Stata Output). Race dummy variables (def and vic) are coded as 1 for white and 0 for black.

y |	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
def |	-.8677969	.3670742	-2.36	0.018	-1.587	-.1483
vic |	2.404444	.600616	4.00	0.000	1.227	3.582
_cons |	-3.596104	.5069138	-7.09	0.000	-4.590	-2.603

The antilog of $\hat{\beta}_1$, namely, $e^{\hat{\beta}_1} = e^{-0.868} = 0.42$, is the estimated odds ratio between defendant's race and the death penalty, controlling for victims' race. The estimated odds of the death penalty for a white defendant equal 0.42 times the estimated odds for a black defendant. We list *white* before *black* in this interpretation, because the dummy variable was set up with $d = 1$ for *white* defendants. If we instead let $d = 1$ for black defendants rather than white, then we get $\hat{\beta}_1 = 0.868$ instead of -0.868. Then, $e^{0.868} = 2.38$, which is 1/0.42; that is, the estimated odds of the death penalty for a black defendant equal 2.38 times the estimated odds for a white defendant, controlling for victims' race.

For victims' race, $e^{2.404} = 11.1$. Since $v = 1$ for *white* victims, the estimated odds of the death penalty when the victims were *white* equal 11.1 times the estimated odds when the victims were black, controlling for defendant's race. This is a very strong effect.

This model assumes that both explanatory variables are associated with the response variable, but with a lack of interaction: The effect of a defendant's race on the death penalty verdict is the same for each victim's race and the effect of victims' race is the same for each defendant's race. This means that the estimated odds ratio between each explanatory variable and the response variable takes the same value at each category of the other predictor. For instance, the estimated odds ratio of 11.1 between victims' race and the death penalty is the same when the defendants were white as when the defendants were black. ∎

MULTIPLICATIVE EFFECTS ON ODDS

The parameter estimates for the logistic regression model are linear effects, but on the scale of the *log* of the odds. It is easier to understand effects on the *odds* scale than the *log odds* scale. The antilogs of the parameter estimates are multiplicative effects on the odds.

To illustrate, for the data on the death penalty, the prediction equation

$$\log\left[\frac{\hat{P}(y = 1)}{1 - \hat{P}(y = 1)}\right] = -3.596 - 0.868d + 2.404v$$

refers to the log odds (i.e., logit). The corresponding prediction equation for the estimated odds is

$$\text{Odds} = e^{-3.596 - 0.868d + 2.404v} = e^{-3.596}e^{-0.868d}e^{2.404v}.$$

For white defendants, $d = 1$, and the estimated odds equal $e^{-3.596}e^{-0.868}e^{2.404v}$. For black defendants, $d = 0$, and the estimated odds equal $e^{-3.596}e^{2.404v}$. The estimated odds for white defendants divided by the estimated odds for black defendants equal $e^{-0.868} = 0.42$. This shows why the antilog of the coefficient for d in the prediction equation is the estimated odds ratio between defendant's race and death penalty verdict, for each fixed victim's race. The effect of the defendant's race being white is to multiply the estimated odds of a *yes* death penalty verdict by $e^{-0.868} = 0.42$ compared to its value for black defendants. The actual values of the odds depend on victims' race, but the ratio of the odds is the same for each.

The logit model expression for the log odds is *additive*, but taking antilogs yields a *multiplicative* expression for the odds. In other words, the antilogs of the parameters are *multiplied* to obtain odds. We can use this expression to calculate odds estimates for any combination of defendant's race and victims' race. For instance, when the defendant is black ($d = 0$) and the victims were white ($v = 1$), the estimated odds of the death penalty are

$$\text{Odds} = e^{-3.596-0.868d+2.404v} = e^{-3.596-0.868(0)+2.404(1)} = e^{-1.192} = 0.304.$$

EFFECTS ON PROBABILITIES ARE SIMPLER TO INTERPRET

We've seen that we can summarize the effects of explanatory variables by estimating *odds ratios*. Many researchers find it easier to get a feel for the effects by viewing summaries that use the *probability* scale. Such summaries can report estimated probabilities at particular values of a variable of interest. This evaluation is done at fixed values of the other variables, such as at their means or at certain values of interest.

The formula for the estimated probability of receiving the death penalty is

$$\hat{P}(y = 1) = \frac{e^{-3.596-0.868d+2.404v}}{1 + e^{-3.596-0.868d+2.404v}}.$$

For instance, when the defendant is white and the victims were white, $d = v = 1$, so

$$\hat{P}(y = 1) = \frac{e^{-3.596-0.868(1)+2.404(1)}}{1 + e^{-3.596-0.868(1)+2.404(1)}} = \frac{e^{-2.059}}{1 + e^{-2.059}} = \frac{0.128}{1.128} = 0.113.$$

When the defendant is black and the victims were white, $d = 0$ and $v = 1$, so

$$\hat{P}(y = 1) = \frac{e^{-3.596-0.868(0)+2.404(1)}}{1 + e^{-3.596-0.868(0)+2.404(1)}} = 0.233.$$

The estimated probability of receiving the death penalty is about twice as high for black defendants. The sample effect is quite strong.

These estimated probabilities are close to the sample proportions (Table 15.3). The estimated probabilities, unlike sample proportions, perfectly satisfy the model. The closer the sample proportions fall to the estimated probabilities, the better the model fits.

The probability modeled relates to the odds by

$$\hat{P}(y = 1) = \frac{\text{Odds}}{1 + \text{Odds}},$$

a formula we first used in Section 8.4. For instance, when the estimated odds $= 0.304$, as we found above when $d = 0$ and $v = 1$, then $\hat{P}(y = 1) = 0.304/(1 + 0.304) = 0.233$, as we just found directly.

With a quantitative explanatory variable x, you can report the change in $\hat{P}(y = 1)$ at the means of the other explanatory variables when x increases by a certain amount, such as (1) by a fixed value (e.g., 1), (2) by a standard deviation, (3) over its range from its lowest to greatest value, or (4) over its interquartile range from the lower quartile to the upper quartile. Approach (4) is, unlike (1), not affected by the choice of scale and, unlike (2) and (3), not affected by outliers.

STANDARDIZED EFFECTS

To compare effects of explanatory variables having different units, you can compare regression parameter estimates after the model has been fitted using standardized explanatory variables. An estimate then refers to the effect on the logit of a standard deviation increase in the value of the explanatory variable. If s_{x_j} is the standard deviation of explanatory variable x_j, the standardized estimate is

$$\hat{\beta}_j^* = \hat{\beta}_j s_{x_j}.$$

Example 15.3

Alcohol Consumption and Unprotected Sex Table 15.5 summarizes results of fitting four logistic regression models, from a study[4] that examined the effects of alcohol consumption and drug use on sexual behavior, for 549 undergraduate students at a university in the southeastern United States. The response variable is whether the subject engaged in unprotected sex (no condom used) in the past three months ($1 =$ yes, $0 =$ no). Participants were asked how much they had used drugs such as marijuana and cocaine in the past three months. (Ecstasy, meth, ketamine, and poppers are not shown in Table 15.5.) Drug use was treated as quantitative in models, with $1 =$ never, $2 =$ once or twice, $3 =$ several times, and $4 =$ at least once a week. Participants were also asked how many times per month they drink alcohol without energy drinks and how many times a month they drink alcohol mixed with energy drinks (AmED). The demographic variables measured included sex, age, race ($1 =$ white, $0 =$ nonwhite), and long-term relationship ($1 =$ yes, $0 =$ no).

TABLE 15.5: Logistic Regression Estimates (with *se* in Parentheses) for Modeling the Probability of Engaging in Unprotected Sex

Variable	Model 1 Coef. (SE)	Model 2 Coef. (SE)	Model 3 Coef. (SE)	Model 4 Coef. (SE)
Sex	0.138 (0.196)			
Age	0.005 (0.073)			
Race	0.369 (0.188)			
Relationship	1.203 (0.220)			
Sexual orientation	−0.540 (0.327)			
Year in school	0.098 (0.149)			
Marijuana use		0.396 (0.104)		
Cocaine use		1.744 (0.676)		
Alcohol use			0.175 (0.030)	
AmED use				0.174 (0.072)

The first model entered six demographic variables as the explanatory variables. The second model added the marijuana use and cocaine use variables, to analyze their effects while controlling for the demographic variables. The authors did not report how the effects of the demographic variables changed after adding those two variables to the model. The third model analyzed the effect of alcohol use, controlling for demographic variables and drug use. The fourth model added AmED use.

From Table 15.5, other things being fixed, the probability of engaging in unprotected sex increases with marijuana use and with cocaine use, controlling for the demographic variables, and it increases with alcohol use and with AmED use, controlling for the variables previously entered in the model. How can we compare effects, such as those of cocaine use in Model 2 and alcohol use in Model 3?

---

[4] By D. Snipes and E. Benotsch, *Addictive Behaviors*, vol. 38 (2013), pp. 1418–1423.

Suppose cocaine use (which is on a scale of 1 to 4) has a standard deviation of 0.50 and monthly alcohol use has a standard deviation of 5.0. Then, the corresponding standardized estimates are $1.744(0.50) = 0.872$ and $0.175(5.0) = 0.875$. Even though the unstandardized parameter estimates are very different, a standard deviation increase in cocaine use has a similar effect on the logit as a standard deviation increase in monthly alcohol use. ∎

PROPENSITY SCORES: SELECTION BIAS IN OBSERVATIONAL STUDIES*

We finish this section by mentioning a use of logistic regression for the very challenging goal of adjusting for *selection bias* in comparing two groups when we want to control for possibly confounding variables. In experimental studies, researchers don't worry about such bias because of the randomized assignment of subjects to the groups. The groups are approximately in balance for all other variables, measured or unmeasured. But for observational studies, subjects typically select the group (e.g., deciding whether to attend college). Researchers face the issue that the group membership may be associated with a confounding variable that is itself associated with the response variable and thus affects summaries such as the difference of means.

Here, *y* could be continuous or categorical, and the focus is on a binary explanatory variable that refers to two groups we want to compare. Suppose we have identified a set of potential confounding variables that we could include as explanatory variables in the model of interest, such as a regression model. The *propensity* is the probability of being in a particular one of the two groups, for a given setting of the explanatory variables. We can use logistic regression to estimate how the propensity depends on the explanatory variables.

In comparing the groups on the response variable *y*, you can control for differing distributions of the groups on the explanatory variables by adjusting for the estimated propensity. This is done by using the propensity (1) to match samples from the groups, or (2) to subclassify subjects into several strata consisting of intervals of propensity scores, or (3) to adjust directly by entering the propensity in the model. With approaches (1) and (2), propensity score matching attempts to mimic randomization by matching a person in one group with a person in the other group who is comparable on the observed explanatory variables. With the matching, one is avoiding the assumption about model structure made in using regression to control for the possible confounding variables. However, some subjects may not be matchable.

This method does not solve the main potential problem in using an observational study. Any study that is observational rather than randomized still has the limitation that propensity score methods adjust only for *observed* confounding variables and not for *unobserved* ones. Also, the methods work better in larger samples, so the observed confounding variables are more truly balanced in the subclassifications. For details about the use of propensity scores, see Guo and Fraser (2014).

15.3 Inference for Logistic Regression Models

As usual, statistical inference assumes randomization for gathering the data. It also assumes a *binomial* distribution for the response variable. The model identifies *y* as having a binomial distribution and uses the logit link function for $P(y = 1)$, which is the mean of *y*.

As in ordinary regression modeling, the logistic regression model

$$\text{logit}[P(y = 1)] = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$$

has three types of hypotheses. The global null hypothesis $H_0: \beta_1 = \cdots = \beta_p = 0$ states that none of the explanatory variables has an effect on $P(y = 1)$. An individual null

hypothesis such as H_0: $\beta_1 = 0$ states that x_1 has no effect on y, controlling for the other explanatory variables. That is, the variables are *conditionally independent*. A third type of test focuses on a subset of the parameters to compare nested models, for example, to test that a set of interaction effects are 0.

LIKELIHOOD-RATIO TEST COMPARING LOGISTIC REGRESSION MODELS

In ordinary regression, we use an F test to compare nested models (Section 11.5). For logistic regression, the analogous test is the **likelihood-ratio test**. This is a general-purpose test in statistical inference that provides a way to compare two models, a full model and a simpler model. It tests that the extra parameters in the full model equal zero. The test uses a key ingredient of maximum likelihood inference, the *likelihood function*. Denoted by ℓ, this gives the probability of the observed data as a function of the parameter values. The maximum likelihood estimates maximize this function. (Specifically, the estimates are the parameter values for which the observed data are most likely; see page 138.)

Let ℓ_0 denote the maximum of the likelihood function when H_0 is true, and let ℓ_1 denote the maximum without that assumption. The formula for the likelihood-ratio test statistic is

$$-2\log\left(\frac{\ell_0}{\ell_1}\right) = (-2\log\ell_0) - (-2\log\ell_1).$$

It compares the maximized values of $(-2\log\ell)$ when H_0 is true and when it need not be true. There is a technical reason for using -2 times the log of this ratio, namely, that the test statistic then has approximately a chi-squared distribution for large samples. The df value equals the number of parameters in the null hypothesis.

To illustrate, for the analyses in Example 15.3 about predictors of unprotected sex (page 481), the authors reported that $(-2\log\ell)$ dropped by 40.5 when they added six drug-use variables (marijuana, cocaine, ecstasy, meth, ketamine, poppers) to the model containing only demographic explanatory variables. This is a chi-squared statistic with $df = 6$, since the more complex model has six additional parameters. This shows extremely strong evidence of a better fit for the more complex model ($P < 0.0001$). So, at least one of these variables provides an improvement in predictive power.

For H_0: $\beta_1 = \cdots = \beta_p = 0$, the likelihood-ratio statistic has $df = p$. The test compares the $(-2\log\ell)$ values for the full model and for the model containing only an intercept term, to test the joint effects of *all* the explanatory variables.

LIKELIHOOD-RATIO AND WALD TESTS ABOUT EFFECTS

The likelihood-ratio test can also be used for an individual parameter such as H_0: $\beta_i = 0$. The estimate $\hat{\beta}$ of β has an approximate normal sampling distribution. So, another possible test for H_0: $\beta_i = 0$ uses as test statistic $z = \hat{\beta}_i/se$. Some software instead reports the square of this statistic, called a **Wald statistic**. The Wald statistic and the likelihood-ratio statistics for H_0: $\beta_i = 0$ have chi-squared null distributions with $df = 1$. The Wald statistic has the same P-value as the z statistic for the two-sided H_a: $\beta_i \neq 0$.

For testing H_0: $\beta_i = 0$ with large samples, the Wald test and likelihood-ratio test usually provide similar results. For small to moderate sample sizes or when the effect is extremely strong, the likelihood-ratio test often tends to be more powerful than the Wald test. The likelihood-ratio and Wald methods also have corresponding confidence intervals.

Example
15.4
Inference for Death Penalty and Racial Predictors   For the death penalty data, Example 15.2 (page 478) used the model

$$\text{logit}[P(y = 1)] = \alpha + \beta_1 d + \beta_2 v,$$

with dummy variables d and v for defendant's race and victims' race, respectively. Software (Stata) shows the results in Table 15.6.

TABLE 15.6: Logistic Regression Inference (Stata Output) for Death Penalty Data of Table 15.3						
				LR chi2(2)	= 21.89	
				Prob > chi2	= 0.0000	
y |	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
def |	-.8677969	.3670742	-2.36	0.018	-1.587249	-.1483447
vic |	2.404444	.600616	4.00	0.000	1.227258	3.581629
_cons |	-3.596104	.5069138	-7.09	0.000	-4.589637	-2.602571

The likelihood-ratio (LR) statistic of 21.89 shown at the top of the output is the test of H_0: $\beta_1 = \beta_2 = 0$, that neither defendant's race nor victims' race has an effect. With $p = 2$ parameters, it is a chi-squared statistic with $df = 2$, and has P-value = 0.0000. We conclude that at least one of the two explanatory variables has an effect.

If $\beta_1 = 0$, the death penalty verdict is independent of defendant's race, controlling for victims' race. The defendant's race effect $\hat{\beta}_1 = -0.868$ has a standard error of 0.367. The z test statistic for H_0: $\beta_1 = 0$ is $z = \hat{\beta}_1/se = -0.868/0.367 = -2.36$. For the two-sided alternative, the P-value is 0.018. Similarly, the test of H_0: $\beta_2 = 0$ has $P = 0.000$. The tests provide strong evidence of individual effects.

The parameter estimates are also the basis of confidence intervals for odds ratios. Since the estimates refer to *log* odds ratios, after constructing the interval for a β_j we take antilogs of the endpoints to form the interval for the odds ratio. For instance, since the estimated log odds ratio of 2.404 between victims' race and the death penalty verdict has a standard error of 0.601, a 95% confidence interval for the true log odds ratio is

$$2.404 \pm 1.96(0.601), \quad \text{or } (1.23, 3.58).$$

The confidence interval for the odds ratio is $(e^{1.23}, e^{3.58}) = (3.4, 35.9)$. For a given defendant's race, when the victims were white, the estimated odds of the death penalty are between 3.4 and 35.9 times the estimated odds when the victims were black.

Most software can also provide confidence intervals for probabilities. Ninety-five percent confidence intervals for the probability of the death penalty are (0.14, 0.37) for black defendants with white victims, (0.01, 0.07) for black defendants with black victims, (0.09, 0.15) for white defendants with white victims, and (0.003, 0.04) for white defendants with black victims. ∎

15.4 Logistic Regression Models for Ordinal Variables*

Many applications have a categorical response variable with more than two categories. For instance, the General Social Survey recently asked subjects whether government spending on the environment should increase, remain the same, or decrease. An extension of logistic regression can handle ordinal response variables.

CUMULATIVE PROBABILITIES AND THEIR LOGITS

For an ordinal response variable y, let $P(y \leq j)$ denote the probability that the response falls in category j or below (i.e., in category 1, 2, ..., or j). This is called a *cumulative probability*. With four categories, for example, the cumulative probabilities are

$$P(y = 1), \quad P(y \leq 2) = P(y = 1) + P(y = 2),$$

$$P(y \leq 3) = P(y = 1) + P(y = 2) + P(y = 3),$$

and the final cumulative probability uses the entire scale, so $P(y \leq 4) = 1$.

A c-category response has c cumulative probabilities. The order of forming the cumulative probabilities reflects the ordering of the response scale. The probabilities satisfy

$$P(y \leq 1) \leq P(y \leq 2) \leq \cdots \leq P(y \leq c) = 1.$$

The odds of response in category j or below is the ratio

$$\frac{P(y \leq j)}{P(y > j)}.$$

For instance, when the odds equal 2.5, the probability of response in category j or below equals 2.5 times the probability of response above category j. Each cumulative probability can convert to an odds.

A popular logistic model for an ordinal response variable uses logits of the cumulative probabilities. With $c = 4$, for example, the logits are

$$\text{logit}[P(y \leq 1)] = \log\left[\frac{P(y = 1)}{P(y > 1)}\right] = \log\left[\frac{P(y = 1)}{P(y = 2) + P(y = 3) + P(y = 4)}\right],$$

$$\text{logit}[P(y \leq 2)] = \log\left[\frac{P(y \leq 2)}{P(y > 2)}\right] = \log\left[\frac{P(y = 1) + P(y = 2)}{P(y = 3) + P(y = 4)}\right],$$

$$\text{logit}[P(y \leq 3)] = \log\left[\frac{P(y \leq 3)}{P(y > 3)}\right] = \log\left[\frac{P(y = 1) + P(y = 2) + P(y = 3)}{P(y = 4)}\right].$$

Since the final cumulative probability necessarily equals 1.0, we exclude it from the model. These logits of cumulative probabilities are called *cumulative logits*. Each cumulative logit regards the response as binary by considering whether the response is at the low end or the high end of the scale, where "low" and "high" have a different definition for each cumulative logit.

CUMULATIVE LOGIT MODELS FOR AN ORDINAL RESPONSE

A model can simultaneously describe the effect of explanatory variables on all the cumulative probabilities for y. For each cumulative probability, the model looks like an ordinary logistic regression, where the two outcomes are low = "category j or below" and high = "above category j." With a single explanatory variable, this model is

$$\text{logit}[P(y \leq j)] = \alpha_j - \beta x, \quad j = 1, 2, \ldots, c - 1.$$

For $c = 4$, for instance, this single model describes three relationships: the effect of x on the odds that $y \leq 1$ instead of $y > 1$, the effect on the odds that $y \leq 2$ instead of $y > 2$, and the effect on the odds that $y \leq 3$ instead of $y > 3$. The model requires a separate intercept parameter α_j for each cumulative probability. Since the cumulative probabilities increase as j increases, so do $\{\alpha_j\}$.

Why is the model written with a minus sign before β? This is not necessary, but it is how the model is parameterized by some software, such as Stata (the *ologit* function) and SPSS. That way, when $\beta > 0$, when x is *higher* cumulative probabilities are *lower*. But cumulative probabilities being lower means it is less likely to observe relatively low values and thus more likely to observe *higher* values of y. So, this parameterization accords with the usual formulation of a positive association, in the sense that a positive β corresponds to a positive association (higher x tending to occur with higher y). Statistical software for fitting the model has no standard convention. Software (such as SAS and the *VGAM* library in R) that specifies the model as

$$\text{logit}[P(y \le j)] = \alpha_j + \beta x$$

will report the opposite sign for $\hat{\beta}$. You should be careful to check how your software defines the model so that you interpret effects properly.

The parameter β describes the effect of x on y. When $\beta = 0$, each cumulative probability does not change as x changes, and the variables are independent. The effect of x increases as $|\beta|$ increases. In this model, β is the same for each cumulative probability. It has the same value for each cumulative logit. In other words, the model assumes that the effect of x is the same for each cumulative probability. This cumulative logit model with this common effect is often called the ***proportional odds*** model.

Figure 15.4 depicts the model for four response categories with a quantitative explanatory variable. The model implies a separate S-shaped curve for each of the three cumulative probabilities. For example, the curve for $P(y \le 2)$ has the appearance of a logistic regression curve for a binary response with the pair of outcomes $(y \le 2)$ and $(y > 2)$. At any fixed x-value, the three curves have the same ordering as the cumulative probabilities, the one for $P(y \le 1)$ being lowest.

FIGURE 15.4: Depiction of Curves for Cumulative Probabilities in a Cumulative Logit Model for a Response Variable with Four Categories
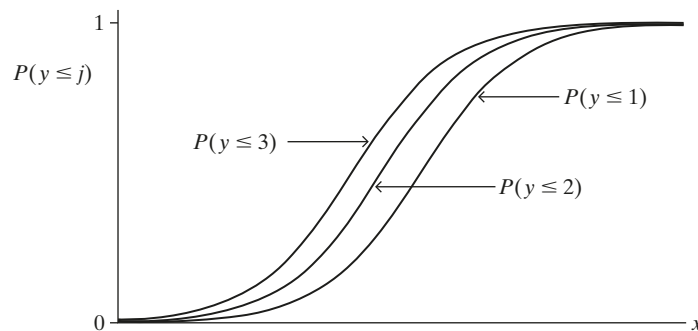
The size of $|\beta|$ determines how quickly the curves climb or drop. The common value for β means that the three response curves have the same shape. In Figure 15.4, the curve for $P(y \le 1)$ is the curve for $P(y \le 2)$ moved to the right and the curve for $P(y \le 3)$ moved even further to the right. To describe the association, e^{β} is a multiplicative effect of x on odds. For each j, the odds that $y > j$ multiply by e^{β} for each one-unit increase in x.

Model fitting treats the observations as independent from a *multinomial distribution*, the generalization of the binomial distribution from two outcome categories to multiple categories. Software estimates the parameters using all the cumulative probabilities at once. This provides a single estimate $\hat{\beta}$ for the effect of x, rather than the three separate estimates we'd get by fitting the model separately for each cumulative probability. If you reverse the order of categories of y (i.e., listing from high to low instead of from low to high), the model fit is the same, but the sign of $\hat{\beta}$ reverses.

Cumulative logit models can handle multiple explanatory variables, which can be quantitative and/or categorical. The model has the form

$$\text{logit}[P(y \leq j)] = \alpha_j - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p, \quad j = 1, 2, \ldots, c - 1.$$

When an explanatory variable is categorical, dummy variables can represent the categories. The model has the proportional odds assumption that the effect of an explanatory variable is the same on each cumulative probability.

Example 15.5

Comparing Political Ideology of Democrats and Republicans In the United States, do Republicans tend to be more conservative than Democrats? Table 15.7, for subjects in the 2014 General Social Survey, relates political ideology (EC = extremely conservative, C = conservative, SC = slightly conservative, M = moderate, SL = slightly liberal, L = liberal, EL = extremely liberal) to political party affiliation. We treat political ideology as the response variable. In Example 15.1, we treated it as a quantitative explanatory variable by assigning scores to categories. Now, we treat it as an ordinal response variable in a cumulative logit model.

TABLE 15.7: Political Ideology by Political Party Affiliation

Political Party	Political Ideology						
	EC	C	SC	M	SL	L	EL
Democratic	16	40	73	330	126	167	60
Republican	59	206	112	124	18	12	2

Let x be a dummy variable for political party affiliation, with $x = 1$ for Democrats and $x = 0$ for Republicans. Table 15.8 shows results of fitting the cumulative logit model. The response variable (political ideology) has seven categories, so the table reports six intercept parameter estimates, referred to as *cuts* because of how their ordered values are cutpoints on the real line. These estimates are not as relevant as the estimated effect of the explanatory variable (party affiliation), which is $\hat{\beta} = 2.527$. Since the dummy variable x is 1 for Democrats and since high values of y represent greater liberalism, the *positive* $\hat{\beta}$-value means that Democrats tend to be *more* liberal than Republicans. Democrats are more likely than Republicans to fall toward the liberal end of the political ideology scale.

TABLE 15.8: Output (Stata) for Cumulative Logit Model Fitted to Table 15.7

```
. ologit response party

Ordered logistic regression                         LR chi2(1) = 506.60
                                                    Prob > chi2 = 0.0000

-----------------------------------------------------------------------
 ideology |     Coef.   Std. Err.     z    P>|z|   [95% Conf. Int.]
    party |   2.527265   .1224756   20.63  0.000    2.2872    2.7673
    /cut1 |  -1.969594   .1249231                  -2.2144   -1.7247
    /cut2 |  -.0430447    .083657                  -.20701    .12092
    /cut3 |   .8684324   .0896656                   .69269    1.0442
    /cut4 |   2.781418   .1172596                   2.5516    3.0112
    /cut5 |   3.481216   .1264142                   3.2334    3.7290
    /cut6 |   5.077736   .1690398                   4.7464    5.4090
```

We can also interpret $\hat{\beta} = 2.527$ by exponentiating $\hat{\beta}$ to form an estimated odds ratio using cumulative probabilities. For any fixed j, the estimated odds that a Democrat's response is in the liberal direction rather than the conservative direction (i.e., $y > j$ rather than $y \leq j$) are $e^{\hat{\beta}} = e^{2.527} = 12.5$ times the estimated odds for Republicans. Specifically, this odds ratio applies to each of the six cumulative probabilities for Table 15.7.

To illustrate, using the third of the six cumulative probabilities, the estimated odds that a Democrat is in a liberal category, rather than a moderate or a conservative category, are 12.5 times the corresponding estimated odds for a Republican. The value 12.5 is far from the no-effect value of 1.0. The sample has a strong association, Democrats tending to be more liberal than Republicans. ∎

INFERENCE FOR EFFECTS ON AN ORDINAL RESPONSE

When $\beta = 0$ in the cumulative logit model, the variables are independent. We test independence by testing H_0: $\beta = 0$. As usual, the z test statistic divides $\hat{\beta}$ by its standard error. The square of that ratio is the Wald statistic, which is chi-squared with $df = 1$. The likelihood-ratio test is based on the difference in $(-2 \log \ell)$ values with and without the explanatory variable in the model. Most software can report both these tests. With multiple explanatory variables, the likelihood-ratio statistic also tests H_0: $\beta_1 = \cdots = \beta_p = 0$.

From Table 15.8, the effect of party affiliation has estimate $\hat{\beta} = 2.527$ and standard error $= 0.122$. So, $z = 2.527/0.122 = 20.63$ and the Wald statistic is $(20.63)^2 = 425.8$. The table also reports the likelihood-ratio test statistic for this hypothesis, which equals 506.6, based on $df = 1$. With either statistic, the P-value is 0.000 for testing H_0: $\beta = 0$ (independence of political ideology and party) against H_a: $\beta \neq 0$. These tests and $\hat{\beta}$ provide extremely strong evidence that Democrats tend to be politically more liberal than Republicans.

These tests of independence take into account the ordering of the response categories. They are usually more powerful than tests of independence that ignore the ordering, such as the Pearson chi-squared test of Section 8.2. When there truly is dependence, the ordinal test usually yields a smaller P-value.

With multiple explanatory variables, to check the fit of the model we can analyze whether extra terms such as interactions provide a significant improvement in the model fit. One way to do this uses a likelihood-ratio test of whether the extra parameters equal 0. Some software also provides a chi-squared test for the *proportional odds* assumption that the β effects are the same for all cumulative probabilities.

As in other statistical endeavors, don't put too much emphasis on statistical tests, whether of effects or of goodness of fit. Results are sensitive to sample size, more significant results tending to occur with larger sample sizes. Test statistics merely indicate the level of parsimony that is possible. It is important to supplement the tests with estimation methods that describe the strength of effects and with residual analyses that detect parts of the data for which the overall trend fails to hold.

A confidence interval for the odds ratio describes the association for the cumulative logit model. Since $\hat{\beta} = 2.527$ with $se = 0.122$, the 95% confidence interval for the population log odds ratio represented by β equals $2.527 \pm 1.96(0.122)$, or $(2.29, 2.77)$. The confidence interval for the odds ratio is $(e^{2.29}, e^{2.77})$, or $(9.8, 15.9)$. The odds that a Democrat's response falls in the liberal direction are more than about 10 times the odds for Republicans. To illustrate, using the fourth of the six cumulative probabilities, we conclude that the odds that a Democrat is in a liberal category,

rather than a moderate or a conservative category, fall between 9.8 and 15.9 times the odds for a Republican.

INVARIANCE TO CHOICE OF RESPONSE CATEGORIES

When the cumulative logit model fits well, it also fits well with similar effects for any collapsing of the response categories. For instance, if a model for categories (extremely conservative, conservative, slightly conservative, moderate, slightly liberal, liberal, extremely liberal) fits well, approximately the same estimated effects result when we fit the model to the data after collapsing the response scale to (conservative, moderate, liberal). This *invariance* to the choice of response categories is a nice feature of the model. Two researchers who use different response categories in studying an association should reach similar conclusions.

To illustrate, we collapse Table 15.7 to a three-category response, combining the three liberal categories and combining the three conservative categories. The estimated effect of party affiliation changes only from 2.527 ($se = 0.122$) to 2.535 ($se = 0.127$). Some slight loss of efficiency occurs in collapsing ordinal scales, resulting in larger standard errors. In practice, when observations are spread fairly evenly among the categories, the efficiency loss is minor unless the collapsing is to a binary response. It is not advisable to collapse ordinal data to binary.

The cumulative logit model implies trends upward or downward among distributions of y at different values of explanatory variables. When x refers to two groups, as in Table 15.7, the model fits well when subjects in one group tend to make higher responses on the ordinal scale than subjects in the other group. The model does not fit well when the response distributions differ in their *variability* rather than their average. If Democrats tended to be primarily moderate in political ideology, while Republicans tended to be both very conservative and very liberal (i.e., at the two extremes of the scale), then the Republicans' responses would show greater variability than the Democrats'. The two political ideology distributions would be quite different, but the model would not detect this if the average responses were similar.

15.5 Logistic Models for Nominal Responses*

For nominal response variables (i.e., *unordered* categories), an extension of the binary logistic regression model provides an ordinary logistic model for each pair of response categories. The models simultaneously use all pairs of categories by specifying the odds of outcome in one category instead of another. The order of listing the categories is irrelevant, because the response scale is nominal.

BASELINE-CATEGORY LOGITS

Logit models for nominal response variables pair each category with a baseline category. Most software uses the last category as the baseline. With three response categories, for example, the *baseline-category logits* are

$$\log\left[\frac{P(y=1)}{P(y=3)}\right] \quad \text{and} \quad \log\left[\frac{P(y=2)}{P(y=3)}\right].$$

For c outcome categories, the baseline-category logit model is

$$\log\left[\frac{P(y=j)}{P(y=c)}\right] = \alpha_j + \beta_j x, \quad j = 1, \dots, c-1.$$

Given that the response falls in category j or the last category, this models the log odds that the response is j. It looks like an ordinary logistic regression model, where the two outcomes are category j and category c.

Each of the $c - 1$ logit equations has its own parameters. With multiple explanatory variables, the model

$$\log\left[\frac{P(y = j)}{P(y = c)}\right] = \alpha_j + \beta_{j1}x_1 + \beta_{j2}x_2 + \cdots + \beta_{jp}x_p, \quad j = 1, \ldots, c - 1,$$

can have a large number of parameters. Software for multicategory logit models fits all the equations *simultaneously*, assuming independent observations from multinomial distributions.

Example 15.6

Belief in Afterlife by Sex and Race   Table 15.9, from a General Social Survey, has y = belief in life after death, with categories (yes, undecided, no), and explanatory variables sex and race. Let $s = 1$ for females and 0 for males, and $r = 1$ for blacks and 0 for whites. With *no* as the baseline category for y, the model is

$$\log\left[\frac{P(y = 1)}{P(y = 3)}\right] = \alpha_1 + \beta_{1S}s + \beta_{1R}r,$$

$$\log\left[\frac{P(y = 2)}{P(y = 3)}\right] = \alpha_2 + \beta_{2S}s + \beta_{2R}r.$$

The S and R subscripts identify the sex and race parameters. For example, β_{1S} compares females and males (controlling for race) on the log odds of responding *yes* rather than *no* to belief in life after death, whereas β_{2S} compares females and males on the log odds of responding *undecided* instead of *no*.

TABLE 15.9: Belief in Afterlife by Sex and Race

Race	Sex	Yes	Undecided	No
Black	Female	64	9	15
	Male	25	5	13
White	Female	371	49	74
	Male	250	45	71

The model assumes a lack of interaction between sex and race in their effects on belief in life after death. Table 15.10 shows the parameter estimates. For the first equation, for the log odds of a *yes* rather than a *no* response on belief in the afterlife, $\hat{\beta}_{1S} = 0.419$ and $\hat{\beta}_{1R} = -0.342$. Since the dummy variables are 1 for females and for blacks, given that a subject's response was *yes* or *no*, the estimated probability of a *yes* response was higher for females than for males (given race) and lower for blacks than for whites (given sex).

The effect parameters represent log odds ratios with the baseline category. For instance, $\hat{\beta}_{1S} = 0.419$ is the conditional log odds ratio between sex and response categories 1 and 3 (*yes* and *no*), given race. For females, the estimated odds of response *yes* rather than *no* on life after death are $e^{0.419} = 1.52$ times those for males, controlling for race. For blacks, the estimated odds of response *yes* rather than *no* on life after death are $e^{-0.342} = 0.71$ times those for whites, controlling for sex. These odds ratios suggest that, conditional on the response being *yes* or *no*, females are more likely than males to respond *yes*, and whites are more likely than blacks to respond *yes*.

ARBITRARY PAIRS OF RESPONSE CATEGORIES, AND FITTED RESPONSE DISTRIBUTIONS

The choice of the baseline category is arbitrary. We can use the equations for a given choice to get equations for *any* pair of categories. For example, using properties of logarithms,

$$\log\left[\frac{P(y=1)}{P(y=2)}\right] = \log\left[\frac{P(y=1)}{P(y=3)}\right] - \log\left[\frac{P(y=2)}{P(y=3)}\right].$$

So, to get the prediction equation for the log odds of belief *yes* instead of *undecided*, we take

$$(1.225 + 0.419s - 0.342r) - (-0.487 + 0.105s - 0.271r) = 1.71 + 0.314s - 0.071r.$$

For example, the odds of a response *yes* instead of *undecided* are higher for females (for whom $s = 1$) than males ($s = 0$).

Since the odds ratios for this model refer to *conditional probabilities*, given outcome in one of two categories, it is also useful to have software report the estimated unconditional probabilities based on the model fit. These *fitted probabilities* for the model sum to 1 over the full response scale. These are also simpler to interpret than odds ratios. Table 15.11 shows how to use R software to enter the four multinomial distributions, fit the model (with the last category as the baseline, by default), and show the fitted probabilities. The estimated probability of saying *yes* for belief in an afterlife varies between 0.622 for black males and 0.755 for white females.

INFERENCE FOR BASELINE-CATEGORY LOGIT MODELS

Inference applies as in ordinary logistic regression, except now to test the effect of an explanatory variable we consider *all* its parameters for the various equations. Likelihood-ratio tests compare the fits of the models with and without the explanatory variable in the model.

For example, for the data on belief in an afterlife, the test of the sex effect has $H_0\colon \beta_1^S = \beta_2^S = 0$. The likelihood-ratio test compares the model to the simpler model dropping sex as an explanatory variable. The test statistic (not shown in Tables 15.10 and 15.11) equals 6.75. It has $df = 2$, since H_0 has two parameters. The P-value of

> **TABLE 15.11:** Output (R) for Baseline-Category Logit Model Fitted to Table 15.9, Also Showing Fitted Probabilities

```
> race <- c(1,1,0,0); sex <- c(1,0,1,0)
> y1 <- c(64,25,371,250); y2 <- c(9,5,49,45); y3 <- c(15,13,74,71)
> library(VGAM)
> fit <- vglm(cbind(y1,y2,y3) ~ race + sex, family=multinomial)
> summary(fit)
               Estimate  Std. Error  z value  Pr(>|z|)
(Intercept):1    1.2248      0.1281    9.561   < 2e-16
(Intercept):2   -0.4870      0.1831   -2.659   0.00783
race:1          -0.3418      0.2370   -1.442   0.14934
race:2          -0.2710      0.3541   -0.765   0.44416
sex:1            0.4186      0.1713    2.444   0.01452
sex:2            0.1051      0.2465    0.426   0.66996
---
> fitted(fit)
        y1          y2          y3
1 0.7073517 0.10018119 0.1924671
2 0.6221640 0.12055943 0.2572766
3 0.7545608 0.09956287 0.1458763
4 0.6782703 0.12244794 0.1992817
```

0.03 shows evidence of a sex effect. By contrast, the effect of race is not significant, the likelihood-ratio statistic equaling 1.99 on $df = 2$. This partly reflects the larger standard errors that the effects of race have, due to a much greater imbalance between sample sizes in the race categories than in the sex categories.

15.6 Loglinear Models for Categorical Variables*

Logistic regression models are similar in structure to ordinary regression models, both types predicting a response variable using explanatory variables. By contrast, *loglinear models* are appropriate for contingency tables in which each classification is a response variable. Loglinear analysis resembles a correlation analysis more than a regression analysis. The loglinear focus is on studying associations between pairs of variables rather than modeling the response on one of them in terms of the others.

Loglinear models are special cases of generalized linear models that assume that each cell count in a contingency table has a *Poisson distribution*. This distribution is defined for discrete variables, such as counts, that can take nonnegative integer values. Equivalently, given the overall sample size, they assume a *multinomial* distribution for the counts in the cells of the contingency table.

The loglinear model formulas express the logs of cell expected frequencies in terms of dummy variables for the categorical variables and interactions between those variables. The actual model formulas can be cumbersome, and this section instead uses a symbolic notation that highlights the pairs of variables that are associated. Exercise 15.39 shows why the models are called *loglinear* models.

Example 15.7

Students' Use of Alcohol, Cigarette, and Marijuana   Table 15.12 is from a survey by the Wright State University School of Medicine and the United Health Services in Dayton, Ohio. The survey asked senior high school students in a nonurban area near Dayton, Ohio, whether they had ever used alcohol, cigarettes, or marijuana. Table 15.12 is a $2 \times 2 \times 2$ contingency table that cross-classifies responses on these three items.

TABLE 15.12: Alcohol, Cigarette, and Marijuana Use for High School Seniors			
Alcohol Use	**Cigarette Use**	**Marijuana Use**	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Source: Thanks to Prof. Harry Khamis, Wright State University, for these data.

In this table, all three variables are response variables, rather than one being a response variable and the others explanatory. The models presented in this section describe their association structure. They analyze whether each pair of variables is associated and whether the association is the same at each category of the third variable. ∎

A HIERARCHY OF LOGLINEAR MODELS FOR THREE VARIABLES

Loglinear models apply to contingency tables with any number of dimensions. We use three-way tables to introduce basic ideas, illustrating for Table 15.12. Denote the three categorical response variables by x, y, and z.

Loglinear models describe ***conditional associations*** in partial tables that relate two of the variables while controlling for the third one. A pair of variables could be statistically independent at each category of the third variable. In other words, the population version of each partial table could satisfy independence. In that case, the variables are said to be ***conditionally independent***, and the odds ratios equal 1 in the partial tables. Or, associations might exist in some or all of the partial tables. We now introduce a hierarchy of five loglinear models, ordered in terms of the extent of association.

1. All three pairs of variables are conditionally independent. That is,
    x is independent of y, controlling for z;
    x is independent of z, controlling for y;
    y is independent of z, controlling for x.

2. Two of the pairs of variables are conditionally independent. For example,
    x is independent of z, controlling for y;
    y is independent of z, controlling for x;
    x and y are associated, controlling for z.

3. One of the pairs of variables is conditionally independent. For example,
    x is independent of z, controlling for y;
    x and y are associated, controlling for z;
    y and z are associated, controlling for x.

4. No pair of variables is conditionally independent, but the association between any two variables is the same at each category of the third. We then say there is ***homogeneous association***.

5. All pairs of variables are associated, but there is *interaction*; that is, the association between each pair varies according to the category of the third variable.

Each model has a symbol that indicates the pairs of variables that are associated. Associated variables appear together in the symbol. For instance, (xy, z) denotes the

model for case 2 (above) in which x and y are associated but the other two pairs are conditionally independent. The symbol (xy, xz, yz) denotes the model for case 4, in which all three pairs are associated but the association is homogeneous. Table 15.13 lists the symbols for the models described above. All the models provide some structure for the pattern of association except for the one symbolized by (xyz). This model fits any sample three-way table perfectly, allowing the associations to be nonhomogeneous. It is called the **saturated model**.

TABLE 15.13: Some Loglinear Models for Three-Dimensional Contingency Tables

Model Symbol	Interpretation
(x, y, z)	All pairs are conditionally independent.
(xy, z)	x and y are the only associated pair.
(xy, yz)	x and z are the only conditionally independent pair.
(xy, yz, xz)	Each pair is associated, controlling for the third variable, but the association is homogeneous.
(xyz)	All pairs are associated, but the association is nonhomogeneous (interaction).

ODDS RATIO INTERPRETATIONS FOR LOGLINEAR MODELS

Interpretations of associations in loglinear models, like those in logistic regression models, can use the odds ratio. In 2×2 contingency tables, independence is equivalent to a population odds ratio of 1.0. In a three-way table, **conditional independence** between x and y means that the population odds ratios in the xy partial tables all equal 1.0. **Homogeneous association** means that the population odds ratios in the xy partial tables are identical at each category of z.

For instance, a $2 \times 2 \times 3$ table consists of three partial tables each of size 2×2, with two categories for x and two categories for y measured at three levels of z. When loglinear model (xz, yz) holds, x and y are conditionally independent, and the xy population odds ratio is 1.0 at the first level of z, 1.0 at the second level of z, and 1.0 at the third level of z. If the population odds ratio $= 2.2$ at the first level of z, 2.2 at the second level of z, and 2.2 at the third level of z, then there is conditional association but it is homogeneous. So, model (xy, xz, yz) holds. When the xy odds ratios are the same at all levels of z, necessarily the xz odds ratios are the same at all levels of y and the yz odds ratios are the same at all levels of x.

In fitting loglinear models, software provides expected frequency estimates (also called *fitted values*) having odds ratios that perfectly satisfy the model. If the model fits well, these odds ratios help us interpret the associations implied by the model.

Example 15.8

Estimated Odds Ratios for Substance Use Data   Denote the variables in Table 15.12 by A for alcohol use, C for cigarette use, and M for marijuana use. Table 15.14 contains estimated expected frequencies for the loglinear model (AC, AM, CM) that permits an association for each pair of variables but assumes homogeneous association, with the odds ratio between two variables being the same at each level of the third variable. The estimated expected frequencies are very close to the observed counts, so the model seems to fit well.

Let's study the estimated association between cigarette use and marijuana use, controlling for alcohol use, using the estimated expected frequencies. For those who have used alcohol, the estimated odds ratio between C and M is

$$\frac{910.4 \times 455.4}{538.6 \times 44.6} = 17.3.$$

TABLE 15.14: Estimated Expected Frequencies for the Loglinear Model (*AC*, *AM*, *CM*) for Alcohol (*A*), Cigarette (*C*), and Marijuana (*M*) Use for High School Seniors		

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	910.4	538.6
	No	44.6	455.4
No	Yes	3.6	42.4
	No	1.4	279.6

Similarly, for those who have not used alcohol, the estimated odds ratio between *C* and *M* is

$$\frac{3.6 \times 279.6}{42.4 \times 1.4} = 17.3.$$

For each category of *A*, students who have smoked cigarettes have estimated odds of having smoked marijuana that are 17.3 times the estimated odds for students who have not smoked cigarettes. The model assumes homogeneous association, so the estimated odds ratio is the same at each category of *A*.

Software for loglinear models provides tables of model parameter estimates from which one can also find estimated odds ratios. Table 15.15 illustrates this. (Software also provides estimates for intercept and main effect parameters. They are not shown here, because interpretations use the interaction terms to describe conditional associations.) For each pair of variables, the association parameter estimate refers to the *log* odds ratio. For the *CM* conditional association, therefore, the estimated odds ratio at each level of *A* equals $e^{2.848} = 17.3$. Similarly, the estimated odds ratio equals $e^{2.054} = 7.8$ between *A* and *C* at each level of *M*, and the estimated odds ratio equals $e^{2.986} = 19.8$ between *A* and *M* at each level of *C*. The estimated conditional association is very strong between each pair of variables. ■

TABLE 15.15: Output of Association Parameter (Log Odds Ratio) Estimates for the Loglinear Model (*AC*, *AM*, *CM*) for Substance Use Data				

Parameter	Estimate	Std. Error	z	Sig.
A*C	2.0545	0.1741	11.80	.000
A*M	2.9860	0.4647	6.43	.000
C*M	2.8479	0.1638	17.38	.000

The model (*AC*, *AM*, *CM*) permits conditional association for each pair of variables. Other possible loglinear models for these data delete at least one of the associations. To illustrate the association patterns implied by some of these models, Table 15.16 presents estimated conditional odds ratios for the estimated expected frequencies for the models. For example, the entry 1.0 for the *AC* conditional association for the model (*AM*, *CM*) is the common value of the estimated *AC* odds ratios at the two categories of *M*. This model implies conditional independence between alcohol use and cigarette use, controlling for marijuana use. It has estimated odds ratios of 1.0 for the *AC* conditional association.

TABLE 15.16: Summary of Estimated Conditional Odds Ratios for Various Loglinear Models Fitted to Substance Use Data

	Conditional Odds Ratio		
Model	AC	AM	CM
(A, C, M)	1.0	1.0	1.0
(AC, M)	17.7	1.0	1.0
(AM, CM)	1.0	61.9	25.1
(AC, AM, CM)	7.8	19.8	17.3
(ACM) Level 1	13.8	24.3	17.5
(ACM) Level 2	7.7	13.5	9.7

Table 15.16 shows that estimated conditional odds ratios equal 1.0 for each pairwise term not appearing in a model, such as the AC association in the model (AM, CM). The odds ratios for the sample data are those reported for the saturated model (ACM), which provides a perfect fit. For that model, the odds ratios between two variables are not the same at each level of the third variable, so they are reported separately for each level. In each case, they are strong at both levels.

We see in Table 15.16 that estimated conditional odds ratios can vary dramatically from model to model. This highlights the importance of good model selection. An estimate from this table is informative only to the extent that its model fits well. The next section shows how to check loglinear model goodness of fit.

15.7 Model Goodness-of-Fit Tests for Contingency Tables*

A *goodness-of-fit test* for a model is a test of the null hypothesis that that model truly holds in the population of interest. Section 8.2 introduced the chi-squared test of independence for contingency tables. That test is a goodness-of-fit test for the loglinear model that states that the two categorical variables are statistically independent. The chi-squared statistic compares the observed frequencies to the estimated expected frequencies that satisfy the independence model. Likewise, logistic regression and loglinear models for multidimensional contingency tables have chi-squared goodness-of-fit tests.

CHI-SQUARED GOODNESS-OF-FIT STATISTICS

Each model for a contingency table has a set of cell estimated expected frequencies, which are numbers that perfectly satisfy the model and give the best fit to the observed counts. The model goodness of fit is tested by comparing the estimated expected frequencies, denoted by $\{f_e\}$, to the observed frequencies $\{f_o\}$. Chi-squared test statistics summarize the discrepancies. Larger differences between the $\{f_o\}$ and $\{f_e\}$ yield larger values of the test statistics and stronger evidence that the model is inadequate.

Two chi-squared statistics, having similar properties, can test goodness of fit. The *Pearson chi-squared statistic*

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

was introduced on page 231 for testing independence. Another statistic, the *likelihood-ratio chi-squared statistic*, is

$$G^2 = 2 \sum f_o \log \left(\frac{f_o}{f_e} \right).$$

It equals the difference between the $(-2 \log \ell)$ values for the model being tested and for the most complex model possible.[5] Both X^2 and G^2 statistics equal 0 when there is a perfect fit (i.e., all $f_o = f_e$). Since large values indicate a poor fit, the *P*-value for testing a model is the right-tail probability above the observed value.

   If the model truly holds, both test statistics have approximate chi-squared distributions. The degrees of freedom (df) for the statistics depend on the model fitted. The df resemble the error df in regression, equaling the number of responses modeled on the left-hand side of the equation minus the number of parameters on the right-hand side of the model. For logistic regression models, for instance, the number of responses modeled is the number of sample logits for the model. This equals the number of combinations of levels of explanatory variables having observations on the binary response, since there is one logit for each combination. Thus, $df =$ number of logits modeled $-$ number of parameters. The simpler the model, in the sense of fewer parameters, the larger the df for the test.

   The chi-squared approximation is better for larger sample sizes. The Pearson statistic is preferred when the expected frequencies average between about 1 and 10, but neither statistic works well if most of the expected frequencies are less than about 5. These tests are not appropriate if any of the explanatory variables are not categorical. The chi-squared sampling distributions result only when they are applied to contingency tables with relatively large counts.

<hr>

Example 15.9

Logistic Model Goodness of Fit for Death Penalty Data   For Table 15.3 on death penalty verdicts, Examples 15.2 (page 478) and 15.4 (page 484) used the model

$$\text{logit}[P(y = 1)] = \alpha + \beta_1 d + \beta_2 v$$

to describe how the probability of receiving the death penalty depends on defendant's race d and victims' race v. A goodness-of-fit test analyzes whether this model with main effects is adequate for describing the data. The more complex model containing the interaction term is necessary if the main effects model fits poorly.

   Software automatically finds the model's estimated expected frequencies and the goodness-of-fit statistics. For instance, this model estimated a probability of 0.233 that a black defendant receives the death penalty for having white victims. Table 15.3 showed there were 48 such black defendants, so the estimated expected number receiving the death penalty is $48(0.233) = 11.2$. This is the estimated expected frequency for the cell in the table having observed frequency 11.

   The df equal the number of logits minus the number of parameters in the model. The death penalty data have four logits, one for each combination of defendant's race and victims' race. The model has three parameters, so both goodness-of-fit statistics have

$$df = \text{Number of logits} - \text{Number of parameters} = 4 - 3 = 1.$$

   Table 15.17 shows software results of chi-squared goodness-of-fit tests for the logistic model for the death penalty data. The null hypothesis for the tests is that the logistic model with main effects truly holds; that is, no interaction occurs between defendant's race and victims' race in their effects on the death penalty verdict. The

<hr>

[5] Software for generalized linear models calls this statistic the ***deviance***.

Pearson test statistic is $X^2 = 0.20$ and the likelihood-ratio test statistic is $G^2 = 0.38$. These test statistic values are small, so neither P-value is small. The model fits the data well. The null hypothesis that the model holds is plausible. ∎

TABLE 15.17: Chi-Squared Goodness-of-Fit Tests for Logistic Model with Main Effects Fitted to Death Penalty Data (Table 15.3)

	Goodness-of-fit Tests		
	Value	df	Sig.
Likelihood Ratio	.380	1	.538
Pearson Chi-Square	.198	1	.656

STANDARDIZED RESIDUALS

The chi-squared goodness-of-fit statistics provide global measures of lack of fit. When the fit is poor, a closer look at the cells of the table may reveal the nature of the lack of fit. Software for logistic and loglinear models can report ***standardized residuals*** (sometimes called *adjusted residuals*), which make a cell-by-cell comparison of f_o and f_e. Each standardized residual has the form

$$\text{Standardized residual} = \frac{f_o - f_e}{\text{Standard error of } (f_o - f_e)}.$$

When the model truly holds, standardized residuals behave like standard normal variables. A large standardized residual (say, exceeding 3 in absolute value) provides strong evidence of lack of fit in that cell. The standardized residuals presented in Section 8.3 (page 237) are special cases for the bivariate model of independence.

For the logistic model for the death penalty data, the standardized residuals all equal ± 0.44. They are small and provide no evidence of lack of fit. This is not surprising, since the goodness-of-fit statistics are small. In fact, when $df = 1$ for the goodness-of-fit test, only one standardized residual is nonredundant, and the square of any of them equals the X^2 test statistic.

LOGLINEAR MODEL GOODNESS OF FIT

The same goodness-of-fit formulas apply to loglinear models. Likewise, standardized residuals compare individual cell counts to expected frequencies satisfying the model.

Examples 15.7 (page 492) and 15.8 (page 494) used loglinear models to describe associations among alcohol use, cigarette use, and marijuana use, for high school students. Table 15.18 displays results of Pearson X^2 and likelihood-ratio G^2 goodness-of-fit tests for various loglinear models, ranging from the model (A, C, M) for which each pair of variables is independent to the model (AC, AM, CM) for which each pair is associated but the association between two variables is the same at each level of the third. The smaller the chi-squared statistics, the better the fit. Small P-values contradict the null hypothesis that the model is adequate. It is usually preferable to select the simplest model that provides a decent fit to the data. If no model fits well, the standardized residuals highlight cells contributing to the lack of fit.

From Table 15.18, the only model that passes the goodness-of-fit test is (AC, AM, CM). This model allows association between all pairs of variables but assumes that the odds ratio between each pair is the same at each category of the third variable. The models that lack any associations fit poorly, having P-values of 0.000.

TABLE 15.18:	Goodness-of-Fit Tests for Loglinear Models of Alcohol (A), Cigarette (C), and Marijuana (M) Use, with Likelihood-Ratio (G^2) and Pearson (X^2) Chi-Squared Test Statistics

Model	G^2	X^2	df	P-Value
(A, C, M)	1286.0	1411.4	4	0.000
(A, CM)	534.2	505.6	3	0.000
(C, AM)	939.6	824.2	3	0.000
(M, AC)	843.8	704.9	3	0.000
(AC, AM)	497.4	443.8	2	0.000
(AC, CM)	92.0	80.8	2	0.000
(AM, CM)	187.8	177.6	2	0.000
(AC, AM, CM)	0.4	0.4	1	0.54

COMPARING MODELS BY COMPARING G^2-VALUES

Table 15.18 illustrates two important properties of the likelihood-ratio G^2 statistic. First, G^2 has similar properties as the SSE (sum of squared residuals) measure in regression. Both compare observed responses to values expected if a model holds, and both cannot increase as the model becomes more complex. For instance, (A, CM) is a more complex model than (A, C, M), since it allows one association. Hence, it provides a better fit and its G^2-value is smaller. Similarly, G^2 drops further for the model (AC, CM) and further yet for (AC, AM, CM). The Pearson X^2 statistic, unlike the likelihood-ratio G^2, does not have this property. It could potentially increase as a model gets more complex, although in practice this rarely happens.

The second property of G^2 refers to model comparison, a topic discussed for logistic models on page 483. Section 11.5 introduced an F test comparing complete and reduced regression models, based on the reduction in SSE. A similar test comparing models for categorical responses uses the reduction in G^2-values. To test the null hypothesis that a model truly holds versus the alternative hypothesis that a more complex model fits better, the test statistic is the difference in G^2-values. This difference is identical to the difference in $(-2 \log \ell)$ values for the two models. It is a chi-squared statistic with degrees of freedom equal to the difference in df values for the two G^2-values. This is the likelihood-ratio test for comparing the models.

To illustrate, we compare loglinear models (AC, CM) and (AC, AM, CM). We test the null hypothesis that the reduced model (AC, CM) is adequate against the alternative that the more complex model (AC, AM, CM) is better. The likelihood-ratio test analyzes whether we can drop the AM association from the model (AC, AM, CM). The test statistic is the difference between their G^2-values, $92.0 - 0.4 = 91.6$, based on $df = 2 - 1 = 1$. This chi-squared statistic has a P-value of $P = 0.000$. So, the model (AC, AM, CM) fits significantly better than (AC, CM).

It is not possible to use G^2 to compare a pair of models such as (A, CM) and (AC, AM). Neither is a special case of the other, since each allows association that the other excludes.

DISTINCTION BETWEEN LOGISTIC AND LOGLINEAR MODELS

Logistic regression models distinguish between a single response variable and a set of explanatory variables. By contrast, loglinear models treat *every* variable as a

response variable. Most applications have a single response variable. It is then more natural to use a logistic regression model than a loglinear model. The logistic analysis focuses on the effects of the explanatory variables on the response, much as in ordinary regression modeling. For that reason, logistic regression has greater scope in practice.

15.8 Chapter Summary

Chapter 8 presented methods for analyzing association between two categorical variables. The methods of this chapter showed how to model a categorical response variable in terms of possibly *several* explanatory variables, which can be categorical or quantitative or both.

- For binary response variables, the ***logistic regression*** model describes how the probability of a particular category depends on explanatory variables. It uses a linear model for the ***logit*** transform of the probability, which is the log of the odds. For a quantitative explanatory variable, an S-shaped curve describes how the probability changes as the explanatory variable changes.

- The antilog of a $\hat{\beta}$ parameter estimate in logistic regression is a multiplicative effect on the odds for the response variable, for each one-unit increase in the explanatory variable of which it is a coefficient. Thus, for logistic regression the ***odds ratio*** is a natural measure of the nature and strength of an association.

- A parameter value of $\beta = 0$ corresponds to a predictor having no effect on the response. To test H_0: $\beta = 0$, we can use the normal test statistic, $z = \hat{\beta}/se$. The ***Wald test*** uses the square of this ratio. The ***likelihood-ratio test*** compares values of $(-2 \log \ell)$ for models with and without that term, where ℓ is the maximized *likelihood* function. The large-sample distribution of these test statistics is chi-squared with $df = 1$. The likelihood-ratio statistic can also test H_0: $\beta_1 = \cdots = \beta_p = 0$ in multiple logistic regression (with $df = p$) or compare nested models.

- For *ordinal* response variables, an extension of logistic regression uses *cumulative logits*, which are logits of *cumulative probabilities*. The model is called a ***cumulative logit model***. The effects of the explanatory variables are the same for each cumulative probability.

- For *nominal* response variables, an extension of logistic regression forms logits by pairing each category with a baseline category. Each logit equation has separate parameters in this ***baseline-category logit model***.

- ***Loglinear*** models are useful for investigating association patterns among a set of categorical response variables. They consider possible conditional independence patterns and use conditional odds ratios to describe association.

- For models for contingency tables, Pearson and likelihood-ratio chi-squared statistics test the ***goodness of fit*** of models to the data.

Karl Pearson introduced the chi-squared test for bivariate contingency tables in 1900. The models presented in this chapter did not become popular until near the end of the 1900s. They are examples of ***generalized linear models***, which apply to discrete as well as continuous response variables. Social scientists now have available a wide variety of tools for analyzing categorical data.

Exercises

Practicing the Basics

15.1. A logistic regression model describes how the probability of voting for Candidate X in an election depends on x = voter's total family income (in thousands of dollars) in the previous year. The sample prediction equation is

$$\log\left[\frac{\hat{P}(y=1)}{1 - \hat{P}(y=1)}\right] = -2.00 + 0.03x.$$

(a) Identify $\hat{\beta}$ and interpret its sign.

(b) Find the estimated probability of voting for the candidate when (i) income = 5000, (ii) income = 10000.

(c) At which income level is the estimated probability of voting for the candidate (i) equal to 0.50? (ii) greater than 0.50?

(d) For the region of x-values for which $P(y = 1)$ is near 0.50, give a linear approximation for the change in the probability for an increase of $1000 in income.

(e) Explain the effect of a $10,000 increase in family income on the odds of voting for the candidate.

15.2. Refer to the previous exercise. When the explanatory variables are x_1 = family income, x_2 = number of years of education, and s = sex (1 = male, 0 = female), the prediction equation is

$$\text{logit}[\hat{P}(y=1)] = -2.40 + 0.03x_1 + 0.05x_2 + 0.40s.$$

For this sample, x_1 ranges from 6 to 157 with a standard deviation of 25, and x_2 ranges from 7 to 20 with a standard deviation of 3.

(a) Find the estimated probability of voting for Candidate X for (i) a man with 10 years of education and income $50,000, (ii) a woman with 10 years of education and income $50,000.

(b) Convert the probabilities in (a) to odds, and find the odds ratio, the odds for men divided by the odds for females. Interpret.

(c) Show how the odds ratio in (b) relates to the sex effect in the prediction equation.

(d) Holding the other variables constant, find the estimated effect on the odds of voting the candidate of (i) a standard deviation change in x_2, (ii) a standard deviation change in x_1. Which explanatory variable has the larger standardized effect? Interpret.

15.3. A sample of 54 elderly men take a psychiatric examination to determine whether symptoms of senility are present. A subtest of the Wechsler Adult Intelligence Scale (WAIS) is the explanatory variable. The WAIS scores range from 4 to 20, with a mean of 11.6. Higher values indicate more effective intellectual functioning. Software shows the following results:

(a) Show (i) $\hat{P}(y = 1) = 0.50$ at $x = 7.2$, (ii) $\hat{P}(y = 1) < 0.50$ for $x > 7.2$.

```
-----------------------------------------------------------
Variable    Coef.    Std. Error   Wald Chi-square     Sig.
INTERCEPT   2.0429    1.0717          3.6338         0.0566
WAIS       -0.2821    0.1007          7.8487         0.0051
-----------------------------------------------------------
```

(b) Estimate the probability of senility at $x = 20$.

(c) The fit of the linear probability model is $\hat{P}(y = 1) = 0.847 - 0.051x$. Estimate the probability of senility at $x = 20$. Does this make sense?

(d) Test $H_0: \beta = 0$ against $H_a: \beta \neq 0$. Report and interpret the P-value.

15.4. Table 15.19, the data file Credit at the text website, shows data for a sample of 100 adults randomly selected for an Italian study on the relation between annual income and having a travel credit card, such as American Express or Diners Club. At each level of annual income (in thousands of euros), the table indicates the number of subjects in the sample and the number of those having at least one travel credit card. Let x = annual income and y = whether have a travel credit card (1 = yes, 0 = no). For instance, for the five observations at $x = 30$, $y = 1$ for two subjects and $y = 0$ for three subjects. Using software, **(a)** find the logistic regression prediction equation, and **(b)** fit the probit model. In each case, interpret the estimated effect and show the change in the estimated probability between the lowest and the highest income values. Attach software output to your solution.

TABLE 15.19

Income	No. Cases	Credit Cards	Income	No. Cases	Credit Cards	Income	No. Cases	Credit Cards
12	1	0	21	2	0	34	3	3
13	1	0	22	1	1	35	5	3
14	8	2	24	2	0	39	1	0
15	14	2	25	10	2	40	1	0
16	9	0	26	1	0	42	1	0
17	8	2	29	1	0	47	1	0
19	5	1	30	5	2	60	6	6
20	7	0	32	6	6	65	1	1

Source: Thanks to R. Piccarreta, Bocconi University, Milan, for original form of data.

15.5. For first-degree murder convictions[6] in East Baton Rouge Parish, Louisiana, between 1990 and 2008, the death penalty was given in 3 out of 25 cases in which a white killed a white, in 0 out of 3 cases in which a white killed a black, in 9 out of 30 cases in which a black killed a white, and in 11 out of 132 cases in which a black killed a black. Table 15.20 shows software output for fitting a logistic regression model, where $d = 1$ ($d = 0$) for black (white) defendants and $v = 1$ ($v = 0$) for black (white) victims. Interpret the estimates and the inference results in this table.

TABLE 15.20

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.0232	0.6137	-3.297	0.000978
d	1.1886	0.7236	1.643	0.100461
v	-1.5713	0.5028	-3.125	0.001778

Residual deviance: 0.16676 on 1 degrees of freedom

15.6. Table 12.1 in Chapter 12 reported GSS data on political ideology (scaled 1 to 7, with 1 being most liberal) by party affiliation of

	1	2	3	4	5	6	7
Democrat	5	18	19	25	7	7	2
Republican	1	3	1	11	10	11	1

Use logistic regression to describe the effect of political ideology on the probability of being a Democrat.

(a) Report the prediction equation, and estimate the probability of Democratic affiliation at ideology level (i) 1 = extremely liberal, (ii) 7 = extremely conservative.

(b) Use the model to test whether the variables are independent. Report the test statistic, P-value, and interpret.

(c) Use the odds ratio to describe the effect on party affiliation of a change in ideology from (i) 1 = extremely liberal to 2 = liberal, (ii) 1 = extremely liberal to 7 = extremely conservative.

(d) Construct and interpret a 95% confidence interval for the population odds ratio in (c), case (i).

15.7. A multination study of whether a country transitioned from autocracy to democracy during the study period[7] reported the prediction equation

$$\text{logit}[\hat{P}(y = 1)] = -3.30 + 0.55t + 1.12(\text{OECD})$$
$$- 1.16m - 0.01f - 0.07g,$$

where $y = 1$ if the nation made that transition, t = number of past transitions, OECD is membership in the Organization for Economic Cooperation and Development

(1 = yes, 0 = no), m = Muslim share of population, f = fuel exports (percentage of merchandise exports), and g = GDP growth (annual percentage).

(a) Interpret the coefficient of OECD.

(b) Interpret the coefficient of t.

(c) The estimated Muslim effect had $se = 0.61$ and P-value of 0.06. When the model was refit excluding countries regarded as the major oil-exporters, the estimated Muslim effect of -1.00 had $se = 0.64$ and P-value $= 0.12$. The authors concluded that the statistical significance of the Muslim effect might be explained by the major oil-exporting countries tending to be high on m. Give an alternative explanation, involving the sample size.

15.8. Let $P(y = 1)$ denote the probability that a randomly selected respondent supports current laws legalizing abortion, estimated using sex of respondent ($s = 0$, male; $s = 1$, female), religious affiliation ($r_1 = 1$, Hindu, 0 otherwise; $r_2 = 1$, Christian, 0 otherwise; $r_1 = r_2 = 0$, Muslim), and political party affiliation ($p_1 = 1$, Party A, 0 otherwise; $p_2 = 1$, Party B, 0 otherwise; $p_1 = p_2 = 0$, No Affiliation). The logistic model with main effects has the prediction equation

$$\text{logit}[\hat{P}(y = 1)] = 0.12 + 0.15s - 0.60r_1 - 0.68r_2 + 0.49p_1.$$

(a) Give the effect of sex on the odds of supporting legalized abortion; that is, if the odds of support for females equal θ times the odds of support for males, report $\hat{\theta}$.

(b) Give the effect of being in support of Party A instead of not being affiliated on the estimated odds of support for legalized abortion.

(c) Give the effect of being in support of Party A instead of in support of Party B on the estimated odds of support for legalized abortion.

(d) Find the estimated probability of supporting legalized abortion, for (i) female Muslim supporters of Party A, (ii) male Christian supporters of Party B.

15.9. Table 15.21 summarizes logistic regression results from a study[8] of how family transitions relate to first home purchase by young married households. The response variable is whether the subject owns a home (1 = yes,

[6] From G. Pierce and M. Radelet, *Louisiana Law Review*, vol. 71 (2011), pp. 647–673.
[7] By M. Gassebner et al., *Journal of Conflict Resolution*, vol. 57 (2013), pp. 171–197.
[8] From J. Henretta, *Social Forces*, vol. 66 (1987), pp. 520–536.

0 = no). Explanatory variables include a categorical variable for marital status two years after the year of observation (categories married, married with a working wife, single, with single the omitted category for the two dummy variables), a dummy variable for whether the family has more children aged 0–17 two years after the year of observation, and a dummy variable for whether the subject's parents owned a home in the last year that the subject lived in the parental home.

TABLE 15.21

Variable	Estimate	Std. Error
Intercept	−2.870	—
Husband's earnings ($10,000)	0.569	0.088
Wife's earnings ($10,000)	0.306	0.140
Number of years married	−0.039	0.042
Married in two years (1 = yes)	0.224	0.304
Working wife in two years (1 = yes)	0.373	0.283
Number of children	0.220	0.101
Additional child in two years (1 = yes)	0.271	0.140
Household head's education (no. of years)	−0.027	0.032
Parents' home ownership (1 = yes)	0.387	0.176

(a) Which explanatory variable seems to have the greatest partial impact on owning a home?

(b) Describe the partial effect of each additional $10,000 increase in (i) husband's income, (ii) wife's income.

(c) To describe the effect of husband's earnings, find the estimated probability of home ownership when wife's earnings = $50,000, years married = 3, not married in two years, the wife is working in two years, number of children = 0, additional child in two years = 0, head's education = 16 years, parents' home ownership is no, when husband's earnings equal (i) $20,000, (ii) $100,000.

15.10. For Table 15.12 on page 493, Table 15.22 shows output for a logistic model treating marijuana use as the response variable and alcohol use and cigarette use as explanatory variables.

(a) Set up dummy variables and report the prediction equation. Interpret the signs of the effects of alcohol use and cigarette use.

(b) Why are the estimates in the table equal to 0 at the second category of each explanatory variable?

15.11. A sample of inmates being admitted to the Rhode Island Department of Corrections were asked whether they ever injected drugs and were tested for hepatitis C virus (HCV). The numbers who reported injecting drugs were 306 of the 887 men who tested HCV positive, 61 of the 3044 men who tested HCV negative, 110 of the 197 women who tested HCV positive, and 13 of the 288 women who tested HCV negative. The authors[9] concluded that the prevalence of HCV may be underestimated by testing only those who reported injecting drugs.

(a) Report the results as a contingency table.

(b) Define dummy variables and specify a model for which the odds ratios between HCV status and whether injected drugs are identical in the population for each sex.

(c) Fit the model in (b), and report the model-based estimate of the odds ratio in (b).

15.12. Table 15.23 refers to individuals who applied for admission into graduate school at the University of California in Berkeley. Data[10] are presented for five of the six largest graduate departments at the university. The variables are

A: Whether admitted (yes, no).
S: Sex of applicant (male, female).
D: Department to which application was sent $(D_1, D_2, D_3, D_4, D_5)$.

(a) Construct the two-way table for sex and whether admitted, collapsing the table over department. Find the odds ratio and interpret.

(b) Treating *A* as the response and *D* and *S* as categorical explanatory variables, fit the logistic model having main effects. Report the prediction equation. Interpret the coefficient of *S* in this equation by finding the estimated conditional odds ratio between *A* and *S*, controlling for *D*.

(c) Contrast the model estimated conditional odds ratio between *A* and *S* in (b) with the odds ratio reported in (a). Explain why they differ so much, satisfying Simpson's paradox.

TABLE 15.22

Parameter		DF	Estimate	Std Err	ChiSquare	Pr > Chi
INTERCEPT		1	−5.309	0.4752	124.820	0.0001
ALCOHOL	yes	1	2.986	0.4647	41.293	0.0001
ALCOHOL	no	0	0.000	0.0000	0.	0.
CIGARETT	yes	1	2.848	0.1638	302.141	0.0001
CIGARETT	no	0	0.000	0.0000	0.	0.

[9] G. Macolino et al., *American Journal of Public Health*, vol. 95 (2005), pp. 1739–1740.
[10] From *Statistics* by D. Freedman, R. Pisani, and R. Purves (W. W. Norton, 1978), p. 14.

TABLE 15.23

Department	Sex	Admitted Yes	Admitted No
D_1	Male	353	207
	Female	17	8
D_2	Male	120	205
	Female	202	391
D_3	Male	138	279
	Female	131	244
D_4	Male	53	138
	Female	94	299
D_5	Male	22	351
	Female	24	317

15.13. Consider Table 8.16 on page 245, treating happiness as the response variable. Table 15.24 shows results of fitting the cumulative logit model $\text{logit}[P(y \leq j)] = \alpha_j + \beta x$, using scores (1, 2, 3) for income, and the chisquared test of independence.

(a) Why does the table report two intercept estimates?

(b) Report and interpret the income effect.

(c) Using the model, test the hypothesis of no income effect. Report the test statistic and P-value, and interpret.

(d) Suppose we instead used the Pearson chi-squared test of independence. Report the test statistic and P-value, and compare results to those in (c). Why are they so different?

TABLE 15.24

	Estimate	Std. Error	z	Sig.
INTERCP1 [happy=1]	-1.102	.275	-4.00	.000
INTERCP2 [happy=2]	1.305	.277	4.70	.000
INCOME	-0.267	.151	-1.77	.077

Statistic	DF	Value	Sig.
Pearson Chi-square	4	4.092	0.394

(b) Report and interpret an odds ratio describing the effect of wearing a seat belt.

(c) Construct a 95% confidence interval for the true odds ratio for the effect of wearing a seat belt. Interpret.

(d) Conduct a test of the hypothesis of no effect of seat belt use on the response, controlling for sex and location. Report the P-value and interpret.

(e) Fit the model that also has the three two-way interactions. Use a likelihood-ratio test to compare this model to the main effects model. Interpret.

TABLE 15.25

Sex	Location	Seat Belt	Response 1	2	3	4	5
Female	Urban	No	7287	175	720	91	10
		Yes	11,587	126	577	48	8
	Rural	No	3246	73	710	159	31
		Yes	6134	94	564	82	17
Male	Urban	No	10,381	136	566	96	14
		Yes	10,969	83	259	37	1
	Rural	No	6123	141	710	188	45
		Yes	6693	74	353	74	12

Source: Dr. Cristanna Cook, Medical Care Development, Augusta, Maine.

15.14. Using software with Table 8.16, replicate the results shown in the previous exercise for the cumulative logit model. Indicate whether the sign for $\hat{\beta}$ agrees with the negative sign for $\hat{\beta}$ in Table 15.24, according to how your software parameterizes the model.

15.15. Table 15.25 refers to passengers in autos and light trucks involved in accidents in the state of Maine. The table, available as the `Accidents` data file at the text website, classifies subjects by sex, location of accident, seat belt use, and a response variable having categories (1) not injured, (2) injured but not transported by emergency medical services, (3) injured and transported by emergency medical services but not hospitalized, (4) injured and hospitalized but did not die, (5) injured and died.

(a) Fit a cumulative logit model having main effects for sex ($s = 1$ for males and 0 for females), location ($r = 1$ for rural and 0 for urban), and seat belt use ($s = 1$ for yes and 0 for no). State the prediction equation. Interpret the sign of the effect for each explanatory variable.

15.16. Explain why the cumulative logit model is not valid with a nominal response variable, but a baseline-category logit model is valid with an ordinal response variable.

15.17. A baseline-category logit model fit predicting preference for U.S. President (Democrat, Republican, Independent) using x = annual income (in \$10,000) is $\log(\hat{\pi}_D/\hat{\pi}_I) = 3.3 - 0.2x$ and $\log(\hat{\pi}_R/\hat{\pi}_I) = 1.0 + 0.3x$.

(a) For each equation, interpret the sign of the estimated effect of x.

(b) Find and interpret the prediction equation for $\log(\hat{\pi}_R/\hat{\pi}_D)$.

(c) Use an estimated odds ratio to describe how the choice between Republican and Democrat depends on income.

15.18. For a sample of people in a developed city, for the most recent time each person shopped for clothes, you plan to model the choice to shop downtown or on the Internet. Explanatory variables include annual income,

whether a student, and distance of residence from downtown. Explain the type of model you would use, and why.

15.19. Refer to the 3×7 table in Table 12.1 (page 364) on party identification and political ideology.

(a) Fit a baseline-category logit model, treating party affiliation as the response and political ideology as a quantitative explanatory variable. Interpret the political ideology effect for the choice between Democrat and Republican.

(b) Fit a cumulative logit model, treating political ideology as the response. Interpret the cumulative odds ratio for comparing Democrats and Republicans on ideology.

15.20. Using software, replicate the results in Example 15.6 (page 490) on belief in an afterlife, sex, and race.

15.21. Consider the fit of the loglinear model (AC, AM, CM) to Table 15.12 for the survey of high school seniors.

(a) Use the estimated expected frequencies in Table 15.14 to estimate the conditional odds ratios between A and M at each level of C.

(b) Show how to obtain the estimated odds ratio in (a) from the parameter estimates for the model in Table 15.15.

(c) By contrast, what is the estimated conditional odds ratio between A and M for the loglinear model denoted by (AC, CM)?

15.22. Refer to the loglinear model analyses reported in Examples 15.7 and 15.8 for use of marijuana, alcohol, and cigarettes. Use software to replicate all the analyses shown there.

15.23. For a four-way cross-classification of variables w, x, y, and z, state the symbol for the loglinear model in which

(a) All pairs of variables are independent.

(b) x and y are associated, but other pairs of variables are independent.

(c) All pairs of variables are associated, but the conditional associations are homogeneous.

15.24. For Table 15.3 on the death penalty, the logistic model that has an effect of victims' race but assumes that the death penalty is independent of defendant's race (given victims' race) has a Pearson goodness-of-fit statistic equal to 5.81 with $df = 2$ (*P*-value 0.055). Specify H_0 for this test, and interpret the *P*-value.

15.25. Refer to the survey data for high school seniors in Table 15.12 and the goodness-of-fit statistics reported in Table 15.18 (page 499). Use these results to illustrate **(a)** when a model fits well and when a model fits poorly, **(b)** how G^2 decreases as the model becomes more complex.

Concepts and Applications

15.26. Refer to the `Students` data file (Exercise 1.11). Using software, conduct and interpret a logistic regression analysis using $y =$ opinion about abortion with explanatory variables

(a) Political ideology.

(b) Sex and political ideology.

15.27. In a one-page report, analyze Table 15.7 by treating party affiliation as the response variable and political ideology as a quantitative explanatory variable. Fit an appropriate model, conduct statistical inference, and interpret results. Attach annotated software output to your report.

15.28. The data shown in Exercise 10.14 in Chapter 10 came from an early study on the death penalty and racial characteristics. Analyze those data using methods of this chapter. Summarize your main findings in a way that you could present to the general public, using as little technical jargon as possible.

15.29. One year, the Metropolitan Police in London, England, reported[11] 30,475 people as missing in the year ending March 1993. For those of age 13 or less, 33 of 3271 missing males and 38 of 2486 missing females were still missing a year later. For ages 14–18, the values were 63 of 7256 males and 108 of 8877 females; for ages 19 and above, the values were 157 of 5065 males and 159 of 3520 females. Analyze and interpret these data.

15.30. In a study of whether an educational program makes sexually active adolescents more likely to obtain condoms, adolescents were randomly assigned to two experimental groups. The educational program, involving a lecture and videotape about transmission of the HIV virus, was provided to one group but not the other. In logistic regression models, factors observed to influence a teenager to obtain condoms were sex, socioeconomic status, lifetime number of partners, and the experimental group. Table 15.26 summarizes results.

TABLE 15.26

Variables	Odds Ratio	95% Confidence Interval
Group (education versus none)	4.04	(1.17, 13.9)
Sex (males versus females)	1.38	(1.23, 12.88)
SES (high versus low)	5.82	(1.87, 18.28)
Lifetime number of partners	3.22	(1.08, 11.31)

Source: V. I. Rickert et al., *Clinical Pediatrics*, Vol. 31 (1992), pp. 205–210.

(a) Find the parameter estimates for the fitted model, using (1, 0) dummy variables for the first three explanatory variables.

---

[11] *The Independent* newspaper, March 8, 1994; thanks to Dr. P. Altham for this.

(b) Explain why either the estimate of 1.38 for the odds ratio for sex or the corresponding confidence interval seems incorrect. The confidence interval is based on taking antilogs of endpoints of a confidence interval for the log odds ratio. Show that if the reported confidence interval is correct, then 1.38 is actually the *log* odds ratio, and the estimated odds ratio equals 3.98.

15.31. A Canadian survey of factors associated with whether a person is a hunter of wildlife showed the results in Table 15.27. Explain how to interpret the results in this table. The study abstract[12] stated, "Men are 10 times more likely to hunt wildlife than females." Comment on how this conclusion was reached, and whether it is correct. Which explanatory variables other than sex seem as if they are important?

TABLE 15.27

	Coef.	S.E.	Wald	Sig	Exp(B)
Constant	-5.04	0.16	943.1	.000	
Male	2.34	0.15	259.9	.000	10.39
Live in rural area	0.98	0.10	106.2	.000	2.67
Not married	-0.04	0.12	0.1	.717	.96
Not employed	-0.36	0.12	8.8	.003	.70
Age: 15 to 29	0.21	0.13	2.5	.113	1.24
Age: 50 or more	-0.27	0.12	4.7	.030	.77
Education up to HS	0.38	0.10	14.9	.000	1.46
Naturalist club member	1.64	0.11	228.6	.000	.42

15.32. A study[13] compared the relative frequency of mental health problems of various types among U.S. Army members before deployment to Iraq, U.S. Army members after serving in Iraq, U.S. Army members after serving in Afghanistan, and U.S. Marines after serving in Iraq. The study stated, "Potential differences in demographic factors among the four study groups were controlled for in our analysis with the use of logistic regression." For this study, identify the response variable, the primary explanatory variable, and likely control variables.

15.33. A report (www.oas.samhsa.gov) by the Office of Applied Studies for the Substance Abuse and Mental Health Services Administration about factors that predict marijuana use stated, "Multiple logistic regression also confirmed that the risk of recent marijuana initiation increased with increasing age among youths aged 12 to 14, but the risk decreased with increasing age among those aged 15 to 25." What does this suggest about the way that age appears in the model used?

15.34. Analyze the data in Exercise 8.16 (page 253) on happiness and marital status using a cumulative logit model. Interpret the results in a report of about 200 words.

15.35. For Table 15.4 (page 479), show that the association between the defendant's race and the death penalty verdict satisfies Simpson's paradox. What causes this?

15.36. For a person, let $y = 1$ represent death during the next year and $y = 0$ represent survival. For adults in the United Kingdon and in the United States, the probability of death is well approximated by the model, logit$[P(y = 1)] = -10.5 + 0.1x$, where $x = $ age in years. Show how the probability of death in the next year increases as x increases from 20 to 60 to 100.

15.37. State the symbols for the loglinear models for categorical variables that are implied by the causal diagrams in Figure 15.5.
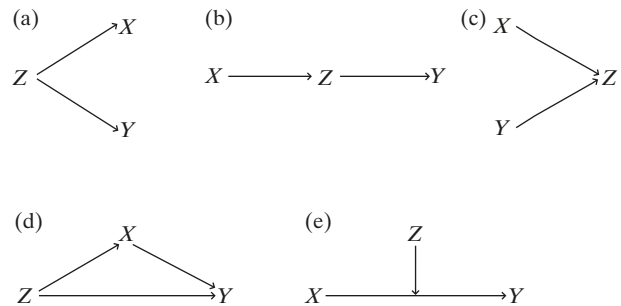
FIGURE 15.5

15.38.* For the logistic regression model, from the linear approximation $\beta/4$ for the rate of change in the probability at the x-value for which $P(y = 1) = 0.50$, show that $1/|\beta|$ is the approximate distance between the x-values at which $P(y = 1) = 1/4$ (or $P(y = 1) = 3/4$) and at which $P(y = 1) = 1/2$. Thus, the larger the value of $|\beta|$, the less the x-distance over which this change in probability occurs.

---

[12] By R. Mitchell, *Crossing Boundaries*, vol. 1 (2001), pp. 107–117.
[13] C. Hoge et al., *New England Journal of Medicine*, vol. 351 (2004), pp. 13–21.

15.39.* For a two-way contingency table, let r_i denote the ith row total, let c_j denote the jth column total, and let n denote the total sample size. Section 8.2 (page 230) stated that the cell in row i and column j has $f_e = r_i c_j / n$ for the independence model. Show that the log of the expected frequency has an additive formula with terms representing the influence of the ith row total, the jth column total, and the sample size. This formula is the loglinear model for independence in two-way contingency tables.

15.40.* Logistic regression has infinite maximum likelihood estimates when the cases with $y = 1$ are separate from the cases with $y = 0$ in the space of explanatory variable values. When this happens, most software merely reports large estimates with huge standard errors. Check what your software does when

(a) $y = 0$ at $x = 1, 2, 3$, but $y = 1$ at $x = 4, 5, 6$.

(b) in a 2×2 table, the numbers of (0, 1) outcomes are (5, 0) for females but (0, 5) for males. (In practice, infinite estimates occur whenever a factor has only $y = 0$ or only $y = 1$ for some category.)

15.41.* Explain what is meant by the absence of statistical interaction in modeling the relationship between a response variable y and two explanatory variables x_1 and x_2 in each of the following cases. Use graphs or tables to illustrate.

(a) y, x_1, and x_2 are quantitative.

(b) y and x_1 are quantitative; x_2 is categorical.

(c) y is quantitative; x_1 and x_2 are categorical.

(d) y, x_1, and x_2 are binary.

(e) y is binary; x_1 and x_2 are quantitative.

This page intentionally left blank

R, Stata, SPSS, and SAS for Statistical Analyses

Major statistical software packages have procedures for the methods presented in this text. This appendix discusses and illustrates the use of R, Stata, SPSS, and SAS for these methods. We deal with basic use of the software rather than the great variety of options provided by the procedures. For ease of reference, the material is organized by chapter of presentation in this text. The full data files for most of the text examples and exercises are available at

`www.pearsonglobaleditions.com/Agresti`.

If you use Stata or SPSS, data files in the respective formats are available at

`www.pearsonglobaleditions.com/Agresti`.

Introduction to R

R is free software maintained and regularly updated by many volunteers. At `www.r-project.org` you can download R and find documentation. The discussion in this appendix refers to R version 3.2.0.

You can get help about R with many books tailored to implementing statistical methods using R and at many sites on the Internet, such as

`www.ats.ucla.edu/stat/R`.

For a particular command, using R you can get help by placing a ? before the name, for example, entering

```
> ?hist
```

for information about the command to construct a histogram. At the end of the help window, you will see an example of the use of the command.

R has a rather steep learning curve, and for use in a basic statistics course, some prefer to use *R Commander*, which is a basic-statistics graphical user interface to R that has a simple menu/dialog-box interface. The menu and dialog-box selections generate R commands. For details and installation notes, see

`http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/`.

Many packages have been created to perform analyses not available in basic R. You can install the package on your computer and then use it. For example, to install the *Rcmdr* (R commander) package, use the command

```
> install.packages("Rcmdr")
```

Once it is installed, to load the package, enter the command

```
> library(Rcmdr)
```

READING DATA FILES AND USING R

This appendix shows R commands for requesting statistical analyses from the command line. A very basic command loads a data file from the text website. To illustrate, to create a data file with the name *Crime* from the data file at the text website that has the name *Crime.dat*, use the command

```
> Crime <- read.table("http://www.stat.ufl.edu/~aa/smss/data/
+                         Crime.dat",  header=TRUE)
```

The *header=TRUE* part of the command tells R that the variable names are at the top of the file.

CHAPTER 2: SAMPLING AND MEASUREMENT

On page 28, we showed how to use R to select a simple random sample, such as

```
-------------------
> sample(1:60, 4)
[1] 22 47 38 44
-------------------
```

to select four people from a population numbered 01 to 60.

CHAPTER 3: DESCRIPTIVE STATISTICS

After loading the data file `Crime`, here is how to request a histogram and a box plot for the violent crime rates in Table 3.2:

```
----------------------------------------------------------------------
> attach(Crime)
> hist(violent)  # histogram
> hist(violent, right=FALSE) # intervals don't include right boundary
> boxplot(violent, xlab="violent", horizontal=TRUE)
----------------------------------------------------------------------
```

The `summary` function provides the five-number summary and the mean. You can also get the mean and standard deviation with the `mean` and `sd` functions:

```
--------------------------------------------------------
> summary(violent)
> mean(violent); sd(violent)
--------------------------------------------------------
```

You can obtain percentiles using the `quantile` function, such as for a variable *y*,

```
> quantile(y, c(.90, .95, .99))
```

for the 90th, 95th, and 99th percentiles.

You can construct scatterplots and find correlations and regression prediction equations using the `plot`, `cor`, and `lm` functions, such as

```
----------------------------------------------------------------
> plot(GDP, Fertility)
> cor(GDP, Fertility); cor(GDP, Internet); cor(GDP, GII)
> lm(Fertility ~ GDP)  # lm is short for 'linear model'
----------------------------------------------------------------
```

CHAPTER 4: PROBABILITY DISTRIBUTIONS

For the normal distribution, the function `pnorm(z)` gives the cumulative probability falling below $\mu + z\sigma$, such as *pnorm(2.0)* for the cumulative probability falling below $\mu + 2.0\sigma$. The function `qnorm(p)` gives the *z*-value for a cumulative probability *p*, such as *qnorm(0.975)* to request the *z*-score for a cumulative probability of 0.975 and a right-tail probability of 0.025.

CHAPTER 5: ESTIMATION

A simple way to construct a confidence interval for the proportion uses the `prop.test` function, such as

```
> prop.test(396, 1200, correct=FALSE)$conf.int
```

for the example on page 119. This confidence interval has a more complex formula than the one in Section 5.2. Called the *score* or *Wilson* confidence interval, Exercise 5.77 explains the idea behind it.

To find the *t*-value from the *t* distribution having a cumulative probability *p*, use the `qt(p, df)` function, such as `qt(0.975, 28)` to find the *t*-score with a right-tail probability of 0.025 when $df = 28$. To find a cumulative probability for a particular *t*-value, use the `pt(t, df)` function.

To construct a confidence interval for the mean, use the `ttest` function, such as for a variable called *y*,

```
> t.test(y, conf.level=0.99)$conf.int
```

For a textbook exercise in which you know n, $\bar{y}$, and s, you can find results yourself, such as for the anorexia example on page 129:

```
----------------------------------------------------------------
> 3.007-qt(0.975,28)*7.309/sqrt(29); 3.007+qt(0.975,28)*7.309/sqrt(29)
[1] 0.2268051
[1] 5.787195
----------------------------------------------------------------
```

CHAPTER 6: SIGNIFICANCE TESTS

To find the *P*-value for a two-sided significance test of whether a variable *y* in the data file has a particular mean of $\mu_0 = 0$,

```
> t.test(y, mu = 0, alternative = "two.sided")$p.value
```

Replace *two-sided* by *greater* or *less* for one-sided alternatives.

For a textbook exercise in which you know n, $\bar{y}$, and s, you can find results yourself, such as for the anorexia example on page 160:

```
------------------------------------------------------------
> t = (3.007 - 0)/(7.309/sqrt(29))
> t
[1] 2.215514
> 2*(1 - pt(t, df=28))  # two-sided P-value when t > 0
[1] 0.03502822
------------------------------------------------------------
```

Here is how to conduct two-sided and one-sided tests for a proportion using summary results, illustrating for the example on page 165:

```
----------------------------------------------------------------------
> prop.test(624,1200,p=0.50, alternative=c("two.sided"), correct=FALSE)
> prop.test(624,1200,p=0.50, alternative=c("greater"), correct=FALSE)
----------------------------------------------------------------------
```

The z test statistic is the square root of the value called "X-squared" on the R output, which is a form of test statistic having a chi-squared null distribution (introduced in Chapter 8) with $df = 1$. The *correct=FALSE* option stops R from using a *continuity correction*, which is not recommended.

CHAPTER 7: COMPARISON OF TWO GROUPS

You can conduct inference comparing proportions by entering the counts in the category of interest and the sample sizes. For example, for Examples 7.2 and 7.3 on the efficacy of prayer,

```
> prop.test(c(315,304),n=c(604,597),conf.level=0.95,correct=FALSE)
```

For a data file with response variable *y* and group variable called *group*, use the command

```
> t.test(y~group,mu=0,alt="two.sided",conf=0.95,var.equal=F,paired=F)
```

to perform a two-sided test and confidence interval comparing means. If *y1* is a variable giving values of a response variable for group 1, and *y2* is a variable for group 2, you can get a confidence interval and test comparing the means by

```
> t.test(y1,y2,alt="two.sided",conf.level=0.95,var.equal=FALSE)
```

In each case, replace *FALSE* by *TRUE* to get the inference that assumes $\sigma_1 = \sigma_2$.

CHAPTER 8: ANALYZING ASSOCIATION BETWEEN CATEGORICAL VARIABLES

With a data file having categorical variables *x* and *y*, you can cross-classify them and obtain the chi-squared statistic using

```
> chisq.test(x,y)
```

If the table is of size 2×2, add `correct=FALSE` to the command so that R does not use the Yates continuity correction. If you already have cell counts, you can proceed as was shown in Table 8.8. To get the standardized residuals, use

```
> chisq.test(x,y)$stdres
```

Use `fisher.test` to perform Fisher's exact test.

The `vcd` package can construct gamma and its standard error, as well as many other measures for contingency tables. See the website

`cran.us.r-project.org/web/packages/vcdExtra/vignettes/vcd-tutorial.pdf`.

CHAPTER 9: LINEAR REGRESSION AND CORRELATION

With a data file having quantitative variables *x* and *y*, you can conduct a basic regression analysis using the `lm` (linear models) function:

```
----------------------------------------------------------------
> plot(x, y) # scatterplot with variable names for x and y
> fit <- lm(y ~ x) # regressing a variable y on a variable x
> summary(fit) #  shows results of the model fit
> cor(x, y) # correlation
----------------------------------------------------------------
```

CHAPTER 11: MULTIPLE REGRESSION AND CORRELATION

Here is how to conduct a multiple regression analysis with the `Mental` data set analyzed in this chapter:

```
----------------------------------------------------------------
> Mental<-read.table("http://www.stat.ufl.edu/~aa/smss/data/
+                     Mental.dat", header=TRUE)
> attach(Mental)
> pairs(~ impair + life + ses)  # scatterplot matrix
> cor(cbind(impair, life, ses)) # correlation matrix
> fit <- lm(impair ~ life + ses) # lm function for regression modeling
> summary(fit) # shows the output
> fit2 <- lm(impair ~ life + ses + life:ses) # permits interaction
----------------------------------------------------------------
```

Put a colon between two variables to represent an interaction term, such as in *fit2*.

The `car` package in R has a function `avPlots` for partial regression (added-variable) plots, with a command of the form

```
> avPlots(fit)
```

applied to the model fit. You can obtain partial correlations with the `ppcor` package. With a command of the form

```
> pcor(datafile)
```

where *datafile* is the name of the data file, it computes the partial correlation between each pair of variables, controlling for the others. Standardized regression coefficients can be found with the `QuantPsyc` package, with a command of the form

```
> lm.beta(fit)
```

applied to the model fit. Alternatively, they can be found by applying the `scale` function to each variable, which standardizes, and using the scaled variables in an ordinary model fit, such as

```
> fit <- lm(scale(impair) ~ scale(life) + scale(ses))
```

CHAPTER 12: REGRESSION WITH CATEGORICAL PREDICTORS: ANALYSIS OF VARIANCE METHODS

To include categorical explanatory variables in a regression model, apply the linear models function by creating factors with dummy variables:

```
> fit <- lm(y ~ factor(x1) + factor(x2))
```

R takes the first category as the baseline that does not have its own dummy variable. If you want to use the last category, then you could create the factor from variable *x*1 with four categories with a command such as

```
> x1fac <- factor(x1, levels = c(4,1,2,3))
```

To conduct ANOVA, use commands such as

```
----------------------------------------------------------------------
> A <- factor(x1); B <- factor(x2) # creates dummy variables
> fit <- aov(y ~ A) # one-way ANOVA
> fit2 <- aov(y ~ A + B) # two-way ANOVA, no interaction
> fit2.int <- aov(y ~ A + B + A:B) # two-way ANOVA, interaction
----------------------------------------------------------------------
```

For the ANOVA fit, Tukey multiple comparisons are available with

```
> TukeyHSD(fit, conf.level = 0.95)
```

To conduct a repeated-measures ANOVA, the data must be in the "long" form in which the repeated measurements are on separate lines of the data file. For example, for Table 12.15 on influences for three types of entertainment, the data file has the form

```
-----------------------------------
  person    type         y
      1       1          -1
      1       2           0
      1       3          -1
 ...

     12       3          -2
-----------------------------------
```

If a data file has all observations for a subject on one row, you can use the `make.rm` command in R to put it in the required form. For a one-way repeated-measures ANOVA, use a command

```
> fit <- aov(y ~ type + Error(person/type))
```

after declaring *type* and *person* as factors, so they are not treated as quantitative. For the anorexia data of Table 12.18 (with *group* as the variable label for the three treatments), the "long" form of the data file is

```
----------------------------------------------------------------------
   person    time     group      y
       1       1         1       80.5
       1       2         1       82.2
 ...
      72       1         3       89.0
      72       2         3       78.8
----------------------------------------------------------------------
```

For two-way ANOVA with repeated measures on a factor such as time and independent samples for different groups, use a command of the form

```
> fit <- aov(y ~ group + time + group:time + Error(person/time))
```

Declare *group*, *time*, and *person* as factors, so they are not treated as quantitative.

CHAPTER 13: MULTIPLE REGRESSION WITH QUANTITATIVE AND CATEGORICAL PREDICTORS

To fit models containing both quantitative and categorical explanatory variables, enter categorical variables as factors, such as

```
> fit <- lm(y ~ x1 + factor(x2))
```

The first category is the baseline without its own dummy variable. Put a colon between variables to create an interaction term, such as

```
> fit <- lm(y ~ x1 + factor(x2) + x1:factor(x2))
```

To construct adjusted means, use the `lsmeans` package or use the `effect` function in the `effects` package, such as

```
---------------------------------------------------------------
> library(effects)
> x2 <- factor(race)
> fit <- lm(y ~ education + x2)
> effect("x2", fit) # finds adjusted means for categories of x2
---------------------------------------------------------------
```

To fit the linear mixed model for clustered data, such as Table 13.13 with the data file containing values for *family*, *y*, *x*1, and *x*2, use the `lme4` package:

```
> fit <- lmer(y ~ x1 + x2 + (1|family))
```

This model assumes compound symmetry structure for correlations within families.

CHAPTER 14: MODEL BUILDING WITH MULTIPLE REGRESSION

Automatic selection methods such as backward elimination and forward selection are available with the `stepAIC` function in the `MASS` library or with the `regsub sets` function in the `leaps` package.

After fitting a model, you can obtain residuals and form a histogram for them and plot them against the predicted values (i.e., fitted values) and find influence diagnostics:

```
----------------------------------------------------------------------
> fit <- lm(y ~ x1 + x2 + x3)
> hist(residuals(fit)) # histogram of residuals
> plot(fitted(fit), residuals(fit), ylab="Residuals", xlab="Predicted y")
> dffits(fit); dfbetas(fit); cooks.distance(fit)
----------------------------------------------------------------------
```

Several packages in R contain functions for finding VIF to assess potential multicollinearity. For example,

```
---------------------------------
> fit <- lm(y ~ x1 + x2 + x3)
> library("car")
> vif(fit)
---------------------------------
```

To fit generalized linear models, use the `glm` function. For instance, to fit a gamma regression model with the identity link, such as done in the text for the home selling price example (page 450), use

```
> fit <- glm(y ~ x1 + x2 + x3, family=Gamma(link = "identity"))
```

To fit a quadratic regression model, you can define a squared term and put it in an ordinary regression model, such as

```
-----------------------------------------------------------------
> x2 = x*x
> fit <- lm(y ~ x + x2)
> summary(fit)
-----------------------------------------------------------------
```

You can fit exponential regression models by fitting the corresponding normal GLM with log link, using the `glm` function, such as

```
> fit <- glm(y ~ x, family = gaussian(link = "log"))
```

CHAPTER 15: LOGISTIC REGRESSION: MODELING CATEGORICAL RESPONSES

You can fit logistic regression models by treating them as generalized linear models using the binomial distribution and logit link. If the data file has a column of 0 and 1 values for a response variable y, use a command such as

```
> fit <- glm(y ~ x1 + x2, family = binomial(link = "logit"))
```

For grouped data, with a column y of numbers of successes and a column n of binomial sample sizes, use a command such as

```
> fit <- glm(y/n ~ x1 + x2, weight=n, family=binomial(link="logit"))
```

To fit a model with probit link, merely change *logit* to *probit* in these statements.

You can fit cumulative logit models to contingency tables using the VGAM package. For Table 15.7, you can enter the cell counts and fit the model as follows:

```
-----------------------------------------------------------------
> party <- c(1,0)
> y1 <- c(16,59); y2 <- c(40,206); y3 <- c(73,112); y4 <- c(330,124);
> y5 <- c(126,18); y6 <- c(167,12); y7 <- c(60,2)
> fit <- vglm(cbind(y1,y2,y3,y4,y5,y6,y7) ~ party,
+             family = cumulative(parallel = TRUE))
> summary(fit)
              Estimate  Std. Error  z value    Pr(>|z|)
party         -2.52727     0.12238  -20.651    <2e-16
-----------------------------------------------------------------
```

The *parallel=TRUE* option requests the same effect for each logit (i.e., proportional odds). Here, $\hat{\beta} = -2.527$ instead of 2.527 because this package uses the parameterization $\text{logit}[P(y \leq j)] = \alpha_j + \beta x$ instead of $\text{logit}[P(y \leq j)] = \alpha_j - \beta x$.

Table 15.11 showed how to use the VGAM package to fit baseline-category logit models, for grouped data in a contingency table.

R can fit loglinear models by regarding them as generalized linear models with response count having a Poisson distribution, using the log link. For example, for a $2\times2\times2$ table such as Table 15.12 constructed from a data file with three columns (x, y, z) of indicators for the variables and a column of cell counts, fit the homogeneous association model as follows:

```
----------------------------------------------------------------
> A <- factor(x); B <- factor(y); C <- factor(z)
> fit <- glm(count ~ A+B+C+A:B+A:C+B:C, family=poisson(link="log"))
----------------------------------------------------------------
```

Introduction to Stata

Basic support information is available from Stata at

`www.stata.com/support`.

Many Internet sites can help you learn how to use Stata, such as the many resources listed at

`www.stata.com/links/resources-for-learning-stata`.

Also, examples are shown for a previous edition of this textbook at

`www.ats.ucla.edu/stat/stata/examples/smss/`.

These tutorials and the discussion below show commands to enter to perform various statistical analyses. Commands are case-sensitive. To get information about a command, use the `help` command, such as

`help histogram`

for help about the command for forming histograms. Click on the first entry in the help file to get much further information from the extensive documentation that Stata includes with every installation. This documentation is also available at the *Help* menu that Stata provides. In this documentation, for each command you will find a large number of options that we do not have space to present in this appendix.

Rather than entering commands in Stata, you can use the *Statistics* and *Graphics* menus. Once you select a particular topic and subtopic from a menu, you get a dialog box in which you can select the particular analyses you would like.

READING DATA FILES AND USING STATA

After starting Stata, it is helpful to create a log file that keeps a record of the commands you enter and the output. To do this, use a command such as

`log using exampleoutput`

which will create this file at the directory Stata tells you.

You can enter data or access a data file in various ways. See, for example,

`www.stata.com/manuals14/u21.pdf`.

The text website has Stata data files (with extension *.dta*) for most examples and data exercises. For example, to load the *Crime* data file that is used extensively in Chapter 3, you can enter the command

`use "http://www.stat.ufl.edu/~aa/smss/data/Stata/Crime.dta"`

CHAPTER 3: DESCRIPTIVE STATISTICS

To form a histogram of a variable named *y*, use the command

`histogram y`

To construct a stem-and-leaf plot, use the command `stem` followed by the variable name.

Typing the command `codebook` shows many summary statistics for each variable in the data file. You can obtain basic descriptive statistics for a variable named *y* with the command

```
summarize y, detail
```

To find the median and other percentiles, use the `centile` command. For example, for a variable called y, we get the quartiles by

```
centile y, centile(25, 50, 75)
```

Construct a box plot with the `graph box` command followed by the variable name.

You can find correlations for each pair of a set of variables with the `corr` command or the `pwcorr` command, such as

```
corr GDP GII Fertility
```

You can find the prediction equation for a regression analysis with the `regress` command, such as

```
regress Fertility GDP
```

to predict Fertility using GDP as the explanatory variable.

CHAPTER 4: PROBABILITY DISTRIBUTIONS

To find a normal cumulative probability for a particular *z*-value, use the `display normal(z)` command, such as

```
display normal(2.0)
```

to find the probability falling below 2.0 for a standard normal curve. To find the *z*-value having a cumulative probability *p*, use the `display invnormal(p)` command, such as

```
display invnormal(0.975)
```

to find the *z*-value having a cumulative probability of 0.975 and thus a right-tail probability of 0.025.

CHAPTER 5: ESTIMATION

To construct confidence intervals for mean and proportions, you can use dialog boxes or use the `ci` command. For example, for the mean of a variable called *y*,

```
ci y
```

If you have only summary statistics, Stata can construct the interval using them. You can use the `cii` command,[1] by entering n, $\bar{y}$, and s (here, $n = 29$, $\bar{y} = 3.007$, $s = 7.309$):

```
cii means 29 3.007 7.309
```

Or, you can launch a dialog box and fill in what you want, using

```
db cii
```

---

[1] Here, *i* following *ci* stands for *immediate*.

or in the Statistics menu, select *Summaries, tables, and tests* and then *Classical tests of hypotheses* and then *t test calculator*.

For the confidence interval for a proportion for a binary variable *y* that is a column in the data file that takes the values 0 and 1,

```
ci proportions y, wald
```

Or, you can find it directly from the sample size and count in the category of interest, such as

```
cii proportions 1200 396, wald
```

for the example in the text with 396 people out of $n = 1200$ sampled who favored restricting access to abortion.

The `mean` command also provides a confidence interval for the mean. For example, for the mean of a variable called *y*,

```
mean y
```

To find the *t*-value having a cumulative probability *p*, use the `display invt(df, p)` command, such as

```
display invt(28, 0.975)
```

to find the *t*-value having a cumulative probability of 0.975 and thus a right-tail probability of 0.025 when $df = 28$. To find a cumulative probability for a particular *t*-value, use the `display tprob(df, t)` command, such as

```
display tprob(28, 2.0)
```

to find the probability falling below 2.0 for a *t* distribution with $df = 28$.

For information on using Stata for the bootstrap, see

```
www.stata.com/features/overview/bootstrap-sampling-and-
estimation.
```

For examples, see Example 5.9 on page 141.

Explicit confidence intervals can be formed for the median (e.g., Exercise 5.79) that do not require the bootstrap. Use the `centile` function as shown above.

CHAPTER 6: SIGNIFICANCE TESTS

To conduct a *t* test of whether a variable *y* in the data file has a mean of 0, use the `ttest` command:

```
ttest y == 0
```

If you already have summary statistics, you can use the `ttesti` command, by entering n, $\bar{y}$, s, and μ_0, such as for the text anorexia example:

```
ttesti 29 3.007 7.309 0
```

To conduct a significance test of whether a categorical variable *y* that takes values 0 and 1 in the data file has a population proportion of 0.50 that take the value 1:

```
prtest y == 0.50
```

If you already have summary statistics, you can use the `prtesti` command, by entering n, $\hat{\pi}$, and π_0, such as for the example on page 165:

```
prtesti 1200 0.52 0.50
```

To find a right-tail probability for a particular t-value with a certain df value, such as to find a one-sided P-value, use

```
display tprob(df, t)
```

to find the cumulative probability, and then subtract from 1.

CHAPTER 7: COMPARISON OF TWO GROUPS

Stata can construct confidence intervals and tests comparing two proportions using the command `prtest`. To test equality of proportions for a variable y between two groups defined by a variable called *group* (such as gender), you can use

```
prtest y, by(group)
```

If you have summary statistics, you can find the inferences directly from the sample size and proportion in the category of interest for each group, such as

```
prtesti 604 0.522 597 0.509
```

for the efficacy of prayer example on pages 194, 196, and 198.

To construct inference for means, use the `ttest` command. For example, to test that the mean of a variable called y is equal between two groups defined by a categorical variable called *group*, use

```
ttest y, by(group)
```

to use the method of Section 7.5 that assumes $\sigma_1 = \sigma_2$. Use

```
ttest y, by(group) unequal
```

to allow unequal population standard deviations as in Section 7.3. The commands also yield confidence intervals comparing the group means.

If you already have summary statistics, you can conduct the inferences with the `ttesti` command, by entering n, $\bar{y}$, and s for each group, such as

```
ttesti 583 8.3 9.4 693 11.9 12.7, unequal
```

for the housework example on pages 200 and 201 in the text.

For the paired-difference t analyses with matched-pairs data in variables called $y1$ and $y2$, use

```
ttest y1 == y2
```

Alternatively, you can create a new variable of difference scores, and use the t methods described for Chapters 5 and 6. When $y1$ and $y2$ are binary, you can get McNemar's test using

```
mcc y1 y2
```

The output shows a *chi-squared statistic* that is the square of the z statistic we present in the text. The P-value for the chi-squared test is the two-sided P-value for the z statistic. Using the summary counts in the contingency table that cross-classifies $y1$ and $y2$, you can get McNemar's test for the example on page 210 using

```
mcci 875 162 9 168
```

For two categorical variables $y1$ and $y2$, you can construct a contingency table and perform Fisher's exact test using the command

```
tab y1 y2, exact
```

You can also enter the counts yourself from the contingency table that cross-classifies *y*1 and *y*2 and request this test. For the example on page 212,

```
tabi 10 18 \ 1 22, exact
```

To conduct the Wilcoxon test with a response variable *y* and groups defined by a variable *x*, use

```
ranksum y, by(x) porder
```

The *porder* option requests an estimate of the probability that one group is higher than the other.

CHAPTER 8: ANALYZING ASSOCIATION BETWEEN CATEGORICAL VARIABLES

With the `tabulate` command (`tab` for short), you can construct contingency tables, find percentages in the conditional distributions (within-row relative frequencies), get expected frequencies for H_0: independence, get the chi-squared statistic and its *P*-value, and conduct Fisher's exact test. For categorical variables *x* and *y* in a data file, for instance, you can use

```
tab x y, row expected chi2 exact gamma
```

If you already have the cell counts, you can enter them by row. For the example on page 234, use

```
tabi 495 590 272 \ 330 498 265, row expected chi2 exact gamma
```

To get standardized residuals, you currently must download a routine written by Nicholas Cox. Within Stata, use the command

```
ssc install tab_chi
```

then followed (if you have the cell counts) by

```
tabchii 495 590 272 \ 330 498 265, adjust
```

to get the standardized (adjusted) residuals.

CHAPTER 9: LINEAR REGRESSION AND CORRELATION

You can conduct a basic linear regression analysis for a response variable *y* and explanatory variable *x* with the `regress` command

```
regress y x
```

For a scatterplot, use

```
scatter y x
```

For the correlation for variables *x* and *y*, use

```
corr y x
```

or use `pwcorr` to have additional options such as *P*-values. List several variables, and you get a correlation matrix. To get a confidence interval for the correlation, use the package `corrci`:

```
ssc install corrci
corrci x y
```

CHAPTER 11: MULTIPLE REGRESSION AND CORRELATION

For a scatterplot matrix, use the `graph matrix` command, such as

```
graph matrix impair life ses
```

for the `Mental` data set at the text website that is analyzed in this chapter. To construct a correlation matrix, use the command of the form

```
corr w x y z
```

entering the variable names. You can conduct a multiple regression analysis for a response variable *y* and explanatory variables with the command of the form

```
regress y x1 x2 x3
```

For a partial regression plot of the response variable *y* against each explanatory variable, follow this by the command `avplots` (*Added-variable plot* is an alternative name for partial regression plot).

To include an interaction term in a model, you can put a # between the variables and add the prefix c to represent that the variable is continuous (actually, quantitative), such as

```
regress y x1 x2 c.x1#c.x2 x3
```

Alternatively, a command such as

```
regress y c.x1##c.x2
```

includes the interaction and the corresponding main effects.

To get partial correlations (and semipartial correlations) of *y* with each explanatory variable, controlling for the others in the model, use a command of the form

```
pcorr y x1 x2 x3
```

To obtain the standardized regression coefficients, use

```
regress impair life ses, beta
```

The name, and the heading *Beta* in the output, reflects the alternative name *beta weights* for these coefficients.

CHAPTER 12: REGRESSION WITH CATEGORICAL PREDICTORS: ANALYSIS OF VARIANCE METHODS

To use the `regress` function with a categorical variable, declaring it to be a factor using the i. prefix to create indicator (dummy) variables, such as for a variable called *party*,

```
regress y i.party
```

The first category is deleted for the dummy variables. To instead use category 3 for the base category, for instance, enter the categorical variable as *b3.party*.

For pairwise multiple comparisons of means for a factor called *party*, follow the `regress` command by

```
pwcompare party, mcompare(bonferroni)
```

substituting *tukey* for *bonferroni* to get the less conservative Tukey intervals.

To conduct a one-way ANOVA with a response variable *y* and a factor *A*, use the command

```
anova y A
```

The variable *A* is assumed to be categorical. You can also do one-way ANOVA with the `oneway` command,

```
oneway y A
```

For a two-way ANOVA with response variable *y* and factors *A* and *B*, without interaction, use

```
anova y A B
```

To allow for interaction, use

```
anova y A B A#B
```

or simply

```
anova y A##B
```

Entering `regress` after requesting an ANOVA fit yields the model fit for the corresponding regression model with dummy variables.

Alternatively, you can do a factorial ANOVA by applying the `regress` function to the factors, declaring them to be factors using the i. prefix to create indicator (dummy) variables, such as

```
regress y i.A i.B
```

To conduct a repeated-measures ANOVA, the data must be in the "long" form with the repeated measurements on separate lines of the data file, as shown above on page 523 for R software. If a data file has all observations for a subject on one row, you can use the `reshape` command in Stata to put it in the required form. For example, if a row of the data file showed all the observations for a particular person, with variable labels *trt*, *y1*, and *y2*, then use the command

```
reshape long y, i(person) j(time)
```

The one-way repeated-measures analysis of Section 12.5 for types of entertainment is obtained by

```
anova y person type, repeated(type)
```

The two-way analysis of Section 12.6, for "long" data file as shown above in the R section, is obtained by

```
anova y group / person|group time group#time, repeated(time)
```

CHAPTER 13: MULTIPLE REGRESSION WITH QUANTITATIVE AND CATEGORICAL PREDICTORS

Stata can fit regression models having both quantitative and categorical explanatory variables using the `regress` function. Prefix a categorical factor with i. to specify indicators for each category of the variable, such as

```
regress y education i.race
```

By default, Stata deletes the first category for the dummy variables.

To interact a quantitative variable with a categorical factor, prefix the quantitative variable with c. (for continuous), such as

```
regress y education i.race c.education#i.race
```

Having fitted a model with no interaction, you can obtain adjusted means by

```
margins i.race, at( (mean) education)
```

To fit the linear mixed model for the clustered family data of Table 13.13, use the command

```
mixed y x1 x2 || family:, residuals(ex, t(family)) reml
```

or else

```
mixed y x1 x2 || family:, covariance(exchangeable) reml
```

both of which yield an exchangeable (compound symmetry) structure for correlations within families.

CHAPTER 14: MODEL BUILDING WITH MULTIPLE REGRESSION

You can conduct automatic variable selection methods using the **stepwise** command. For backward elimination, with 0.10 as the α-level in tests, use a command such as (with five potential explanatory variables)

```
stepwise, pr(0.10): regress y x1 x2 x3 x4 x5
```

where *pr* stands for the probability needed to be exceeded for removal. For forward selection, use

```
stepwise, pe(0.10): regress y x1 x2 x3 x4 x5
```

where *pe* stands for the probability needed to be below to be eligible for addition. The command

```
stepwise, pr(0.10) pe(0.10) forward: regress y x1 x2 x3 x4 x5
```

uses the stepwise variation of forward selection that removes a previously entered term if it is no longer significant.

After fitting a model with the **regress** command, to obtain the residuals and plot them against the model's fitted values, use

```
rvfplot, yline(0)
```

(Here, *rvf* stands for *residual-versus-fitted* plot.) Use **rvpplot** to plot them against a predictor *x*,

```
rvpplot x, yline(0)
```

Use the **predict** command with the **rstudent** option to generate the studentized residuals. Here, we name them *r* and then form a histogram and plot them against a predictor.

```
------------------------
. predict r, rstudent
. histogram r
. scatter r x1
------------------------
```

With the `dfbeta` and `dfits` commands, Stata will form DFBETA values for all the model parameters and DFFIT values for all the observations. Use `dfbeta(x1)` with the variable name in parentheses to inspect DFBETA values for a particular parameter.

After fitting a model, to assess multicollinearity you can obtain VIF values with the command

```
vif
```

To fit GLMs, use the `glm` command. For instance, to fit a gamma regression model with the identity link, use a command such as

```
glm y  x1  x2, family(gamma) link(identity)
```

To fit a quadratic regression model, you can use the command

```
regress y x c.x#c.x
```

You can fit exponential regression models by fitting the normal GLM with log link, using the `glm` command, such as

```
glm y  x, family(gaussian) link(log)
```

CHAPTER 15: LOGISTIC REGRESSION: MODELING CATEGORICAL RESPONSES

Stata can fit logistic regression models with the `logit` command, for which the standard output is the model parameter estimates, or the `logistic` command, for which the standard output is the odds ratios obtained by exponentiating the estimates. For example, the `logit` command for a binary response variable with three explanatory variables is

```
logit y x1 x2 x3
```

Adding the *or* option to this command requests the odds ratio form of estimate. Stata can also fit the model with the `glm` command, treating the model as a generalized linear model for a binomial distribution with logit link, such as

```
glm evolved ideology, family(binomial) link(logit)
```

If the data are counts in a contingency table, and each row of the data file has a value for each explanatory variable, the 0 or the 1 value for *y*, and a variable (say, called *count*) containing the cell counts, you can use the command

```
logit y x1 x2 x3 [fweight = count]
```

Here, *fweight = count* indicates that the data file has data grouped according to the variable called *count*.

To do a likelihood-ratio test about an individual explanatory variable, store the results for the full model, fit the simpler model without that variable, and then request the likelihood-ratio test comparing the models. For example, to test the effect of defendant's race for the death penalty data of Table 15.3,

```
-----------------------------------------------------------------
. logit y d v [fweight = count]
. estimates store full
. logit y v [fweight = count]
. lrtest full
-----------------------------------------------------------------
```

A command such as

```
test race
```

conducts the Wald test about an explanatory variable (i.e., the square of the *z* test statistic), which is not as reliable a test as the likelihood-ratio test.

Linear probability models can be fitted with the `glm` command, treating the model as a generalized linear model for a binomial distribution with identity link, such as

```
glm evolved ideology, family(binomial) link(identity)
```

Probit models are fitted like logistic regression models, merely substituting *probit* for *logit* in the command. Propensity-score matching is obtained with the command `teffects psmatch`, such as

```
teffects psmatch (y) (group x1 x2 x3 x4)
```

to compare two groups identified by the variable *group* in their response on *y* after using logistic regression to get propensity scores for predicting *group* using *x*1, *x*2, *x*3, and *x*4.

Stata fits the cumulative logit model with the `ologit` (ordinal logit) command, such as

```
ologit response party
```

If the data file contains grouped data (i.e., cell counts in the response categories), such as columns labeled *party* (a 1/0 indicator), *response* (giving the response category), and *count*, fit the model with the command

```
ologit response party [fweight = count]
```

Stata fits the baseline-category logit model with the `mlogit` (multinomial logit) command, such as

```
mlogit response sex race, base(3)
```

where *base(3)* indicates the baseline category for the logits. If the data file contains grouped data (i.e., cell counts in the response categories), such as columns labeled *race*, *sex*, *response* (giving the response category), and *count*, fit the model with the command

```
mlogit response sex race [fweight = count], base(3)
```

Stata can fit loglinear models by regarding them as generalized linear models with response count having a Poisson distribution, using the log link. For example, for a $2\times2\times2$ table such as Table 15.12 constructed from a data file with three columns of levels for the variables and a column of cell counts, fit the homogeneous association model with the command

```
glm count i.a i.c i.m i.a#i.c i.a#i.m i.c#i.m, family(poisson)
                                                   link(log)
```

Introduction to SPSS

SPSS has a windows-with-menus structure that makes requesting statistical procedures simple. Our discussion below applies to version 23. It can help to look at the online manual at

```
www.spss-tutorials.com
```

and view sites such as

`www.ats.ucla.edu/stat/spss`

that have sample analyses.

Various versions of SPSS have student discounts. A *Base Grad Pack* has basic statistical analyses, including basic regression and contingency table analysis. A *Standard Grad Pack* has some advanced methods such as for repeated measures and MANOVA, logistic regression and the ordinal extension and loglinear models, and generalized linear models such as gamma regression. A *Premium Grad Pack* is more advanced yet, having capabilities such as handling missing data and doing some exact small-sample analyses for contingency tables.

READING DATA FILES AND USING SPSS

When you start a session, you see a *Data Editor* window that contains a menu bar with a wide variety of separate menus. These include a FILE menu for creating a new file or opening an existing one, an ANALYZE menu that displays options for selecting a statistical method, a GRAPHS menu for creating a graph of some type, and menus for choosing other special features. The *Data Editor* window displays the contents of the data file. The *Output* window shows results of the analyses after you request them from the menus. You can edit and save this for later use or printing. The text website has SPSS data files (which have extensions *.sav*) for most text examples. For example, to load the *Crime* data file that is used extensively in Chapter 3, click the FILE menu and choose the OPEN option and INTERNET DATA suboption. Enter the web location and file name

`www.stat.ufl.edu/~aa/smss/data/SPSS/Crime.sav`.

Select files of type *SPSS* and click on *ok*.

In the *Variable View* of the *Data Editor* window, SPSS should identify quantitative variables as NUMERIC and categorical variables (with labels for the categories) as STRING. You can re-define names and characteristics for each variable. In the Measure column, make sure SPSS has not inappropriately labeled a variable as NOMINAL that should be SCALE (interval) or ORDINAL.

You can select a statistical procedure from the ANALYZE menu on the *Data Editor*. When you do so, a *dialog box* opens that shows you the source variables in your data set. You highlight the ones you want to use currently and click on the arrow to the right of the list to move them to the selected variables list further to the right. You then click on OK and the procedure runs, showing results in the output window. For many procedures, you can click on *Options* and an additional *subdialog box* will open that displays extra available options for the method.

CHAPTER 3: DESCRIPTIVE STATISTICS

To construct frequency distributions, histograms, and basic summary statistics, on the ANALYZE menu select the DESCRIPTIVE STATISTICS option with the FREQUENCIES suboption. A FREQUENCIES dialog box will open. Select the variables you want from the list for your file. Then, clicking on OK provides a frequency distribution in the *Output* window. Clicking on CHARTS in the FREQUENCIES dialog box presents you with a FREQUENCIES: CHARTS dialog box containing a histogram option for quantitative variables and a bar chart option for categorical variables. You can also construct a histogram from the GRAPHS menu on the *Data Editor* window by selecting CHART BUILDER and then HISTOGRAM. To construct a stem-and-leaf plot, from the DESCRIPTIVE STATISTICS option in the

ANALYZE menu, select the EXPLORE suboption. The EXPLORE dialog box contains a *Plots* option; clicking on it reveals stem-and-leaf and histogram plot options.

To obtain basic measures of central tendency, variability, and position, select the DESCRIPTIVE STATISTICS option with the FREQUENCIES suboption, and click on STATISTICS, which presents you with a FREQUENCIES: STATISTICS dialog box containing options.

To construct a box plot, on the GRAPHS menu in the *Data Editor* window select CHART BUILDER and then select BOXPLOT and drag the box plot icon into the open canvas. Select the variable for the box plot and click OK. CHART BUILDER also has options for side-by-side box plots according to the value of a categorical variable.

To obtain correlation and regression results, on the ANALYZE menu select the REGRESSION option with the LINEAR suboption. You will then see a LINEAR REGRESSION dialog box in which you can identify response (dependent) and explanatory (independent) variables. To construct a scatterplot, enter the GRAPH menu and select REGRESSION VARIABLE PLOTS, which has the option of showing the prediction equation line over the plot. Or, you can select CHART BUILDER, and in the dialog box, drag the appropriate variables to the *y* and *x* axes.

CHAPTERS 5 AND 6: ESTIMATION AND SIGNIFICANCE TESTS

The ANALYZE menu has a COMPARE MEANS option with a ONE-SAMPLE T TEST suboption. The default with that option is a 95% confidence interval for the mean and a two-sided *t* test that the true mean equals 0. The options permit you to select a different confidence level. To test that the mean equals a constant μ_0, put that number in the Test Value box on the ONE-SAMPLE T TEST dialog box. Options also allow you to use the bootstrap to obtain inference for the mean.

CHAPTER 7: COMPARISON OF TWO GROUPS

The ANALYZE menu has a COMPARE MEANS option with an INDEPENDENT-SAMPLES T TEST suboption. Select the response variable (labeled the *Test variable*) and the variable that defines the two groups to be compared (labeled the *Grouping variable*), which can be a numeric or a string variable. With *Define Groups* under the *Grouping Variable* label, identify the two levels of the grouping variable that specify the groups to be compared. In the *Output* window, in the *Equal variances* row, this procedure provides the results of the *t* test assuming the two populations have $\sigma_1 = \sigma_2$. The procedure also provides the 95% confidence interval for comparing the means. The output also shows results for the method that does not assume equal variances, in the *Unequal variances* row.

The COMPARE MEANS option also has a SUMMARY INDEPENDENT-SAMPLES T TEST suboption, in which you can input n, $\bar{y}$, and s for each group, and then get results of the significance test and confidence interval. The COMPARE MEANS option also has a PAIRED-SAMPLES T TEST suboption, which supplies the dependent-samples comparisons of means. For Fisher's exact test, see the description for the following chapter.

CHAPTER 8: ANALYZING ASSOCIATION BETWEEN CATEGORICAL VARIABLES

The DESCRIPTIVE STATISTICS option on the ANALYZE menu has a suboption called CROSSTABS, which provides several methods for contingency tables.

After identifying the row and column variables in CROSSTABS, clicking on STATISTICS provides a wide variety of options, including the chi-squared test. The output lists the Pearson chi-squared statistic (X^2), its degrees of freedom, and its *P*-value (labeled *Asymptotic Significance*). It also reports an alternative test statistic, called the *likelihood-ratio* statistic. Chapter 15 (page 497) introduces this statistic. You can request Fisher's exact test by clicking on EXACT in the CROSSTABS dialog box and selecting the exact test.

In CROSSTABS, clicking on CELLS provides options for displaying observed and expected frequencies, as well as the standardized residuals, labeled as *Adjusted standardized*.

In CROSSTABS, the options in STATISTICS include measures of association. One option, labeled *Risk*, provides as output for 2×2 tables the odds ratio and its confidence interval. For the ordinal measure gamma, the output includes a test statistic that the true measure equals zero, which is the ratio of the estimate to its standard error. This test uses a simpler standard error that only applies under independence and is inappropriate for confidence intervals.

Suppose you enter the data as cell counts for the various combinations of the two variables, rather than as responses on the two variables for individual subjects; for instance, perhaps you call COUNT the variable that contains these counts. Then, select the WEIGHT CASES option on the DATA menu in the *Data Editor* window, which instructs SPSS to weight cases by COUNT.

CHAPTER 9: LINEAR REGRESSION AND CORRELATION

To construct a scatterplot, enter the GRAPH menu and select REGRESSION VARIABLE PLOTS, which has the option of showing the prediction equation line over the plot. Or, you can select CHART BUILDER, and in the dialog box, drag the appropriate variables to the *y* and *x* axes.

To fit the regression line, on the ANALYZE menu select REGRESSION and then LINEAR. You identify the response (Dependent) variable and the explanatory (Independent) variable. Various options are available by clicking on *Statistics* in the LINEAR REGRESSION dialog box, including estimates of the model parameters, confidence intervals for the parameters, and model fit statistics. After selecting what you want, click on CONTINUE and then back in the LINEAR REGRESSION dialog box click on OK. Output for the *Estimates* option includes the estimates for the prediction equation (labeled *B*), their standard errors, the *t* statistic for testing that a regression parameter equals 0 and the associated *P*-value (labeled *Sig*), and a standardized regression coefficient (labeled as *Beta*) that in this bivariate model is merely the correlation.

Output for the model fit option in a *Model Summary* table includes the correlation (labeled as *R*), the r^2-value, and the estimate *s* of the conditional standard deviation (labeled *Std. Error of the Estimate*). In the LINEAR REGRESSION box, when you click on *Save* you can request unstandardized predicted values and residuals and studentized residuals. After running the regression, they appear in your saved data file (with .sav extension).

CHAPTER 11: MULTIPLE REGRESSION AND CORRELATION

For a multiple regression analysis, choose REGRESSION from the ANALYZE menu with the LINEAR suboption, and add additional variables to the list of independent variables. Among the options provided by clicking on *Statistics* in the dialog

box are estimates of the coefficients and confidence intervals based on them and detail about the model fit. For the *Estimates* option, the output includes standard errors of the estimates, the *t* statistic for testing that the regression parameter equals zero and its associated two-sided *P*-value, and the estimated standardized regression coefficient (labeled *Beta*).

Requesting the *Model fit* option in the STATISTICS sub-dialog box provides additional information. For instance, *F* in the *ANOVA* table is the test statistic for the hypothesis that the coefficients of the explanatory variables all equal 0. Also provided in a *Model Summary* table are the multiple correlation R, R^2, and the estimate *s* of the conditional standard deviation (labeled as *Std. Error of the Estimate*).

In the ANALYZE menu, selecting CORRELATE and then BIVARIATE gives a BIVARIATE CORRELATIONS sub-dialog box. You select the variables you want, check PEARSON CORRELATION, and then click OK, and the output window shows a correlation matrix with the *P*-values for testing the significance of each.

To construct a scatterplot matrix, from the GRAPHS menu in the *Data Editor* choose the CHART BUILDER option. Then click the GALLERY tab and select SCATTER/DOT in the *Choose From* list. Drag the Scatterplot Matrix icon onto the blank canvas. Drag the wanted variables to the Scattermatrix drop zone, and then click on OK. You'll see the graph in the *Output* window.

To produce all partial regression plots, click on PLOTS in the LINEAR REGRESSION dialog window for the REGRESSION option and LINEAR suboption and then click on *Produce all partial plots* in the LINEAR REGRESSION: PLOTS dialog box.

To obtain a partial correlation analysis, choose the PART AND PARTIAL CORRELATIONS option in the STATISTICS option box in the LINEAR REGRESSION window for the REGRESSION option and LINEAR suboption. Or, in the ANALYZE menu choose the CORRELATE option with the PARTIAL suboption. In the resulting PARTIAL CORRELATIONS dialog box, select the variables to correlate and select at least one variable to control.

To model interaction, you can construct an interaction variable within the SPSS data editor by selecting the COMPUTE VARIABLE option on the TRANSFORM menu. Provide a name for the new variable in the Target Variable box. Create the mathematical formula for the interaction term in the Numeric Expressions box, such as LIFE*SES for the explanatory variables LIFE and SES (the * symbol represents multiplication). Click OK, and in the data file in the *Data Editor* you will see a new column of observations for the new variable. This variable can then be entered into the model formula when requesting a regression equation.

Here's a second way to build an interaction term in a model, one that is especially useful for models in following chapters that also have categorical predictors. This second method is well suited for forming multiple interaction terms but presents output in a slightly different form and offers fewer options for data analysis. Choose the GENERAL LINEAR MODEL in the ANALYZE menu and select the UNI-VARIATE suboption. Enter the response variable into the Dependent Variable box and the explanatory variables into the Covariate(s) box. Now click on the *Model* box and select the *Custom* option. Using the Build Term(s) arrow, enter the covariates as *Main effects*. Highlight a pair of variables for which you want a cross product and enter them by selecting *Interaction* on the *Build Term(s)* arrow. Or, you can select the *All 2-way* option for the *Build Term(s)* arrow to request interaction terms for all pairs of variables. After specifying the terms for the model, click *Continue* and return to the UNIVARIATE dialog box. To display model parameter estimates, select the *Options* box and check the *Parameter Estimates* option. Click *Continue* to return to the UNIVARIATE dialog box and then click *OK* to perform the regression analysis.

CHAPTER 12: REGRESSION WITH CATEGORICAL PREDICTORS: ANALYSIS OF VARIANCE METHODS

To conduct a one-way ANOVA, on the ANALYZE menu select the COMPARE MEANS option with the ONE-WAY ANOVA suboption. Select the dependent variable and select the factor that defines the groups to be compared. (This must be coded numerically for SPSS to display it as a potential factor, even though it is treated as nominal scale. Otherwise, use the approach in the following paragraph.) Results provided include the *F* test statistic and its *P*-value, and sums of squares and mean squares for between-groups and within-groups variation. Clicking on *Post Hoc* in the ONE-WAY ANOVA dialog box provides a variety of options for multiple comparison procedures, including the Bonferroni and Tukey methods. The LSD (least significant difference) option provides ordinary confidence intervals with the confidence level applying to each interval. Clicking on *Options* in the ONE-WAY ANOVA dialog box provides the *Descriptive statistics* option of additional descriptive statistics, including the mean, standard deviation, standard error, and a 95% confidence interval for each group.

You can also conduct a one-way ANOVA on the ANALYZE menu by selecting the GENERAL LINEAR MODEL option with the UNIVARIATE suboption. With this approach, the categorical variable that is selected as the Fixed Factor can be coded with labels rather than numerically (i.e., a *string* variable in SPSS). In the UNIVARIATE dialog box, click on *Options* and you can request Descriptive statistics and Parameter estimates for displaying the regression parameter estimates from viewing the analysis as a special case of a regression analysis. Return to the UNIVARIATE dialog box and click on *Post Hoc* to select ordinary confidence intervals for comparing means (LSD) or multiple comparison intervals such as Bonferrroni or Tukey.

To conduct a two-way or higher-way factorial ANOVA, on the ANALYZE menu select the GENERAL LINEAR MODEL option with the UNIVARIATE suboption. Select the dependent variable and select the Fixed Factor(s) that define the cross-classification for the means. (If you have set up dummy variables yourself, they would be entered as Covariates.) The default model is a full factorial model containing all interactions. Click on *Model* to build a customized model that contains only some or none of the interactions. Highlight variables, select *Interaction* or *Main Effects* from the Build Term(s) list, and click on the arrow to move the terms to the model list on the right. Return to the UNIVARIATE dialog box and click on *Options*. You can request Descriptive statistics, Parameter estimates, and you can select particular factors and request Display Means to see the observed and predicted means for subgroups defined by the factors. Return to the UNIVARIATE dialog box and click on *Contrasts* to display parameter estimates with standard errors, *t* statistics, and confidence intervals for comparing means for levels of each factor. Change the contrast type to *Simple* to compare each level to a baseline level, either the last (such as in setting up (1, 0) dummy variables for all categories but the last one) or the first. Return to the UNIVARIATE dialog box and click on *Post Hoc* to select confidence intervals for comparing means (LSD) or multiple comparison intervals such as Bonferrroni or Tukey.

Alternatively for ANOVA, you could set up dummy variables in your data file and then use ordinary regression. On the ANALYZE menu, you would then select the REGRESSION option and LINEAR suboption, as in Chapter 11.

You can conduct repeated-measures ANOVA using the GENERAL LINEAR MODEL option on the ANALYZE menu, with the REPEATED MEASURES suboption. This assumes that for each subject, the data file has the "short" form in which each outcome for the response falls in a different column. For Example 12.8 on three

influences, in a given row you would put the response for Movies in one column, for TV in a separate column, and for Rock in a third column. In the REPEATED MEASURES DEFINE FACTOR(S) dialog window, type the name and number of levels of the within-subjects factor (such as *influence* and 3) and click on *Add*. Then click on *Define* to define the model. Now, in the REPEATED MEASURES dialog box, select the between-subjects factors (if there are any), and select the response variable for each level of the within-subjects factor (such as Movies, TV, Rock). The default is a model containing all the factor interactions. Click on *Model*, and customize the model if you want to delete an interaction. Return to the REPEATED MEASURES dialog box and click on *Contrasts*, and options are provided for displaying parameter estimates and confidence intervals for contrasts comparing means in different factor levels, and for individual or Bonferroni confidence intervals. Change the contrast type to *Simple* for estimates of the between-subjects factors to refer to comparing each factor level to the first or last level. Return to the REPEATED MEASURES dialog box and click on *Options*, and you can request between-subjects observed and estimated means and various model diagnostics.

For repeated-measures analyses, SPSS also reports results of standard multivariate (MANOVA) tests that do not make the assumption of sphericity for the joint distribution of the repeated responses (see Section 16.1). They are less powerful than the repeated-measures ANOVA methods when the sphericity assumption is not violated.

CHAPTER 13: MULTIPLE REGRESSION WITH QUANTITATIVE AND CATEGORICAL PREDICTORS

To fit an analysis of covariance model, you can set up dummy variables for categorical predictors and use ordinary regression procedures, such as described earlier for Chapter 11. To create cross-product terms for interactions, after creating the data file, you can select COMPUTE VARIABLE on the TRANSFORM menu and create products of appropriate variables.

Alternatively, on the ANALYZE menu select the GENERAL LINEAR MODEL option with the UNIVARIATE suboption. Proceed as described above for Chapter 11, now adding quantitative covariates in the Covariate(s) box. As in ANOVA, add categorical predictors to the Fixed Factor(s) box. Click on *Model* to build a custom model that contains only some or none of the interactions. Select *Interaction* or *Main Effects* from the Build Term(s) list, and click on the arrow to move the terms to the model list on the right. Click on *Options* in the UNIVARIATE dialog box and under *Estimated Marginal Means* you can select a factor on which to find adjusted means. The output also includes a table with F tests for the between-subjects effects.

In the ANALYZE menu, there is a MIXED MODELS option with a LINEAR suboption. At that menu, you choose the variable that identifies the subjects (clusters) who have random effects and the variable on which the observations are repeated at the subject (cluster) level. After clicking on *Continue*, you identify the dependent variable, factors, and covariates.

CHAPTER 14: MODEL BUILDING WITH MULTIPLE REGRESSION

In the LINEAR REGRESSION dialog window for the REGRESSION choice on the ANALYZE menu, you can select a *Method* for fitting the model, among which are options such as BACKWARD, FORWARD, and STEPWISE for selecting predictors in the model (or ENTER for adding all of them).

In the LINEAR REGRESSION dialog window, you can plot studentized residuals (labeled SRESID) and request all partial regression plots by clicking on *Plots* and then making appropriate selections in the PLOTS dialog box. To obtain predicted values, residuals, studentized residuals, leverage values, and influence diagnostics, click on *Save* in the LINEAR REGRESSION dialog box. The resulting LINEAR REGRESSION: SAVE dialog box contains options for these, such as *Standardized DfBeta(s)* for DFBETAS and *Standardized DfFit* for DFFITS. To find VIF (variance inflation factors), click on *Statistics* in the LINEAR REGRESSION dialog box and select *Collinearity diagnostics*.

To fit generalized linear models, on the ANALYZE menu select the GENERALIZED LINEAR MODELS option and the GENERALIZED LINEAR MODELS suboption. Click on the *Type of Model* tab at the top of the dialog box and make a selection. If you want to use the identity link function with the gamma distribution or the Poisson distribution for the response variable, you must select a Custom model and then select the desired combination of distribution for *y* and link function. Then, click on *Response* and select the Dependent Variable and click on the *Predictors* tab and enter quantitative variables as Covariates and categorical variables as Factors. Click on the *Model* tab and enter these variables as main effects, and construct any interactions that you want in the model. Click on OK to run the model. (If you build a model assuming the gamma distribution, with the Estimation tab you can select *Maximum Likelihood Estimate* or *Pearson chi-square* for the Scale Parameter Method.) At the *Statistics* tab, you can select likelihood-ratio test statistics and profile likelihood confidence intervals, which are preferable to the Wald method. With the *Save* tab, you can request predicted values and diagnostics such as standardized Pearson residuals and Cook's distance values.

To fit a quadratic regression model, on the ANALYZE menu select the REGRESSION option with the CURVE ESTIMATION suboption. Then, in the CURVE ESTIMATION dialog box, select the variables and choose the *Quadratic* model. The PLOT MODELS option provides a plot of the fitted curve. It can be useful to choose the Linear and Quadratic models so that this plot shows the comparison.

To obtain a LOESS smoothing curve, on the GRAPHS menu select the REGRESSION VARIABLE PLOTS, pick the variables, and then under OPTIONS select the LOESS option.

To fit the exponential regression model, on the ANALYZE menu select the GENERALIZED LINEAR MODELS option and the GENERALIZED LINEAR MODELS suboption. With the Type of Model tab, select custom and pick the normal distribution with the log link function. Use the Response and Predictor tabs to select those variables.

There is also an option for an exponential regression model by selecting the CURVE ESTIMATION suboption under the REGRESSION option in the ANALYZE menu. However, this provides a somewhat different fit than using GLM software, since it assumes that the log of *y*, rather than *y*, is normally distributed with constant variance. As discussed following Example 14.9 (page 459), it fits the model $E[\log(y)] = \alpha + \beta x$ rather than the model $\log[E(y)] = \alpha + \beta x$.

CHAPTER 15: LOGISTIC REGRESSION

To fit logistic regression models, on the ANALYZE menu select the REGRESSION option and the BINARY LOGISTIC suboption. In the LOGISTIC REGRESSION dialog box, identify the binary response (dependent) variable and the explanatory predictors (covariates). Highlight variables in the source list and click on $a * b$ to

create an interaction term. Identify the explanatory variables that are categorical and for which you want dummy variables by clicking on Categorical and declaring such a covariate to be a Categorical Covariate in the LOGISTIC REGRESSION: DEFINE CATEGORICAL VARIABLES dialog box. Highlight the categorical covariate and under Change Contrast you will see several options for setting up dummy variables. The *Simple* contrast constructs them with the final category as the baseline.

In the LOGISTIC REGRESSION dialog box, click on *Method* for stepwise model selection procedures, such as backward and forward selection. Click on *Save* to save predicted probabilities, measures of influence such as Cook's distance and DFBETAS, and standardized residuals. Click on *Options* to open a dialog box that contains an option to construct confidence intervals for exponentiated parameters.

Another way to fit logistic regression models is with the GENERALIZED LINEAR MODELS option and suboption on the ANALYZE menu. You pick the Binary response and Binary logistic model. With the *Response* tab, you can also enter the data as the number of successes out of a certain number of trials, which is useful when the data are in contingency table form such as with the death penalty example in Table 15.3 on page 478. For example, suppose in one column you have the number of successes at each particular setting of predictors, and in a separate column you have the sample size that number of successes is based on. Then, you identify the dependent variable as the variable listing the number of successes, you click the box "Variable represents binary response or number of events," and then "Number of events occurring in a set of trials," entering the variable listing the sample sizes as the "Trials variable."

SPSS can also fit logistic models for categorical response variables having several response categories. On the ANALYZE menu, choose the REGRESSION option and then the ORDINAL suboption for a cumulative logit model. (This model is also available under the GENERALIZED LINEAR MODELS option on the ANALYZE menu.) Select the MULTINOMIAL LOGISTIC suboption for a baseline-category logit model. In the latter, click on *Statistics* and check Likelihood-ratio tests under Parameters to obtain results of likelihood-ratio tests for the effects of the predictors.

For loglinear models, use the LOGLINEAR option with GENERAL suboption in the ANALYZE menu. (You can also select a Poisson loglinear model with the GENERALIZED LINEAR MODELS option on the ANALYZE menu.) You enter the factors for the model. The default is the saturated model, so click on *Model* and select a *Custom* model. Enter the factors as terms in a customized (unsaturated) model and then select additional interaction effects. Click on *Options* to show options for displaying observed and expected frequencies and adjusted residuals. When the data file contains the data as cell counts for the various combinations of factors rather than as responses listed for individual subjects, weight each cell by the cell count using the WEIGHT CASES option in the DATA menu.

Introduction to SAS

In learning SAS, you can get help at sites such as

`www.ats.ucla.edu/stat/sas`,

`support.sas.com/documentation/onlinedoc/stat`.

In the **S**AS language, all statements must end with a semicolon. The data follow the DATALINES statement, one line per subject, unless the INPUT statement ends with @@. After the data lines, a line containing only a semicolon ends the data set. Following the data entry, PROC statements invoke the statistical procedures. A typical

PROC statement lists the procedure, such as MEANS, and then also may select some options.

CHAPTER 3: DESCRIPTIVE STATISTICS

Table A.1 shows the format for entering the data and performing basic analyses, using the data set in Table 3.2 on violent crime rates for the 50 states. When you input characters rather than numbers for a variable, such as the state labels, the variable has an accompanying $ label in the INPUT statement, such as state does in Table A.1.

TABLE A.I: SAS for Printing Data, Computing Basic Summary Statistics, and Preparing Plots

```
data crime ;
input  state $  violent;
datalines;
  AL  43
  AK  64
  ...
;
proc print;  var  state  violent;
proc freq;  tables  violent;
proc chart;  vbar  violent;
proc means;  var  violent;
proc univariate   plot;  var  violent;  id  state;
run ;
```

PROC FREQ provides a frequency distribution for the variable listed following TABLES. PROC CHART provides a histogram of the variable listed in the VBAR statement. Options exist for choosing the number of bars (e.g., VBAR VIOLENT / LEVELS = 5) or their midpoints and for forming horizontal rather than vertical bars (HBAR instead of VBAR). PROC MEANS provides the mean and standard deviation. The PROC UNIVARIATE statement requests a greater variety of basic statistics, including the quartiles. The ID statement, which is optional, names STATE as the variable to identify some of the extreme observations in part of the display from this procedure. Listing the PLOT option in PROC UNIVARIATE requests stem-and-leaf and box plots for the variables listed.

CHAPTERS 5 AND 6: ESTIMATION AND SIGNIFICANCE TESTS

Use PROC FREQ for a table of counts of two types to get confidence intervals for proportions and a test of H_0: $\pi = 0.50$. Table A.2 shows code for the example on page 137 about estimating the proportion of vegetarians. The `binomial(ac)` option gives the Agresti–Coull confidence interval, which is appropriate with small samples.

The estimated standard error for a sample mean is provided by PROC UNIVARIATE. Table A.3 shows how to obtain the standard error and the t-score for the data from Example 6.2. The two arguments for the TINV function are half the error probability and the df value. For instance, the statement in Table A.2 requests the t-score with left-tail probability equal to 0.025 (for a 95% confidence interval) when $df = 28$, which equals -2.048. That table also shows how to input data for two

TABLE A.2: SAS Code for Confidence Intervals for a Proportion

```
data veg;
input response $ count;
datalines;
no   25
yes  0
;
proc freq data=veg; weight count;
tables response / binomial(ac) alpha=.05;
run;
```

TABLE A.3: SAS for Obtaining Standard Errors, *t*-Scores, and *t* Test

```
data anorexia ;
input weight1 weight 2 ;
change = weight2 - weight1;
datalines;
  80.5  82.2
  84.9  85.6
  ...
;
proc univariate ; var diff ;
data findt;
       tvalue = tinv(.025, 28) ;
proc print    data = findt ;
proc ttest h0=0 sides=2 alpha=0.05; var change;
run ;
```

dependent samples (WEIGHT1 and WEIGHT2 being the weights of anorexic girls at two times) and create a new variable (CHANGE) that is the difference between WEIGHT2 and WEIGHT1 and perform the *t* test and 95% confidence interval.

CHAPTER 7: COMPARISON OF TWO GROUPS

Table A.4 performs a two-sample *t* test for comparing two means (Section 7.3), using the data in the anorexia example. The input variables are THERAPY, the levels of which are the two groups to be compared, and CHANGE, the change in weight (the response variable). PROC SORT sorts the data into groups, according to the levels of therapy, and then PROC MEAN finds means and standard deviations for the observations in each group, when you use BY followed by the group variable. SAS uses the BY statement to do an analysis separately for each level of the variable specified in the BY statement.

PROC TTEST is a procedure for two-sample *t* tests with independent samples. The CLASS statement names the variable that identifies the groups to be compared, and the VAR statement identifies the response variable.

With PROC FREQ, for contingency tables the RISKDIFF option provides confidence intervals for the proportions and their difference. For contingency tables having small cell counts, the EXACT statement in PROC FREQ can provide Fisher's exact test, with the keyword FISHER.

TABLE A.4: SAS for Two-Sample *t* Test for Example 7.8 (See Table 12.18 for the Data)

```
data depress;
input therapy $ change ;
datalines;
 cogbehav 1.7
 cogbehav 0.7
...
 control -0.5
 control -9.3
...
;
proc sort; by therapy ;
proc means; by therapy ; var change ;
proc ttest; class therapy ; var change ;
run;
```

CHAPTER 8: ANALYZING ASSOCIATION BETWEEN CATEGORICAL VARIABLES

Table A.5 illustrates the analysis of two-way contingency tables. PROC FREQ conducts chi-squared tests of independence using the CHISQ option and provides expected frequencies with the EXPECTED option. The MEASURES option provides measures of association (including gamma) and their standard errors. For 2×2 tables, this option provides confidence intervals for the odds ratio (labeled "case-control" on output). The EXACT option provides Fisher's exact test. SAS lists the category levels in alphanumeric order unless you state ORDER=DATA in the PROC directive, in which case the levels have the order in which they occur in the input data.

TABLE A.5: SAS for Chi-Squared Test with Table 8.1

```
data politics;
input  gender  $  party  $  count  @@;
datalines;
  Female  Democ  495   Female  Indep  590  Female  Repub  272
  Male    Democ  330   Male    Indep  498  Male    Repub  265
;
proc freq;  weight count ;
     tables  gender*party / chisq  expected  measures ;
proc genmod;  class gender party;
 model count = gender party / dist=poi link=log obstats residuals;
run;
```

You can also perform chi-squared tests using PROC GENMOD. This procedure, discussed in greater detail below for Chapter 14, uses a generalized linear modeling approach introduced in Section 14.4. (The code in Table A.5 views the independence hypothesis as a "loglinear model" for Poisson counts with main effects of gender and party but no interaction.) The OBSTATS and RESIDUALS options in GENMOD provide cell residuals; the output labeled *StReschi* is the standardized residual.

CHAPTER 9: LINEAR REGRESSION AND CORRELATION

Table A.6 performs linear regression, with data as shown in Table 9.1. The PROC PLOT statement requests a scatterplot for murder rate and poverty rate; the first variable listed goes on the *y*-axis. The PROC REG statement requests a regression analysis, predicting murder rate using poverty rate. The P option following this model statement requests the predicted values and residuals for all observations. The PROC CORR statement requests the correlation between each pair of variables listed in the VAR list.

TABLE A.6: SAS for Regression Analysis with Table 9.1

```
data crime ;
input state $  violent  murder  metro  white  hs poverty single ;
datalines;
  AK    761     9.0  41.8    75.2    86.6    9.1   14.3
  AL    780    11.6  67.4    73.5    66.9   17.4   11.5
  ....
;
proc plot; plot murder*poverty ;
proc reg;  model  murder = poverty / p;
proc corr;  var  violent  murder  metro  white hs poverty single ;
run;
```

CHAPTER 11: MULTIPLE REGRESSION AND CORRELATION

Table A.7 performs multiple regression. You list every explanatory variable in the model to the right of the equal sign in the model statement. The PARTIAL option provides partial regression scatterplots, PCORR2 provides squared partial correlations, and STB provides standardized regression coefficients. We create centered variables by subtracting the means from the predictors and then define a variable life_ses to be the cross product of centered life events and ses. We enter that variable in the second regression model to permit interaction in the model.

TABLE A.7: SAS for Multiple Regression Analysis with Table 11.1

```
data mental ;
input  impair  life   ses ;
life_cen = life - 44.425;  ses_cen = ses - 56.60;
life_ses = life_cen*ses_cen;
datalines;
  17   46   84
  19   39   97
  ....
;
proc plot ;  plot  impair*life  impair*ses ;
proc reg;  model  impair = life  ses  / partial  stb  pcorr2 ;
proc reg;  model  impair = life  ses  life_ses ;
run;
```

CHAPTER 12: REGRESSION WITH CATEGORICAL PREDICTORS: ANALYSIS OF VARIANCE METHODS

Table A.8 performs one-way ANOVA with Table 12.1 and two-way ANOVA with Table 12.9 . The first PROC MEANS statement requests sample means on ideology for the data grouped by party. PROC GLM is a procedure for *general linear models*. It is similar in many ways to PROC REG except that PROC GLM can use CLASS statements to create dummy variables for categorical predictors.

TABLE A.8: SAS for One-Way ANOVA with Table 12.1 and Two-Way ANOVA with Table 12.9

```
data anova;
input party  $  sex  $   ideology ;
datalines;
  Dem  F  1
   ...
  Rep  M  7
;
proc means;   by party;   var ideology;
proc glm;   class party ;
    model ideology = party   /   solution;
    means party  /   bon   tukey   alpha=.10;
proc means;  by   party   sex;   var   ideology;
proc glm;  class   party   sex;
    model ideology = party sex   /   solution;
    means party /   bon   tukey;
proc glm;   class   party   sex;
    model ideology = party sex party*sex;
run;
```

The first GLM statement requests a one-way ANOVA. The CLASS statement requests dummy variables for party. The MEANS option provides multiple comparison confidence intervals. Here, we request the Bonferroni and Tukey methods and specify alpha = 0.10 for overall 90% confidence. The SOLUTION option requests the estimates for the prediction equation.

The second PROC MEANS requests sample means on ideology for each combination of party and gender. A GLM statement then conducts a two-way ANOVA, setting up dummy variables for party and gender with the CLASS statement. A MEANS option then requests multiple comparisons across levels of party, assuming a lack of interaction. The final GLM statement adds an interaction term to the model.

Table A.9 shows SAS for the repeated-measures ANOVA with Table 12.15. You can use PROC REG or else PROC ANOVA. The latter applies for "balanced" analyses having the same number of responses at each level of a factor. The analysis is followed by a multiple comparison of means across the levels of type of entertainment.

Table A.10 shows an alternative way of inputting data for a repeated-measures ANOVA. The model statement indicates that the repeated responses are modeled as a function of *therapy* and that the levels at which the repeated measurements occur refer to a variable labeled as *occasion*. The analysis is followed by a Bonferroni multiple comparison of the response means by category of therapy.

TABLE A.9: SAS for Repeated-Measures ANOVA with Table 12.15

```
data repeat;
input subject  $ type  $   opinion @@;
datalines;
  1  M  -1     1  T   0      1  R  -1
 ....
 12  M  -1    12  T  -1     12  R  -2
;
proc anova;    classes subject type;
    model opinion = type subject ;
    means type   /   tukey bon;
run;
```

TABLE A.10: SAS for Two-Way Repeated-Measures ANOVA with Table 12.18

```
data repeat2;
input subject $ therapy $ weight1-weight2;
datalines;
  1    CB    80.5    82.2
  2    CB    84.9    85.6
  3    CB    81.5    81.4
 ....
 72    C    89.0    78.8
;
proc anova; class therapy ;
model weight1-weight2 = therapy ;
  repeated occasion  /  short printe;
means therapy / bon ;
run;
```

You can also conduct repeated-measures ANOVA in SAS using PROC MIXED, to have additional options for the covariance structure of the random effect (see Section 13.7). This procedure, unlike PROC ANOVA or GLM, can use data from subjects that have missing observations. Other advantages of PROC MIXED are that you can use continuous variables in within-subject effects, instead of only classification variables, and you can omit the between–within interaction effects from the model. See Littell et al. (2006) for details.

CHAPTER 13: MULTIPLE REGRESSION WITH QUANTITATIVE AND CATEGORICAL PREDICTORS

Table A.11 fits models to Table 13.1. The PLOT statement requests a plot of income by education, with symbols indicating which race each observation has. The first GLM statement fits the model, assuming no interaction, using a CLASS statement to provide dummy variables for race. This is followed by a request for adjusted means (also called "least squares means" and abbreviated by SAS as LSMEANS) on the response for the different levels of race, with Bonferroni multiple comparisons. The second GLM statement adds an interaction of race and education.

TABLE A.11: SAS for Regression Models with Table 13.1

```
data ancova ;
input  income  educ  race $ ;
datalines;
 16  10  black
 18   7  black
 26   9  black
 ....
 56  20  white
;
proc plot;   plot income*educ = race;
proc glm; class race;   model income = educ race / solution;
lsmeans race   adjust=bon ;
proc glm; class race; model income = educ race educ*race/solution;
run;
```

Use PROC MIXED to fit linear mixed models, as shown in Table A.12.

TABLE A.12: SAS for Linear Mixed Model for Table 13.13, with Compound Symmetry Correlation Structure

```
data smss;
  input family opinion party sex;
datalines;
 1      8    8    1
 1      7    9    0
 1      7    7    1
 2      4    6    0
...
;
proc mixed data=smss;
   class family;
   model opinion = party sex / solution; random family / vcorr;
   repeated/type=cs subject=family r rcorr;
run;
```

CHAPTER 14: MODEL BUILDING WITH MULTIPLE REGRESSION

Table A.13 analyzes the house sales data. The BACKWARD, FORWARD, and STEPWISE choices for the SELECTION option yield these selection procedures. The P option yields predicted values and the PRESS model diagnostic. The INFLU-ENCE option yields studentized residuals, leverage values, and measures of influence such as DFFITS and DFBETAS. The PLOT option following the second model statement requests plots of residuals against the predicted values and against size of home. The code sets up an artificial variable *size_2* that is the square of size. Entering it in the model, as in the third regression statement, provides a quadratic regression model.

　　PROC GENMOD fits generalized linear models. GENMOD specifies the distribution in the DIST option ("nor" for normal, "gam" for gamma, "poi" for Poisson,

TABLE A.13: SAS for Various Analyses Conducted with House Sales Data

```
data housing ;
input  price  size   bed   bath   new;
size_2 = size*size;
datalines;
  279900   2048   4   2   0
  146500    912   2   2   0
  ....
;
proc reg;  model price = size bed bath new  / selection=backward;
proc reg;  model price = size bath new   /   p influence partial;
plot r.*p. r.*size ;
proc reg;  model price = size size_2 ;
proc genmod;  model price = size  /  dist = nor   link = identity;
proc genmod;  model price = size  /  dist = gam   link = identity;
run;
```

"bin" for binomial) and has a LINK option (including "log," "identity," and "logit"). The first GENMOD statement in Table A.13 gives the same results as using least squares with PROC REG or GLM. The second GENMOD statement assumes a gamma distribution for price.

Table A.14 uses PROC GENMOD to fit an exponential regression model to the population growth data of Table 14.8.

TABLE A.14: SAS for Fitting Exponential Regression Model as a Generalized Linear Model to Table 14.8

```
data growth ;
input   decade    popul ;
datalines;
 0    62.95
 1    75.99
 ....
;
proc genmod;  model popul = decade  /  dist = nor  link = log ;
run;
```

CHAPTER 15: LOGISTIC REGRESSION

For logistic regression, Table A.15 applies PROC GENMOD and PROC LOGIS-TIC to Table 15.1. These procedures order the levels of the response variable alphanumerically, forming the logit, for instance, as $\log[P(Y = 0)/P(Y = 1)]$. The DESCENDING option reverses the order. Following the LOGISTIC model fit, Table A.15 requests predicted probabilities and lower and upper 95% confidence limits for the true probabilities.

For PROC GENMOD and PROC LOGISTIC with binomial models, the response in the model statements can have the form of the number of successes divided by the number of cases. Table A.16 fits a logistic model with categorical predictors to the death penalty data in Table 15.3. The OBSTATS option in GENMOD provides

TABLE A.15: SAS for Fitting Logistic Regression Model to Table 15.1

```
data binary ;
input   ideology opinion ;
datalines;
 4 1
 3 0
 ....
;
proc genmod descending;
    model opinion = ideology  /  dist = bin  link = logit ;
proc logistic descending; model opinion = ideology  / influence;
    output  out=predict   p=pi_hat   lower=LCL  upper=UCL;
proc print  data = predict;
run;
```

predicted probabilities and their confidence limits, and the RESIDUALS option pro-
vides standardized residuals (labeled *StReschi*). In models with multiple predictors,
the TYPE3 option in GENMOD provides likelihood-ratio tests.

TABLE A.16: SAS for Fitting Logistic Model to Table 15.3

```
data death ;
input   vic   def   yes   n ;
datalines;
  1   1   53   467
  1   0   11    48
  0   1    0    16
  0   0    4   143
;
proc genmod; model yes/n = def vic / dist=bin link=logit residuals
  obstats type3;
proc logistic;  model yes/n = def  vic;
```

Table A.17 fits the loglinear model (AC, AM, CM) to the student survey data
of Table 15.12. The AM association is represented by $A * M$. The OBSTATS and

TABLE A.17: SAS for Fitting Loglinear Models to Table 15.12

```
data drugs ;
input   a   $   c   $   m   $   count @@ ;
datalines;
 yes  yes  yes  911   yes  yes  no  538
 yes  no   yes   44   yes   no  no  456
 no   yes  yes    3    no  yes  no   43
 no   no   yes    2    no   no  no  279
;
proc genmod; class  a  c  m ;
 model count = a  c  m  a*c  a*m  c*m / dist=poi link=log obstats
                                                      residuals;
run;
```

RESIDUALS options provide expected frequencies (predicted values) and diagnostics, including standardized residuals.

For ordinal responses, PROC LOGISTIC provides ML fitting of the proportional odds version of cumulative logit models. PROC GENMOD fits this model using options DIST=MULTINOMIAL and LINK=CLOGIT. PROC LOGISTIC fits the baseline-category logit model for nominal responses with the option LINK=GLOGIT. For examples of these and other methods for categorical data, see `www.stat.ufl.edu/~aa/cda/Sas_web.pdf`.

Bibliography

Afifi, A. A., May, S., and Clark, V. (2011). *Practical Multivariate Analysis*, 5th ed. Chapman & Hall.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*, 2nd ed. Wiley.

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*, 2nd ed. Wiley.

Agresti, A., Franklin, C., and Klingenberg, B. (2017). *Statistics: The Art and Science of Learning from Data*, 4th ed. Pearson.

Allison, P. (2002). *Missing Data*. Sage.

Allison, P. D. (2014). *Event History and Survival Analysis*, 2nd ed. Sage.

Bartholomew, D. J. (1982). *Stochastic Models for Social Processes*, 3rd ed. Wiley.

Bartholomew, D. J., Steele, F., Galbraith, J., and Moustaki, I. (2008). *Analysis of Multivariate Social Science Data*, 2nd ed. Chapman and Hall/CRC.

Berk, R. A. (2004). *Regression Analysis: A Constructive Critique*. Sage.

Boker, S. M., and McArdle, J. J. (2014). Path analysis and path diagrams. In *Wiley StatsRef: Statistics Reference Online*.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Wiley.

Carpenter, J. R., and Kenward, M. G. (2013). *Multiple Imputation and Its Application*. Wiley.

Collins, L. M., and Lanza, S. T. (2009). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. Wiley.

Cook, G. (2014). *The Best American Infographics*. Houghton Mifflin Harcourt.

Crossen, C. (1994). *Tainted Truth: The Manipulation of Fact in America*. Simon & Schuster.

DeMaris, A. (1992). *Logit Modeling: Practical Applications*. Sage.

DeMaris, A. (2002). Covariance structure models. In *Handbook for Conducting Research on Human Sexuality*, eds. M. Wiederman and B. E. Bradley. Erlbaum.

DeMaris, A. (2004). *Regression with Social Data: Modeling Continuous and Limited Response Variables*. Wiley.

Duncan, O. D. (1966). Path analysis: Sociological examples. *American Journal of Sociology*, 72, 1–16.

Eliason, S. (1993). *Maximum Likelihood Estimation: Logic and Practice*. Sage.

Ellenberg, J. (2014). *How Not to Be Wrong: The Power of Mathematical Thinking*. Penguin.

Enders, C. (2010). *Applied Missing Data Analysis*. Guilford Press.

Fitzmaurice, G., Laird, N., and Ware, J. (2011). *Applied Longitudinal Analysis*, 2nd ed. Wiley.

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models*, 3rd ed. Sage.

Freedman, D. A. (2005). *Statistical Models: Theory and Practice*. Cambridge University Press.

Gelman, A., and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gill, J. (2000). *Generalized Linear Models: A Unified Approach*. Sage.

Gill, J. (2014). *Bayesian Methods: A Social and Behavioral Sciences Approach*, 3rd ed. CRC Press.

Goodman, L. A. (1962). Statistical methods for analyzing processes of change. *American Journal of Sociology*, 68, 57–78.

Gould, S. J. (1981). *The Mismeasure of Man*. W. W. Norton.

Gueorguieva, R., and Krystal, J. H. (2004). Move over ANOVA. *Archives of General Psychiatry*, 61, 310–317.

Guo, S., and Fraser, M. W. (2014). *Propensity Score Analysis*, 2nd ed. Sage.

Hagenaars, J., and McCutcheon, A. (editors). (2006). *Applied Latent Class Analysis*. Cambridge University Press.

Harman, H. (1967). *Modern Factor Analysis*, 2nd ed. University of Chicago Press.

Hedeker, D., and Gibbons, R. D. (2006). *Longitudinal Data Analysis*. Wiley.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer.

Hollander, M., Wolfe, D., and Chicken, E. (2013). *Nonparametric Statistical Methods*, 3rd ed. Wiley.

Holzer, C. E., III (1977). *The Impact of Life Events on Psychiatric Symptomatology*. Ph.D. dissertation, University of Florida, Gainesville.

Howell, D. C. (2012). *Statistical Methods for Psychology*, 8th ed. Cengage.

Kennedy, P. (2008). *A Guide to Econometrics*, 6th ed. Wiley-Blackwell.

King, G. (1989). *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Cambridge University Press.

Kirk, R. E. (2012). *Experimental Design: Procedures for the Behavioral Sciences*, 4th ed. Brooks/Cole.

Kruschke, J. (2014). *Doing Bayesian Data Analysis*, 2nd ed. Academic Press.

Kutner, M. H. (2013). *Applied Linear Statistical Models*, 5th ed. McGraw-Hill.

Littell, R., Stroup, W., and Freund, R. (2002). *SAS for Linear Models*, 4th ed. SAS Institute.

Lohr, S. L. (2009). *Sampling: Design and Analysis*, 2nd ed. Duxbury.

Long, J. S. (1983). *Confirmatory Factor Analysis* and *Covariance Structure Models, An Introduction to LISREL*. Sage.

Morgan, S. L., and Winship, C. (2007). *Counterfactuals and Causal Inference*. Cambridge University Press.

Moustaki, I., Knott, M., and Bartholomew, D. J. (2014). Latent-variable modeling. In *Wiley StatsRef: Statistics Reference Online*.

Mulaik, S. A. (2011). *Foundations of Factor Analysis*, 2nd ed. CRC Press.

Pedhazur, E. J. (1997). *Multiple Regression in Behavioral Research*, 3rd ed. Wadsworth.

Privault, N. (2013). *Understanding Markov Chains: Examples and Applications*. Springer.

Raudenbush, S., and Bryk, A. 2002. *Hierarchical Linear Models*, 2nd ed. Sage.

Scheaffer, R. L., Mendenhall, W., Ott, L., and Gerow, K. G. (2011). *Elementary Survey Sampling*, 7th ed. Cengage Learning.

Snijders, T. A. B., and Bosker, R. J. (2011). *Multilevel Analysis*, 2nd ed. Sage.

Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis*. American Psychological Association.

Thompson, S. K. (2012). *Sampling*, 3rd ed. Wiley.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*, 2nd ed. Graphics Press.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Ullman, J. B., and Bentler, P. M. (2013). Structural equation modeling. Chapter 23 in *Handbook of Psychology*, 2nd ed., ed. I. B. Weiner. Wiley.

Vermunt, J. K., and Moors, G. (2014). Event history analysis: Models. In *Wiley StatsRef: Statistics Reference Online*.

Weisberg, S. (2005). *Applied Linear Regression*, 3rd ed. Wiley.

Winer, B. J., Brown, D. R., and Michels, K. M. (1991). *Statistical Principles in Experimental Design*, 3rd ed. McGraw-Hill.

This page intentionally left blank

CREDITS

Chapter 1
Page 19: Screenshot from Stata Statistical Software. Copyright © by StataCorp LP. Used by permission of StataCorp LP. **Page 20:** R Foundation.

Chapter 2
Page 27: Constructed using sample function in R. **Page 38:** `www.realclearpolitics.com`. **Page 40:** Crossen, C. (1994). *Tainted Truth: The Manipulation of Fact in America*. Simon & Schuster, p. 168.

Chapter 3
Page 42: U.S. Census Bureau. **Page 43:** Federal Bureau of Investigation. **Page 47:** `www.socialwatch.org`. **Page 65:** `http://hdr.undp.org/en/data` and `http://data.worldbank.org`; complete data file UN (n = 42) is at text website. **Page 69:** From Statistical Abstract of the United States. Published by Unites States Census Bureau. **Page 70:** `stats.oecd.org`, `hdr.undp.org/en/data`, and `www.pewresearch.org`; complete data file is at text website. **Page 71:** `www.socialwatch.org`. **Page 72:** Human Development Report 2014. **Page 72:** `www.stateofworkingamerica.org`.

Chapter 4
Screenshots of Shiny App. Copyright © by RStudio. Used by permission of RStudio.
Page 87: Screenshot of Art of Stat. Copyright © by Department of Mathematics and Statistics. Used by permission of Department of Mathematics and Statistics.

Chapter 5
Screenshots of Shiny App. Copyright © by RStudio. Used by permission of RStudio.
Page 129: From Anorexia CB Data File. Copyright © by Brian Everitt. Used by permission of Brian Everitt. **Page 141:** From the Library data file at the text website.

Chapter 7
Page 203: Data courtesy of David Strayer, University of Utah. See D. Strayer and W. Johnston, Psychological Science, vol. 21 (2001), pp. 462–466. **Page 215:** Table 2, Presence or Absence of Congenital Sex Organ Malformation Categorized by Alcohol Consumption of the Mother from "Choice of Column Scores for Testing Independence in Ordered 2 × K Contingency Tables" by Barry I. Graubard and Edward L. Korn in Biometrics Vol. 43, No. 2, pp. 471–476. Copyright © 1987 by The International Biometric Society. **Page 220:** `www.statistics.gov.uk`. **Page 224:** E. Diener et al., Journal of Personality & Social Psychology, vol. 69 (1995), pp. 120–129.

Chapter 8
Screenshots of Shiny App. Copyright © by RStudio. Used by permission of RStudio. **Page 242:** Federal Bureau of Investigation. **Page 252:** By D. Snipes and E. Benotsch, Addictive

Behaviors, Volume 38 (2013), pp. 1418–1423. **Page 254:** www.bjs.gov. **Page 254:** Thanks to Dr. Cristanna Cook, Medical Care Development, Augusta, Maine, for supplying these data.

Chapter 9

Page 270: https://www.iusb.edu/ugr-journal/static/2002/index.php. **Page 290:** Figure 8H at www.stateofworkingamerica.org. **Page 295:** Dr. Larry Winner, University of Florida. Complete data are in Florida data file at the text website. **Page 295:** The Economist, January 31, 2015.

Chapter 10

Page 315: Data from M. Radelet, Amer. Sociol. Rev., Volume 46 (1981), pp. 918–927. **Page 316:** National Center for Education Statistics, Digest of Education Statistics, 2015, Table 316.10.

Chapter 11

Page 324: Dr Charles Holzer. **Page 330:** ANOVA. **Page 354:** Thanks to Barbara Finlay for these results.

Chapter 12

Pages 367, 372, 376, 381, 383, 387, 391, 392, 394, 395: ANOVA. **Page 385:** From Anorexia Data File. Copyright © by Brian Everitt. Used by permission of Brian Everitt.

Chapter 13

Page 411: Published in Sociological Inquiry, vol. 83 (2013), pp. 537–569; you may be able to access a pdf file of this article through your university's library at http://onlinelibrary.wiley.com/doi/10.1111/soin.12019/abstract. **Page 412:** Published in Psychiatric Service, vol. 52 (2001), pp. 1621–1626; you can access a pdf file of this article at http://ps.psychiatryonline.org/doi/pdf/10.1176/appi.ps.52.12.1621. **Page 427:** ANOVA.

Chapter 14

Page 453: hdr.undp.org/en/data. **Page 457:** U.S. Census Bureau. **Page 467:** U.S. Census Bureau.

Chapter 15

Page 473: Complete data file Evolution for n = 1064 is at the text website. **Page 481:** By D. Snipes and E. Benotsch, Addictive Behaviors, vol. 38 (2013), pp. 1418–1423. **Page 493:** Thanks to Prof. Harry Khamis, Wright State University, for these data. **Page 501:** Thanks to R. Piccarreta, Bocconi University, Milan, for original form of data. **Page 503:** G. Macolino et al., *American Journal of Public Health*, vol. 95 (2005), pp. 1739–1740. From *Statistics* by D. Freedman, R. Pisani, and R. Purves (W.W. Norton, 1978), p. 14. **Page 504:** Dr. Cristanna Cook, Medical Care Development, Augusta, Maine. **Page 505:** V. I. Rickert et al., Clinical Pediatrics, Vol. 31 (1992), pp. 205–210. *The Independent* newspaper, March 8, 1994; thanks to Dr. P. Altham for this. **Page 506:** By R. Mitchell, Crossing Boundaries, vol. 1 (2001), pp. 107–117. C. Hoge et al., *New England Journal of Medicine*, vol. 351 (2004), pp. 13–21.

INDEX

Key Formulas for Statistical Methods

Chapter 3   Descriptive Statistics

Mean $\bar{y} = \dfrac{\sum y_i}{n}$ Standard deviation $s = \sqrt{\dfrac{\sum(y_i - \bar{y})^2}{n-1}}$

Chapter 4   Probability Distributions

z-score $z = \dfrac{y - \mu}{\sigma}$ Standard error $\sigma_{\bar{y}} = \dfrac{\sigma}{\sqrt{n}}$

Chapter 5   Statistical Inference: Estimation

Confidence interval for mean $\bar{y} \pm z(se)$ with $se = \dfrac{s}{\sqrt{n}}$

Confidence interval for proportion $\hat{\pi} \pm z(se)$ with $se = \sqrt{\dfrac{\hat{\pi}(1-\hat{\pi})}{n}}$

Chapter 6   Statistical Inference: Significance Tests

$H_0 : \mu = \mu_0$ test statistic $t = \dfrac{\bar{y} - \mu_0}{se}$ with $se = \dfrac{s}{\sqrt{n}}, df = n - 1$

$H_0 : \pi = \pi_0$ test statistic $z = \dfrac{\hat{\pi} - \pi_0}{se_0}$ with $se_0 = \sqrt{\dfrac{\pi_0(1 - \pi_0)}{n}}$

Chapter 7   Comparison of Two Groups

Compare means: $(\bar{y}_2 - \bar{y}_1) \pm t(se)$ with $se = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$

Test $H_0 : \mu_1 = \mu_2$ using $t = \dfrac{\bar{y}_2 - \bar{y}_1}{se}$

Compare proportions: $(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se)$ with $se = \sqrt{\dfrac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \dfrac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$

Chapter 8   Analyzing Association Between Categorical Variables

Chi-squared test of H_0: Independence, $X^2 = \sum \dfrac{(f_o - f_e)^2}{f_e}, \;\; df = (r-1)(c-1)$

Ordinal measure $\hat{\gamma} = \dfrac{C - D}{C + D}, \; -1 \le \hat{\gamma} \le 1, \; z = \dfrac{\hat{\gamma}}{\hat{\sigma}_{\hat{\gamma}}}, \; \hat{\gamma} \pm z\hat{\sigma}_{\hat{\gamma}}$

Chapter 9   Linear Regression and Correlation

Linear regression model $E(y) = \alpha + \beta x$, prediction equation $\hat{y} = a + bx$

Pearson correlation $r = \left(\dfrac{s_x}{s_y}\right) b, \; -1 \le r \le 1$

$r^2 = \dfrac{\text{TSS} - \text{SSE}}{\text{TSS}}, \; \text{TSS} = \sum(y - \bar{y})^2, \; \text{SSE} = \sum(y - \hat{y})^2, \; 0 \le r^2 \le 1$

Test of independence $H_0 : \beta = 0, \quad t = \dfrac{b}{se}, \quad df = n - 2$

Chapter 11 Multiple Regression and Correlation

Multiple regression model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

Global test $H_0 : \beta_1 = \cdots = \beta_p = 0$

$$\text{Test statistic } F = \frac{\text{Model mean square}}{\text{Error mean square}} = \frac{R^2/p}{(1 - R^2)/[n - (p + 1)]}$$

$df_1 = p, df_2 = n - (p + 1)$

Partial test $H_0 : \beta_i = 0$, test statistic $t = \dfrac{b_i}{se}$, $df = n - (p + 1)$

Chapter 12 Regression with Categorical Predictors: Analysis of Variance Methods

$H_0 : \mu_1 = \cdots = \mu_g$, One-way ANOVA test statistic

$$F = \frac{\text{Between-groups sum of squares}/(g - 1)}{\text{Within-groups sum of squares}/(N - g)}, \; df_1 = g - 1, df_2 = N - g$$

Chapter 13 Multiple Regression with Quantitative and Categorical Predictors

$E(y) = \alpha + \beta x + \beta_1 z_1 + \cdots + \beta_{g-1} z_{g-1}$, $z_i = 1$ or 0 is dummy variable for group i

Chapter 14 Model Building with Multiple Regression

Quadratic regression $E(y) = \alpha + \beta_1 x + \beta_2 x^2$
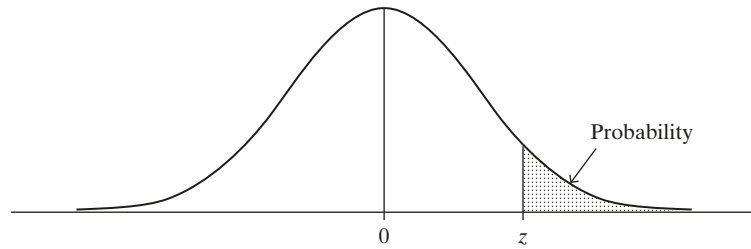
Exponential regression $E(y) = \alpha \beta^x$ (log of mean is linear in x)

Chapter 15 Logistic Regression: Modeling Categorical Responses

Logistic regression $\text{logit} = \log(\text{odds}) = \log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \alpha + \beta x$

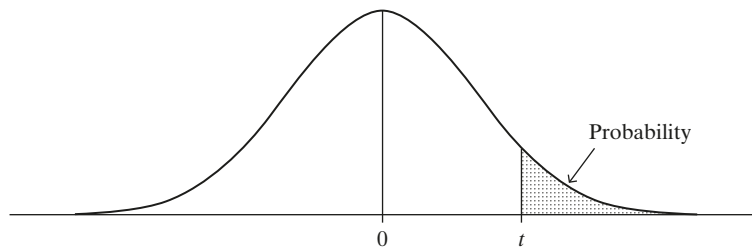$P(y = 1) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} = \frac{\text{odds}}{1 + \text{odds}}$

TABLE A: Normal curve tail probabilities. Standard normal probability in right-hand tail (for negative values of z, probabilities are found by symmetry).

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
					Second Decimal Place of z					
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.00135									
3.5	.000233									
4.0	.0000317									
4.5	.00000340									
5.0	.000000287									

Source: R. E. Walpole, *Introduction to Statistics* (New York: Macmillan, 1968).

TABLE B: *t* Distribution Critical Values

	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.611
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
80	1.292	1.664	1.990	2.374	2.639	3.195
100	1.290	1.660	1.984	2.364	2.626	3.174
∞	1.282	1.645	1.960	2.326	2.576	3.091

Source: "Table of Percentage Points of the *t*-Distribution." Computed by Maxine Merrington, Biometrika, 32 (1941): 300. Reproduced by permission of the Biometrika trustees.

A GUIDE TO CHOOSING A STATISTICAL METHOD

Quantitative Response Variable (Analyzing Means)

1. If no other variables, use descriptive methods of Chapter 3 and inferential methods of Section 5.3 (confidence interval) and Section 6.2 (significance test) for a mean.
2. Categorical explanatory variable: If two levels, use methods for comparing two means from Section 7.3 (two independent samples) or Section 7.4 (two dependent samples). If several levels, use ANOVA methods for comparing several means from Sections 12.1–3 (several independent samples) or Section 12.6 (several dependent samples). These are equivalent to regression methods with dummy variables for predictors. If several categorical variables, use ANOVA methods of Sections 12.4 or 12.6 or use regression with dummy variables.
3. Quantitative explanatory variable: Use regression and correlation methods of Chapter 9. If several quantitative predictors, use multiple regression methods of Chapters 11 and 14.
4. Quantitative and categorical explanatory variables: Use analysis of covariance methods of Chapter 13, which are regression methods with dummy variables for categorical predictors.

Categorical Response Variable (Analyzing Proportions)

1. If no other variable, use descriptive methods of Section 3.1 and inferential methods of Section 5.2 (confidence interval) and Section 6.3 (significance test) for proportions.
2. Categorical explanatory variable: Use contingency table methods of Chapter 8, with Section 7.2 for special case of comparing proportions for two groups and Section 8.5 for ordinal classifications.
3. If binary response with quantitative explanatory variable or multiple quantitative and/or categorical predictors, use logistic regression methods of Chapter 15.
4. If ordinal response with quantitative and/or categorical predictors, use ordinal logit model of Section 15.4.

GLOBAL EDITION

For these Global Editions, the editorial team at Pearson has collaborated with educators across the world to address a wide range of subjects and requirements, equipping students with the best possible learning tools. This Global Edition preserves the cutting-edge approach and pedagogy of the original, but also features alterations, customization, and adaptation from the North American version.

Pearson Global Edition