Essentials of

Business Statistics

Communicating with Numbers



Jaggia | Kelly

Second Edition

Mc Graw Hill Education

Essentials of Business Statistics

The McGraw-Hill Education Series in Operations and Decision Sciences

Supply Chain Management

Benton
Purchasing and Supply Chain Management
Third Edition

Bowersox, Closs, Cooper, and Bowersox
Supply Chain Logistics Management
Fifth Edition

Burt, Petcavage, and Pinkerton
Supply Management
Eighth Edition

Johnson
Purchasing and Supply Management
Sixteenth Edition

Simchi-Levi, Kaminsky, and Simchi-Levi
Designing and Managing the Supply Chain: Concepts, Strategies, Case Studies
Third Edition

Stock and Manrodt
Fundamentals of Supply Chain Management

Project Management

Brown and Hyer
Managing Projects: A Team-Based Approach

Larson and Gray
Project Management: The Managerial Process
Seventh Edition

Service Operations Management

Bordoloi, Fitzsimmons, and Fitzsimmons
Service Management: Operations, Strategy, Information Technology
Ninth Edition

Management Science

Hillier and Hillier
Introduction to Management Science: A Modeling and Case Studies Approach with Spreadsheets
Sixth Edition

Business Research Methods

Cooper and Schindler
Business Research Methods
Thirteenth Edition

Business Forecasting

Keating, Wilson, and John Galt Solutions, Inc.
Business Forecasting
Seventh Edition

Linear Statistics and Regression

Kutner, Nachtsheim, and Neter
Applied Linear Regression Models
Fourth Edition

Business Systems Dynamics

Sterman
Business Dynamics: Systems Thinking and Modeling for a Complex World

Operations Management

Cachon and Terwiesch
Operations Management
Second Edition

Cachon and Terwiesch
Matching Supply with Demand: An Introduction to Operations Management
Fourth Edition

Jacobs and Chase
Operations and Supply Chain Management: The Core
Fifth Edition

Jacobs and Chase
Operations and Supply Chain Management
Fifteenth Edition

Schroeder and Goldstein
Operations Management in the Supply Chain: Decisions and Cases
Seventh Edition

Stevenson
Operations Management
Thirteenth Edition

Swink, Melnyk, Hartley, and Cooper
Managing Operations across the Supply Chain
Fourth Edition

Business Math

Slater and Wittry
Practical Business Math Procedures
Thirteenth Edition

Slater and Wittry
Math for Business and Finance: An Algebraic Approach
Second Edition

Business Statistics

Bowerman, O'Connell, and Murphree
Business Statistics in Practice
Ninth Edition

Doane and Seward
Applied Statistics in Business and Economics
Sixth Edition

Doane and Seward
Essential Statistics in Business and Economics
Third Edition

Jaggia and Kelly
Business Statistics: Communicating with Numbers
Third Edition

Jaggia and Kelly
Essentials of Business Statistics: Communicating with Numbers
Second Edition

Lind, Marchal, and Wathen
Basic Statistics for Business and Economics
Ninth Edition

Lind, Marchal, and Wathen
Statistical Techniques in Business and Economics
Seventeenth Edition

McGuckian
Connect Master: Business Statistics

2e

Essentials of Business Statistics

Communicating with Numbers

SANJIV JAGGIA
California Polytechnic State University

ALISON KELLY
Suffolk University

Mc Graw Hill Education



ESSENTIALS OF BUSINESS STATISTICS: COMMUNICATING WITH NUMBERS, SECOND EDITION

Published by McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121. Copyright © 2020 by McGraw-Hill Education. All rights reserved. Printed in the United States of America. Previous editions © 2014. No part of this publication may be reproduced or distributed in any form or by any means, or stored in a database or retrieval system, without the prior written consent of McGraw-Hill Education, including, but not limited to, in any network or other electronic storage or transmission, or broadcast for distance learning.

Some ancillaries, including electronic and print components, may not be available to customers outside the United States.

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 LWI 21 20 19

ISBN 978-1-260-23951-5
MHID 1-260-23951-9

Portfolio Manager: *Noelle Bathurst*
Product Developers: *Ryan McAndrews*
Marketing Manager: *Harper Christopher*
Content Project Managers: *Pat Frederickson* and *Jamie Koch*
Buyer: *Laura Fuller*
Design: *Egzon Shaqiri*
Content Licensing Specialist: *Ann Marie Jannette*
Cover Design: *Beth Blech*
Compositor: *SPi Global*

All credits appearing on page or at the end of the book are considered to be an extension of the copyright page.

Library of Congress Cataloging-in-Publication Data

Names: Jaggia, Sanjiv, 1960- author. | Hawke, Alison Kelly, author.
Title: Essentials of business statistics : communicating with numbers/Sanjiv Jaggia, California Polytechnic State University, Alison Kelly, Suffolk University.
Description: Second Edition. | Dubuque : McGraw-Hill Education, [2018] | Revised edition of the authors' Essentials of business statistics, c2014.
Identifiers: LCCN 2018023099 | ISBN 9781260239515 (alk. paper)
Subjects: LCSH: Commercial statistics.
Classification: LCC HF1017 .J343 2018 | DDC 519.5-dc23
LC record available at https://lccn.loc.gov/2018023099

The Internet addresses listed in the text were accurate at the time of publication. The inclusion of a website does not indicate an endorsement by the authors or McGraw-Hill Education, and McGraw-Hill Education does not guarantee the accuracy of the information presented at these sites.

mheducation.com/highered

Dedicated to Chandrika, Minori, John, Megan, and Matthew

Sanjiv Jaggia



Courtesy of Sanjiv Jaggia

Sanjiv Jaggia is the associate dean of graduate programs and a professor of economics and finance at California Polytechnic State University in San Luis Obispo, California. After earning a Ph.D. from Indiana University, Bloomington, in 1990, Dr. Jaggia spent 17 years at Suffolk University, Boston. In 2003, he became a Chartered Financial Analyst (CFA®). Dr. Jaggia's research interests include empirical finance, statistics, and econometrics. He has published extensively in research journals, including the *Journal of Empirical Finance, Review of Economics and Statistics, Journal of Business and Economic Statistics, Journal of Applied Econometrics,* and *Journal of Econometrics.* Dr. Jaggia's ability to communicate in the classroom has been acknowledged by several teaching awards. In 2007, he traded one coast for the other and now lives in San Luis Obispo, California, with his wife and daughter. In his spare time, he enjoys cooking, hiking, and listening to a wide range of music.

Alison Kelly



Courtesy of Alison Kelly

Alison Kelly is a professor of economics at Suffolk University in Boston, Massachusetts. She received her B.A. degree from the College of the Holy Cross in Worcester, Massachusetts; her M.A. degree from the University of Southern California in Los Angeles; and her Ph.D. from Boston College in Chestnut Hill, Massachusetts. Dr. Kelly has published in journals such as the *American Journal of Agricultural Economics, Journal of Macroeconomics, Review of Income and Wealth, Applied Financial Economics,* and *Contemporary Economic Policy.* She is a Chartered Financial Analyst (CFA®) and teaches review courses in quantitative methods to candidates preparing to take the CFA exam. Dr. Kelly has also served as a consultant for a number of companies; her most recent work focused on how large financial institutions satisfy requirements mandated by the Dodd-Frank Act. She resides in Hamilton, Massachusetts, with her husband, daughter, and son.

A Unique Emphasis on Communicating with Numbers Makes Business Statistics Relevant to Students

We wrote *Essentials of Business Statistics: Communicating with Numbers* because we saw a need for a contemporary, core statistics text that sparked student interest and bridged the gap between how statistics is taught and how practitioners think about and apply statistical methods. Throughout the text, the emphasis is on communicating with numbers rather than on number crunching. In every chapter, students are exposed to statistical information conveyed in written form. By incorporating the perspective of practitioners, it has been our goal to make the subject matter more relevant and the presentation of material more straightforward for students. Although the text is application-oriented and practical, it is also mathematically sound and uses notation that is generally accepted for the topic being covered.

From our years of experience in the classroom, we have found that an effective way to make statistics interesting is to use timely applications. For these reasons, examples in *Essentials of Business Statistics* come from all walks of life, including business, economics, sports, health, housing, the environment, polling, and psychology. By carefully matching examples with statistical methods, students learn to appreciate the relevance of statistics in our world today, and perhaps, end up learning statistics without realizing they are doing so.

This is probably the ***best book*** *I have seen in terms of explaining concepts.*

Brad McDonald, *Northern Illinois University*

The book is ***well written, more readable and interesting than most stats texts,*** *and effective in explaining concepts. The examples and cases are particularly good and effective teaching tools.*

Andrew Koch, *James Madison University*

Clarity and brevity are the most important things I look for—*this text has both in abundance.*

Michael Gordinier, *Washington University, St. Louis*

Continuing Key Features

The second edition of *Essentials of Business Statistics* reinforces and expands six core features that were well-received in the first edition.

Integrated Introductory Cases. Each chapter begins with an interesting and relevant introductory case. The case is threaded throughout the chapter, and once the relevant statistical tools have been covered, a synopsis—a short summary of findings—is provided. The introductory case often serves as the basis of several examples in other chapters.

Writing with Statistics. Interpreting results and conveying information effectively is critical to effective decision making in virtually every field of employment. Students are taught how to take the data, apply it, and convey the information in a meaningful way.

Unique Coverage of Regression Analysis. Relevant and extensive coverage of regression without repetition is an important hallmark of this text.

Written as Taught. Topics are presented the way they are taught in class, beginning with the intuition and explanation and concluding with the application.

Integration of Microsoft Excel®. Students are taught to develop an understanding of the concepts and how to derive the calculation; then Excel is used as a tool to perform the cumbersome calculations. In addition, guidelines for using Minitab, SPSS, JMP, and now R are provided in chapter appendices.

***Connect*®.** *Connect* is an online system that gives students the tools they need to be successful in the course. Through guided examples and LearnSmart adaptive study tools, students receive guidance and practice to help them master the topics.

I really like the case studies and the emphasis on writing. We are making a big effort to incorporate more business writing in our core courses, so that meshes well.

Elizabeth Haran, *Salem State University*

For a statistical analyst, your analytical skill is only as good as your communication skill. Writing with statistics reinforces the importance of communication and provides students with concrete examples to follow.

Jun Liu, *Georgia Southern University*

Features New to the Second Edition

The second edition of *Essentials of Business Statistics* features a number of improvements suggested by many reviewers and users of the first edition. The following are the major changes.

We focus on the *p*-Value Approach. We have found that students often get confused with the mechanics of implementing a hypothesis test using both the *p*-value approach and the critical value approach. While the critical value approach is attractive when a computer is unavailable and all calculations must be done by hand, most researchers and practitioners favor the *p*-value approach since virtually every statistical software package reports *p*-values. Our decision to focus on the *p*-value approach was further supported by recommendations set forth by the *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016* published by the American Statistical Association (http://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf). *The GAISE Report* recommends that 'students should be able to interpret and draw conclusions from standard output from statistical software' (page 11) and that instructors should consider shifting away from the use of tables (page 23). Finally, we surveyed users of *Essentials of Business Statistics*, and they unanimously supported our decision to focus on the *p*-value approach. For those instructors interested in covering the critical value approach, it is discussed in the appendix to Chapter 9.

We added dozens of applied exercises with varying levels of difficulty. Many of these exercises include new data sets that encourage the use of the computer; however, just as many exercises retain the flexibility of traditional solving by hand.

We streamlined the Excel instructions. We feel that this modification provides a more seamless reinforcement for the relevant topic. For those instructors who prefer to omit the Excel parts so that they can use a different software, these sections can be easily skipped.

We completely revised Chapter 13 (More on Regression Analysis). Recognizing the importance of regression analysis in applied work, we have made major enhancements to Chapter 13. The chapter now contains the following sections: Dummy Variables, Interaction with Dummy Variables, Nonlinear Relationships, Trend Forecasting Models, and Forecasting with Trend and Seasonality.

In addition to the Minitab, SPSS, and JMP instructions that appear in chapter appendices, we now include instructions for R. The main reason for this addition is that R is an easy-to-use and wildly popular software that merges the convenience of statistical packages with the power of coding.

We reviewed every Connect exercise. Since both of us use Connect in our classes, we have attempted to make the technology component seamless with the text itself. In addition to reviewing every Connect exercise, we have added more conceptual exercises, evaluated rounding rules, and revised tolerance levels. The positive feedback from users of the first edition has been well worth the effort. We have also reviewed every LearnSmart probe. Instructors who teach in an online or hybrid environment will especially appreciate our Connect product.

Here are other noteworthy changes:

- For the sake of simplicity and consistency, we have streamlined or rewritten many Learning Outcomes.
- In Chapter 1 (Statistics and Data), we introduce structured data, unstructured data, and big data; we have also revised the section on online data sources.
- In Chapter 4 (Introduction to Probability), we examine marijuana legalization in the United States in the Writing with Statistics example.
- In Chapter 6 (Continuous Probability Distributions), we cover the normal distribution in one section, rather than two sections.
- In Chapter 7 (Sampling and Sampling Distributions), we added a discussion of the Trump election coupled with social-desirability bias.
- We have moved the section on "Model Assumptions and Common Violations" from Chapter 13 (More on Regression Analysis) to Chapter 12 (Basics of Regression Analysis).

Students Learn Through Real-World Cases and Business Examples . . .

Integrated Introductory Cases

Each chapter opens with a real-life case study that forms the basis for several examples within the chapter. The questions included in the examples create a roadmap for mastering the most important learning outcomes within the chapter. A synopsis of each chapter's introductory case is presented when the last of these examples has been discussed. Instructors of distance learners may find these introductory cases particularly useful.



Introductory Case

Investment Decision

Jacqueline Brennan works as a financial advisor at a large investment firm. She meets with an inexperienced investor who has some questions regarding two approaches to mutual fund investing: growth investing versus value investing. The investor has heard that growth funds invest in companies whose stock prices are expected to grow at a faster rate, relative to the overall stock market, and value funds invest in companies whose stock prices are below their true worth. The investor has also heard that the main component of investment return is through capital appreciation in growth funds and through dividend income in value funds. The investor shows Jacqueline the annual return data for Vanguard's Growth Index mutual fund (henceforth, Growth) and Vanguard's Value Index mutual fund (henceforth, Value). Table 3.1 shows the annual return data for these two mutual funds for the years 2007–2016.

SYNOPSIS OF INTRODUCTORY CASE


©Ingram Publishing/Getty Images

Growth and value are two fundamental styles in stock and mutual fund investing. Proponents of growth investing believe that companies that are growing faster than their peers are trendsetters and will be able to maintain their superior growth. By investing in the stocks of these companies, they expect their investment to grow at a rate faster than the overall stock market. By comparison, value investors focus on the stocks of companies that are trading at a discount relative to the overall market or a specific sector. Investors of value stocks believe that these stocks are undervalued and that their price will increase once their true value is recognized by other investors. The debate between growth and value investing is age-old, and which style dominates depends on the sample period used for the analysis.

An analysis of annual return data for Vanguard's Growth Index mutual fund (Growth) and Vanguard's Value Index mutual fund (Value) for the years 2007 through 2016 provides important information for an investor trying to determine whether to invest in a growth mutual fund, a value mutual fund, or both types of mutual funds. Over this period, the mean return for the Growth fund of 10.09% is greater than the mean return for the Value fund of 7.56%. While the mean return typically represents the reward of investing, it does not incorporate the risk of

In all of these chapters, ***the opening case leads directly into the application questions*** *that students will have regarding the material. Having a strong and related case will certainly provide more benefit to the student, as context leads to improved learning.*

Alan Chow, *University of South Alabama*

This is an excellent approach. *The student gradually gets the idea that he can look at a problem—one which might be fairly complex—and break it down into root components. He learns that a little bit of math could go a long way, and even more math is even more beneficial to evaluating the problem.*

Dane Peterson, *Missouri State University*

and Build Skills to Communicate Results

Writing with Statistics

One of our most important innovations is the inclusion of a sample report within every chapter (except Chapter 1). Our intent is to show students how to convey statistical information in written form to those who may not know detailed statistical methods. For example, such a report may be needed as input for managerial decision making in sales, marketing, or company planning. Several similar writing exercises are provided at the end of each chapter. Each chapter also includes a synopsis that addresses questions raised from the introductory case. This serves as a shorter writing sample for students. Instructors of large sections may find these reports useful for incorporating writing into their statistics courses.

Writing with statistics shows that statistics ***is more than number crunching.***

Greg Cameron, *Brigham Young University*

These technical writing examples provide a ***very useful example of how to make statistics work and turn it into a report*** *that will be useful to an organization. I will strive to have my students learn from these examples.*

Bruce P. Christensen, *Weber State University*

This is an ***excellent approach. . . .*** *The ability to translate numerical information into words that others can understand is critical.*

Scott Bailey, *Troy University*

Excellent. *Students need to become better writers.*

Bob Nauss, *University of Missouri, St. Louis*

WRITING WITH STATISTICS

Professor Lang is a professor of economics at Salem State University. She has been teaching a course in Principles of Economics for over 25 years. Professor Lang has never graded on a curve since she believes that relative grading may unduly penalize (benefit) a good (poor) student in an unusually strong (weak) class. She always uses an absolute scale for making grades, as shown in the two left columns of Table 6.5.



©Image Source, all rights reserved.

TABLE 6.5 Grading Scales with Absolute Grading versus Relative Grading

Absolute Grading		Relative Grading	
Grade	Score	Grade	Probability
A	92 and above	A	0.10
B	78 up to 92	B	0.35
C	64 up to 78	C	0.40
D	58 up to 64	D	0.10
F	Below 58	F	0.05

A colleague of Professor Lang's has convinced her to move to relative grading, since it corrects for unanticipated problems. Professor Lang decides to experiment with grading based on the relative scale as shown in the two right columns of Table 6.5. Using this relative grading scheme, the top 10% of students will get A's, the next 35% B's, and so on. Based on her years of teaching experience, Professor Lang believes that the scores in her course follow a normal distribution with a mean of 78.6 and a standard deviation of 12.4.

Professor Lang wants to use the above information to

1. Calculate probabilities based on the absolute scale. Compare these probabilities to the relative scale.
2. Calculate the range of scores for various grades based on the relative scale. Compare these ranges to the absolute scale.
3. Determine which grading scale makes it harder to get higher grades.

Sample Report—Absolute Grading versus Relative Grading

Many teachers would confess that grading is one of the most difficult tasks of their profession. Two common grading systems used in higher education are relative and absolute. Relative grading systems are norm-referenced or curve-based, in which a grade is based on the student's relative position in class. Absolute grading systems, on the other hand, are criterion-referenced, in which a grade is related to the student's absolute performance in class. In short, with absolute grading, the student's score is compared to a predetermined scale, whereas with relative grading, the score is compared to the scores of other students in the class.

Let X represent a grade in Professor Lang's class, which is normally distributed with a mean of 78.6 and a standard deviation of 12.4. This information is used to derive the grade probabilities based on the absolute scale. For instance, the probability of receiving an A is derived as $P(X \geq 92) = P(Z \geq 1.08) = 0.14$. Other probabilities, derived similarly, are presented in Table 6.A.

TABLE 6.A Probabilities Based on Absolute Scale and Relative Scale

Grade	Probability Based on Absolute Scale	Probability Based on Relative Scale
A	0.14	0.10
B	0.38	0.35
C	0.36	0.40
D	0.07	0.10
F	0.05	0.05

Unique Coverage and Presentation . . .

By comparing this chapter with other books, I think that this is one of the best explanations about regression I have seen.
Cecilia Maldonado, *Georgia Southwestern State University*

Unique Coverage of Regression Analysis

We combine simple and multiple regression in one chapter, which we believe is a seamless grouping and eliminates needless repetition. This grouping allows more coverage of regression analysis than the vast majority of *Essentials* texts. This focus reflects the topic's growing use in practice. However, for those instructors who prefer to cover only simple regression, doing so is still an option.

The authors have put forth a novel and innovative way to present regression which in and of itself should make instructors take a long and hard look at this book. ***Students should find this book very readable and a good companion for their course.***
Harvey A. Singer, *George Mason University*

This is ***easy for students to follow*** *and I do get the feeling . . . the sections are spoken language.*
Zhen Zhu, *University of Central Oklahoma*

Written as Taught

We introduce topics just the way we teach them; that is, the relevant tools follow the opening application. Our roadmap for solving problems is

1. Start with intuition
2. Introduce mathematical rigor, and
3. Produce computer output that confirms results.

We use worked examples throughout the text to illustrate how to apply concepts to solve real-world problems.

that Make the Content More Effective

Integration of Microsoft Excel®

We prefer that students first focus on and absorb the statistical material before replicating their results with a computer. Solving each application manually provides students with a deeper understanding of the relevant concept. However, we recognize that, primarily due to cumbersome calculations or the need for statistical tables, embedding computer output is necessary. Microsoft Excel is the primary software package used in this text. We chose Excel over other statistical packages based on reviewer feedback and the fact that students benefit from the added spreadsheet experience. We provide instructions for using Minitab, SPSS, JMP, and R in chapter appendices.

Using Excel to Obtain Binomial Probabilities

We use Excel's **BINOM.DIST** function to calculate binomial probabilities. In order to find $P(X = x)$, we enter "=BINOM.DIST(x, n, p, 0)" where x is the number of successes, n is the number of trials, and p is the probability of success. If we enter a "1" for the last argument in the function, then Excel returns $P(X \leq x)$.

a. In order to find the probability that exactly 70 American adults are Facebook users, $P(X = 70)$, we enter "=BINOM.DIST(70, 100, 0.68, 0)" and Excel returns 0.0791.

b. In order to find the probability that no more than 70 American adults are Facebook users, $P(X \leq 70)$, we enter "=BINOM.DIST(70, 100, 0.68, 1)" and Excel returns 0.7007.

c. In order to find the probability that at least 70 American adults are Facebook users, $P(X \geq 70) = 1 - P(X \leq 69)$, we enter "=1−BINOM.DIST(69, 100, 0.68, 1)" and Excel returns 0.3784.

. . . does a solid job of building the intuition behind the concepts and ***then adding mathematical rigor to these ideas*** *before finally verifying the results with Excel.*

Matthew Dean, *University of Southern Maine*

Real-World Exercises and Case Studies that Reinforce the Material

Mechanical and Applied Exercises

Chapter exercises are a well-balanced blend of mechanical, computational-type problems followed by more ambitious, interpretive-type problems. We have found that simpler drill problems tend to build students' confidence prior to tackling more difficult applied problems. Moreover, we repeatedly use many data sets—including house prices, rents, stock returns, salaries, and debt—in various chapters of the text. For instance, students first use these real data to calculate summary measures, make statistical inferences with confidence intervals and hypothesis tests, and finally, perform regression analysis.

Applied exercises from *The Wall Street Journal, Kiplinger's, Fortune, The New York Times, USA Today;* various websites—Census.gov, Zillow.com, Finance.yahoo.com, ESPN.com; and more.

18. Consider the following hypothesis test:

$$H_0: \mu \leq -5$$
$$H_A: \mu > -5$$

A random sample of 50 observations yields a sample mean [illegible]. The population standard deviation is 10. Calculate the [illegible]alue. What is the conclusion to the test if $\alpha = 0.05$?

[illegible]nsider the following hypothesis test:

$$H_0: \mu \leq 75$$
$$H_A: \mu > 75$$

A random sample of 100 observations yields a sample mean of 80. The population standard deviation is 30. Calculate the p-value. What is the conclusion to the test if $\alpha = 0.10$?

20. Consider the following hypothesis test:

$$H_0: \mu = -100$$
$$H_A: \mu \neq -100$$

A random sample of 36 observations yields a sample mean of −125. The population standard deviation is 42. Conduct the test at $\alpha = 0.01$.

21. Consider the following hypotheses:

$$H_0: \mu = 120$$
$$H_A: \mu \neq 120$$

The population is normally distributed with a population standard deviation of 46.

a. If $\bar{x} = 132$ and $n = 50$, what is the conclusion at the 5% significance level?

b. If $\bar{x} = 108$ and $n = 50$, what is the conclusion at the 10% significance level?

22. FILE ***Excel_1.*** Given the accompanying sample data, use Excel's formula options to determine if the population mean is less than 125 at the 5% significance level. Assume that the population is normally distributed and that the population standard deviation equals 12.

23. FILE ***Excel_2.*** Given the accompanying sample data, use

25. Customers at Costco spend an average of $130 per trip (*The Wall Street Journal,* October 6, 2010). One of Costco's rivals would like to determine whether its customers spend more per trip. A survey of the receipts of 25 customers found that the sample mean was $135.25. Assume that the population standard deviation is $10.50 and that spending follows a normal distribution.

a. Specify the null and alternative hypotheses to test whether average spending at the rival's store is more than $130.

b. Calculate the value of the test statistic and the p-value.

c. At the 5% significance level, what is the conclusion to the test?

26. In May 2008, CNN reported that sports utility vehicles (SUVs) are plunging toward the "endangered" list. Due to the uncertainty of oil prices and environmental concerns, consumers are replacing gas-guzzling vehicles with fuel-efficient smaller cars. As a result, there has been a big drop in the demand for new as well as used SUVs. A sales manager of a used car dealership for SUVs believes that it takes more than 90 days, on average, to sell an SUV. In order to test his claim, he samples 40 recently sold SUVs and finds that it took an average of 95 days to sell an SUV. He believes that the population standard deviation is fairly stable at 20 days.

a. State the null and the alternative hypotheses for the test.

b. What is the p-value?

c. Is the sales manager's claim justified at $\alpha = 0.01$?

27. According to the *Centers for Disease Control and Prevention* (February 18, 2016), 1 in 3 American adults do not get enough sleep. A researcher wants to determine if Americans are sleeping less than the recommended 7 hours of sleep on weekdays. He takes a random sample of 150 Americans and computes the average sleep time of 6.7 hours on weekdays. Assume that the population is normally distributed with a known standard deviation of 2.1 hours. Test the researcher's claim at $\alpha = 0.01$.

I especially like the introductory cases, the ***quality of the end-of-section problems,*** *and the writing examples.*

Dave Leupp, *University of Colorado at Colorado Springs*

Their exercises and problems are excellent!

Erl Sorensen, *Bentley University*

Features that Go Beyond the Typical

Conceptual Review

At the end of each chapter, we present a conceptual review that provides a more holistic approach to reviewing the material. This section revisits the learning outcomes and provides the most important definitions, interpretations, and formulas.

CONCEPTUAL REVIEW

LO 5.1 Describe a discrete random variable and its probability distribution.

A **random variable** summarizes outcomes of an experiment with numerical values. A **discrete random variable** assumes a countable number of distinct values, whereas a **continuous random variable** is characterized by uncountable values in an interval.

The **probability mass function** for a discrete random variable X is a list of the values of X with the associated probabilities; that is, the list of all possible pairs $(x, P(X = x))$. The **cumulative distribution function** of X is defined as $P(X \leq x)$.

LO 5.2 Calculate and interpret summary measures for a discrete random variable.

For a discrete random variable X with values $x_1, x_2, x_3, \ldots$, which occur with probabilities $P(X = x_i)$, the **expected value** of X is calculated as $E(X) = \mu = \Sigma x_i P(X = x_i)$. We interpret the expected value as the long-run average value of the random variable over infinitely many independent repetitions of an experiment. Measures of dispersion indicate whether the values of X are clustered about μ or widely scattered from μ. The variance of X is calculated as $Var(X) = \sigma^2 = \Sigma(x_i - \mu)^2 P(X = x_i)$. The standard deviation of X is $SD(X) = \sigma = \sqrt{\sigma^2}$.

In general, a **risk-averse consumer** expects a reward for taking risk. A risk-averse consumer may decline a risky prospect even if it offers a positive expected gain. A **risk-neutral consumer** completely ignores risk and always accepts a prospect that offers a positive expected gain.

LO 5.3 Calculate and interpret probabilities for a binomial random variable.

A **Bernoulli process** is a series of n independent and identical trials of an experiment such that on each trial there are only two possible outcomes, conventionally labeled "success" and "failure." The probabilities of success and failure, denoted p and $1 - p$, remain the same from trial to trial.

For a **binomial random variable** X, the probability of x successes in n Bernoulli trials is $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x}$ for $x = 0, 1, 2, \ldots, n$.

The expected value, the variance, and the standard deviation of a binomial random variable are $E(X) = np$, $Var(X) = \sigma^2 = np(1 - p)$, and $SD(X) = \sigma = \sqrt{np(1 - p)}$, respectively.

Most texts basically list what one should have learned but don't add much to that. You do a good job of reminding the reader of what was covered and what was most important about it.

Andrew Koch, *James Madison University*

They have gone beyond the typical [summarizing formulas] and I like the structure. This is a very strong feature of this text.

Virginia M. Miori, *St. Joseph's University*

Students—study more efficiently, retain more and achieve better outcomes. Instructors—focus on what you love—teaching.

SUCCESSFUL SEMESTERS INCLUDE CONNECT

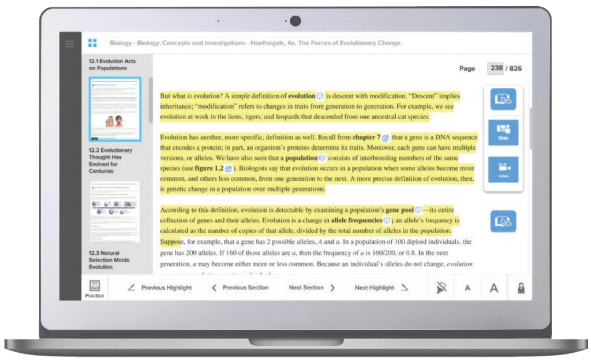
FOR INSTRUCTORS

You're in the driver's seat.

Want to build your own course? No problem. Prefer to use our turnkey, prebuilt course? Easy. Want to make changes throughout the semester? Sure. And you'll save time with Connect's auto-grading too.

65%

Less Time Grading



They'll thank you for it.

Adaptive study resources like SmartBook® help your students be better prepared in less time. You can transform your class time from dull definitions to dynamic debates. Hear from your peers about the benefits of Connect at **www.mheducation.com/highered/connect**

Make it simple, make it affordable.

Connect makes it easy with seamless integration using any of the major Learning Management Systems—Blackboard®, Canvas, and D2L, among others—to let you organize your course in one convenient location. Give your students access to digital materials at a discount with our inclusive access program. Ask your McGraw-Hill representative for more information.



©Hill Street Studios/Tobin Rogers/Blend Images LLC



Solutions for your challenges.

A product isn't a solution. Real solutions are affordable, reliable, and come with training and ongoing support when you need it and how you want it. Our Customer Experience Group can also help you troubleshoot tech problems—although Connect's 99% uptime means you might not need to call them. See for yourself at **status.mheducation.com**

For Students



©Shutterstock/wavebreakmedia

Effective, efficient studying.

Connect helps you be more productive with your study time and get better grades using tools like SmartBook, which highlights key concepts and creates a personalized study plan. Connect sets you up for success, so you walk into class with confidence and walk out with better grades.

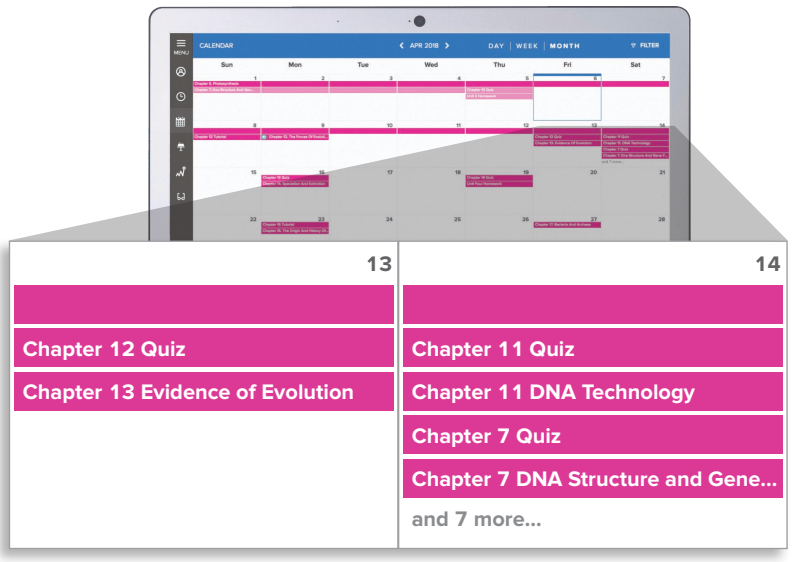
> "I really liked this app—it made it easy to study when you don't have your textbook in front of you."
>
> - Jordan Cunningham,
> Eastern Washington University

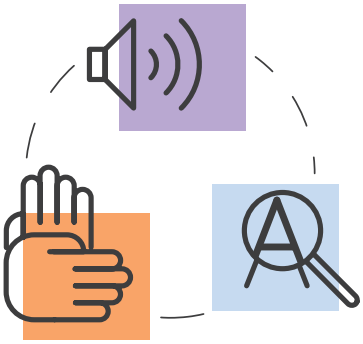
Study anytime, anywhere.

Download the free ReadAnywhere app and access your online eBook when it's convenient, even if you're offline. And since the app automatically syncs with your eBook in Connect, all of your notes are available every time you open it. Find out more at **www.mheducation.com/readanywhere**

No surprises.

The Connect Calendar and Reports tools keep you on track with the work you need to get done and your assignment scores. Life gets busy; Connect tools help you keep learning through it all.





Learning for everyone.

McGraw-Hill works directly with Accessibility Services Departments and faculty to meet the learning needs of all students. Please contact your Accessibility Services office and ask them to email accessibility@mheducation.com, or visit **www.mheducation.com/accessibility** for more information.

What Resources are Available for Instructors?

Instructor Library

The *Connect* Instructor Library is your repository for additional resources to improve student engagement in and out of class. You can select and use any asset that enhances your lecture. The *Connect* Instructor Library includes:

- PowerPoint presentations
- Excel Data Files
- Test Bank
- Instructor's Solutions Manual
- Digital Image Library

Tegrity Campus: Lectures 24/7

Tegrity Campus is integrated in *Connect* to help make your class time available 24/7. With Tegrity, you can capture each one of your lectures in a searchable format for students to review when they study and complete assignments using *Connect.* With a simple one-click start-and-stop process, you can capture everything that is presented to students during your lecture from your computer, including audio. Students can replay any part of any class with easy-to-use browser-based viewing on a PC or Mac.

Educators know that the more students can see, hear, and experience class resources, the better they learn. In fact, studies prove it. With *Tegrity Campus,* students quickly recall key moments by using *Tegrity Campus*'s unique search feature. This search helps students efficiently find what they need, when they need it, across an entire semester of class recordings. Help turn all your students' study time into learning moments immediately supported by your lecture. To learn more about *Tegrity,* watch a two-minute Flash demo at http://tegritycampus.mhhe.com.

ALEKS

ALEKS is an assessment and learning program that provides individualized instruction in Business Statistics, Business Math, and Accounting. Available online in partnership with McGraw-Hill Education, ALEKS interacts with students much like a skilled human tutor, with the ability to assess precisely a student's knowledge and provide instruction on the exact topics the student is most ready to learn. By providing topics to meet individual students' needs, allowing students to move between explanation and practice, correcting and analyzing errors, and defining terms, ALEKS helps students to master course content quickly and easily.

ALEKS also includes an instructor module with powerful, assignment-driven features and extensive content flexibility. ALEKS simplifies course management and allows instructors to spend less time with administrative tasks and more time directing student learning. To learn more about ALEKS, visit www.aleks.com.

MegaStat® for Microsoft Excel®

MegaStat® by J. B. Orris of Butler University is a full-featured Excel add-in that is available online through the *MegaStat* website at **www.mhhe.com/megastat** or through an access card packaged with the text. It works with Excel 2016, 2013, and 2010 (and Excel: Mac 2016). On the website, students have 10 days to successfully download and install *MegaStat* on their local computer. Once installed, *MegaStat* will remain active in Excel with no expiration date or time limitations. The software performs statistical analyses within an Excel workbook. It does basic functions, such as descriptive statistics, frequency distributions, and probability calculations, as well as hypothesis testing, ANOVA, and regression. *MegaStat* output is carefully formatted, and its ease-of-use features include Auto Expand for quick data selection and Auto Label detect. Since *MegaStat* is easy to use, students can focus on learning statistics without being distracted by the software. *MegaStat* is always available from Excel's main menu. Selecting a menu item pops up a dialog box. Screencam tutorials are included that provide a walkthrough of major business statistics topics. Help files are built in, and an introductory user's manual is also included.

What Resources are Available for Students?

deviation is $100 (in $1,000s). What is the value of the test statistic and the *p*-value?

c. At $\alpha = 0.05$, what is the conclusion to the test? Is the realtor's claim supported by the data?

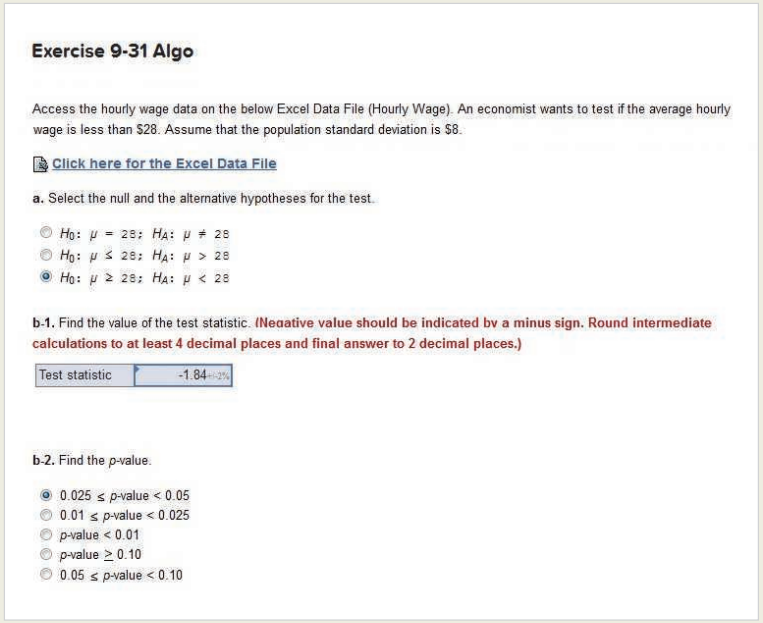
30. FILE ***Home_Depot.*** The data accompanying this exercise show the weekly stock price for Home Depot. Assume that stock prices are normally distributed with a population standard deviation of $3.
    a. State the null and the alternative hypotheses in order to test whether or not the average weekly stock price differs from $30.
    b. Find the value of the test statistic and the *p*-value.
    c. At $\alpha = 0.05$, can you conclude that the average weekly stock price does not equal $30?
31. FILE ***Hourly_Wage.*** An economist wants to test if the average hourly wage is less than $22. Assume that the population standard deviation is $6.
    a. State the null and the alternative hypotheses for the test.
    b. The data accompanying this exercise show hourly wages. Find the value of the test statistic and the *p*-value.
    c. At $\alpha = 0.05$, what is the conclusion to the test? Is the average hourly wage less than $22?
32. FILE ***CT_Undergrad_Debt.*** On average, a college student graduates with $27,200 in debt (*The Boston Globe,* May 27, 2012). The data accompanying this exercise show the debt for 40 recent undergraduates from Connecticut. Assume that the population standard deviation is $5,000.
    a. A researcher believes that recent undergraduates from Connecticut have less debt than the national average. Specify the competing hypotheses to test this belief.
    b. Find the value of the test statistic and the *p*-value.
    c. Do the data support the researcher's claim, at $\alpha = 0.10$?

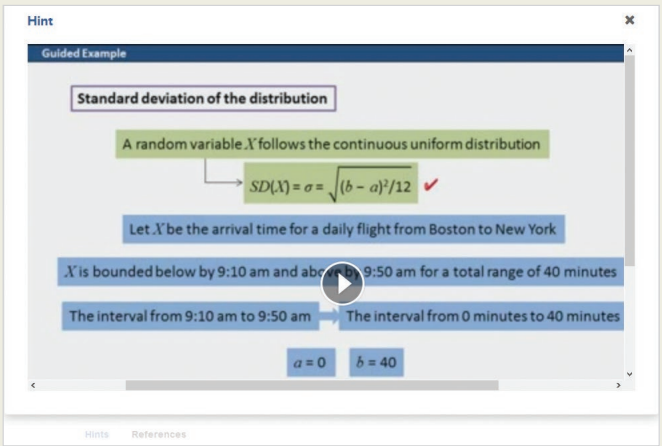
Integration of Excel Data Sets. A convenient feature is the inclusion of an Excel data file link in many problems using data files in their calculation. The link allows students to easily launch into Excel, work the problem, and return to *Connect* to key in the answer and receive feedback on their results.

15. FILE ***CT_Undergrad_Debt.*** A study reports that recent college graduates from New Hampshire face the highest average debt of $31,048 (*The Boston Globe,* May 27, 2012). A researcher from Connecticut wants to determine how recent undergraduates from that state fare. He collects data on debt from 40 recent undergraduates. A portion of the data is shown in the accompanying table. Assume that the population standard deviation is $5,000.

Debt
24040
19153
⋮
29329

a. Construct the 95% confidence interval for the mean debt of all undergraduates from Connecticut.
b. Use the 95% confidence interval to determine if the debt of Connecticut undergraduates differs from that of New Hampshire undergraduates.





Guided Examples. These narrated video walkthroughs provide students with step-by-step guidelines for solving selected exercises similar to those contained in the text. The student is given personalized instruction on how to solve a problem by applying the concepts presented in the chapter. The video shows the steps to take to work through an exercise. Students can go through each example multiple times if needed.

The *Connect* Student Resource page is the place for students to access additional resources. The Student Resource page offers students quick access to the recommended study tools, data files, and helpful tutorials on statistical programs.

McGraw-Hill Customer Care Contact Information

At McGraw-Hill, we understand that getting the most from new technology can be challenging. That's why our services don't stop after you purchase our products. You can e-mail our product specialists 24 hours a day to get product training online. Or you can search our knowledge bank of frequently asked questions on our support website.

For customer support, call **800-331-5094** or visit **www.mhhe.com/support**. One of our technical support analysts will be able to assist you in a timely fashion.

ACKNOWLEDGMENTS

We would like to acknowledge the following people for providing useful comments and suggestions for past and present editions of all aspects of *Business Statistics.*

John Affisco
Hofstra University
Mehdi Afiat
College of Southern Nevada
Mohammad Ahmadi
University of Tennessee—Chattanooga
Sung Ahn
Washington State University
Mohammad Ahsanullah
Rider University
Imam Alam
University of Northern Iowa
Mostafa Aminzadeh
Towson University
Ardavan Asef-Vaziri
California State University
Antenah Ayanso
Brock University
Scott Bailey
Troy University
Jayanta Bandyopadhyay
Central Michigan University
Samir Barman
University of Oklahoma
Douglas Barrett
University of North Alabama
John Beyers
University of Maryland
Arnab Bisi
Purdue University—West Lafayette
Gary Black
University of Southern Indiana
Randy Boan
Aims Community College
Matthew Bognar
University of Iowa
Juan Cabrera
Ramapo College of New Jersey
Scott Callan
Bentley University
Gregory Cameron
Brigham Young University
Kathleen Campbell
St. Joseph's University
Alan Cannon
University of Texas—Arlington
Michael Cervetti
University of Memphis
Samathy Chandrashekar
Salisbury University
Gary Huaite Chao
University of Pennsylvania—Kutztown
Sangit Chatterjee
Northeastern University
Leida Chen
California Polytechnic State University
Anna Chernobai
Syracuse University
Alan Chesen
Wright State University
Juyan Cho
Colorado State University—Pueblo
Alan Chow
University of South Alabama
Bruce Christensen
Weber State University
Howard Clayton
Auburn University
Robert Collins
Marquette University
M. Halim Dalgin
Kutztown University
Tom Davis
University of Dayton
Matthew Dean
University of Maine
Jason Delaney
University of Arkansas—Little Rock
Ferdinand DiFurio
Tennessee Tech University
Matt Dobra
UMUC
Luca Donno
University of Miami
Joan Donohue
University of South Carolina
David Doorn
University of Minnesota
James Dunne
University of Dayton
Mike Easley
University of New Orleans
Erick Elder
University of Arkansas—Little Rock
Ashraf ElHoubi
Lamar University
Roman Erenshteyn
Goldey-Beacom College
Grace Esimai
University of Texas—Arlington
Soheila Fardanesh
Towson University
Carol Flannery
University of Texas—Dallas
Sydney Fletcher
Mississippi Gulf Coast Community College
Andrew Flight
Portland State University
Samuel Frame
Cal Poly San Luis Obispo
Priya Francisco
Purdue University
Vickie Fry
Westmoreland County Community College
Ed Gallo
Sinclair Community College
Glenn Gilbreath
Virginia Commonwealth University
Robert Gillette
University of Kentucky
Xiaoning Gilliam
Texas Tech University
Mark Gius
Quinnipiac University
Malcolm Gold
Saint Mary's University of Minnesota
Michael Gordinier
Washington University

Deborah Gougeon
University of Scranton
Don Gren
Salt Lake Community College
Thomas G. Groleau
Carthage College
Babita Gupta
CSU Monterey Bay
Robert Hammond
North Carolina State University
Jim Han
Florida Atlantic University
Elizabeth Haran
Salem State University
Jack Harshbarger
Montreat College
Edward Hartono
University of Alabama—Huntsville
Clifford Hawley
West Virginia University
Santhi Heejebu
Cornell College
Paul Hong
University of Toledo
Ping-Hung Hsieh
Oregon State University
Marc Isaacson
Augsburg College
Mohammad Jamal
Northern Virginia Community College
Robin James
Harper College
Molly Jensen
University of Arkansas
Craig Johnson
Brigham Young University—Idaho
Janine Sanders Jones
University of St. Thomas
Vivian Jones
Bethune—Cookman University
Jerzy Kamburowski
University of Toledo
Howard Kaplon
Towson University
Krishna Kasibhatla
North Carolina A&T State University
Mohammad Kazemi
University of North Carolina—Charlotte
Ken Kelley
University of Notre Dame
Lara Khansa
Virginia Tech
Esther C. Klein
St. Francis College
Ronald Klimberg
St. Joseph's University
Andrew Koch
James Madison University
Subhash Kochar
Portland State University
Brandon Koford
Weber University
Randy Kolb
St. Cloud State University
Vadim Kutsyy
San Jose State University
Francis Laatsch
University of Southern Mississippi
David Larson
University of South Alabama
John Lawrence
California State University—Fullerton
Shari Lawrence
Nicholls State University
Radu Lazar
University of Maryland
David Leupp
University of Colorado—Colorado Springs
Carel Ligeon
Auburn University—Montgomery
Carin Lightner
North Carolina A&T State University
Constance Lightner
Fayetteville State University
Scott Lindsey
Dixie State College of Utah
Ken Linna
Auburn University—Montgomery
Andy Litteral
University of Richmond
Jun Liu
Georgia Southern University
Chung-Ping Loh
University of North Florida
Salvador Lopez
University of West Georgia
John Loucks
St. Edward's University
Cecilia Maldonado
Georgia Southwestern State University
Farooq Malik
University of Southern Mississippi
Ken Mayer
University of Nebraska—Omaha
Bradley McDonald
Northern Illinois University
Elaine McGivern
Duquesne University
John McKenzie
Babson University
Norbert Michel
Nicholls State University
John Miller
Sam Houston State University
Virginia Miori
St. Joseph's University
Prakash Mirchandani
University of Pittsburgh
Jason Molitierno
Sacred Heart University
Elizabeth Moliski
University of Texas—Austin
Joseph Mollick
Texas A&M University—Corpus Christi
James Moran
Oregon State University
Khosrow Moshirvaziri
California State University—Long Beach
Tariq Mughal
University of Utah
Patricia Mullins
University of Wisconsin—Madison
Kusum Mundra
Rutgers University—Newark
Anthony Narsing
Macon State College
Robert Nauss
University of Missouri—St. Louis
Satish Nayak
University of Missouri—St. Louis
Thang Nguyen
California State University—Long Beach
Mohammad Oskoorouchi
California State University—San Marcos
Barb Osyk
University of Akron
Bhavik Pathak
Indiana University South Bend
Scott Paulsen
Illinois Central College
James Payne
Calhoun Community College
Norman Pence
Metropolitan State College of Denver

Dane Peterson
Missouri State University
Joseph Petry
University of Illinois—Urbana/Champaign
Courtney Pham
Missouri State University
Martha Pilcher
University of Washington
Cathy Poliak
University of Wisconsin—Milwaukee
Simcha Pollack
St. John's University
Hamid Pourmohammadi
California State University—Dominguez Hills
Tammy Prater
Alabama State University
Zbigniew H. Przasnyski
Loyola Marymount University
Manying Qiu
Virginia State University
Troy Quast
Sam Houston State University
Michael Racer
University of Memphis
Srikant Raghavan
Lawrence Technological University
Bharatendra Rai
University of Massachusetts—Dartmouth
Michael Aaron Ratajczyk
Saint Mary's University of Minnesota
Tony Ratcliffe
James Madison University
David Ravetch
University of California
Bruce Reinig
San Diego State University
Darlene Riedemann
Eastern Illinois University
David Roach
Arkansas Tech University
Carolyn Rochelle
East Tennessee State University
Alfredo Romero
North Carolina A&T State University
Ann Rothermel
University of Akron
Jeff Rummel
Emory University
Deborah Rumsey
The Ohio State University
Stephen Russell
Weber State University
William Rybolt
Babson College
Fati Salimian
Salisbury University
Fatollah Salimian
Perdue School of Business
Samuel Sarri
College of Southern Nevada
Jim Schmidt
University of Nebraska—Lincoln
Patrick Scholten
Bentley University
Bonnie Schroeder
Ohio State University
Pali Sen
University of North Florida
Donald Sexton
Columbia University
Vijay Shah
West Virginia University—Parkersburg
Dmitriy Shaltayev
Christopher Newport University
Soheil Sibdari
University of Massachusetts—Dartmouth
Prodosh Simlai
University of North Dakota
Harvey Singer
George Mason University
Harry Sink
North Carolina A&T State University
Don Skousen
Salt Lake Community College
Robert Smidt
California Polytechnic State University
Gary Smith
Florida State University
Antoinette Somers
Wayne State University
Ryan Songstad
Augustana College
Erland Sorensen
Bentley University
Arun Kumar Srinivasan
Indiana University—Southeast
Scott Stevens
James Madison University
Alicia Strandberg
Temple University
Linda Sturges
Suny Maritime College
Wendi Sun
Rockland Trust
Bedassa Tadesse
University of Minnesota
Pandu Tadikamalta
University of Pittsburgh
Roberto Duncan Tarabay
University of Wisconsin—Madison
Faye Teer
James Madison University
Deborah Tesch
Xavier University
Patrick Thompson
University of Florida
Satish Thosar
University of Redlands
Ricardo Tovar-Silos
Lamar University
Quoc Hung Tran
Bridgewater State University
Elzbieta Trybus
California State University—Northridge
Fan Tseng
University of Alabama—Huntsville
Silvanus Udoka
North Carolina A&T State University
Shawn Ulrick
Georgetown University
Bulent Uyar
University of Northern Iowa
Ahmad Vakil
Tobin College of Business
Tim Vaughan
University of Wisconsin—Eau Claire
Raja Velu
Syracuse University
Holly Verhasselt
University of Houston—Victoria
Zhaowei Wang
Citizens Bank
Rachel Webb
Portland State University
Kyle Wells
Dixie State College
Alan Wheeler
University of Missouri—St. Louis
Mary Whiteside
University of Texas—Arlington
Blake Whitten
University of Iowa
Rick Wing
San Francisco State University

Jan Wolcott
Wichita State University
Rongning Wu
Baruch College
John Yarber
Northeast Mississippi Community College
John C. Yi
St. Joseph's University
Kanghyun Yoon
University of Central Oklahoma
Mark Zaporowski
Canisius College
Ali Zargar
San Jose State University
Dewit Zerom
California State University
Eugene Zhang
Midwestern State University
Ye Zhang
Indiana University—Purdue University—Indianapolis
Yi Zhang
California State University—Fullerton
Yulin Zhang
San Jose State University
Wencang Zhou
Baruch College
Zhen Zhu
University of Central Oklahoma

The editorial staff of McGraw-Hill Education are deserving of our gratitude for their guidance throughout this project, especially Noelle Bathurst, Pat Frederickson, Ryan McAndrews, Harper Christopher, Daryl Horrocks, and Egzon Shaqiri. We would also like to thank Eric Kambestad and Matt Kesselring for their outstanding research assistance.

BRIEF CONTENTS

CHAPTER 1 Statistics and Data *2*

CHAPTER 2 Tabular and Graphical Methods *18*

CHAPTER 3 Numerical Descriptive Measures *60*

CHAPTER 4 Introduction to Probability *104*

CHAPTER 5 Discrete Probability Distributions *144*

CHAPTER 6 Continuous Probability Distributions *182*

CHAPTER 7 Sampling and Sampling Distributions *218*

CHAPTER 8 Interval Estimation *258*

CHAPTER 9 Hypothesis Testing *292*

CHAPTER 10 Comparisons Involving Means *328*

CHAPTER 11 Comparisons Involving Proportions *370*

CHAPTER 12 Basics of Regression Analysis *402*

CHAPTER 13 More on Regression Analysis *456*

APPENDIXES

APPENDIX A Tables *510*

APPENDIX B Answers to Selected Even-Numbered Exercises *520*

Glossary *537*

Index *I-1*

Mc Graw Hill Education
5011333

CONTENTS

CHAPTER 1

STATISTICS AND DATA 2

1.1 **The Relevance of Statistics** 4

1.2 **What is Statistics?** 5
The Need for Sampling 6
Cross-Sectional and Time Series Data 6
Structured and Unstructured Data 7
Big Data 8
Data on the Web 8

1.3 **Variables and Scales of Measurement** 10
The Nominal Scale 11
The Ordinal Scale 12
The Interval Scale 13
The Ratio Scale 14
Synopsis of Introductory Case 15

Conceptual Review 16

CHAPTER 2

TABULAR AND GRAPHICAL METHODS 18

2.1 **Summarizing Qualitative Data** 20
Pie Charts and Bar Charts 21
Cautionary Comments When Constructing or Interpreting Charts or Graphs 24
Using Excel to Construct a Pie Chart and a Bar Chart 24
A Pie Chart 24
A Bar Chart 25

2.2 **Summarizing Quantitative Data** 27
Guidelines for Constructing a Frequency Distribution 28
Synopsis Of Introductory Case 32
Histograms, Polygons, and Ogives 32
Using Excel to Construct a Histogram, a Polygon, and an Ogive 36
A Histogram Constructed from Raw Data 36
A Histogram Constructed from a Frequency Distribution 37
A Polygon 38
An Ogive 38

2.3 **Stem-and-Leaf Diagrams** 42

2.4 **Scatterplots** 44
Using Excel to Construct a Scatterplot 46
Writing with Statistics 47
Conceptual Review 49

Additional Exercises And Case Studies 50
Exercises 50
Case Studies 53

Appendix 2.1 **Guidelines for Other Software Packages** 55

CHAPTER 3

NUMERICAL DESCRIPTIVE MEASURES 60

3.1 **Measures of Central Location** 62
The Mean 62
The Median 64
The Mode 65
The Weighted Mean 66
Using Excel to Calculate Measures of Central Location 67
Using Excel's Function Option 67
Using Excel's Data Analysis Toolpak Option 68
Note on Symmetry 69

3.2 **Percentiles and Boxplots** 71
Calculating the *p*th Percentile 72
Note on Calculating Percentiles 73
Constructing and Interpreting a Boxplot 73

3.3 **Measures of Dispersion** 76
Range 76
The Mean Absolute Deviation 77
The Variance and the Standard Deviation 78
The Coefficient of Variation 79
Using Excel to Calculate Measures of Dispersion 80
Using Excel's Function Option 80
Using Excel's Data Analysis Toolpak Option 80

3.4 **Mean-Variance Analysis and the Sharpe Ratio** 81
Synopsis of Introductory Case 83

3.5 **Analysis of Relative Location** 84
Chebyshev's Theorem 85
The Empirical Rule 85
z-Scores 86

3.6 **Summarizing Grouped Data** 89

3.7 **Measures of Association** 92
Using Excel to Calculate Measures of Association 94
Writing with Statistics 95
Conceptual Review 97

Additional Exercises and Case Studies 98
Exercises 98
Case Studies 101

Appendix 3.1: **Guidelines for Other Software Packages** 102

CHAPTER 4

INTRODUCTION TO PROBABILITY 104

4.1 **Fundamental Probability Concepts** 106
Events 107
Assigning Probabilities 109

4.2 Rules of Probability *113*
The Complement Rule *113*
The Addition Rule *114*
The Addition Rule for Mutually Exclusive Events *115*
Conditional Probability *116*
Independent and Dependent Events *118*
The Multiplication Rule *119*
The Multiplication Rule for Independent Events *119*

4.3 Contingency Tables and Probabilities *123*
A Note on Independence *126*
Synopsis of Introductory Case *126*

4.4 The Total Probability Rule and Bayes' Theorem *128*
The Total Probability Rule *128*
Bayes' Theorem *131*
Writing With Statistics *135*
Conceptual Review *137*

Additional Exercises and Case Studies *138*
Exercises 138
Case Studies *142*

CHAPTER 5

DISCRETE PROBABILITY DISTRIBUTIONS *144*

5.1 Random Variables and Discrete Probability Distributions *146*
The Discrete Probability Distribution *147*

5.2 Expected Value, Variance, and Standard Deviation *151*
Expected Value *152*
Variance and Standard Deviation *152*
Risk Neutrality and Risk Aversion *153*

5.3 The Binomial Distribution *156*
Using Excel to Obtain Binomial Probabilities *161*

5.4 The Poisson Distribution *164*
Synopsis of Introductory Case *167*
Using Excel to Obtain Poisson Probabilities *167*

5.5 The Hypergeometric Distribution *169*
Using Excel to Obtain Hypergeometric Probabilities *171*
Writing with Statistics *173*
Conceptual Review *175*

Additional Exercises and Case Studies *176*
Exercises 176
Case Studies *178*

Appendix 5.1: Guidelines for Other Software Packages *179*

CHAPTER 6

CONTINUOUS PROBABILITY DISTRIBUTIONS *182*

6.1 Continuous Random Variables and the Uniform Distribution *184*
The Continuous Uniform Distribution *185*

6.2 The Normal Distribution *188*
Characteristics of the Normal Distribution *189*
The Standard Normal Distribution *190*
Finding a Probability for a Given *z* Value *191*
Finding a *z* Value for a Given Probability *193*
The Transformation of Normal Random Variables *195*
Synopsis of Introductory Case *199*
A Note on the Normal Approximation of the Binomial Distribution *199*
Using Excel for the Normal Distribution *199*

6.3 The Exponential Distribution *204*
Using Excel for the Exponential Distribution *207*
Writing with Statistics *209*
Conceptual Review *210*

Additional Exercises and Case Studies *211*
Exercises 211
Case Studies *214*

Appendix 6.1: Guidelines for Other Software Packages *215*

CHAPTER 7

SAMPLING AND SAMPLING DISTRIBUTIONS *218*

7.1 Sampling *220*
Classic Case of a "Bad" Sample: The *Literary Digest* Debacle of 1936 *220*
Trump's Stunning Victory in 2016 *221*
Sampling Methods *222*
Using Excel to Generate a Simple Random Sample *224*

7.2 The Sampling Distribution of the Sample Mean *225*
The Expected Value and the Standard Error of the Sample Mean *226*
Sampling from a Normal Population *227*
The Central Limit Theorem *228*

7.3 The Sampling Distribution of the Sample Proportion *232*
The Expected Value and the Standard Error of the Sample Proportion *232*
Synopsis of Introductory Case *236*

7.4 The Finite Population Correction Factor *237*

7.5 Statistical Quality Control *240*
Control Charts *241*
Using Excel to Create a Control Chart *244*
Writing with Statistics *247*
Conceptual Review *248*
Additional Exercises and Case Studies *250*
Exercises 250
Case Studies *252*

Appendix 7.1: Derivation of the Mean and the Variance for $\overline{X}$ and $\overline{P}$ *253*

Appendix 7.2: Properties of Point Estimators *254*

Appendix 7.3: Guidelines for Other Software Packages *255*

CHAPTER 8

INTERVAL ESTIMATION *258*

8.1 Confidence Interval for the Population Mean when σ is Known *260*
Constructing a Confidence Interval for μ When σ Is Known *261*
The Width of a Confidence Interval *263*
Using Excel to Construct a Confidence Interval for μ When σ Is Known *265*

8.2 Confidence Interval for the Population Mean When σ is Unknown *268*
The *t* Distribution *268*
Summary of the t_{df} Distribution *268*
Locating t_{df} Values and Probabilities *269*
Constructing a Confidence Interval for μ When σ Is Unknown *270*
Using Excel to Construct a Confidence Interval for μ When σ Is Unknown *271*

8.3 Confidence Interval for the Population Proportion *275*

8.4 Selecting the Required Sample Size *278*
Selecting *n* to Estimate μ *279*
Selecting *n* to Estimate *p* *280*
Synopsis of Introductory Case *281*
Writing with Statistics *282*
Conceptual Review *284*

Additional Exercises and Case Studies *285*
Exercises 285
Case Studies *288*

Appendix 8.1: Guidelines for Other Software Packages *290*

CHAPTER 9

HYPOTHESIS TESTING *292*

9.1 Introduction to Hypothesis Testing *294*
The Decision to "Reject" or "Not Reject" the Null Hypothesis *294*
Defining the Null and the Alternative Hypotheses *295*
Type I and Type II Errors *297*

9.2 Hypothesis Test for the Population Mean When σ is Known *300*
The *p*-Value Approach *300*
Confidence Intervals and Two-Tailed Hypothesis Tests *304*
Using Excel to Test μ When σ Is Known *305*
One Last Remark *306*

9.3 Hypothesis Test for the Population Mean When σ is Unknown *308*
Using Excel to Test μ When σ is Unknown *309*
Synopsis of Introductory Case *310*

9.4 Hypothesis Test for the Population Proportion *313*
Writing with Statistics *317*
Conceptual Review *318*

Additional Exercises and Case Studies *320*
Exercises 320
Case Studies *322*

Appendix 9.1: The Critical Value Approach *324*

Appendix 9.2: Guidelines for Other Software Packages *326*

CHAPTER 10

COMPARISONS INVOLVING MEANS *328*

10.1 Inference Concerning the Difference Between Two Means *330*
Confidence Interval for $\mu_1 - \mu_2$ *330*
Hypothesis Test for $\mu_1 - \mu_2$ *332*
Using Excel for Testing Hypotheses about $\mu_1 - \mu_2$ *334*

10.2 Inference Concerning Mean Differences *340*
Recognizing a Matched-Pairs Experiment *341*
Confidence Interval for μ_D *341*
Hypothesis Test for μ_D *342*
Using Excel for Testing Hypotheses about μ_D *344*
Synopsis of Introductory Case *345*

10.3 Inference Concerning Differences Among Many Means *349*
The *F* Distribution *349*
Finding $F_{(df_1,df_2)}$ Values and Probabilities *349*
One-Way ANOVA Test *350*
Between-Treatments Estimate of σ^2: *MSTR* *352*
Within-Treatments Estimate of σ^2: *MSE* *353*
The One-Way ANOVA Table *355*
Using Excel to Construct a One-Way ANOVA Table *355*
Writing with Statistics *359*
Conceptual Review *360*

Additional Exercises and Case Studies *362*
Exercises 362
Case Studies *366*

Appendix 10.1: Guidelines for Other Software Packages *367*

CHAPTER 11

COMPARISONS INVOLVING PROPORTIONS *370*

11.1 Inference Concerning the Difference Between Two Proportions *372*
Confidence Interval for $p_1 - p_2$ *372*
Hypothesis Test for $p_1 - p_2$ *373*

11.2 Goodness-Of-Fit Test for a Multinomial Experiment *378*
The χ^2 Distribution *378*
Finding χ^2_{df} Values and Probabilities *379*

11.3 Chi-Square Test For Independence *385*
Calculating Expected Frequencies *386*
Synopsis of Introductory Case *389*
Writing with Statistics *392*
Conceptual Review *393*

Additional Exercises and Case Studies *394*
Exercises 394
Case Studies *398*

Appendix 11.1: Guidelines for Other Software Packages *399*

CHAPTER 12

BASICS OF REGRESSION ANALYSIS *402*

12.1 The Simple Linear Regression Model *404*
Determining the Sample Regression Equation *406*
Using Excel *408*
Constructing a Scatterplot with Trendline *408*
Estimating a Simple Linear Regression Model *408*

12.2 The Multiple Linear Regression Model *411*
Using Excel to Estimate a Multiple Linear Regression Model *413*

12.3 Goodness-of-Fit Measures *416*
The Standard Error of the Estimate *416*
The Coefficient of Determination, R^2 *417*
The Adjusted R^2 *419*

12.4 Tests of Significance *422*
Tests of Individual Significance *422*
A Test for a Nonzero Slope Coefficient *425*
Test of Joint Significance *427*
Reporting Regression Results *429*
Synopsis of Introductory Case *429*

12.5 Model Assumptions and Common Violations 433
Common Violation 1: Nonlinear Patterns *435*
Detection *435*
Remedy *436*
Common Violation 2: Multicollinearity *436*
Detection *437*
Remedy *438*
Common Violation 3: Changing Variability *438*
Detection *438*
Remedy *439*
Common Violation 4: Correlated Observations *440*
Detection *440*
Remedy *441*
Common Violation 5: Excluded Variables *441*
Remedy *441*
Summary *441*
Using Excel to Construct Residual Plots *442*
Writing with Statistics *444*
Conceptual Review *446*
Additional Exercises and Case Studies *448*
Case Studies *451*
Appendix 12.1: Guidelines for Other Software Packages *453*

CHAPTER 13

MORE ON REGRESSION ANALYSIS *456*

13.1 Dummy Variables *458*
A Qualitative Explanatory Variable with Two Categories *458*
A Qualitative Explanatory Variable with Multiple Categories *461*
13.2 Interactions with Dummy Variables 467
Synopsis of Introductory Case *471*
13.3 Regression Models for Nonlinear Relationships *473*
Quadratic Regression Models *473*
Regression Models with Logarithms *478*
The Log-Log Model *478*
The Logarithmic Model *479*
The Exponential Model *480*
13.4 Trend Forecasting Models *487*
The Linear and the Exponential Trend *487*
Polynomial Trends *490*
13.5 Forecasting with Trend and Seasonality *495*
Seasonal Dummy Variables *495*
Writing with Statistics *499*
Conceptual Review *501*
Additional Exercises and Case Studies *503*
Case Studies *507*

APPENDIXES

APPENDIX A Tables *510*
APPENDIX B Answers to Selected Even-Numbered Exercises *520*

Glossary *537*
Index *I-1*

Essentials of Business Statistics

1 Statistics and Data

Learning Objectives

After reading this chapter you should be able to:

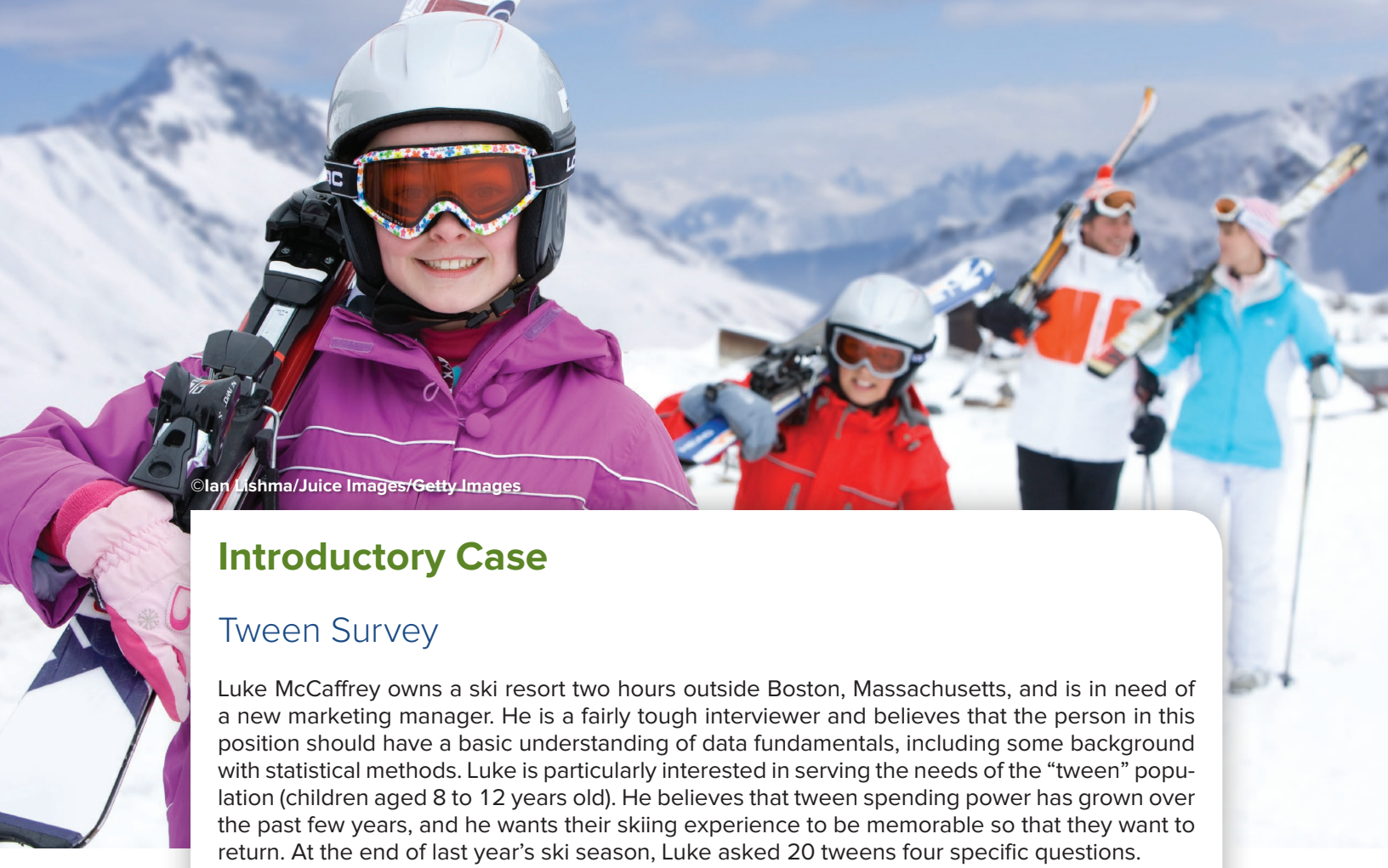
LO **1.1** **Describe the importance of statistics.**

LO **1.2** **Differentiate between descriptive statistics and inferential statistics.**

LO **1.3** **Explain the various data types.**

LO **1.4** **Describe variables and types of measurement scales.**

Every day we are bombarded with data and claims. The analysis of data and the conclusions made from data are part of the field of statistics. A proper understanding of statistics is essential in understanding more of the real world around us, including business, sports, politics, health, social interactions—just about any area of contemporary human activity. In this first chapter, we will differentiate between sound statistical conclusions and questionable conclusions. We will also introduce some important terms that will help us describe different aspects of statistics and their practical importance. You are probably familiar with some of these terms already, from reading or hearing about opinion polls, surveys, and the all-pervasive product ads. Our goal is to place what you already know about these uses of statistics within a framework that we then use for explaining where they came from and what they really mean. A major portion of this chapter is also devoted to a discussion of variables and types of measurement scales. As we will see in later chapters, we need to distinguish between different variables and measurement scales in order to choose the appropriate statistical methods for analyzing data.



Introductory Case

Tween Survey

Luke McCaffrey owns a ski resort two hours outside Boston, Massachusetts, and is in need of a new marketing manager. He is a fairly tough interviewer and believes that the person in this position should have a basic understanding of data fundamentals, including some background with statistical methods. Luke is particularly interested in serving the needs of the "tween" population (children aged 8 to 12 years old). He believes that tween spending power has grown over the past few years, and he wants their skiing experience to be memorable so that they want to return. At the end of last year's ski season, Luke asked 20 tweens four specific questions.

Q1. On your car drive to the resort, which radio station was playing?

Q2. On a scale of 1 to 4, rate the quality of the food at the resort (where 1 is poor, 2 is fair, 3 is good, and 4 is excellent).

Q3. Presently, the main dining area closes at 3:00 pm. What time do you think it should close?

Q4. How much of your *own* money did you spend at the lodge today?

The responses to these questions are shown in Table 1.1

TABLE 1.1 Tween Responses to Resort Survey

Tween	Q1	Q2	Q3	Q4	Tween	Q1	Q2	Q3	Q4
1	JAMN94.5	4	5:00 pm	20	11	JAMN94.5	3	3:00 pm	0
2	MIX104.1	2	5:00 pm	10	12	JAMN94.5	4	4:00 pm	5
3	KISS108	2	4:30 pm	10	13	KISS108	2	4:30 pm	5
4	JAMN94.5	3	4:00 pm	0	14	KISS108	2	5:00 pm	10
5	KISS108	1	3:30 pm	0	15	KISS108	3	4:00 pm	5
6	JAMN94.5	1	6:00 pm	25	16	JAMN94.5	3	6:00 pm	20
7	KISS108	2	6:00 pm	15	17	KISS108	2	5:00 pm	15
8	KISS108	3	5:00 pm	10	18	MIX104.1	4	6:00 pm	15
9	KISS108	2	4:30 pm	10	19	KISS108	1	5:00 pm	25
10	KISS108	3	4:30 pm	20	20	KISS108	2	4:30 pm	10

FILE
Tween_Survey

Luke asks each job applicant to use the information to

1. Summarize the results of the survey.
2. Provide management with suggestions for improvement.

A synopsis from the job applicant with the best answers is provided at the end of Section 1.3.

LO **1.1**

Describe the importance of statistics.

1.1 THE RELEVANCE OF STATISTICS

In order to make intelligent decisions in a world full of uncertainty, we all have to understand statistics—the language of data. Data are usually compilations of facts, figures, or other contents, both numerical and nonnumerical. Insights from data enable businesses to make better decisions, such as deepening customer engagement, optimizing operations, preventing threats and fraud, and capitalizing on new sources of revenue. We must understand statistics or risk making uninformed decisions and costly mistakes. A knowledge of statistics also provides the necessary tools to differentiate between sound statistical conclusions and questionable conclusions drawn from an insufficient number of data points, "bad" data points, incomplete data points, or just misinformation. Consider the following examples.

Example 1. After Washington, DC, had record amounts of snow in the winter of 2010, the headline of a newspaper asked, "What global warming?"

Problem with conclusion: The existence or nonexistence of climate change cannot be based on one year's worth of data. Instead, we must examine long-term trends and analyze decades' worth of data.

Example 2. A gambler predicts that his next roll of the dice will be a lucky 7 because he did not get that outcome on the last three rolls.

Problem with conclusion: As we will see later in the text when we discuss probability, the probability of rolling a 7 stays constant with each roll of the dice. It does not become more likely if it did not appear on the last roll or, in fact, any number of preceding rolls.

Example 3. On January 10, 2010, nine days prior to a special election to fill the U.S. Senate seat that was vacated due to the death of Ted Kennedy, a *Boston Globe* poll gave the Democratic candidate, Martha Coakley, a 15-point lead over the Republican candidate, Scott Brown. On January 19, 2010, Brown won 52% of the vote, compared to Coakley's 47%, and became a U.S. senator for Massachusetts.

Problem with conclusion: Critics accused the *Globe,* which had endorsed Coakley, of purposely running a bad poll to discourage voters from coming out for Brown. In reality, by the time the *Globe* released the poll, it contained old information from January 2–6, 2010. Even more problematic was that the poll included people who said that they were unlikely to vote!

Example 4. Starbucks Corp., the world's largest coffee-shop operator, reported that sales at stores open at least a year climbed 4% at home and abroad in the quarter ended December 27, 2009. Chief Financial Officer Troy Alstead said that "the U.S. is back in a good track and the international business has similarly picked up. . . . Traffic is really coming back. It's a good sign for what we're going to see for the rest of the year." (www.bloomberg.com, January 20, 2010)

Problem with conclusion: In order to calculate same-store sales growth, which compares how much each store in the chain is selling compared with a year ago, we remove stores that have closed. Given that Starbucks closed more than 800 stores over the past few years to counter large sales declines, it is likely that the sales increases in many of the stores were caused by traffic from nearby, recently closed stores. In this case, same-store sales growth may overstate the overall health of Starbucks.

Example 5. Researchers at the University of Pennsylvania Medical Center found that infants who sleep with a nightlight are much more likely to develop myopia later in life (*Nature,* May 1999).

Problem with conclusion: This example appears to commit the *correlation-to-causation fallacy.* Even if two variables are highly correlated, one does not necessarily cause the other. *Spurious correlation* can make two variables appear closely related when no causal relation exists. Spurious correlation between two variables is not based on any demonstrable relationship, but rather can be explained by confounding factors. For instance, the fact that the cost of a hamburger is correlated with how much people spend on a computer is explained by a confounding factor called inflation, which makes both the hamburger and the computer costs grow over time. In a follow-up study regarding myopia, researchers at The Ohio State University found no link between infants who sleep with a nightlight and the development of myopia (*Nature,* March 2000). They did, however, find strong links between parental myopia and the development of child myopia, and between parental myopia and the parents' use of a nightlight in their children's room. So the confounding factor for both conditions (the use of a nightlight and the development of child myopia) is parental myopia. See www.tylervigen.com/spurious-correlations for some outrageous examples of spurious correlation.

Note the diversity of the sources of these examples—the environment, psychology, polling, business, and health. We could easily include others, from sports, sociology, the physical sciences, and elsewhere. Data and data interpretation show up in virtually every facet of life, sometimes spuriously. All of the preceding examples basically misuse data to add credibility to an argument. A solid understanding of statistics provides you with tools to react intelligently to information that you read or hear.

1.2 WHAT IS STATISTICS?

LO **1.2**

Differentiate between descriptive statistics and inferential statistics.

In the broadest sense, we can define the study of statistics as the methodology of extracting useful information from a data set. Three steps are essential for doing good statistics. First, we have to find the right data, which are both complete and lacking any misrepresentation. Second, we must use the appropriate statistical tools, depending on the data at hand. Finally, an important ingredient of a well-executed statistical analysis is to clearly communicate numerical information into written language.

We generally divide the study of statistics into two branches: descriptive statistics and inferential statistics. **Descriptive statistics** refers to the summary of important aspects of a data set. This includes collecting data, organizing the data, and then presenting the data in the form of charts and tables. In addition, we often calculate numerical measures that summarize, for instance, the data's typical value and the data's variability. Today, the techniques encountered in descriptive statistics account for the most visible application of statistics—the abundance of quantitative information that is collected and published in our society every day. The unemployment rate, the president's approval rating, the Dow Jones Industrial Average, batting averages, the crime rate, and the divorce rate are but a few of the many "statistics" that can be found in a reputable newspaper on a frequent, if not daily, basis. Yet, despite the familiarity of descriptive statistics, these methods represent only a minor portion of the body of statistical applications.

The phenomenal growth in statistics is mainly in the field called inferential statistics. Generally, **inferential statistics** refers to drawing conclusions about a large set of data—called a **population**—based on a smaller set of **sample** data. A population is defined as all members of a specified group (not necessarily people), whereas a sample is a subset of that particular population. The individual values contained in a population or a sample are often referred to as **observations**. In most statistical applications, we must rely on sample data in order to make inferences about various characteristics of the population. For example, a 2016 Gallup survey found that only 50% of Millennials plan to be with their current job for more than a year. Researchers use this sample result, called a

sample statistic, in an attempt to estimate the corresponding unknown **population parameter**. In this case, the parameter of interest is the percentage of *all* Millennials who plan to be with their current job for more than a year. It is generally not feasible to obtain population data and calculate the relevant parameter directly, due to prohibitive costs and/or practicality, as discussed next.

POPULATION VERSUS SAMPLE

A population consists of all items of interest in a statistical problem. A sample is a subset of the population. We analyze sample data and calculate a sample statistic to make inferences about the unknown population parameter.

The Need for Sampling

A major portion of inferential statistics is concerned with the problem of estimating population parameters or testing hypotheses about such parameters. If we have access to data that encompass the entire population, then we would know the values of the parameters. Generally, however, we are unable to use population data for two main reasons.

- **Obtaining information on the entire population is expensive**. Consider how the monthly unemployment rate in the United States is calculated by the Bureau of Labor Statistics (BLS). Is it reasonable to assume that the BLS counts every unemployed person each month? The answer is a resounding NO! In order to do this, every home in the country would have to be contacted. Given that there are approximately 160 million individuals in the labor force, not only would this process cost too much, it would take an inordinate amount of time. Instead, the BLS conducts a monthly sample survey of about 60,000 households to measure the extent of unemployment in the United States.
- **It is impossible to examine every member of the population**. Suppose we are interested in the average length of life of a Duracell AAA battery. If we tested the duration of each Duracell AAA battery, then in the end, all batteries would be dead and the answer to the original question would be useless.

LO **1.3**

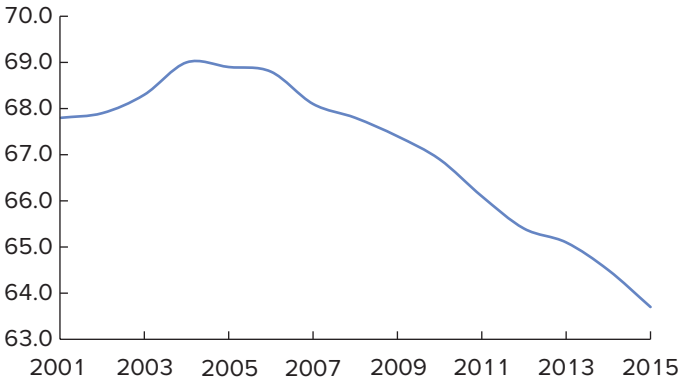
Explain the various data types.

Cross-Sectional and Time Series Data

Sample data are generally collected in one of two ways. **Cross-sectional data** refer to data collected by recording a characteristic of many subjects at the same point in time, or without regard to differences in time. Subjects might include individuals, households, firms, industries, regions, and countries. The tween data set presented in Table 1.1 in the introductory case is an example of cross-sectional data because it contains tween responses to four questions at the end of the ski season. It is unlikely that all 20 tweens took the questionnaire at exactly the same time, but the differences in time are of no relevance in this example. Other examples of cross-sectional data include the recorded scores of students in a class, the sale prices of single-family homes sold last month, the current price of gasoline in different states in the United States, and the starting salaries of recent business graduates from The Ohio State University.

Time series data refer to data collected over several time periods focusing on certain groups of people, specific events, or objects. Time series can include hourly, daily, weekly, monthly, quarterly, or annual observations. Examples of time series data include the hourly body temperature of a patient in a hospital's intensive care unit, the daily price of General Electric stock in the first quarter of 2015, the weekly exchange rate between the U.S. dollar and the euro over the past six months, the monthly sales of cars at a dealership in 2016, and the annual growth rate of India in the last decade.

FILE
Homeownership



Source: Federal Reserve Bank of St. Louis

FIGURE 1.1 Homeownership Rate (%) in the United States from 2001 through 2015

Figure 1.1 shows a plot of the national homeownership rate in the United States from 2001 through 2015. According to the U.S. Census Bureau, the national homeownership rate in the first quarter of 2016 plummeted to 63.6% from a high of 69.4% in 2004. An obvious explanation for the decline in homeownership is the stricter lending practices caused by the housing market crash in 2007 that precipitated a banking crisis and the Great Recession. This decline can also be attributed to home prices outpacing wages in the sample period.

CROSS-SECTIONAL DATA AND TIME SERIES DATA

Cross-sectional data contain values of a characteristic of many subjects at the same point or approximately the same point in time. Time series data contain values of a characteristic of a subject over time.

Structured and Unstructured Data

As mentioned earlier, consumers and businesses are increasingly turning to data to make decisions. When you hear the word "data," you probably imagine lots of numbers and perhaps some charts and graphs as well. In reality, data can come in multiple forms. For example, information exchange in social networking services such as Facebook, LinkedIn, Twitter, YouTube, and blogs also constitutes data. In order to better understand the various forms of data, we make a distinction between structured data and unstructured data.

The term **structured data** generally refers to data that has a well-defined length and format. Structured data reside in a predefined row-column format. Examples of structured data include numbers, dates, and groups of words and numbers called strings. Structured data generally consist of numerical information that is objective. In other words, structured data are not open to interpretation. The data set that appears in Table 1.1 from the introductory case is an example of structured data.

Unlike structured data, **unstructured data** (or unmodeled data) do not conform to a predefined row-column format. They tend to be textual (e.g., written reports, e-mail messages, doctor's notes, or open-ended survey responses) or have multimedia contents (e.g., photographs, videos, and audio data). Even though these data may have some implied structure (e.g., a report title, an e-mail's subject line, or a time stamp on a photograph), they are still considered unstructured because they do not conform to a row-column model required in most database systems. Social media data, such as those that appear on Facebook, LinkedIn, Twitter, YouTube, and blogs, are examples of unstructured data.

Big Data

Nowadays, businesses and organizations generate and gather more and more data at an increasing pace. The term **big data** is a catchphrase, meaning a massive volume of both structured and unstructured data that are extremely difficult to manage, process, and analyze using traditional data processing tools. Despite the challenges, big data present great opportunities to glean intelligence from data that affects company revenues, margins, and organizational efficiency.

Big data, however, do not necessarily imply complete (population) data. Take, for example, the analysis of all Facebook users. It certainly involves big data, but if we consider all Internet users in the world, Facebook data are only a very large sample. There are many Internet users who do not use Facebook, so the data on Facebook do not represent the population. Even if we define the population as pertaining to those who use online social media, Facebook is still one of many social media services that consumers use. Therefore, Facebook data would still just be considered a large sample.

In addition, we may choose not to use a big data set in its entirety even when it is available. Sometimes it is just inconvenient to analyze a very large data set because it is computationally burdensome, even with a modern, high-capacity computer system. Other times, the additional benefits of working with a big data set may not justify its associated additional resource costs. In sum, we often choose to work with a small data set, which, in a sense, is a sample drawn from big data.

STRUCTURED DATA, UNSTRUCTURED DATA, AND BIG DATA

Structured data reside in a predefined row-column format, while unstructured data do not conform to a predefined row-column format. The term big data is used to describe a massive volume of both structured and unstructured data that are extremely difficult to manage, process, and analyze using traditional data processing tools. The availability of big data, however, does not necessarily imply complete (population) data.

In this textbook, we will not cover specialized tools to manage, process, and analyze big data. Instead, we will focus on structured data. Text analytics and other sophisticated tools to analyze unstructured data are beyond the scope of this textbook.

Data on the Web



©Comstock Images/Jupiter Images

At every moment, data are being generated at an increasing velocity from countless sources in an overwhelming volume. Many experts believe that 90% of the data in the world today were created in the last two years alone. Not surprisingly, businesses continue to grapple with how to best ingest, understand, and operationalize large volumes of data. We access much of the data in this text by simply using a search engine like Google. These search engines direct us to data-providing websites. For instance, searching for economic data may lead you to the Bureau of Economic Analysis (www.bea.gov), the Bureau of Labor Statistics (www.bls.gov/data), the Federal Reserve Economic Data (research.stlouisfed.org), and the U.S. Census Bureau (www.census.gov/data.html). These websites provide data on inflation, unemployment, GDP, and much more, including useful international data. The National Climatic Data Center (www.ncdc.noaa.gov/data-access) provides a large collection of environmental, meteorological, and climate data. Similarly, transportation data can be found at www.its-rde.net. The University of Michigan has compiled sentiment data found at www.sca.isr.umich.edu. Several cities in the United States have publicly available data in categories such as finance, community and economic development, education, and crime. For example, the Chicago data portal data.cityofchicago.org provides a

large volume of city-specific data. Excellent world development indicator data are available at data.worldbank.org. The happiness index data for most countries are available at www.happyplanetindex.org/data.

Private corporations also make data available on their websites. For example, Yahoo Finance (www.finance.yahoo.com) and Google Finance (www.google.com/finance) list data such as stock prices, mutual fund performance, and international market data. Zillow (www.zillow.com/) supplies data for recent home sales, monthly rent, mortgage rates, and so forth. Similarly, www.espn.go.com offers comprehensive sports data on both professional and college teams. Finally, *The Wall Street Journal, The New York Times, USA Today, The Economist, Business Week, Forbes,* and *Fortune* are all reputable publications that provide all sorts of data. We would like to point out that all of the above data sources represent only a fraction of publicly available data.

EXERCISES 1.2

1. It came as a big surprise when Apple's touch screen iPhone 4, considered by many to be the best smartphone ever, was found to have a problem (*The New York Times,* June 24, 2010). Users complained of weak reception, and sometimes even dropped calls, when they cradled the phone in their hands in a particular way. A quick survey at a local store found that 2% of iPhone 4 users experienced this reception problem.
   a. Describe the relevant population.
   b. Does 2% denote the population parameter or the sample statistic?
2. Many people regard video games as an obsession for youngsters, but, in fact, the average age of a video game player is 35 years (Telegraph.co.uk, July 4, 2013). Is the value 35 likely the actual or the estimated average age of the population? Explain.
3. An accounting professor wants to know the average GPA of the students enrolled in her class. She looks up information on Blackboard about the students enrolled in her class and computes the average GPA as 3.29.
   a. Describe the relevant population.
   b. Does the value 3.29 represent the population parameter or the sample statistic?
4. Business graduates in the United States with a marketing concentration earn high salaries. According to the Bureau of Labor Statistics, the average annual salary for marketing managers was $140,660 in 2015.
   a. What is the relevant population?
   b. Do you think the average salary of $140,660 was computed from the population? Explain.
5. Research suggests that depression significantly increases the risk of developing dementia later in life (*BBC News,* July 6, 2010). In a study involving 949 elderly persons, it was reported that 22% of those who had depression went on to develop dementia, compared to only 17% of those who did not have depression.
   a. Describe the relevant population and the sample.
   b. Do the numbers 22% and 17% represent population parameters or sample statistics?
6. Go to www.finance.yahoo.com/ to get a current stock quote for General Electric, Co. (ticker symbol = GE). Then, click on historical prices to record the monthly adjusted close price of General Electric stock in 2016. Create a table that uses this information. What type of data do these numbers represent? Comment on the data.
7. Ask 20 of your friends whether they live in a dormitory, a rental unit, or other form of accommodation. Also find out their approximate monthly lodging expenses. Create a table that uses this information. What type of data do these numbers represent? Comment on the data.
8. Go to www.zillow.com/ and find the sale price data of 20 single-family homes sold in Las Vegas, Nevada, in the last 30 days. In the data set, include the sale price, the number of bedrooms, the square footage, and the age of the house. What type of data do these numbers represent? Comment on the data.
9. The Federal Reserve Bank of St. Louis is a good source for downloading economic data. Go to research.stlouisfed.org/fred2/ to extract quarterly data on gross private saving (GPSAVE) from 2012 to 2015 (16 observations). Create a table that uses this information. Plot the data over time and comment on the savings trend in the United States.
10. Go to the U.S. Census Bureau website at www.census.gov/ and extract the most recent median household income for Alabama, Arizona, California, Florida, Georgia, Indiana, Iowa, Maine, Massachusetts, Minnesota, Mississippi, New Mexico, North Dakota, and Washington. What type of data do these numbers represent? Comment on the regional differences in income.
11. Go to *The New York Times* website at www.nytimes.com/ and review the front page. Would you consider the data on the page to be structured or unstructured? Explain.

12. Conduct an online search to compare price and fuel economy of small hybrid vehicles such as Toyota Prius, Ford Fusion, and Chevrolet Volt. Would the resulting data be structured or unstructured? Explain.
13. Ask your peers about their online social media usage. In particular collect information on (a) whether they use Facebook, Instagram, and Snapchat, (b) how often they use each social media service, and (c) their overall satisfaction with each of these services using a 5-point numerical scale where 1 represents totally unsatisfied and 5 represents totally satisfied. Are the resulting data structured or unstructured? Are the data cross-sectional or time series?
14. Conduct an online search for a weekly car rental in Seattle, Washington, and Portland, Oregon, for different car types and rental car companies for the year 2017. Are the data structured or unstructured? Are the data cross-sectional or time series?

LO **1.4**

Describe variables and types of measurement scales.

1.3 VARIABLES AND SCALES OF MEASUREMENT

When we conduct a statistical investigation, we invariably focus on people, objects, or events with particular characteristics. When a characteristic of interest differs in kind or degree among various observations, then the characteristic can be termed a **variable**. We further categorize a variable as either qualitative or quantitative. For a **qualitative variable**, we use labels or names to identify the distinguishing characteristic of each observation. For instance, the 2010 Census asked each respondent to indicate gender on the form. Each respondent chose either male or female. Gender is a qualitative variable. Other examples of qualitative variables include race, profession, type of business, the manufacturer of a car, and so on.

A variable that assumes meaningful numerical values is called a **quantitative variable**. Quantitative variables, in turn, are either discrete or continuous. A **discrete variable** assumes a countable number of values. Consider the number of children in a family or the number of points scored in a basketball game. We may observe values such as 3 children in a family or 90 points being scored in a basketball game, but we will not observe 1.3 children or 92.5 scored points. The values that a discrete variable assumes need not be whole numbers. For example, the price of a stock for a particular firm is a discrete variable. The stock price may take on a value of \$20.37 or \$20.38, but it cannot take on a value between these two points. Finally, a discrete variable may assume an infinite number of values, but these values are countable; that is, they can be presented as a sequence x_1, x_2, x_3, and so on. The number of cars that cross the Golden Gate Bridge on a Saturday is a discrete variable. Theoretically, this variable assumes the values 0, 1, 2, . . .

A **continuous variable** is characterized by uncountable values within an interval. Weight, height, time, and investment return are all examples of continuous variables. For example, an unlimited number of values occur between the weights of 100 and 101 pounds, such as 100.3, 100.625, 100.8342, and so on. In practice, however, continuous variables may be measured in discrete values. We may report a newborn's weight (a continuous variable) in discrete terms as 6 pounds 10 ounces and another newborn's weight in similar discrete terms as 6 pounds 11 ounces.

QUALITATIVE VARIABLES VERSUS QUANTITATIVE VARIABLES

A variable is a general characteristic being observed on a set of people, objects, or events, where each observation varies in kind or degree. Labels or names are used to categorize the distinguishing characteristics of a qualitative variable; eventually, these attributes may be coded into numbers for purposes of data processing. A quantitative variable assumes meaningful numerical values, and can be further categorized as either discrete or continuous. A discrete variable assumes a countable number of values, whereas a continuous variable is characterized by uncountable values within an interval.

In order to choose the appropriate statistical methods for summarizing and analyzing data, we need to distinguish between different measurement scales. All data measurements can be classified into one of four major categories: nominal, ordinal, interval, and ratio. Nominal and ordinal scales are used for qualitative variables, whereas interval and ratio scales are used for quantitative variables. We discuss these scales in ascending order of sophistication.

The Nominal Scale

The **nominal scale** represents the least sophisticated level of measurement. If we are presented with nominal data, all we can do is categorize or group the data. The values in the data set differ merely by name or label. Consider the following example.



©ymgerman/Shutterstock

Each company listed in Table 1.2 is a member of the Dow Jones Industrial Average (DJIA). The DJIA is a stock market index that shows how 30 large, publicly owned companies based in the United States have traded during a standard trading session in the stock market. Table 1.2 also shows where stocks of these companies are traded: on either the National Association of Securities Dealers Automated Quotations (Nasdaq) or the New York Stock Exchange (NYSE). These data are classified as nominal scale since we are simply able to group or categorize them. Specifically, only four stocks are traded on Nasdaq, whereas the remaining 26 are traded on the NYSE.

TABLE 1.2 Companies of the DJIA and Exchange Where Stock Is Traded

Company	Exchange	Company	Exchange
3M (MMM)	NYSE	Intel (INTC)	Nasdaq
American Express (AXP)	NYSE	Johnson & Johnson (JNJ)	NYSE
Apple (AAPL)	Nasdaq	JPMorgan Chase (JPM)	NYSE
Boeing (BA)	NYSE	McDonald's (MCD)	NYSE
Caterpillar (CAT)	NYSE	Merck (MRK)	NYSE
Chevron (CVX)	NYSE	Microsoft (MFST)	Nasdaq
Cisco (CSCO)	Nasdaq	Nike (NKE)	NYSE
Coca-Cola (KO)	NYSE	Pfizer (PFE)	NYSE
Disney (DIS)	NYSE	Procter & Gamble (PG)	NYSE
Dupont (DD)	NYSE	Travelers (TRV)	NYSE
ExxonMobil (XOM)	NYSE	United Health (UNH)	NYSE
General electric (GE)	NYSE	United Tech. Corp. (UTX)	NYSE
Goldman Sachs (GS)	NYSE	Verizon (VZ)	NYSE
Home Depot (HD)	NYSE	Visa (V)	NYSE
IBM (IBM)	NYSE	Walmart (WMT)	NYSE

Source: Money.CNN.com information retrieved March 21, 2015.

Often we substitute *numbers* for the particular qualitative characteristic or trait that we are grouping. One reason why we do this is for ease of exposition; always referring to the National Association of Securities Dealers Automated Quotations, or even Nasdaq, becomes awkward and unwieldy. In addition, as we will see later in the text, statistical analysis is greatly facilitated by using numbers instead of names. For example, we might use the number 0 to show that a company's stock is traded on Nasdaq and the number 1 to show that a company's stock is traded on the NYSE, or in tabular form:

Exchange	Number of Companies Trading on Exchange
0	4
1	26

The Ordinal Scale

Compared to the nominal scale, the **ordinal scale** reflects a stronger level of measurement. With ordinal data we are able to both *categorize* and *rank* the data with respect to some characteristic or trait. The weakness with ordinal data is that we cannot interpret the difference between the ranked values because the actual numbers used are arbitrary. For example, suppose you are asked to classify the service at a particular hotel as excellent, good, fair, or poor. A standard way to record the ratings is

Category	Rating
Excellent	4
Good	3
Fair	2
Poor	1

Here the value attached to excellent (4) is higher than the value attached to good (3), indicating that the response of excellent is preferred to good. However, another representation of the ratings might be

Category	Rating
Excellent	100
Good	80
Fair	70
Poor	40

Excellent still receives a higher value than good, but now the difference between the two categories is 20 (100 – 80), as compared to a difference of 1 (4 – 3) when we use the first classification. In other words, *differences between categories are meaningless with ordinal data.* (We also should note that we could reverse the ordering so that, for instance, excellent equals 40 and poor equals 100; this renumbering would not change the nature of the data.)

EXAMPLE 1.1

FILE
Tween_Survey

In the introductory case, four questions were posed to tweens. The first question (Q1) asked tweens to name the radio station that was playing on the ride to the resort, and the second question (Q2) asked tweens to rate the food quality at the resort on a scale of 1 to 4. The tweens' responses to these questions are shown in Table 1.1 in the introductory case.

a. What is the scale of measurement of the radio station data?

b. How are the data based on the ratings of the food quality similar to the radio station data? How are the data different?

c. Summarize the tweens' responses to Q1 and Q2 in tabular form. How can the resort use the information from these responses?

SOLUTION:

a. When asked which radio station played on the car ride to the resort, tweens responded with one of the following answers: JAMN94.5, MIX104.1, or KISS108. These are nominal data—the values in the data differ merely in name or label.

b. Since we can both categorize and rank the food quality data, we classify these responses as ordinal data. Ordinal data are similar to nominal data in the sense that we can categorize the data. The main difference between ordinal

and nominal data is that the categories of ordinal data are ranked. A rating of 4 is better than a rating of 3. With the radio station data, we cannot say that KISS108 is ranked higher than MIX104.1; some tweens may argue otherwise, but we simply categorize nominal data without ranking.

c. With respect to the radio station data (Q1), we can assign 1 to JAMN94.5, 2 to MIX104.1, and 3 to KISS108. Counting the responses that fall into each category, we find that six tweens listened to 1, two listened to 2, and 12 listened to 3, or in tabular form:

Radio Station	Number of Tweens
1	6
2	2
3	12

Twelve of the 20 tweens, or 60%, listened to KISS108. This information could prove useful to the management of the resort as they make decisions as to where to allocate their advertising dollars. If the resort could only choose to advertise at one radio station, it would appear that KISS108 would be the wise choice.

Given the food quality responses (Q2), we find that three of the tweens rated food quality with a 4, six tweens rated food quality with a 3, eight tweens rated food quality with a 2, and three tweens rated food quality with a 1, or in tabular form:

Rating	Number of Tweens
4	3
3	6
2	8
1	3

The food quality results may be of concern to management. Just as many tweens rated the food quality as excellent as compared to poor. Moreover, the majority [(8 + 3)/20 = 55%] felt that the food was, at best, fair. Perhaps a more extensive survey that focuses solely on food quality would reveal the reason for their apparent dissatisfaction.

As mentioned earlier, nominal and ordinal scales are used for *qualitative variables.* Values corresponding to a qualitative variable are typically expressed in words but are coded into numbers for purposes of data processing. When summarizing the results of a qualitative variable, we typically count the number or calculate the percentage of persons or objects that fall into each possible category. With a qualitative variable, we are unable to perform meaningful arithmetic operations such as adding and subtracting.

The Interval Scale

With data that are measured on an **interval scale**, not only can we categorize and rank the data, we are also assured that the differences between scale values are meaningful. Thus, the arithmetic operations of addition and subtraction are meaningful. The Fahrenheit scale for temperatures is an example of an interval scale. Not only is 60 degrees Fahrenheit hotter than 50 degrees Fahrenheit, the same difference of 10 degrees also exists between 90 and 80 degrees Fahrenheit.

The main drawback of data on an interval scale is that the value of zero is arbitrarily chosen; the zero point of an interval scale does not reflect a complete absence of what is being measured. No specific meaning is attached to zero degrees Fahrenheit other than to say it is 10 degrees colder than 10 degrees Fahrenheit. With an arbitrary zero point, meaningful ratios cannot be constructed. For instance, it is senseless to say that 80 degrees is twice as hot as 40 degrees; in other words, the ratio 80/40 has no meaning.

The Ratio Scale

The **ratio scale** represents the strongest level of measurement. Ratio data have all the characteristics of interval data as well as a *true zero* point, which allows us to interpret the ratios of values. A ratio scale is used to measure many types of data in business analysis. Variables such as sales, profits, and inventory levels are expressed as ratio data. A meaningful zero allows us to state, for example, that profits for firm A are double those of firm B. Measurements such as weight, time, and distance are also measured on a ratio scale since zero is meaningful.

Unlike qualitative data, arithmetic operations are valid on interval- and ratio-scaled values. In later chapters, we will calculate summary measures for the typical value and the variability of quantitative variables; we cannot calculate these measures if the variable is qualitative in nature.

FILE
Tween_Survey

EXAMPLE 1.2

In the last two questions from the introductory case's survey (Q3 and Q4), the 20 tweens were asked: "What time should the main dining area close?" and "How much of your *own* money did you spend at the lodge today?" Their responses appear in Table 1.1 in the introductory case.

a. How are the time data classified? In what ways do the time data differ from ordinal data? What is a potential weakness of this measurement scale?

b. What is the measurement scale of the money data? Why is it considered the strongest form of data?

c. In what ways is the information from Q3 and Q4 useful for the resort?

SOLUTION:

a. Clock time responses, such as 3:00 pm and 3:30 pm, or 5:30 pm and 6:00 pm, are on an interval scale. Interval data are a stronger measurement scale than ordinal data because differences between interval-scaled values are meaningful. In this particular example, we can say that 3:30 pm is 30 minutes later than 3:00 pm and 6:00 pm is 30 minutes later than 5:30 pm. The weakness with interval data is that the value of zero is arbitrary. Here, with the clock time responses, we have no apparent zero point; however, we could always arbitrarily define a zero point, say, at 12:00 am. Thus, although differences are comparable with interval data, ratios are meaningless due to the arbitrariness of the zero point. In other words, it is senseless to form the ratio 6:00 pm/3:00 pm and conclude that 6:00 pm is twice as long a time period as 3:00 pm.

b. Since the tweens' responses are in dollar amounts, this is ratio data. The ratio scale is the strongest form of data because we can categorize and rank values as well as calculate meaningful differences. Moreover, since there is a natural zero point, valid ratios can also be calculated. For example, the data show that three tweens spent \$20. These tweens spent four times as much as the three tweens that spent \$5 (\$20/\$5 = 4).

c. A review of the clock time responses (Q3) in Table 1.1 shows that the vast majority of the tweens would like the dining area to remain open later. In fact, only one tween feels that the dining area should close at 3:00 pm. An inspection of the money responses (Q4) in Table 1.1 indicates that only three of the 20 tweens did not spend any of his/her own money. This is very important information. It does appear that the discretionary spending of this age group is significant. The resort would be wise to cater to some of their preferences.

SYNOPSIS OF INTRODUCTORY CASE

A preliminary survey of tween preferences conducted by the management of a ski resort two hours outside Boston, Massachusetts, revealed some interesting information.

- Tweens were first asked to name the radio station that they listened to on the way to the resort. The responses show that 60% of the tweens listened to KISS108. If the resort wishes to contact tweens using this medium, it may want to direct its advertising dollars to this station.
- Next, the tweens were asked to rate the food quality at the resort on a scale of 1 to 4 (where 1 is poor, 2 is fair, 3 is good, and 4 is excellent). The survey results with respect to food quality are disturbing. The majority of the tweens, 55% (11/20), felt that the food was, at best, fair. A more extensive study focusing on food quality appears necessary.
- Tweens were then asked what time the main dining area should close, given that it presently closes at 3:00 pm. The data suggest that the vast majority of the tweens (19 out of 20) would like the dining area to remain open later.



©Dennis Welsh/UpperCut Images/Getty Images

- Finally, the tweens were asked to report the amount of their *own* money that they spent at the lodge. The resort is likely pleased with the responses to this question because 17 of the 20 tweens spent their own money at the lodge. This finding appears consistent with the belief that tween spending is growing.

EXERCISES 1.3

15. Which of the following variables are qualitative and which are quantitative? If the variable is quantitative, then specify whether the variable is discrete or continuous.
    a. Points scored in a football game.
    b. Racial composition of a high school classroom.
    c. Heights of 15-year-olds.
16. Which of the following variables are qualitative and which are quantitative? If the variable is quantitative, then specify whether the variable is discrete or continuous.
    a. Colors of cars in a mall parking lot.
    b. Time it takes each student to complete a final exam.
    c. The number of patrons who frequent a restaurant.
17. In each of the following scenarios, define the type of measurement scale.
    a. A kindergarten teacher marks whether each student is a boy or a girl.
    b. A ski resort records the daily temperature during the month of January.
    c. A restaurant surveys its customers about the quality of its waiting staff on a scale of 1 to 4, where 1 is poor and 4 is excellent.
18. In each of the following scenarios, define the type of measurement scale.
    a. An investor collects data on the weekly closing price of gold throughout a year.

b. An analyst assigns a sample of bond issues to one of the following credit ratings, given in descending order of credit quality (increasing probability of default): AAA, AA, BBB, BB, CC, D.

c. The dean of the business school at a local university categorizes students by major (i.e., accounting, finance, marketing, etc.) to help in determining class offerings in the future.

19. In each of the following scenarios, define the type of measurement scale.
    a. A meteorologist records the amount of monthly rainfall over the past year.
    b. A sociologist notes the birth year of 50 individuals.
    c. An investor monitors the daily stock price of BP following the 2010 oil disaster in the Gulf of Mexico.

20. A professor records the majors of her 30 students as follows:

Accounting	Economics	Undecided	Finance	Management
Management	Finance	Marketing	Economics	Management
Marketing	Finance	Marketing	Accounting	Finance
Finance	Undecided	Management	Undecided	Economics
Economics	Accounting	Management	Undecided	Economics
Accounting	Economics	Management	Accounting	Economics

    a. What is the measurement scale of these data?
    b. Summarize the results in tabular form.
    c. What information can be extracted from the data?

21. FILE ***DOW_Characteristics.*** The accompanying table shows a portion of the 30 companies that comprise the Dow Jones Industrial Average (DJIA). The second column shows the year that the company joined the DJIA (Year). The third column shows each company's Morningstar rating (Rating). (Five stars is the best rating that a company can receive, indicating that the company's stock price is undervalued and thus a very good buy. One star is the worst rating a company can be given, implying that the stock price is overvalued and a bad buy.) Finally, the fourth column shows each company's stock price as of March 17, 2017 (Price in $).

Company	Year	Rating	Price
3M (MMM)	1976	**	192.36
American Express (AMX)	1982	***	79.25
⋮	⋮	⋮	⋮
Wal-Mart (WMT)	1991	****	69.89

Source: Morningstar ratings retrieved from www.morningstar.com on March 17, 2017; stock prices retrieved from www.finance.yahoo.com.

    a. What is the measurement scale of the Year data? What are the strengths of this type of data? What are the weaknesses?
    b. What is the measurement scale of Morningstar's star-based rating system? Summarize Morningstar's star-based rating system for the companies in tabular form. Let 5 denote *****, 4 denote ****, and so on. What information can be extracted from these data?
    c. What is the measurement scale of the Stock Price data? What are its strengths?

CONCEPTUAL REVIEW

LO 1.1 Describe the importance of statistics.

A proper understanding of statistical ideas and concepts helps us understand more of the real world around us, including issues in business, sports, politics, health, and social interactions. We must understand statistics or risk making bad decisions and costly mistakes. A knowledge of statistics also provides the necessary tools to differentiate between sound statistical conclusions and questionable conclusions drawn from an insufficient number of data points, "bad" data points, incomplete data points, or just misinformation.

LO 1.2 Differentiate between descriptive statistics and inferential statistics.

The study of statistics is generally divided into two branches: descriptive statistics and inferential statistics. **Descriptive statistics** refers to the summary of a data set in the form of tables, graphs, and/or the calculation of numerical measures. **Inferential statistics** refers to extracting useful information from a sample to draw conclusions about a population. A **population** consists of all items of interest in a statistical problem; a **sample** is a subset of that population.

In general, we use sample data rather than population data for two main reasons: (1) obtaining information on the entire population is expensive and/or (2) it is impossible to examine every item of the population.

LO 1.3 Explain the various data types.

Cross-sectional data contain values of a characteristic of many subjects at the same point in time or without regard to differences in time. **Time series data** contain values of a characteristic of a subject over time.

Structured data conform but **unstructured data** do not conform to a predefined row-column format.

Big data are a massive volume of both structured and unstructured data that are extremely difficult to manage, process, and analyze using traditional data processing tools. Big data, however, do not necessarily imply complete (population) data.

LO 1.4 Describe variables and types of measurement scales.

A variable is categorized as either qualitative or quantitative. For a **qualitative variable**, we use labels or names to identify the distinguishing characteristic of each observation. A **quantitative variable** assumes meaningful numerical values and can be further categorized as either **discrete** or **continuous**. A discrete variable assumes a countable number of values, whereas a continuous variable is characterized by uncountable values within an interval.

All data measurements can be classified into one of four major categories.

- The values on a **nominal scale** differ merely by name or label. These values are then simply categorized or grouped by name.
- The values on an **ordinal scale** can be categorized *and* ranked; however, differences between the ranked values are meaningless.
- Values on the **interval scale** can be categorized and ranked, and differences between values are meaningful. The main drawback of the interval scale is that the value of zero is arbitrarily chosen; this implies that ratios constructed from interval-scaled values bear no significance.
- Ratio data have all the characteristics of interval data as well as a true zero point; thus, as its name implies, meaningful ratios can be calculated with values on the ratio scale.

Nominal and ordinal scales are used for qualitative variables. When summarizing the results of qualitative data, we typically count the number or calculate the percentage of persons or objects that fall into each possible category. Interval and ratio scales are used for quantitative variables. Unlike qualitative variables, arithmetic operations are valid on quantitative variables.

2 Tabular and Graphical Methods

Learning Objectives

After reading this chapter you should be able to:

LO **2.1** **Summarize qualitative data by constructing a frequency distribution.**

LO **2.2** **Construct and interpret a pie chart and a bar chart.**

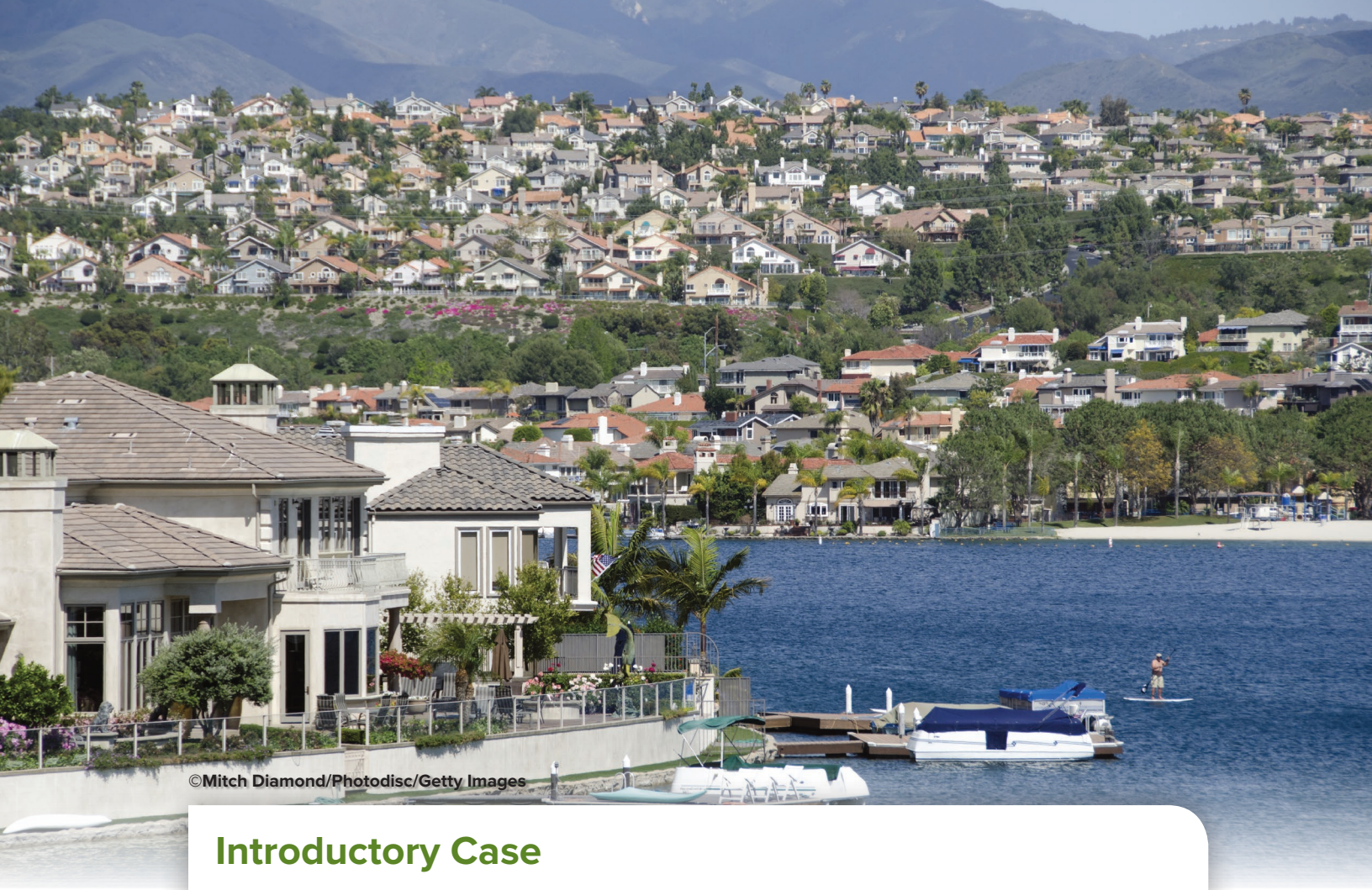
LO **2.3** **Summarize quantitative data by constructing a frequency distribution.**

LO **2.4** **Construct and interpret a histogram, a polygon, and an ogive.**

LO **2.5** **Construct and interpret a stem-and-leaf diagram.**

LO **2.6** **Construct and interpret a scatterplot.**

People often have difficulty processing information provided by data in its raw form. A useful way of interpreting data effectively is through data visualization. In this chapter, we present several tabular and graphical tools that can help us organize and present data. We first make a table referred to as a frequency distribution using qualitative data. For visual representations of qualitative data, we construct a pie chart or a bar chart. For quantitative data, we again make a frequency distribution. In addition to giving us an overall picture of where the data tend to cluster, a frequency distribution using quantitative data also shows us how the data are spread out from the lowest value to the highest value. For visual representations of quantitative data, we construct a histogram, a polygon, an ogive, and a stem-and-leaf diagram. Finally, we show how to construct a scatterplot, which graphically depicts the relationship between two quantitative variables. We will find that a scatterplot is a very useful tool when conducting correlation and regression analysis, topics discussed in depth later in the text.



©Mitch Diamond/Photodisc/Getty Images

Introductory Case

House Prices in Southern California

Mission Viejo, a city located in Southern California, was named the safest city in California and the third-safest city in the nation (CQPress.com, November 23, 2009). Matthew Edwards, a relocation specialist for a real estate firm in Mission Viejo, often relays this piece of information to clients unfamiliar with the many benefits that the city offers. Recently, a client from Seattle, Washington, asked Matthew for a summary of recent sales. The client is particularly interested in the availability of houses in the $500,000 range. Table 2.1 shows the sale price for 36 single-family houses in Mission Viejo during June 2010.

TABLE 2.1 Recent Sale Price of Houses in Mission Viejo, CA, for June 2010 (data in $1,000s)

430	670	530	521	669	445
520	417	525	350	660	412
460	533	430	399	702	735
475	525	330	560	540	537
670	538	575	440	460	630
521	370	555	425	588	430

Source: www.zillow.com.

FILE
MV_Houses

Matthew wants to use the sample information to

1. Make summary statements concerning the range of house prices.
2. Comment on where house prices tend to cluster.
3. Calculate appropriate percentages in order to compare house prices in Mission Viejo, California, to those in Seattle, Washington.

A synopsis of this case is provided in Section 2.2.

LO **2.1**

Summarize qualitative data by constructing a frequency distribution.

2.1 SUMMARIZING QUALITATIVE DATA

As we discussed in Chapter 1, nominal and ordinal data are types of qualitative data. Nominal data typically consist of observations that represent labels or names; information related to gender or race are examples. Nominal data are considered the least sophisticated form of data since all we can do with the data is categorize it. Ordinal data are stronger in the sense that we can categorize and order the data. Examples of ordinal data include the ratings of a product or a professor, where 1 represents the worst and 4 represents the best. In order to organize qualitative data, it is often useful to construct a **frequency distribution**.

FREQUENCY DISTRIBUTION FOR QUALITATIVE DATA

A frequency distribution for qualitative data groups data into categories and records the number of observations that fall into each category.

To illustrate the construction of a frequency distribution with nominal data, Table 2.2 shows the weather for the month of February (2010) in Seattle, Washington.

TABLE 2.2 Seattle Weather, February 2010

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
	1 Rainy	2 Rainy	3 Rainy	4 Rainy	5 Rainy	6 Rainy
7 Rainy	8 Rainy	9 Cloudy	10 Rainy	11 Rainy	12 Rainy	13 Rainy
14 Rainy	15 Rainy	16 Rainy	17 Sunny	18 Sunny	19 Sunny	20 Sunny
21 Sunny	22 Sunny	23 Rainy	24 Rainy	25 Rainy	26 Rainy	27 Rainy
28 Sunny						

Source: www.wunderground.com.



©sbk_20d pictures/Moment/Getty Images

We first note that the weather in Seattle is categorized as cloudy, rainy, or sunny. The first column in Table 2.3 lists these categories. Initially, we use a "tally" column to record the number of days that fall into each category. Since the first eight days of February were rainy days, we place the first eight tally marks in the rainy category; the ninth day of February was cloudy, so we place one tally mark in the cloudy category, and so on. Finally, we convert each category's total tally count into its respective numerical value in the frequency column. Since only one tally mark appears in the cloudy category, we record the value 1 as its frequency. Note that if we sum the frequency column, we obtain the sample size. A frequency distribution in its final form does not include the tally column.

TABLE 2.3 Frequency Distribution for Seattle Weather, February 2010

Weather	Tally	Frequency
Cloudy	|	1
Rainy	𝍸 𝍸 𝍸 𝍸	20
Sunny	𝍸 ||	7
		Total = 28 days

From the frequency distribution, we can now readily observe that the most common type of day in February was rainy since this type of day occurs with the highest frequency. In many applications, we want to compare data sets that differ in size. For

example, we might want to compare the weather in February to the weather in March. However, February has 28 days (except during a leap year) and March has 31 days. In this instance, we would convert the frequency distribution to a **relative frequency distribution**. We calculate each category's relative frequency by dividing the respective category's frequency by the total number of observations. The sum of the relative frequencies should equal one, or a value very close to one due to rounding.

In Table 2.4, we convert the frequency distribution from Table 2.3 into a relative frequency distribution. Similarly, we obtain the relative frequency distribution for the month of March; the raw data for March are not shown. March had 3 cloudy days, 18 rainy days, and 10 sunny days. Each of these frequencies was then divided by 31, the number of days in the month of March.

TABLE 2.4 Relative Frequency Distribution for Seattle Weather

Weather	February 2010: Relative Frequency	March 2010: Relative Frequency
Cloudy	1/28 = 0.036	3/31 = 0.097
Rainy	20/28 = 0.714	18/31 = 0.581
Sunny	7/28 = 0.250	10/31 = 0.323
	Total = 1	Total = 1 (subject to rounding)

Source: www.wunderground.com.

We can easily convert relative frequencies into percentages by multiplying by 100. For instance, the percent of cloudy days in February and March equals 3.6% and 9.7%, respectively. From the relative frequency distribution, we can now conclude that the weather in Seattle in both February and March was predominantly rainy. However, the weather in March was a bit nicer in that approximately 32% of the days were sunny, as opposed to only 25% of the days in February.

CALCULATING RELATIVE AND PERCENT FREQUENCIES

The relative frequency for each category of a qualitative variable equals the proportion (fraction) of observations in each category. A category's relative frequency is calculated by dividing its frequency by the total number of observations. The sum of the relative frequencies should equal one.

The percent frequency for each category of a qualitative variable equals the percent (%) of observations in each category; it equals the relative frequency of the category multiplied by 100.

Pie Charts and Bar Charts

LO **2.2**

Construct and interpret a pie chart and a bar chart.

We can visualize the information found in frequency distributions by constructing various graphs. Graphical representations often portray the data more dramatically, as well as simplify interpretation. A **pie chart** and a **bar chart** are two widely used graphical representations for qualitative data.

GRAPHICAL DISPLAY OF QUALITATIVE DATA: A PIE CHART

A pie chart is a segmented circle whose segments portray the relative frequencies of the categories of some qualitative variable.

A pie chart is best explained by using an example. Consider Example 2.1.

EXAMPLE 2.1

Is America having a "marriage crisis?" The answer depends on whom you ask, but nearly every study focuses on the women's liberation movement of the late 1960s and 1970s. As more and more women earned college degrees, they entered the workforce and delayed motherhood. Marriage became less necessary for their economic survival. No matter what the reason for the decline in marriage, here are some facts. In 1960, 71% of all adults in the United States were married. Today, barely half of all adults are married, just 52%. Table 2.5 shows the proportions of all adults who were married, widowed, divorced, or single in 1960 compared to those same proportions in 2010. Construct pie charts to graphically depict marital status in the United States in these two time periods.

FILE
Marital_Status

TABLE 2.5 Marital Status, 1960 versus 2010

Marital Status	1960	2010
Married	0.71	0.52
Single	0.15	0.28
Divorced	0.05	0.14
Widowed	0.09	0.06

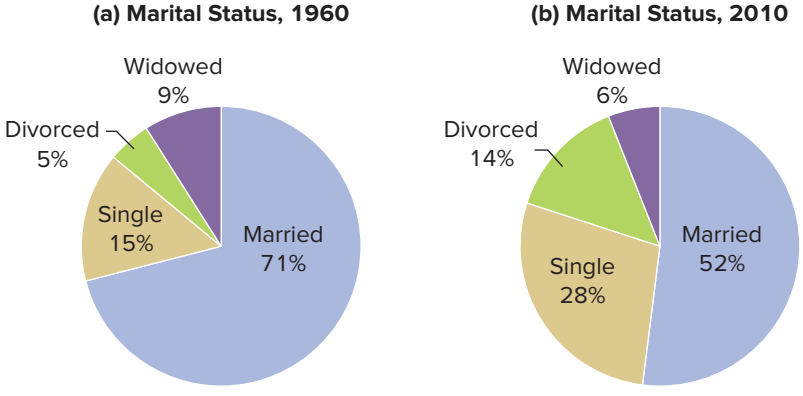
Note: Proportions for each year rounded so that they summed to one.
Source: Pew Research Center analysis of Decennial Census (1960–2000) and American Community Survey data (2008, 2010).

SOLUTION: In order to construct a pie chart, we first draw a circle. We then cut the circle into slices, or sectors such that each sector is proportional to the size of the category we wish to display. For instance, Table 2.5 shows that married adults accounted for 0.71 of all adults in 1960. Since a circle contains 360 degrees, the portion of the circle representing married adults encompasses $0.71 \times 360 = 255.6$ degrees. Similar calculations for the other three categories yield:

Single: $0.15 \times 360 = 54$ degrees
Divorced: $0.05 \times 360 = 18$ degrees
Widowed: $0.09 \times 360 = 32.4$ degrees

The same methodology can be used to construct a pie chart for marital status in 2010. Figure 2.1 shows the resulting pie charts.

FIGURE 2.1 Pie charts for marital status



Another way to graphically depict qualitative data is to construct a **bar chart**.

GRAPHICAL DISPLAY OF QUALITATIVE DATA: A BAR CHART

A bar chart depicts the frequency or the relative frequency for each category of the qualitative variable as a series of horizontal or vertical bars, the lengths of which are proportional to the values that are to be depicted.

We first discuss a vertical bar chart, sometimes referred to as a column chart. Here, we place each category on the horizontal axis and mark the vertical axis with an appropriate range of values for either frequency or relative frequency. The height of each bar is equal to the frequency or the relative frequency of the corresponding category. Typically, we leave space between categories to improve clarity.

Figure 2.2 shows a relative frequency bar chart for the marital status example. It is particularly useful because we can group marital status by year, emphasizing the decline in the proportion of U.S. adults who are married and the rise in the proportion of U.S. adults who are single.

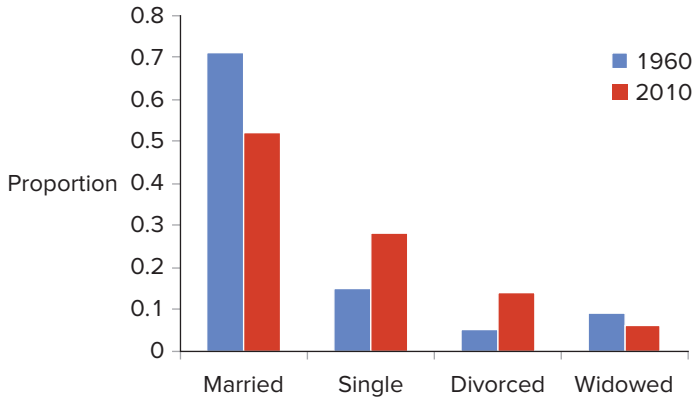


FIGURE 2.2
Marital status of U.S. adults, 1960 versus 2010

For a horizontal bar chart, we simply place each category on the vertical axis and mark the horizontal axis with an appropriate range of values for either frequency or relative frequency. For example, a recent poll asked more than 1,000 Americans: "When traveling in a non-English-speaking country, which word or phrase is most important to know in that country's language?" (Source: *Vanity Fair,* January 2, 2012). Figure 2.3 shows the results of the poll. The phrase "Thank you" earned the largest percentage of votes (38%). Fortunately, only 1% of Americans believed that the phrase "Where is McDonald's?" was of vital importance. The proportions in Figure 2.3 do not sum to one because we exclude responses with uncommon words or phrases.

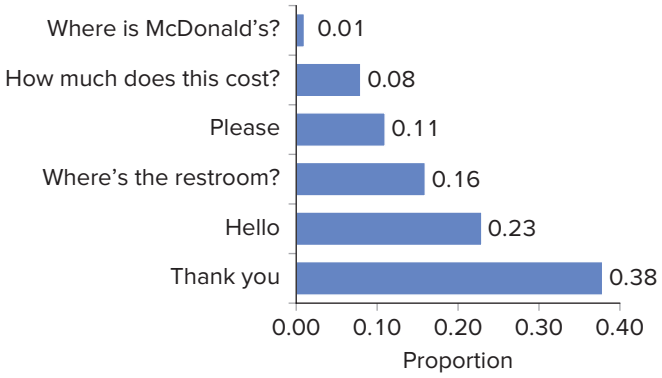


FIGURE 2.3
Results to question: "When traveling in a non-English-speaking country, which word or phrase is most important to know in that country's language?"

Cautionary Comments When Constructing or Interpreting Charts or Graphs

As with many of the statistical methods that we examine throughout this text, the possibility exists for unintentional, as well as purposeful, distortions of graphical information. As a careful researcher, you should follow these basic guidelines:

- The simplest graph should be used for a given set of data. Strive for clarity and avoid unnecessary adornments.
- Axes should be clearly marked with the numbers of their respective scales; each axis should be labeled.
- When creating a bar chart, each bar should be of the same width. Differing bar widths create distortions. The same principle holds in the next section when we discuss histograms.
- The vertical axis should not be given a very high value as an upper limit. In these instances, the data may appear compressed so that an increase (or decrease) of the data is not as apparent as it perhaps should be. For example, Figure 2.4(a) plots the daily price for a barrel of crude oil for the first quarter of 2011. Due to Middle East unrest, the price of crude oil rose from a low of \$83.13 per barrel to a high of \$106.19 per barrel, or approximately 28% $\left(= \frac{106.19 - 83.13}{83.13}\right)$. However, since Figure 2.4(a) uses a high value as an upper limit on the vertical axis (\$325), the rise in price appears dampened.

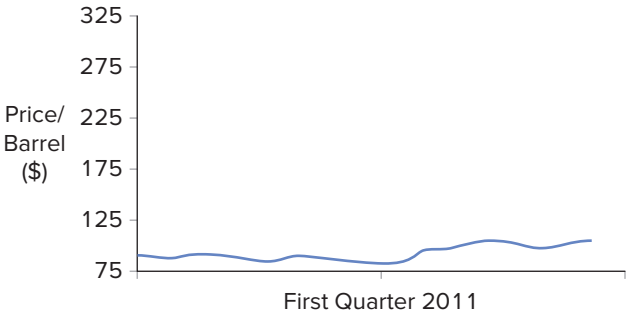
FILE *Crude_Oil*

- The vertical axis should not be stretched so that an increase (or decrease) of the data appears more pronounced than warranted. For example, Figure 2.4(b) charts the daily closing stock price for Johnson & Johnson (JNJ) for the week of April 4, 2011. It is true that the stock price declined over the week from a high of \$60.15 to a low of \$59.46; this amounts to a \$0.69 decrease or an approximate 1% decline. However, since the vertical axis is stretched, the drop in stock price appears more dramatic.

FILE *JNJ*

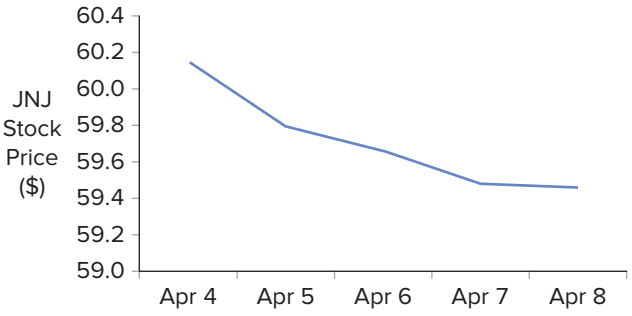
FIGURE 2.4 Misleading scales on vertical axes

(a) Vertical axis with high upper limit



Source: U.S. Energy Information Administration.

(b) Stretched vertical axis



Source: www.finance.yahoo.com.

Using Excel to Construct a Pie Chart and a Bar Chart

Excel offers various options for displaying a pie chart. To replicate the pie chart in Figure 2.1(a), follow these steps:

A Pie Chart

FILE *Marital_Status*

A. FILE Open ***Marital_Status*** (Table 2.5).

B. First, select the category names and respective relative frequencies from the year 1960. Leave out the heading (top row). Then, from the menu, choose **Insert > Pie > 2-D Pie**. From the options given, choose the graph on the top left. (If you are having trouble finding this option after selecting **Insert**, look for the circle above **Charts**.)

C. In order to add category names and their respective percentages, select the '+' sign that appears to the right of the pie chart.

A Bar Chart

Excel provides many options for showing a bar chart. To replicate the bar chart in Figure 2.2, follow these steps:

A. FILE Open ***Marital_Status*** (Table 2.5).

FILE
Marital_Status

B. First, select the category names and respective relative frequencies for the years 1960 and 2010. Leave out the heading (top row). Then, from the menu, choose **Insert > Column > 2-D Column**. From the options given, choose the graph on the top left. (If you are having trouble finding this option after selecting **Insert**, look for the vertical bars above **Charts**.)

C. In the legend to the right of the bar chart, Excel labels the data for the year 1960 as "Series 1" and the data for the year 2010 as "Series 2" by default. In order to edit the legend, select the legend and choose **Design > Select Data**. From the *Legend Entries,* select "Series 1," then select *Edit,* and under *Series Name,* type the new name of 1960. Follow the same steps to rename "Series 2" to 2010.

EXERCISES 2.1

1. A local restaurant is committed to providing its patrons with the best dining experience possible. On a recent survey, the restaurant asked patrons to rate the quality of their entrées. The responses ranged from 1 to 5, where 1 indicated a disappointing entrée and 5 indicated an exceptional entrée. The results of the survey are as follows:

3	5	4	4	3	2	3	3	2	5	5	5
5	3	3	2	1	4	5	5	4	2	5	5
5	4	4	3	1	5	2	1	5	4	4	4

   a. Construct frequency and relative frequency distributions that summarize the survey's results.
   b. Are patrons generally satisfied with the quality of their entrées? Explain.

2. First-time patients at North Shore Family Practice are required to fill out a questionnaire that gives the doctor an overall idea of each patient's health. The first question is: "In general, what is the quality of your health?" The patient chooses Excellent, Good, Fair, or Poor. Over the past month, the responses to this question from first-time patients were:

Fair	Good	Fair	Excellent
Good	Good	Good	Poor
Excellent	Excellent	Poor	Good
Fair	Good	Good	Good
Good	Poor	Fair	Excellent
Excellent	Good	Good	Good

   a. Construct frequency and relative frequency distributions that summarize the responses to the questionnaire.
   b. What is the most common response to the questionnaire? How would you characterize the health of first-time patients at this medical practice?

3. A survey asked chief executives at leading U.S. firms the following question: "Where do you expect the U.S. economy to be 12 months from now?" A representative sample of their responses appears below:

Same	Same	Same	Better	Worse
Same	Same	Better	Same	Worse
Same	Better	Same	Better	Same
Worse	Same	Same	Same	Worse
Same	Same	Same	Better	Same

   a. Construct frequency and relative frequency distributions that summarize the responses to the survey. Where did most chief executives expect the U.S. economy to be in 12 months?
   b. Construct a pie chart and a bar chart to summarize your results.

4. AccuWeather.com reported the following weather delays at these major U.S. airline hubs for July 21, 2010:

City	Delay	City	Delay
Atlanta	PM Delays	Mpls./St. Paul	None
Chicago	None	New York	All Day Delays
Dallas/Ft. Worth	None	Orlando	None
Denver	All Day Delays	Philadelphia	All Day Delays
Detroit	AM Delays	Phoenix	None
Houston	All Day Delays	Salt Lake City	None
Las Vegas	All Day Delays	San Francisco	AM Delays
Los Angeles	AM Delays	Seattle	None
Miami	AM Delays	Washington	All Day Delays

a. Construct frequency and relative frequency distributions that summarize the delays at major U.S. hubs. What was the most common type of delay? Explain.

b. Construct a pie chart and a bar chart to summarize your results.

5. Fifty pro-football rookies were rated on a scale of 1 to 5, based on performance at a training camp as well as on past performance. A ranking of 1 indicated a poor prospect whereas a ranking of 5 indicated an excellent prospect. The following frequency distribution was constructed.

Rating	Frequency
1	4
2	10
3	14
4	18
5	4

a. How many of the rookies received a rating of 4 or better? How many of the rookies received a rating of 2 or worse?

b. Construct the corresponding relative frequency distribution. What percent received a rating of 5?

c. Construct a bar chart for these data.

6. A recent survey asked 5,324 individuals: "What's most important to you when choosing where to live?" The responses are shown in the following relative frequency distribution.

Response	Relative Frequency
Good jobs	0.37
Affordable homes	0.15
Top schools	0.11
Low crime	0.23
Things to do	0.14

Source: Turner, Inc.

a. Construct the corresponding frequency distribution. How many of the respondents chose "low crime" as the most important criterion when choosing where to live?

b. Construct a bar chart for the frequency distribution found in part a.

7. What is the perfect summer trip? A National Geographic Kids survey (*AAA Horizons,* April 2007) asked this question to 316 children ages 8 to 14. Their responses are given in the following frequency distribution.

Top Vacation Choice	Frequency
Cruises	140
Beaches	68
Amusement Parks	68
Big Cities	20
Lakes	12
Summer Camp	8

a. Construct the relative frequency distribution. What percentage of the responses cited "Cruises" as the perfect summer trip?

b. Construct a bar chart for these data.

8. The following table lists U.S. revenue (in $ billions) of the major car-rental companies.

Car-Rental Company	Revenue in 2009
Enterprise	10.7
Hertz	4.7
Avis Budget	4.0
Dollar Thrifty	1.5
Other	1.0

Source: *The Wall Street Journal*, July 30, 2010.

a. Compute the relative market share of the car-rental companies.

b. Hertz accounted for what percentage of sales?

c. Construct a pie chart for these data.

9. In a CBS News survey, 829 respondents were provided with a list of major events and asked which event would happen first. The following percent frequency distribution was constructed.

Event	Percent Frequency
Cure for cancer found	40
End of dependence on oil	27
Signs of life in outer space	12
Peace in Middle East	8
Other	6
None will happen	7

Source: *Vanity Fair,* December 2009.

a. Construct a pie chart and a bar chart for these data.

b. How many people think that a cure for cancer will be found first?

10. A 2010 poll conducted by NBC asked respondents who would win Super Bowl XLV in 2011. The responses by 20,825 people are summarized in the following table.

Team	Number of Votes
Atlanta Falcons	4,040
New Orleans Saints	1,880
Houston Texans	1,791
Dallas Cowboys	1,631
Minnesota Vikings	1,438
Indianapolis Colts	1,149
Pittsburgh Steelers	1,141
New England Patriots	1,095
Green Bay Packers	1,076
Others	

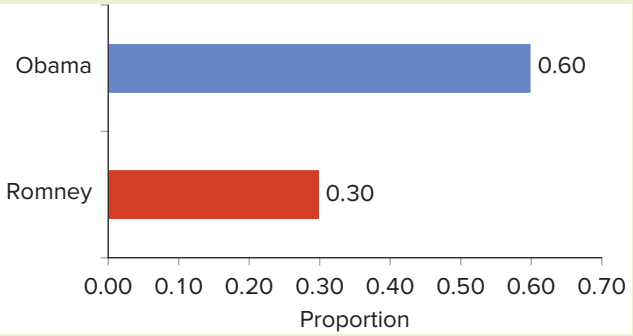
a. How many responses were for "Others"?

b. The Green Bay Packers won Super Bowl XLV, defeating the Pittsburgh Steelers by the score of 31–25. What

proportion of respondents felt that the Green Bay Packers would win?

c. Construct a bar chart for these data using relative frequencies.

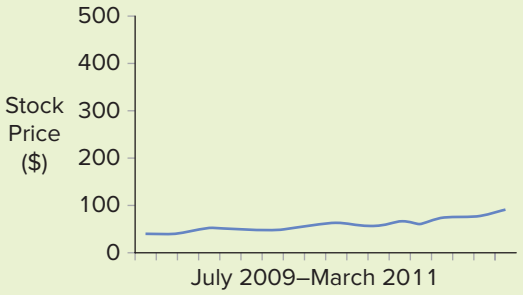
11. In a USA TODAY/Gallup Poll, respondents favored Barack Obama over Mitt Romney in terms of likeability, 60% to 30% (*Los Angeles Times,* July 28, 2012). The following bar chart summarizes the responses.



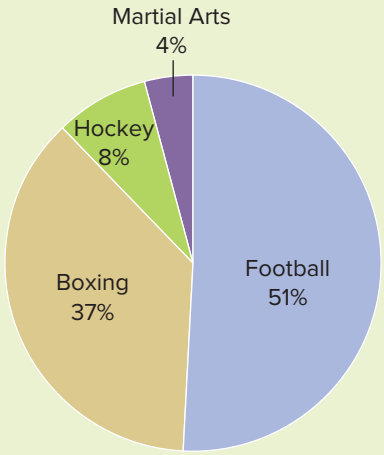
a. What percentage of respondents favored neither Obama nor Romney in terms of likeability?

b. Suppose this survey was based on 500 respondents. How many respondents favored Obama over Romney?

12. The accompanying figure plots the monthly stock price of Caterpillar, Inc., from July 2009 through March 2011. The stock has experienced tremendous growth over this time period, almost tripling in price. Does the figure reflect this growth? If not, why not?



13. A recent survey of 992 people asked: In which professional sport—football, boxing, hockey, or martial arts—is an athlete most likely to sustain an injury that will affect the athlete after he or she retires? (*Vanity Fair,* January 29, 2012.) The following pie chart summarizes the responses.

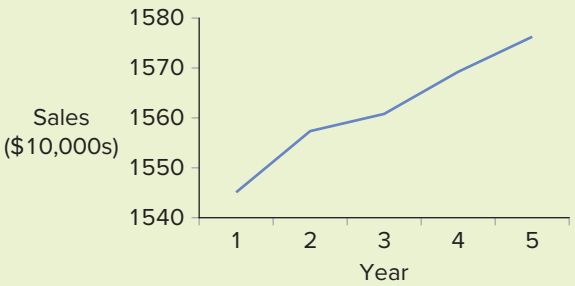


Source: Conde Nast.

a. According to this survey, in which sport was an athlete most likely to sustain an injury with lifelong consequences? In which sport was an athlete least likely to sustain an injury with lifelong consequences?

b. How many respondents believed that professional hockey players were most likely to sustain an injury with lifelong consequences?

14. Annual sales at a small pharmaceutical firm have been rather stagnant over the most recent five-year period, exhibiting only 1.2% growth over this time frame. A research analyst prepares the accompanying graph for inclusion in a sales report.



Does this graph accurately reflect what has happened to sales over the last five years? If not, why not?

2.2 SUMMARIZING QUANTITATIVE DATA

LO 2.3

Summarize quantitative data by constructing a frequency distribution.

With quantitative data, each observation represents a meaningful amount or count. The number of patents held by pharmaceutical firms (count) and household incomes (amount) are examples of quantitative data. Although different in nature from qualitative data, we still use a **frequency distribution** to summarize quantitative data.

Before discussing the mechanics of constructing a frequency distribution, we find it useful to first examine one in its final form, using the house-price data from the introductory case to this chapter. We take the raw data (the observations) from Table 2.1 and construct a frequency distribution with five intervals or **classes**, each of width 100, as shown

in Table 2.6. We see, for instance, that four houses sold in the first class, where prices ranged from $300,000 up to $400,000. The data are more manageable using a frequency distribution, but some detail is lost because we no longer see the observations.

TABLE 2.6 Frequency Distribution for House-Price Data

Class (in $1,000s)	Frequency
300 up to 400	4
400 up to 500	11
500 up to 600	14
600 up to 700	5
700 up to 800	2
	Total = 36

EXAMPLE 2.2

Based on the frequency distribution in Table 2.6, what is the price range over this time period? What price range exhibited the highest frequency?

SOLUTION: The frequency distribution shows that house prices ranged from $300,000 up to $800,000 over this time period. The most houses (14) sold in the $500,000 up to $600,000 range. Note that only four houses sold in the lowest price range and only two houses sold at the highest price range.

It turns out that reading and understanding a frequency distribution is actually easier than forming one. When we constructed a frequency distribution with qualitative data, the raw data could be categorized in a well-defined way. With quantitative data, we must make certain decisions about the number of classes, as well as the width of each class. We do not apply concrete rules when we define the classes in Table 2.6; however, we are able to follow several guidelines.

TABLE 2.7 Too Many Classes in a Distribution

Class (in $1,000s)	Frequency
325 up to 350	2
350 up to 375	1
375 up to 400	1
400 up to 425	3
425 up to 450	5
450 up to 475	3
475 up to 500	0
500 up to 525	5
525 up to 550	5
550 up to 575	3
575 up to 600	1
600 up to 625	0
625 up to 650	1
650 up to 675	4
675 up to 700	0
700 up to 725	1
725 up to 750	1
	Total = 36

Guidelines for Constructing a Frequency Distribution

- *Classes are mutually exclusive.* In other words, classes do not overlap. Each observation falls into one, and only one, class. For instance, suppose a value of 400 appeared in Table 2.1. Given the class divisions in Table 2.6, we would have included this observation in the second class interval. Mathematically, the second class interval is expressed as $400 \leq \text{Price} < 500$. Alternatively, we can define the second interval as $400 < \text{Price} \leq 500$, in which case the value 400 is included in the previous class interval. In short, no matter the specification of the classes, the observation is included in only one of the classes.
- *Classes are exhaustive.* The total number of classes covers the entire sample (or population). In Table 2.6, if we had left off the last class, 700 up to 800, then we would be omitting two observations from the sample.
- *The total number of classes in a frequency distribution usually ranges from 5 to 20.* Smaller data sets tend to have fewer classes than larger data sets. Recall that the goal of constructing a frequency distribution is to summarize the data in a form that accurately depicts the group as a whole. If we have too many classes, then this advantage of the frequency distribution is lost. For instance, suppose we create a frequency distribution for the house-price data with 17 classes, each of width 25, as shown in Table 2.7. Technically, this is a valid frequency distribution, but the summarization advantage of the frequency distribution is lost because there are too

many class intervals. Similarly, if the frequency distribution has too few classes, then considerable accuracy and detail are lost. Consider a frequency distribution of the house-price data with three classes, each of width 150, as shown in Table 2.8.

TABLE 2.8 Too Few Classes in a Distribution

Class (in $1,000s)	Frequency
300 up to 450	12
450 up to 600	17
600 up to 750	7
	Total = 36

Again, this is a valid frequency distribution. However, we cannot tell whether the 17 houses that sold for $450,000 up to $600,000 fall closer to the price of $450,000, fall closer to the price of $600,000, or are evenly spread within the interval. With only three classes in the frequency distribution, too much detail is lost.

- Once we choose the number of classes for a raw data set, we can then *approximate the width of each class* by using the formula

$$\frac{\text{Largest value} - \text{Smallest value}}{\text{Number of classes}}.$$

Generally, the width of each class is the same for each class interval. If the class width varied, comparisons between the numbers of observations in different intervals would be misleading.

- It is preferable to *define class limits that are easy to recognize and interpret.* Suppose we conclude, as we do in Table 2.6, that we should have five classes in the frequency distribution for the house-price data. Applying the class-width formula with the largest value of 735 and the smallest value of 330 (from Table 2.1) yields $\frac{735 - 330}{5} = 81$. Table 2.9 shows the frequency distribution with five classes and a class width of 81.

TABLE 2.9 Cumbersome Class Width in a Distribution

Class (in $1,000s)	Frequency
330 up to 411	4
411 up to 492	11
492 up to 573	12
573 up to 654	3
654 up to 735	6
	Total = 36

Again, this is a valid frequency distribution, but it proves unwieldy. Recall that one major goal in forming a frequency distribution is to provide more clarity in interpreting the data. Grouping the data in this manner actually makes analyzing the data more difficult. In order to facilitate interpretation of the frequency distribution, it is best to define class limits with ease of recognition in mind. To this end, and as initially shown in Table 2.6, we set the lower limit of the first class at 300 (rather than 330) and obtain the remaining class limits by successively adding 100 (rather than 81).

Once we have clearly defined the classes for a particular data set, the next step is to count and record the number of data points that fall into each class. As we did with the construction of a qualitative frequency distribution, we usually include a tally column to aid in counting (see Table 2.10), but then we remove this column in the final presentation of the frequency distribution. For instance, in Table 2.1, the first observation, 430, falls in the second class, so we place a tally mark in the second class; the next observation, 520, falls in the third class, so we place a tally

mark in the third class, and so on. The frequency column shows the numerical value of the respective tally count. Since four tally marks appear in the first class, we record the value 4 as its frequency—the number of observations that fall within the first class. One way to ensure that we have included all the observations in the frequency distribution is to sum the frequency column. This sum should always equal the population or sample size.

TABLE 2.10 Constructing Frequency Distributions for the House-Price Data

Class (in $1,000s)	Tally	Frequency	Cumulative Frequency
300 up to 400	IIII	4	4
400 up to 500	𝍸 𝍸 I	11	4 + 11 = 15
500 up to 600	𝍸 𝍸 IIII	14	4 + 11 + 14 = 29
600 up to 700	𝍸	5	4 + 11 + 14 + 5 = 34
700 up to 800	II	2	4 + 11 + 14 + 5 + 2 = 36
		Total = 36	

A frequency distribution indicates how many observations fall within some range. However, we might want to know how many observations fall below the upper limit of a particular class. In these cases, our needs are better served with a **cumulative frequency distribution**.

The last column of Table 2.10 shows values for cumulative frequency. The cumulative frequency of the first class is the same as the frequency of the first class—that is, the value 4. However, the interpretation is different. With respect to the frequency column, the value 4 tells us that four of the houses sold in the $300,000 up to $400,000 range. For the cumulative frequency column, the value 4 tells us that four of the houses sold for less than $400,000. To obtain the cumulative frequency for the second class we add its frequency, 11, with the preceding frequency, 4, and obtain 15. This tells us that 15 of the houses sold for less than $500,000. We solve for the cumulative frequencies of the remaining classes in a like manner. Note that the cumulative frequency of the last class is equal to the sample size of 36. This indicates that all 36 houses sold for less than $800,000.

FREQUENCY AND CUMULATIVE FREQUENCY DISTRIBUTIONS FOR QUANTITATIVE DATA

For quantitative data, a frequency distribution groups data into intervals called classes and records the number of observations that falls within each class. A cumulative frequency distribution records the number of observations that fall below the upper limit of each class.

EXAMPLE 2.3

Using Table 2.10, how many of the houses sold in the $500,000 up to $600,000 range? How many of the houses sold for less than $600,000?

SOLUTION: From the frequency distribution, we find that 14 houses sold in the $500,000 up to $600,000 range. In order to find the number of houses that sold for less than $600,000, we use the cumulative frequency distribution. We readily observe that 29 of the houses sold for less than $600,000.

Suppose we want to compare house prices in Mission Viejo, California, to house prices in another region of the United States. Just as for qualitative data, when making comparisons between two quantitative data sets—especially if the data sets differ in

size—a **relative frequency distribution** tends to provide more meaningful information than a frequency distribution.

The second column of Table 2.11 shows the construction of a relative frequency distribution from the frequency distribution in Table 2.10. We take each class's frequency and divide by the total number of observations. For instance, we observed four houses that sold in the lowest range of $300,000 up to $400,000. We take the class frequency of 4 and divide by the sample size, 36, and obtain 0.11. Equivalently, we can say 11% of the houses sold in this price range. We make similar calculations for each class and note that when we sum the column of relative frequencies, we should get a value of one (or, due to rounding, a number very close to one).

TABLE 2.11 Constructing Relative Frequency Distributions for House-Price Data

Class (in $1,000s)	Relative Frequency	Cumulative Relative Frequency
300 up to 400	4/36 = 0.11	0.11
400 up to 500	11/36 = 0.31	0.11 + 0.31 = 0.42
500 up to 600	14/36 = 0.39	0.11 + 0.31 + 0.39 = 0.81
600 up to 700	5/36 = 0.14	0.11 + 0.31 + 0.39 + 0.14 = 0.95
700 up to 800	2/36 = 0.06	0.11 + 0.31 + 0.39 + 0.17 + 0.06 ≈ 1
	Total = 1 (subject to rounding)	

The last column of Table 2.11 shows the **cumulative relative frequency distribution**. The cumulative relative frequency for a particular class indicates the proportion (fraction) of the observations that falls below the upper limit of that particular class. We can calculate the cumulative relative frequency of each class in one of two ways: (1) We can sum successive relative frequencies, or (2) we can divide each class's cumulative frequency by the sample size. In Table 2.11 we show the first way. The value for the first class is the same as the value for its relative frequency—that is, 0.11. For the second class, we add 0.31 to 0.11 and obtain 0.42; this value indicates that 42% of the house prices were less than $500,000. We continue calculating cumulative relative frequencies in this manner until we reach the last class. Here, we get the value one, which means that 100% of the houses sold for less than $800,000.

RELATIVE AND CUMULATIVE RELATIVE FREQUENCY DISTRIBUTIONS

For quantitative data, a relative frequency distribution identifies the proportion (or the fraction) of observations that falls within each class—that is,

$$\text{Class relative frequency} = \frac{\text{Class frequency}}{\text{Total number of observations}}.$$

A cumulative relative frequency distribution records the proportion (or the fraction) of observations that fall below the upper limit of each class.

EXAMPLE 2.4

Using Table 2.11, what percent of the houses sold for at least $500,000 but not more than $600,000? What percent of the houses sold for less than $600,000? What percent of the houses sold for $600,000 or more?

SOLUTION: The relative frequency distribution indicates that 39% of the houses sold for at least $500,000 but not more than $600,000. Further, the cumulative relative frequency distribution indicates that 81% of the houses sold for less than $600,000. This result implies that 19% sold for $600,000 or more.

SYNOPSIS OF INTRODUCTORY CASE



©Brand X Pictures/Stockbyte/Getty Images

During June 2010, Matthew Edwards reviewed the selling prices of 36 house sales in Mission Viejo, California, for a client from Seattle, Washington. After constructing various frequency distributions, he is able to make the following summary conclusions. House prices ranged from $300,000 up to $800,000 over this time period. Most of the houses (14) sold in the $500,000 up to $600,000 range, which is, more or less, the client's price range. Twenty-nine of the houses sold for less than $600,000. Converting the data into percentages so the client can make comparisons with home sales in the Seattle area, Matthew found that 39% of the houses sold for $500,000 up to $600,000. Moreover, 81% of the houses sold for less than $600,000, which implies that 19% sold for $600,000 or more.

LO **2.4**

Construct and interpret a histogram, a polygon, and an ogive.

Histograms, Polygons, and Ogives

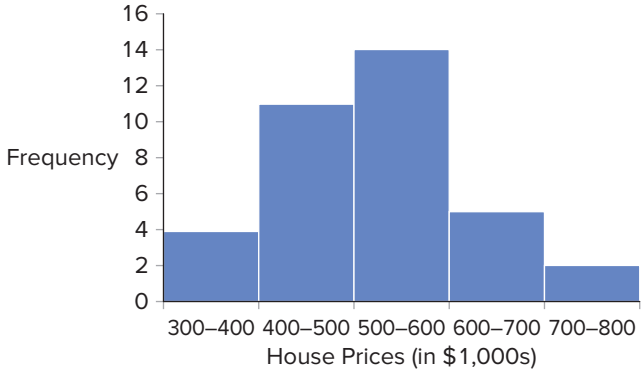
A **histogram** and a **polygon** are graphical depictions of frequency and relative frequency distributions. The advantage of a visual display is that we can quickly see where most of the observations tend to cluster, as well as the spread and shape of the data. For instance, a histogram and a polygon may reveal whether or not the distribution has a symmetric shape.

GRAPHICAL DISPLAY OF QUANTITATIVE DATA: A HISTOGRAM

A histogram is a series of rectangles where the width and height of each rectangle represent the class width and frequency (or relative frequency) of the respective class.

For quantitative data, a histogram is essentially the counterpart to the vertical bar chart we use for qualitative data. When constructing a histogram, we mark off the class limits along the horizontal axis. The height of each bar represents either the frequency or the relative frequency for each class. No gaps appear between the interval limits. Figure 2.5 shows a histogram for the frequency distribution of house prices shown in Table 2.6. A casual inspection of the histogram reveals that the selling price of houses in this sample ranged from $300,000 to $800,000; however, the most common house price fell in the $500,000 to $600,000 range.

FIGURE 2.5
Frequency histogram for house prices



The only difference between a frequency histogram and a relative frequency histogram is the unit of measurement on the vertical axis. For the frequency histogram, we use the frequency of each class to represent the height; for the relative frequency histogram we use the proportion (or the fraction) of each class to represent the height. In a relative frequency histogram, the area of any rectangle is proportional to the relative frequency of observations falling into that class. Figure 2.6 shows the relative frequency histogram for house prices.

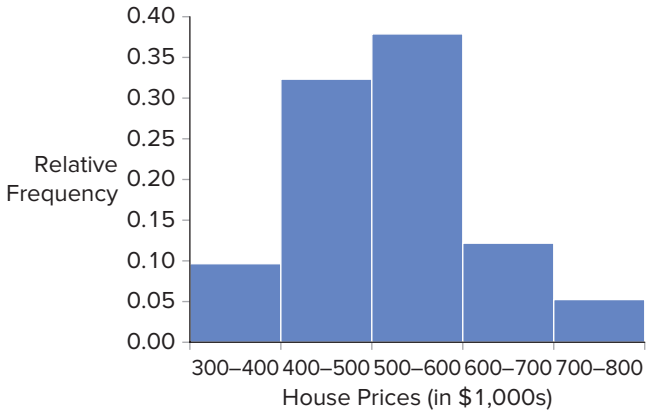
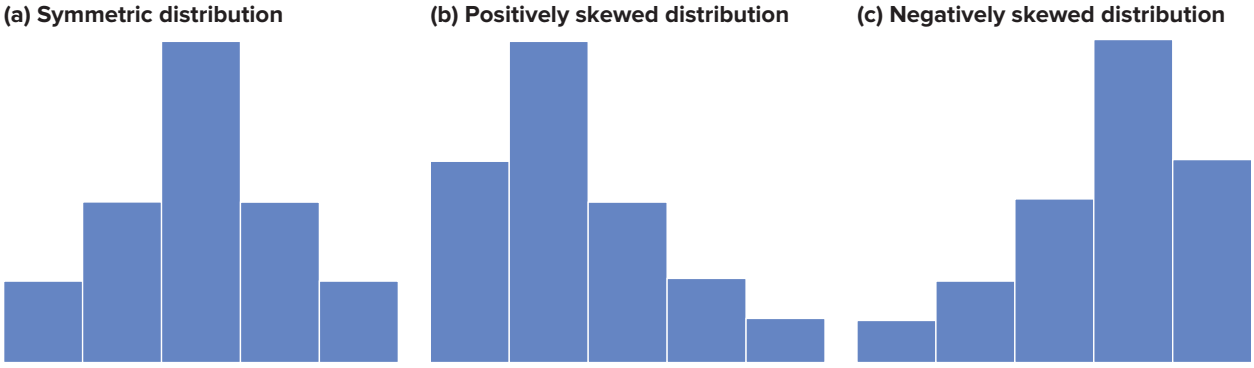


FIGURE 2.6 Relative frequency histogram for house prices

In general, the shape of most data distributions can be categorized as either symmetric or skewed. A symmetric distribution is one that is a mirror image of itself on both sides of its center. That is, the location of values below the center correspond to those above the center. As we will see in later chapters, the smoothed histogram for many data sets approximates a bell-shaped curve, which is indicative of the well-known normal distribution. If the distribution is not symmetric, then it is either positively skewed or negatively skewed.

FIGURE 2.7 Histograms with differing shapes



The histogram in Figure 2.7(a) shows a symmetric distribution. If the edges were smoothed, this histogram would look somewhat bell-shaped. In Figure 2.7(b), the histogram shows a positively skewed, or skewed to the right, distribution with a long tail extending to the right. This attribute reflects the presence of a small number of relatively large values. Finally, the histogram in Figure 2.7(c) indicates a negatively skewed, or skewed to the left, distribution since it has a long tail extending off to the left. Data that follow a negatively skewed distribution have a small number of relatively small values.

Though not nearly as skewed as the data exhibited in Figure 2.7(b), the house-price data in Figure 2.6 exhibit slight positive skew. This is the result of a few, relatively expensive homes in the city. It is common for distributions of house prices and incomes to exhibit positive skewness.

A polygon provides another convenient way of depicting a frequency distribution. It too gives a general idea of the shape of a distribution. Like the histogram, we place either the frequency or the relative frequency of the distribution on the *y*-axis, and the upper and lower limits of each class on the *x*-axis. We plot the midpoint of each class with its corresponding frequency or relative frequency. We then connect neighboring points with a straight line.

GRAPHICAL DISPLAY OF QUANTITATIVE DATA: A POLYGON

A polygon connects a series of neighboring points where each point represents the midpoint of a particular class and its associated frequency or relative frequency.

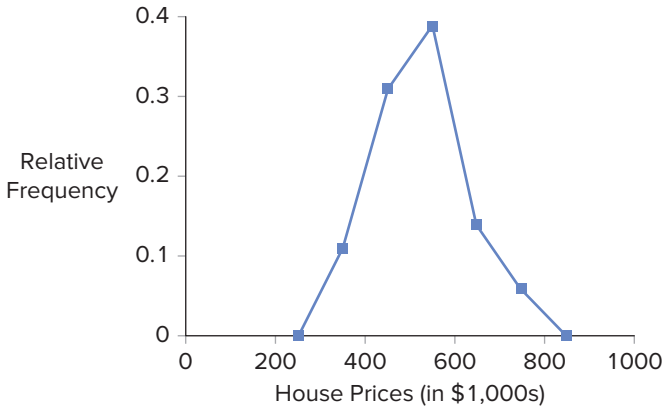
If we choose to construct a polygon for the house-price data, we first calculate the midpoint of each interval; thus, the midpoint for the first interval is $\frac{300 + 400}{2} = 350$, and similarly, the midpoints for the remaining intervals are 450, 550, 650, and 750. We treat each midpoint as the *x*-coordinate and the respective frequency (or relative frequency) as the *y*-coordinate. After plotting the points, we connect neighboring points. In order to close off the graph at each end, we add one interval below the lowest interval (so, 200 up to 300 with midpoint 250) and one interval above the highest interval (so, 800 up to 900 with midpoint 850) and assign each of these classes zero frequencies. Table 2.12 shows the relevant coordinates for plotting a polygon using the house-price data. Here we use relative frequency to represent the *y*-coordinate.

TABLE 2.12 Coordinates for Plotting Relative Frequency Polygon

Classes	x-coordinate (midpoint)	y-coordinate (relative frequency)
(Lower end)	250	0
300–400	350	0.11
400–500	450	0.31
500–600	550	0.39
600–700	650	0.14
700–800	750	0.06
(Upper end)	850	0

Figure 2.8 plots a relative frequency polygon for the house-price data. The distribution appears to approximate the bell-shaped distribution discussed earlier. Only a careful inspection of the right tail suggests that the data are slightly positively skewed.

FIGURE 2.8
Polygon for the house-price data



In many instances, we might want to convey information by plotting an **ogive** (pronounced "ojive").

GRAPHICAL DISPLAY OF QUANTITATIVE DATA: AN OGIVE

An ogive connects a series of neighboring points where each point represents the upper limit of a particular class and its associated cumulative frequency or cumulative relative frequency.

An ogive differs from a polygon in that we use the upper limit of each class as the x-coordinate and the cumulative frequency or cumulative relative frequency of the corresponding class as the y-coordinate. After plotting the points, we connect neighboring points. Lastly, we close the ogive only at the lower end by intersecting the x-axis at the lower limit of the first class. Table 2.13 shows the relevant coordinates for plotting an ogive using the house-price data. Here we use the cumulative relative frequency as the y-coordinate since the resulting graph tends to have more interpretive appeal. The use of cumulative frequency would not change the shape of the ogive, just the unit of measurement on the y-axis.

TABLE 2.13 Coordinates for the Ogive for the House-Price Data

Classes	x-coordinate (upper limit)	y-coordinate (cumulative relative frequency)
(Lower end)	300	0
300–400	400	0.11
400–500	500	0.42
500–600	600	0.81
600–700	700	0.95
700–800	800	1

Figure 2.9 plots the ogive for the house-price data. In general, we can use an ogive to approximate the proportion of values that are less than a specified value on the horizontal axis. Consider an application to the house-price data in Example 2.5.

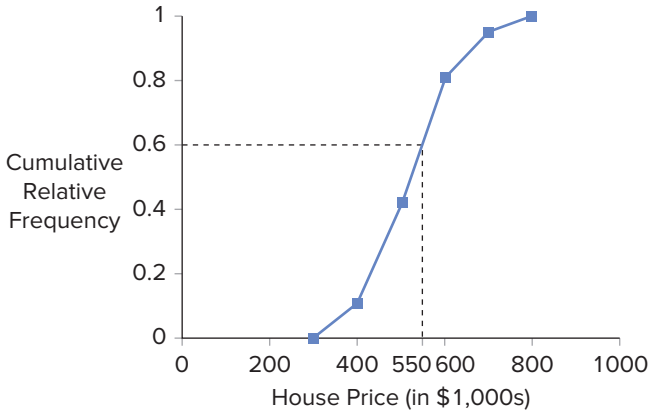


FIGURE 2.9
Ogive for the house-price data

EXAMPLE 2.5

Using Figure 2.9, approximate the percentage of houses that sold for less than $550,000.

SOLUTION: We first draw a vertical line that starts at 550 and intersects the ogive. Then we follow the line to the vertical axis and read the value. We can conclude that approximately 60% of the houses sold for less than $550,000.

Using Excel to Construct a Histogram, a Polygon, and an Ogive

A Histogram Constructed from Raw Data

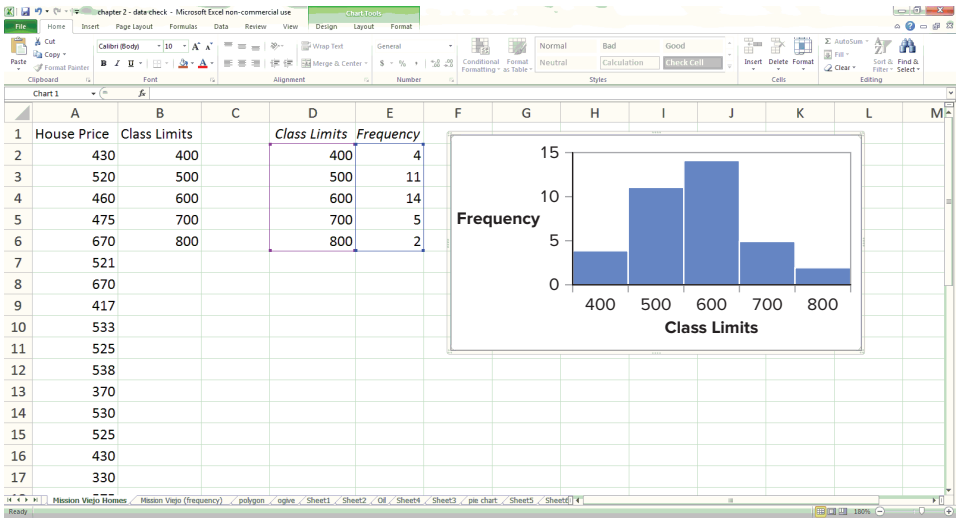
In general, Excel offers two different ways to construct a histogram, depending on whether we have access to the raw data or the frequency distribution. We first replicate Figure 2.5 using the house-price data from Table 2.1 where the data are in raw form. We then replicate Figure 2.5 using the house-price data from Table 2.6 where the data have been converted to a frequency distribution.

FILE
MV_Houses

A. Open ***MV_Houses*** (Table 2.1).

B. See Figure 2.10. In a column next to the data, enter the values of the upper limits for each class, or in this example, 400, 500, 600, 700, and 800; label this column as Class Limits. The reason for these entries is explained in step C. From the menu choose **Data > Data Analysis > Histogram > OK**. (*Note:* If you do not see the **Data Analysis** option under **Data**, you must *add in* this option. From the menu choose **File > Options > Add-Ins** and choose **Go** at the bottom of the dialog box. Select the box to the left of **Analysis Toolpak**, and then click **OK**. If you have installed this option properly, you should now see **Data Analysis** under **Data**.)

FIGURE 2.10 Constructing a histogram from raw data with Excel

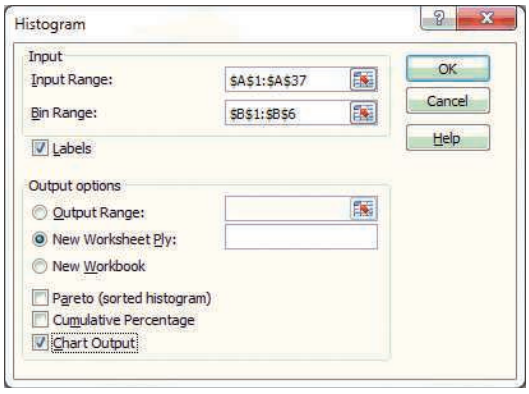


Source: Microsoft Excel

C. See Figure 2.11. In the *Histogram* dialog box, under *Input Range*, select the House Price data. Excel uses the term "bins" for the class limits. If we leave the *Bin Range* box empty, Excel creates evenly distributed intervals using the minimum and maximum values of the input range as end points. This methodology is rarely satisfactory. In order to construct a histogram that is more informative, we use the upper limit of each class as the bin values. Under *Bin Range*, we select the Class Limits data. (Check

the *Labels* box if you have included the names House Price and Class Limits as part of the selection.) Under *Output Options,* we choose **Chart Output**, then click **OK**.

FIGURE 2.11 Excel's dialog box for a histogram



Source: Microsoft Excel

D. Since Excel leaves spaces between the rectangles, we right-click on any of the rectangles, choose **Format Data Series** and change the *Gap Width* to 0, then choose **Close**. In the event that the given class limits do not include all the data points, Excel automatically adds another interval labeled "More" to the resulting frequency distribution and histogram. Since we observe zero observations in this interval for this example, we delete this interval for expositional purposes. Excel also defines its classes by excluding the value of the lower limit and including the value of the upper class limit for each interval. For example, if the value 400 appeared in the house-price data, Excel would have accounted for this observation in the first class. If any upper-limit value appeared in the house-price data, we would have adjusted the class limits in the *Bin Range* to 399, 499, and so on, so that Excel's frequency distribution and histogram would be consistent with those that we constructed in Table 2.10 and Figure 2.5. Formatting (regarding axis titles, gridlines, etc.) can be done by selecting **Format > Add Chart Element** from the menu.

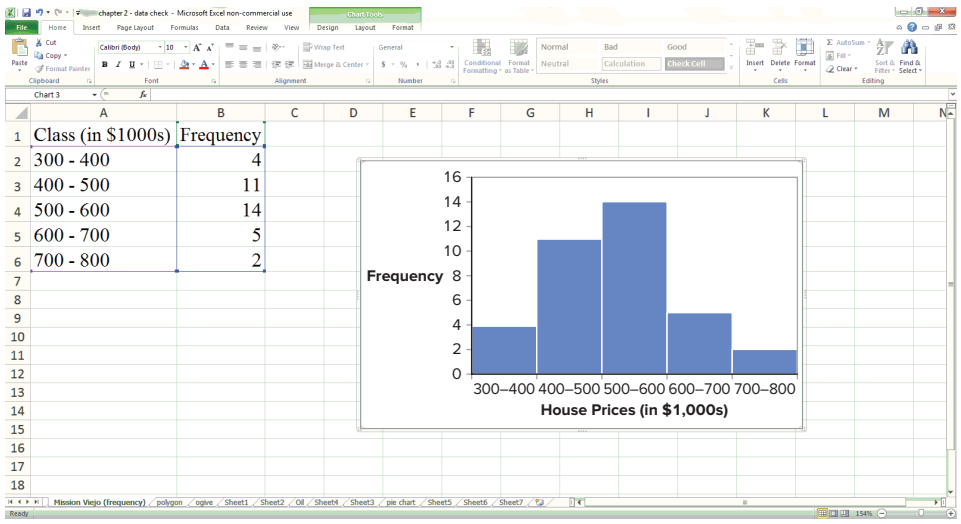
A Histogram Constructed from a Frequency Distribution

Suppose we do not have the raw data for house prices, but we have the frequency distribution reported in Table 2.6.

A. Open ***MV_Frequency*** (Table 2.6).

FILE
MV_Frequency

B. See Figure 2.12. First, select the classes and respective frequencies. Then, from the menu, choose **Insert > Column > 2-D Column**. From the options, choose the



Source: Microsoft Excel

FIGURE 2.12
Constructing a histogram from a frequency distribution with Excel

graph on the top left. (If you are having trouble finding this option after selecting **Insert**, look for the vertical bars above **Charts**.)

C. In order to remove the spaces between the rectangles, right-click on any of the rectangles, choose **Format Data Series** and change the *Gap Width* to 0, then choose **Close**.

D. Formatting (regarding axis titles, gridlines, etc.) can be done by selecting **Format > Add Chart Element** from the menu.

A Polygon

We replicate the polygon in Figure 2.8.

FILE *Polygon*

A. Open ***Polygon*** (this is a simplified version of the data in Table 2.12).

B. Select the values in the *x* and *y* columns, and choose **Insert > Scatter (X, Y) or Bubble Chart**. From the options given, select the box at the middle right. (If you are having trouble finding this option after selecting **Insert**, look for the graph with data points above **Charts**.)

C. Formatting (regarding axis titles, gridlines, etc.) can be done by selecting **Format > Add Chart Element** from the menu.

An Ogive

We replicate the polygon in Figure 2.9.

FILE *Ogive*

A. Open ***Ogive*** (this is a simplified version of the data in Table 2.13).

B. Select the values in the *x* and *y* columns, and choose **Insert > Scatter (X, Y) or Bubble Chart**. From the options given, select the box at the middle right. (If you are having trouble finding this option after selecting **Insert**, look for the graph with data points above **Charts**.)

C. Formatting (regarding axis titles, gridlines, etc.) can be done by selecting **Format > Add Chart Element** from the menu.

EXERCISES 2.2

Mechanics

15. Consider the following data set:

4.3	10.9	8.7	7.6	6.5	10.9	11.1	14.3	13.2	14.3
3.2	9.8	8.7	5.4	7.6	6.5	10.9	3.2	11.1	11.1
8.7	8.7	4.3	5.4	5.4	12.1	12.1	3.2	8.7	8.7

a. Construct the frequency distribution using classes of 3 up to 5, 5 up to 7, etc.

b. Construct the relative frequency, the cumulative frequency, and the cumulative relative frequency distributions.

c. How many of the observations are at least 7 but less than 9? How many of the observations are less than 9?

d. What percentage of the observations are at least 7 but less than 9? What percentage of the observations are less than 9?

e. Graph the relative frequency histogram.

f. Graph the ogive.

16. Consider the following data set:

4.3	10.9	8.7	7.6	6.5	10.9	11.1	14.3	13.2	14.3
3.2	9.8	8.7	5.4	7.6	6.5	10.9	3.2	11.1	11.1
8.7	8.7	4.3	5.4	5.4	12.1	12.1	3.2	8.7	8.7
10.9	−8.8	28.7	14.3	−4.8	9.8	11.1	5.4	8.7	−2.8
33.2	−3.9	2.1	3.2	22.1	25.4	5.4	29.8	26.5	0.1
−7.6	−4.8	0.1	15.4	−3.8	35.4	21.1	15.4	19.8	23.2
4.3	6.5	−1.9	12.1	24.3	36.5	15.4	3.2	−4.8	2.1

a. Construct the frequency distribution using classes of −10 up to 0, 0 up to 10, etc. How many of the observations are at least 10 but less than 20?

b. Construct the relative frequency distribution and the cumulative relative frequency distribution. What percent of the observations are at least 10 but less than 20? What percent of the observations are less than 20?

c. Graph the relative frequency polygon. Is the distribution symmetric? If not, then how is it skewed?

17. Consider the following frequency distribution:

Class	Frequency
10 up to 20	12
20 up to 30	15
30 up to 40	25
40 up to 50	4

a. Construct the relative frequency distribution. Graph the relative frequency histogram.
b. Construct the cumulative frequency distribution and the cumulative relative frequency distribution.
c. What percent of the observations are at least 30 but less than 40? What percent of the observations are less than 40?

18. Consider the following frequency distribution:

Class	Frequency
1,000 up to 1,100	2
1,100 up to 1,200	7
1,200 up to 1,300	3
1,300 up to 1,400	4

a. Construct the relative frequency distribution. What percent of the observations are at least 1,100 but less than 1,200?
b. Construct the cumulative frequency distribution and the cumulative relative frequency distribution. How many of the observations are less than 1,300?
c. Graph the frequency histogram.

19. Consider the following cumulative frequency distribution:

Class	Cumulative Frequency
15 up to 25	30
25 up to 35	50
35 up to 45	120
45 up to 55	130

a. Construct the frequency distribution. How many observations are at least 35 but less than 45?
b. Graph the frequency histogram.
c. What percent of the observations are less than 45?

20. Consider the following relative frequency distribution:

Class	Relative Frequency
−20 up to −10	0.04
−10 up to 0	0.28
0 up to 10	0.26
10 up to 20	0.22
20 up to 30	0.20

a. Suppose this relative frequency distribution is based on a sample of 50 observations. Construct the frequency distribution. How many of the observations are at least −10 but less than 0?
b. Construct the cumulative frequency distribution. How many of the observations are less than 20?
c. Graph the relative frequency polygon.

21. Consider the following cumulative relative frequency distribution.

Class	Cumulative Relative Frequency
150 up to 200	0.10
200 up to 250	0.35
250 up to 300	0.70
300 up to 350	1

a. Construct the relative frequency distribution. What percent of the observations are at least 250 but less than 300?
b. Graph the ogive.

Applications

22. *Kiplinger's* (August 2007) lists the assets (in billions of $) for the 20 largest stock mutual funds (ranked by size) as follows:

99.8	49.7	86.3	109.2	56.9
88.2	44.1	58.8	176.7	49.9
61.4	128.8	53.6	95.2	92.5
55.0	96.5	45.3	73.0	70.9

a. Construct the frequency distribution using classes of 40 up to 70, 70 up to 100, etc.
b. Construct the relative frequency distribution, the cumulative frequency distribution, and the cumulative relative frequency distribution.
c. How many of the funds had assets of at least $100 but less than $130 (in billions)? How many of the funds had assets less than $160 (in billions)?
d. What percent of the funds had assets of at least $70 but less than $100 (in billions)? What percent of the funds had assets less than $130 (in billions)?
e. Construct the frequency histogram. Comment on the shape of the distribution.

23. The number of text messages sent by 25 13-year-olds over the past month are as follows:

630	516	892	643	627	510	937	909	654
817	760	715	605	975	888	912	952	701
744	793	852	504	562	670	685		

a. Construct the frequency distribution using classes of 500 up to 600, 600 up to 700, etc.
b. Construct the relative frequency distribution, the cumulative frequency distribution, and the cumulative relative frequency distribution.

c. How many of the 13-year-olds sent at least 600 but less than 700 text messages? How many sent less than 800 text messages?

d. What percent of the 13-year-olds sent at least 500 but less than 600 text messages? What percent of the 13-year-olds sent less than 700 text messages?

e. Graph the relative frequency polygon. Comment on the shape of the distribution.

24. AccuWeather.com listed the following high temperatures (in degrees Fahrenheit) for 33 European cities on July 21, 2010.

75	92	81	85	91	73	94	95	81	64	85
62	84	85	81	86	91	79	74	92	91	95
88	87	81	73	76	86	92	83	75	92	83

a. Construct the frequency distribution using classes of 60 up to 70, 70 up to 80, etc.

b. Construct the relative frequency, the cumulative frequency, and the cumulative relative frequency distributions.

c. How many of the cities had high temperatures less than 80°?

d. What percent of the cities had high temperatures of at least 80° but less than 90°? What percent of the cities had high temperatures less than 90°?

e. Construct the relative frequency polygon. Comment on the shape of the distribution.

25. The following relative frequency distribution summarizes the ages of women who had a child in the last year.

Ages	Relative Frequency
15 up to 20	0.10
20 up to 25	0.25
25 up to 30	0.28
30 up to 35	0.24
35 up to 40	0.11
40 up to 45	0.02

Source: *The Statistical Abstract of the United States, 2010.*

a. Assume the relative frequency distribution is based on a sample of 2,000 women. Construct the frequency, the cumulative frequency, and the cumulative relative frequency distributions.

b. What percent of the women were at least 25 but less than 30 years old? What percent of the women were younger than 35 years old?

c. Graph the relative frequency polygon. Comment on the shape of the distribution.

d. Graph the ogive. Using the graph, approximate the age of the middle 50% of the distribution.

26. Fifty cities provided information on vacancy rates (in percent) in local apartments in the following frequency distribution.

Vacancy Rate	Frequency
0 up to 3	5
3 up to 6	10
6 up to 9	20
9 up to 12	10
12 up to 15	5

a. Construct the relative frequency distribution, the cumulative frequency distribution, and the cumulative relative frequency distribution.

b. How many of the cities had a vacancy rate less than 12%? What percent of the cities had a vacancy rate of at least 6% but less than 9%? What percent of the cities had a vacancy rate of less than 9%?

c. Graph the frequency histogram. Comment on the shape of the distribution.

27. The manager of a nightclub near a local university recorded the ages of the last 100 guests in the following cumulative frequency distribution.

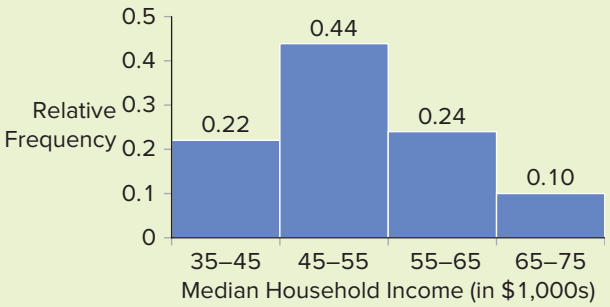
Ages	Cumulative Frequency
18 up to 22	45
22 up to 26	70
26 up to 30	85
30 up to 34	96
34 up to 38	100

a. Construct the frequency, the relative frequency, and the cumulative relative frequency distributions.

b. How many of the guests were at least 26 but less than 30 years old? What percent of the guests were at least 22 but less than 26 years old? What percent of the guests were younger than 34 years old? What percent were 34 years or older?

c. Graph the frequency histogram. Comment on the shape of the distribution.

28. The following relative frequency histogram summarizes the median household income for the 50 states in the United States (*U.S. Census,* 2010).

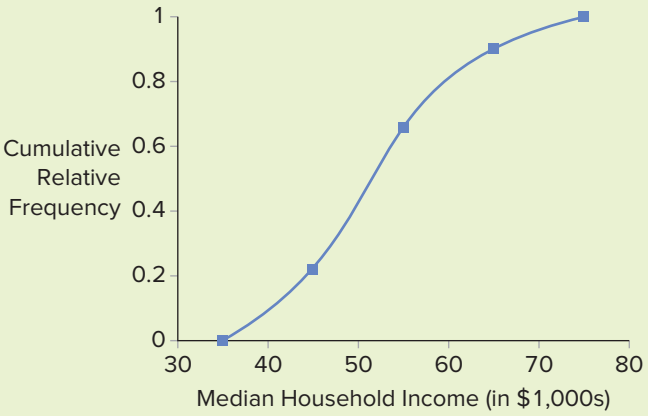


a. Is the distribution symmetric? If not, is it positively or negatively skewed?

b. What percentage of the states had median household income between $45,000 and $55,000?

c. What percentage of the states had median household income between $35,000 and $55,000?

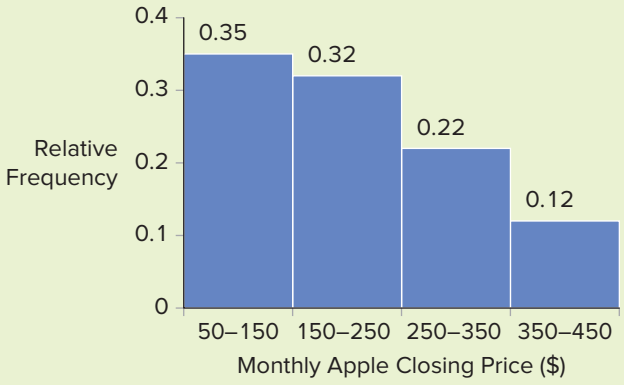
29. The following ogive summarizes the median household income for the 50 states in the United States (*U.S. Census,* 2010).



a. Approximate the percentage of states with median household income less than $50,000.

b. Approximate the percentage of states with median household income more than $60,000.

30. The following histogram summarizes Apple Inc.'s monthly stock price for the years 2007 through 2011 (http://finance.yahoo.com, data retrieved April 20, 2012).

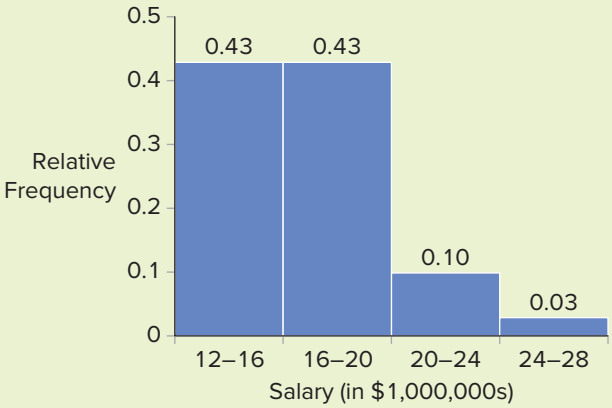


a. Is the distribution symmetric? If not, is it positively or negatively skewed?

b. Over this five-year period, approximate the minimum monthly stock price and the maximum monthly stock price.

c. Over this five-year period, which class had the highest relative frequency?

31. The following histogram summarizes the salaries (in $1,000,000s) for the 30 highest-paid players in the National Basketball Association (NBA) for the 2012 season (www.nba.com, data retrieved March 2012).

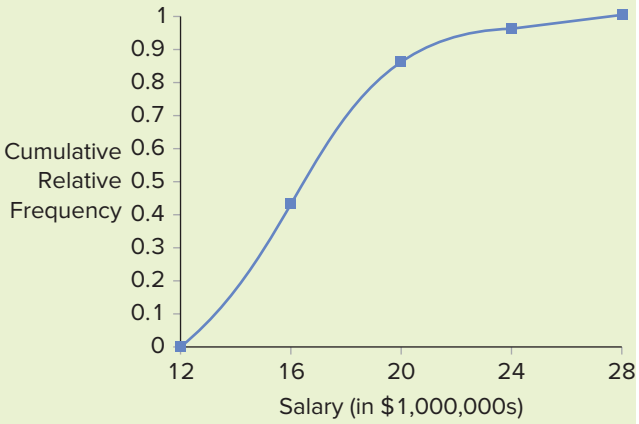


a. Is the distribution symmetric? If not, is it positively or negatively skewed?

b. How many NBA players earned between $20,000,000 and $24,000,000?

c. Approximately how many NBA players earned between $12,000,000 and $20,000,000?

32. The following ogive summarizes the salary (in $1,000,000s) for the 30 highest-paid players in the National Basketball Association (NBA) for the 2012 season (www.nba.com, data retrieved March 2012).



a. Approximate the percentage of salaries that were less than $18,000,000.

b. Approximate the number of salaries that were more than $14,000,000.

33. FILE ***Math_SAT.*** The following table lists a portion of the average math SAT scores for each state for the year 2009.

State	SAT
Alabama	552
Alaska	516
⋮	⋮
Wyoming	568

Source: www.collegeboard.com.

a. Construct the frequency distribution and graph the frequency histogram using classes of 450 to 500, 501 to 550, etc. Comment on the shape of the distribution. How many of the states had scores between 551 and 600?
b. Construct the relative frequency, the cumulative frequency, and the cumulative relative frequency distributions.
c. How many of the states had average math SAT scores of 550 or less?
d. What percent of the states had average math SAT scores between 551 and 600? What percent of the states had average math SAT scores of 550 or less?

34. FILE ***Census.*** The accompanying table shows a portion of median house values (in $) for the 50 states as reported by the U.S. Census Bureau in 2010.

State	House Value
Alabama	117600
Alaska	229100
⋮	⋮
Wyoming	174000

a. Construct the frequency distribution and graph the frequency histogram for the median house values. Use six classes with upper limits of $100,000, $200,000, etc.
b. Is the distribution symmetric? If not, is it positively or negatively skewed?
c. Which class interval had the highest frequency?
d. What percentage of the states had median house values between $300,000 and $400,000?
e. How many of the states had median house values less than $300,000?

35. FILE ***Gas_Prices.*** The accompanying table shows a portion of the average price (in $) for a gallon of gas for the 50 states during April 2012.

State	Price
Alabama	4.36
Alaska	3.79
⋮	⋮
Wyoming	3.63

Source: www.AAA.com, data retrieved April 16, 2012.

a. Construct the frequency distribution and graph the frequency histogram for the average gas price. Use six classes with upper limits of $3.70, $3.90, etc.
b. Is the distribution symmetric? If not, is it positively or negatively skewed?
c. Which class interval had the highest frequency?
d. Graph the ogive. Approximate the percentage of states that had an average gas price of $3.90 or less. Approximate the number of states that had an average gas price greater than $3.90.

36. FILE ***DJIA_2012.*** For the first three months of 2012, the stock market put up its best first-quarter performance in over a decade (Money.cnn.com, April 9, 2012). The accompanying table shows a portion of the daily price index for the Dow Jones Industrial Average (DJIA) over this period.

Day	DJIA
January 3, 2012	12397
January 4, 2012	12418
⋮	⋮
March 31, 2012	13212

Source: Finance.yahoo.com, data retrieved April 20, 2012.

a. Construct the frequency distribution and the frequency histogram for the DJIA price index. Use five classes with upper limits of 12,500, 12,750, etc. On how many days during this quarter was the DJIA less than 12,500?
b. Graph the relative frequency polygon. Is the distribution symmetric? If not, is it positively or negatively skewed?
c. Graph the ogive. Approximate the percentage of days that the DJIA was less than 13,000.

LO **2.5**

Construct and interpret a stem-and-leaf diagram.

2.3 STEM-AND-LEAF DIAGRAMS

John Tukey (1915–2000), a well-known statistician, provided another visual method for displaying quantitative data. A **stem-and-leaf diagram** is often a preliminary step when analyzing a data set. It is useful in that it gives an overall picture of where the data are centered and how the data are dispersed from the center.

GRAPHICAL DISPLAY OF QUANTITATIVE DATA: A STEM-AND-LEAF DIAGRAM

A stem-and-leaf diagram is constructed by separating each value of a data set into two parts: a *stem,* which consists of the leftmost digits, and a *leaf,* which consists of the last digit.

The best way to explain a stem-and-leaf diagram is to show an example.

EXAMPLE 2.6

Table 2.14 shows the ages of the 25 wealthiest people in the world in 2010. Construct and interpret a stem-and-leaf diagram.

FILE
Wealthiest_People

TABLE 2.14 Wealthiest People in the World, 2010

Name	Age	Name	Age
Carlos Slim Helu	70	Li Ka-shing	81
William Gates III	54	Jim Walton	62
Warren Buffet	79	Alice Walton	60
Mukesh Ambani	52	Liliane Bettencourt	87
Lakshmi Mittal	59	S. Robson Walton	66
Lawrence Ellison	65	Prince Alwaleed Alsaud	54
Bernard Arnault	61	David Thomson	52
Eike Batista	53	Michael Otto	66
Amancio Ortega	74	Lee Shau Kee	82
Karl Albrecht	90	Michael Bloomberg	68
Ingvar Kamprad	83	Sergey Brin	36
Christy Walton	55	Charles Koch	74
Stefan Persson	62		

Reprinted by permission of Forbes Media LLC © 2011.

SOLUTION: For each age, we first decide that the number in the tens spot will denote the stem, thus leaving the number in the ones spot as the leaf. We then identify the lowest and highest values in the data set. Sergey Brin is the youngest member of this group at 36 years of age (stem: 3, leaf: 6) and Karl Albrecht is the oldest at 90 years of age (stem: 9, leaf: 0). These values give us the first and last values in the stem. This means the stems will be 3, 4, 5, 6, 7, 8, and 9, as shown in Panel A of Table 2.15.

TABLE 2.15 Constructing a Stem-and-Leaf Diagram for Example 2.6

Panel A		Panel B		Panel C	
Stem	Leaf	Stem	Leaf	Stem	Leaf
3		3	6	3	6
4		4		4	
5		5	4 2 9 3 5 4 2	5	2 2 3 4 4 5 9
6		6	5 1 2 2 0 6 6 8	6	0 1 2 2 5 6 6 8
7	0	7	0 9 4 4	7	0 4 4 9
8		8	3 1 7 2	8	1 2 3 7
9		9	0	9	0

We then begin with the wealthiest man in the world, Carlos Slim Helu, whose age of 70 gives us a stem of 7 and a leaf of 0. We place a 0 in the row corresponding to a stem of 7, as shown in Panel A of Table 2.15. We continue this process with all the other ages and obtain the values in Panel B. Finally, in Panel C we arrange each individual leaf row in ascending order; this is the stem-and-leaf diagram in its final form.

The stem-and-leaf diagram (Panel C) presents the original 25 values in a more organized form. From the diagram we can readily observe that the ages range from

36 to 90. Wealthy individuals in their sixties make up the largest group in the sample with eight members, while those in their fifties place a close second, accounting for seven members. We also note that the distribution is not perfectly symmetric. A stem-and-leaf diagram is similar to a histogram turned on its side with the added benefit of retaining the original values.

EXERCISES 2.3

Mechanics

37. Consider the following data set:

5.4	4.6	3.5	2.8	2.6	5.5	5.5	2.3	3.2	4.2
4.0	3.0	3.6	4.5	4.7	4.2	3.3	3.2	4.2	3.4

Construct a stem-and-leaf diagram. Is the distribution symmetric? Explain.

38. Consider the following data set:

−64	−52	−73	−82	−85	−80	−79	−65	−50	−71
−80	−85	−75	−65	−77	−87	−72	−83	−73	−80

Construct a stem-and-leaf diagram. Is the distribution symmetric? Explain.

Applications

39. A sample of patients arriving at Overbrook Hospital's emergency room recorded the following body temperature readings over the weekend:

100.4	99.6	101.5	99.8	102.1	101.2	102.3	101.2	102.2	102.4
101.6	101.5	99.7	102.0	101.0	102.5	100.5	101.3	101.2	102.2

Construct and interpret a stem-and-leaf diagram.

40. Suppose the following high temperatures were recorded for major cities in the contiguous United States for a day in July.

84	92	96	91	96	94	93	82	81	76
90	95	84	90	84	98	94	90	83	78
88	96	106	78	92	98	91	84	80	94
94	93	107	87	77	99	94	73	74	92

Construct and interpret a stem-and-leaf diagram.

41. A police officer is concerned with excessive speeds on a portion of Interstate 90 with a posted speed limit of 65 miles per hour. Using his radar gun, he records the following speeds for 25 cars and trucks:

66	72	73	82	80	81	79	65	70	71
80	75	75	65	67	67	72	73	73	80
81	78	71	70	70					

Construct a stem-and-leaf diagram. Are the officer's concerns warranted?

42. Spain was the winner of the 2010 World Cup, beating the Netherlands by a score of 1–0. The ages of the players from both teams were as follows:

Spain									
29	25	23	30	32	25	29	30	26	29
21	28	24	21	27	22	25	21	23	24
Netherlands									
27	22	26	30	35	33	29	25	27	25
35	27	27	26	23	25	23	24	26	39

Construct a stem-and-leaf diagram for each country. Comment on similarities and differences between the two data sets.

LO **2.6**

Construct and interpret a scatterplot.

2.4 SCATTERPLOTS

All of the tabular and graphical tools presented thus far have focused on describing one variable. However, in many instances we are interested in the relationship between two variables. People in virtually every discipline examine how one variable may systematically influence another variable. Consider, for instance, how

- Incomes vary with education.
- Sales vary with advertising expenditures.

- Stock prices vary with corporate profits.
- Crop yields vary with the use of fertilizer.
- Cholesterol levels vary with dietary intake.
- Price varies with reliability.

When examining the relationship between two quantitative variables, a **scatterplot** often proves to be a powerful first step in any analysis.

GRAPHICAL DISPLAY OF TWO QUANTITATIVE VARIABLES: A SCATTERPLOT

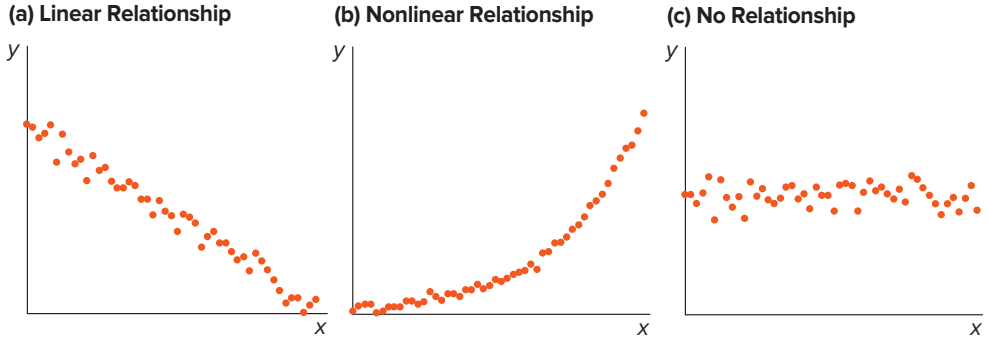
A scatterplot is a graphical tool that helps in determining whether or not two quantitative variables are related in some systematic way. Each point in the diagram represents a pair of observed values of the two variables.

When constructing a scatterplot, we generally refer to one of the variables as x and represent it on the horizontal axis and the other variable as y and represent it on the vertical axis. We then plot each pairing: (x_1, y_1), (x_2, y_2), and so on. Once the data are plotted, the graph may reveal that

- A linear relationship exists between the two variables;
- A nonlinear relationship exists between the two variables; or
- No relationship exists between the two variables.

For example, Figure 2.13(a) shows points on a scatterplot clustered together along a line with a negative slope; we infer that the two variables have a negative linear relationship. Figure 2.13(b) depicts a positive nonlinear relationship; as x increases, y tends to increase at an increasing rate. The points in Figure 2.13(c) are scattered with no apparent pattern; thus, there is no relationship between the two variables.

FIGURE 2.13 Scatterplots depicting relationships between two variables



In order to illustrate a scatterplot, consider the following example.

EXAMPLE 2.7

A social scientist wants to analyze the relationship between educational attainment and income. He collects the data shown in Table 2.16, where Education and Income refer to an individual's years of higher education and annual income (in $1,000s), respectively. Construct and interpret a scatterplot.

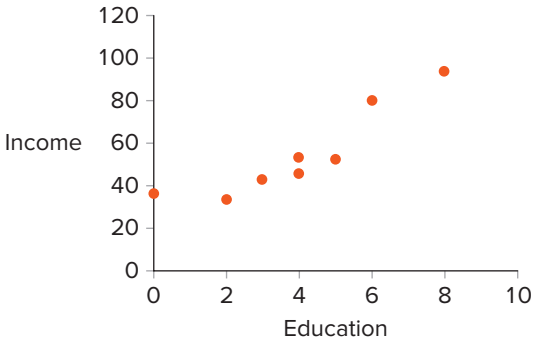
TABLE 2.16 Education and Income for Eight Individuals

Individual	Education	Income
1	3	45
2	4	56
3	6	85
4	2	35
5	5	55
6	4	48
7	8	100
8	0	38

FILE
Edu_Inc

SOLUTION: We let x and y denote Education and Income, respectively. We plot the first individual's pairing as (3, 45), the second individual's pairing as (4, 56), and so on. Figure 2.14 shows the scatterplot.

FIGURE 2.14
Scatterplot of Income versus Education



As expected, we observe a positive relationship between the two variables; that is, when Education increases, Income tends to increase.

Using Excel to Construct a Scatterplot

In order to demonstrate the construction of a scatterplot, we replicate Figure 2.14.

FILE
Edu_Inc

A. Open ***Edu_Inc***.

B. Simultaneously select the values in the Education and Income columns, and then choose **Insert > Scatter (X, Y) or Bubble Chart**. From the options, choose the graph at the top left. (If you are having trouble finding this option after selecting **Insert**, look for the graph with data points above **Charts**.)

C. Formatting (regarding axis titles, gridlines, etc.) can be done by selecting **Format > Add Chart Element** from the menu.

EXERCISES 2.4

Mechanics

43. Construct a scatterplot with the following data. Describe the relationship between x and y.

x	3	7	12	5	6
y	22	10	5	14	12

44. Construct a scatterplot with the following data. Does a linear relationship exist between x and y?

x	10	4	6	3	7
y	3	2	6	6	4

45. Construct a scatterplot with the following data. Describe the relationship between x and y.

x	1	2	3	4	5	6	7	8
y	22	20	18	10	5	4	3	2

Applications

46. A statistics instructor wants to examine whether a relationship exists between the hours a student spends studying for the final exam (Hours) and a student's grade on the final exam (Grade). She takes a sample of eight students.

Hours	8	2	3	8	10	15	25	5
Grade	75	47	50	80	85	88	93	55

Construct a scatterplot. What conclusions can you draw from the scatterplot?

47. A study offers evidence that the more weight a woman gains during pregnancy, the higher the risk of having a high-birth-weight baby, defined as at least 8 pounds, 13 ounces, or 4 kilograms (*The Wall Street Journal,* August 5, 2010). High-birth-weight babies are more likely to be obese in adulthood. The weight gain (in kilograms) of eight mothers and the birth weight of their newborns (in kilograms) are recorded in the accompanying table.

Mother's Weight Gain	Newborn's Birth Weight
18	4.0
7	2.5
8	3.0
22	4.5
21	4.0
9	3.5
8	3.0
10	3.5

Construct a scatterplot. Do the results support the findings of the study?

48. In order to diversify risk, investors are often encouraged to invest in assets whose returns have either a negative relationship or no relationship. The annual return data (in %) on two assets is shown in the accompanying table.

Return A	Return B
−20	8
−5	5
18	−1
15	−2
−12	2

Construct a scatterplot. In order to diversify risk, would the investor be wise to include both of these assets in her portfolio? Explain.

49. In an attempt to determine whether a relationship exists between the price of a home (in $1,000s) and the number of days it takes to sell the home, a real estate agent collects data on the recent sales of eight homes.

Price	Days to Sell Home
265	136
225	125
160	120
325	140
430	145
515	150
180	122
423	145

Construct a scatterplot. What can the realtor conclude?

WRITING WITH STATISTICS

The tabular and graphical tools introduced in this chapter are the starting point for most studies and reports that involve statistics. They can help you organize data so you can see patterns and trends in the data, which can then be analyzed by the methods described in later chapters of this text. In this section, we present an example of using tabular and graphical methods in a sample report. Each of the remaining chapters contains a sample report incorporating the concepts developed in that respective chapter.

Camilla Walford is a newly hired journalist for a national newspaper. One of her first tasks is to analyze gas prices in the United States during the week of the Fourth of July holiday.



©Rubberball/Getty Images

She collects average gas prices (in $ per gallon) for the 48 contiguous states and the District of Columbia (DC), a portion of which is shown in Table 2.17.

FILE
Gas_Prices_2010

TABLE 2.17 U.S. Gas Prices, July 2, 2010

State	Price
Alabama	2.59
Arkansas	2.60
⋮	⋮
Wyoming	2.77

Source: *AAA's Daily Fuel Gauge Report*, July 2, 2010.

Camilla wants to use the sample information to

1. Construct frequency distributions to summarize the data.
2. Make summary statements concerning gas prices.
3. Convey the information from the distributions into graphical form.

Sample Report—Gas Prices across the United States

Historically, in the United States, many people choose to take some time off during the Fourth of July holiday period and travel to the beach, the lake, or the mountains. The roads tend to be heavily traveled, making the cost of gas a concern. The following report provides an analysis of gas prices across the nation over this holiday period.

The analysis focuses on the average gas price for the 48 contiguous states and the District of Columbia (henceforth, referenced as 49 states for ease of exposition). The range of gas prices is from a low of $2.52 per gallon (South Carolina) to a high of $3.15 per gallon (California). To find out how gas prices are distributed between these extremes, the data have been organized into several frequency distributions as shown in Table 2.A. The frequency distribution shows that 17 of the 49 states have an average gas price between $2.70 and $2.80 per gallon; or comparably, 35% of the states have an average price in this range as shown by the relative frequency distribution. The cumulative frequency distribution indicates that 35 states have an average price less than $2.80 per gallon. Finally, the cumulative relative frequency distribution shows that the average price in 72% of the states (approximately three-quarters of the sample) is less than $2.80 per gallon.

TABLE 2.A Frequency Distributions for Gas Prices in the United States, July 2, 2010

Average Price ($ per gallon)	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
2.50 up to 2.60	5	0.10	5	0.10
2.60 up to 2.70	13	0.27	18	0.37
2.70 up to 2.80	17	0.35	35	0.72
2.80 up to 2.90	8	0.16	43	0.88
2.90 up to 3.00	4	0.08	47	0.96
3.00 up to 3.10	1	0.02	48	0.98
3.10 up to 3.20	1	0.02	49	1.00
	Sample Size = 49			

Figure 2.A shows a histogram for gas prices. This graph reinforces the fact that the average price of gas nationwide is between $2.50 and $3.20 per gallon. Moreover, gas prices are positively skewed since the distribution runs off to the right; only two states (California and Washington) have gas prices that are more than $3.00 per gallon.

Another useful visual representation of the data is an ogive, shown in Figure 2.B. The ogive is useful for approximating the "middle" price. If we draw a horizontal line to the ogive at the 0.5 relative frequency mark, it intersects the plot at a point corresponding on the horizontal axis to a "middle price" of approximately $2.75. This indicates that gas stations in approximately half of the states charged below this price and half charged above it.

FIGURE 2.A Histogram of average gas prices nationwide

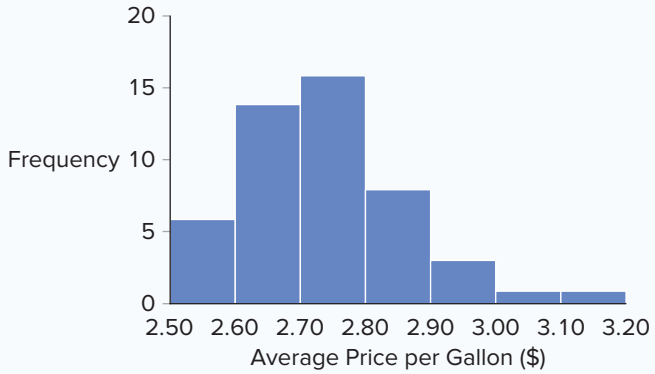
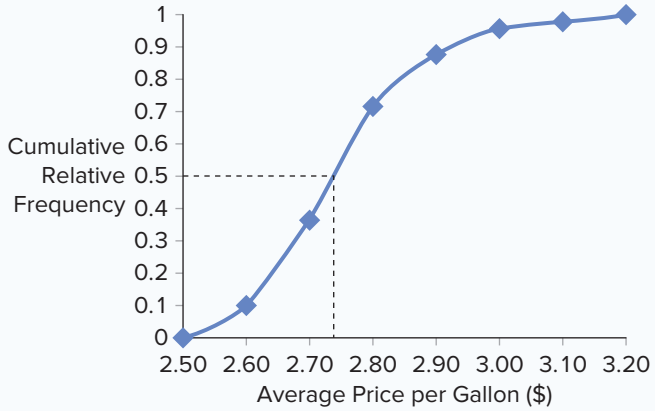


FIGURE 2.B Ogive of average gas prices nationwide



CONCEPTUAL REVIEW

LO 2.1 Summarize qualitative data by constructing a frequency distribution.

For qualitative data, a **frequency distribution** groups data into categories and records the number of observations that fall into each category. A **relative frequency distribution** shows the proportion (or the fraction) of observations in each category.

LO 2.2 Construct and interpret a pie chart and a bar chart.

Graphically, we can show a frequency distribution for qualitative data by constructing a **pie chart** or a **bar chart**. A pie chart is a segmented circle that clearly portrays the categories of some qualitative variable. A bar chart depicts the frequency or the relative frequency for each category of the qualitative variable as a series of horizontal or vertical bars, the lengths of which are proportional to the values that are to be depicted.

LO 2.3 Summarize quantitative data by constructing a frequency distribution.

For quantitative data, a **frequency distribution** groups data into intervals called classes, and records the number of observations that falls within each class. A **cumulative frequency distribution** records the number of observations that falls below the upper limit of each class. A **relative frequency distribution** identifies the proportion (or the fraction) of observations that falls within each class. A **cumulative relative frequency distribution** shows the proportion (or the fraction) of observations that falls below the upper limit of each class.

LO 2.4 **Construct and interpret a histogram, a polygon, and an ogive.**

A **histogram** and a **polygon** are graphical representations of a frequency or a relative frequency distribution for quantitative data. An inspection of these graphs reveals where most of the observations tend to cluster, as well as the general shape and spread of the data. An **ogive** is a graphical representation of a cumulative frequency or cumulative relative frequency distribution.

LO 2.5 **Construct and interpret a stem-and-leaf diagram.**

A **stem-and-leaf diagram** is another visual method of displaying quantitative data. It is constructed by separating each value of a data set into a *stem,* which consists of the left-most digits, and a *leaf,* which consists of the last digit. Like a histogram and a polygon, a stem-and-leaf diagram gives an overall picture of where the data are centered and how the data are dispersed from the center.

LO 2.6 **Construct and interpret a scatterplot.**

A **scatterplot** is a graphical tool that helps in determining whether or not two quantitative variables are related in some systematic way. Each point in the diagram represents a pair of observed values of the two variables.

ADDITIONAL EXERCISES AND CASE STUDIES

Exercises

50. A 2003 survey by the Centers for Disease Control and Prevention concluded that smoking is forbidden in nearly 75% of U.S. households (*The Boston Globe,* May 25, 2007). The survey gathered responses from at least 900 households in each state. When residents of Utah were asked whether or not smoking was allowed in their households, a representative sample of responses was as follows:

No	No	No	No	No	No	Yes	No	No	No
No	Yes	No	No	No	No	No	No	No	No

When a similar survey was taken in Kentucky, a representative sample of responses was as follows:

No	No	Yes	No	Yes	No	Yes	Yes	No	No
No	Yes	Yes	No	Yes	No	No	Yes	Yes	No

a. Construct a relative frequency distribution that summarizes the responses of residents from Utah and Kentucky. Comment on the results.
b. Construct a bar chart that summarizes the results for each state.

51. Patrons at a local restaurant were asked to rate their recent experience at the restaurant with respect to its advertised atmosphere of upbeat, comfortable, and clean. Possible responses included Outstanding, Good, OK, and Horrible. The following table shows the responses of 28 patrons:

Horrible	OK	Horrible	Horrible
OK	OK	Horrible	Horrible
Horrible	OK	Horrible	Good
Horrible	Good	Good	Good
Horrible	OK	Horrible	OK
Good	Good	Horrible	Good
Horrible	OK	Horrible	Good

a. Construct a relative frequency distribution that summarizes the responses of the patrons. Briefly summarize your findings. What recommendations would you make to the owner of the restaurant?
b. Construct a pie chart and a bar chart for these data.

52. A survey conducted by CBS News asked parents about the professions they would want their children to pursue. Parents' preferences (in %) are summarized in the following table.

Profession	Parents' Preference
Doctor, banker, lawyer, or president	65
Internet mogul	13
Humanitarian-aid worker	6
Athlete	9
Movie star, rock star	2
Other	5

Source: *Vanity Fair*, December 2009.

a. Construct a pie chart and a bar chart for these data.
b. How many parents wanted their children to become athletes if the results were based on 550 responses?

53. The one-year return (in %) for 24 mutual funds is as follows:

−14.5	−5.0	−3.7	2.5	−7.9	−11.2
4.8	−16.8	9.0	6.5	8.2	5.3
−12.2	15.9	18.2	25.4	3.4	−1.4
5.5	−4.2	−0.5	6.0	−2.4	10.5

a. Construct the frequency distribution using classes of −20 up to −10, −10 up to 0, etc.
b. Construct the relative frequency, the cumulative frequency, and the cumulative relative frequency distributions.
c. How many of the funds had returns of at least 0% but less than 10%? How many of the funds had returns of 10% or more?
d. What percentage of the funds had returns of at least 10% but less than 20%? What percentage of the funds had returns less than 20%?

54. *The Statistical Abstract of the United States, 2010* provided the following frequency distribution of the number of people (in 1,000s) who live below the poverty level by region.

Region	Number of People
Northeast	6,166
Midwest	7,237
South	15,501
West	8,372

a. Construct the relative frequency distribution. What percentage of people who live below the poverty level live in the Midwest?
b. Construct a pie chart and a bar chart for these data.

55. *Money* magazine (January 2007) reported that an average of 77 million adults in the United States make financial resolutions at the beginning of a new year. Consider the following frequency distribution, which reports the top financial resolutions of 1,026 Americans (MONEY/ICR poll conducted November 8–12, 2006).

Financial Resolution	Frequency
Saving more	328
Paying down debt	257
Making more income	154
Spending less	133
Investing more	103
Saving for a large purchase	41
Don't know	10

a. Construct the relative frequency distribution. What percentage of the respondents indicated that paying down debt was their top financial resolution?
b. Construct the bar chart.

56. A recent poll of 3,057 individuals asked: "What's the longest vacation you plan to take this summer?" The following relative frequency distribution summarizes the results.

Response	Relative Frequency
A few days	0.21
A few long weekends	0.18
One week	0.36
Two weeks	0.25

a. Construct the frequency distribution. How many people are going to take a one-week vacation this summer?
b. Construct the pie chart.

57. A survey conducted by CBS News asked 1,026 respondents: "What would you do with an unexpected tax refund?" The responses (in %) are summarized in the following table;

Response	Percent Frequency
Pay off debts	47
Put it in the bank	30
Spend it	11
I never get a refund	10
Other	2

Source: CBS News Archives.

a. Construct the bar chart.
b. How many people will spend the tax refund?

58. The following table reports the number of people residing in regions in the U.S. as well as the number

of people living below the poverty level in these regions for the year 2013. (All numbers are in 1,000s.)

Region	Total	Below Poverty Level
Northeast	55,478	7,046
Midwest	66,785	8,590
South	116,961	18,870
West	73,742	10,812

Source: www.census.gov/hhes/www/poverty/data/incpovhlth/2013/table3.pdf, data retrieved March 23, 2015.

a. Graph and interpret the pie chart that summarizes the proportion of people who reside in each region.
b. Graph and interpret the pie chart that summarizes the proportion of people living below the poverty level in each region. Is this pie chart consistent with the one you constructed in part (a); that is, in those regions that are relatively less populated, is the proportion of people living below the poverty level less?

59. The manager at a water park constructed the following frequency distribution to summarize attendance in July and August.

Attendance	Frequency
1,000 up to 1,250	5
1,250 up to 1,500	6
1,500 up to 1,750	10
1,750 up to 2,000	20
2,000 up to 2,250	15
2,250 up to 2,500	4

a. Construct the relative frequency, the cumulative frequency, and the cumulative relative frequency distributions.
b. What is the most likely attendance range? How many times was attendance less than 2,000 people?
c. What percentage of the time was attendance at least 1,750 but less than 2,000 people? What percentage of the time was attendance less than 1,750 people? What percentage of the time was attendance 1,750 or more?
d. Construct the relative frequency histogram. Comment on the shape of the distribution.

60. *The Wall Street Journal* (August 28, 2006) asked its readers: "Ideally, how many days a week, if any, would you work from home?" The following relative frequency distribution summarizes the responses from 3,478 readers.

Days Working from Home	Relative Frequency
0	0.12
1	0.18
2	0.30
3	0.15
4	0.07
5	0.19

Construct the pie chart and the bar chart to summarize the data.

61. A researcher conducts a mileage economy test involving 80 cars. The frequency distribution describing average miles per gallon (mpg) appears in the following table.

Average mpg	Frequency
15 up to 20	15
20 up to 25	30
25 up to 30	15
30 up to 35	10
35 up to 40	7
40 up to 45	3

a. Construct the relative frequency, the cumulative frequency, and the cumulative relative frequency distributions.
b. How many of the cars got less than 30 mpg? What percentage of the cars got at least 20 but less than 25 mpg? What percentage of the cars got less than 35 mpg? What percent got 35 mpg or more?
c. Construct the relative frequency histogram. Comment on the shape of the distribution.

62. FILE ***Wealthiest_Americans.*** The accompanying table lists a portion of the ages and net worth (in $ billions) of the wealthiest people in America.

Name	Age	Net Worth
William Gates III	53	50.0
Warren Buffet	79	40.0
⋮	⋮	⋮
Philip Knight	71	9.5

Source: *Forbes*, Special Report, September 2009.

a. What percentage of the wealthiest people in America had net worth more than $20 billion?
b. What percentage of the wealthiest people in America had net worth between $10 billion and $20 billion?
c. Construct a stem-and-leaf diagram on age. Comment on the shape of the distribution.

63. FILE ***DOW_PEG.*** The price-to-earnings growth ratio, or PEG ratio, is the market's valuation of a company relative to its earnings prospects. A PEG ratio of 1 indicates that the stock's price is in line with growth expectations. A PEG ratio less than 1 suggests that the stock of the company is undervalued (typical of value stocks), whereas a PEG ratio greater than 1 suggests the stock is overvalued (typical of growth stocks). The accompanying table shows a portion of PEG ratios of companies listed on the Dow Jones Industrial Average.

Company	PEG Ratio
3M (MMM)	1.4
Alcoa (AA)	0.9
⋮	⋮
Walt Disney (DIS)	1.2

Source: www.finance.yahoo, data retrieved April 13, 2011.

Construct the stem-and-leaf diagram on the PEG ratio. Interpret your findings.

64. The following table lists the sale price (in $1,000s) and house type of 20 houses that recently sold in New Jersey.

Price	Type	Price	Type
305	Ranch	568	Colonial
450	Colonial	385	Other
389	Contemporary	310	Contemporary
525	Other	450	Colonial
300	Ranch	400	Other
330	Contemporary	359	Ranch
355	Contemporary	379	Ranch
405	Colonial	509	Colonial
365	Ranch	435	Colonial
415	Ranch	510	Other

a. Construct a frequency distribution for types of houses. Interpret the results.
b. Construct a frequency distribution for house prices. Use six classes, starting with 300, each with a width of 50. Interpret the results.

65. A manager of a local retail store analyzes the relationship between Advertising (in $100s) and Sales (in $1,000s) by reviewing the store's data for the previous six months. Construct a scatterplot and comment on whether or not a relationship exists.

Advertising	Sales
20	15
25	18
30	20
22	16
27	19
26	20

66. The following table lists the National Basketball Association's (NBA's) leading scorers, their average minutes per game (MPG), and their average points per game (PPG) for 2008:

Player	MPG	PPG
D. Wade	38.6	30.2
L. James	37.7	28.4
K. Bryant	36.1	26.8
D. Nowitzki	37.3	25.9
D. Granger	36.2	25.8
K. Durant	39.0	25.3
C. Paul	38.5	22.8
C. Anthony	34.5	22.8
C. Bosh	38.0	22.7
B. Roy	37.2	22.6

Source: www.espn.com.

Construct and interpret a scatterplot of PPG against MPG. Does a relationship exist between the two variables?

CASE STUDIES

CASE STUDY 2.1 There are six broad sectors that comprise the Dow Jones Industrial Average (DJIA). The following table shows a portion of the 30 companies that comprise the DJIA and each company's sector.

FILE
DJIA_Sector

Data for Case Study 2.1 Companies and Sectors of the DJIA

Company	Sector
3M (MMM)	Manufacturing
American Express (AXP)	Finance
⋮	⋮
Walmart (WMT)	Consumer

Source: www.money.cnn.com/data/dow30/, information retrieved March 21, 2015.

In a report, use the sample information to

1. Construct the frequency distribution and the relative frequency distribution for the sectors that comprise the DJIA. Use pie charts for data visualization.
2. Discuss how the various sectors are represented in the DJIA.

CASE STUDY 2.2 When reviewing the overall strength of a particular firm, financial analysts typically examine the net profit margin. This statistic is generally calculated as the ratio of a firm's net profit after taxes (net income) to its revenue, expressed as a percentage. For example, a 20% net profit margin means that a firm has a net income of $0.20 for each dollar of sales. A net profit margin can even be negative if the firm has a negative net income. In general, the higher the net profit margin, the more effective the firm is at converting revenue into actual profit. The net profit margin serves as a good way of comparing firms in the same industry, since such firms generally are subject to the same business conditions. However, financial analysts also use the net profit margin to compare firms in different industries in order to gauge which firms are relatively more profitable. The accompanying table shows a portion of net profit margins (in %) for a sample of clothing retailers.

FILE
Net_Profit_Margins

Data for Case Study 2.2 Net Profit Margin for Clothing Retailers

Firm	Net Profit Margin
Abercrombie & Fitch	1.58
Aéropostale	10.64
⋮	⋮
Wet Seal	16.15

Source: www.finance.yahoo.com, data retrieved July 2010.

In a report, use the sample information to

1. Provide a brief definition of net profit margin and explain why it is an important statistic.
2. Construct appropriate tables (frequency distribution, relative frequency distribution, etc.) and graphs that summarize the clothing industry's net profit margin. Use −5, 0, 5, and so on, for the upper limits of the classes for the distributions.
3. Discuss where the data tend to cluster and how the data are spread from the lowest value to the highest value.
4. Comment on the net profit margin of the clothing industry, as compared to the beverage industry's net profit margin of approximately 10.9% (Source: biz.yahoo, July 2010).

CASE STUDY 2.3 The following table lists a portion of U.S. life expectancy (in years) for the 50 states.

Data for Case Study 2.3 Life Expectancy by State, 2010–2011

FILE
Life_Expectancy

Rank	State	Life Expectancy
1	Hawaii	81.5
2	Minnesota	80.9
⋮	⋮	⋮
50	Mississippi	74.8

Source: en.wikipedia.org/wiki/List_of_U.S._states_by_life_expectancy, data retrieved April 25, 2012.

In a report, use the sample information to

1. Construct appropriate tables (frequency distribution, relative frequency distribution, etc.) and graphs to summarize life expectancy in the United States. Use 75, 76.5, 78, and so on, for the upper limits of the classes for the distributions.
2. Discuss where the data tend to cluster and how the data are spread from the lowest value to the highest value.
3. Comment on the shape of the distribution.

APPENDIX 2.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, JMP, and R. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands. When importing data into R, use the menu-driven option File > Import Dataset > From Excel.

Minitab

Pie Chart

FILE
Marital_Status

A. (Replicating Figure 2.1) From the menu, choose **Graph > Pie Chart**. Select **Chart values from a table**, select Marital Status as the **Categorical variable**, and 1960 and 2010 as the **Summary variables**.

B. Choose **Labels**. Select **Titles/Footnotes** and enter Marital Status, 1960 versus 2010. Then select **Slice Labels** and select **Category name** and **Percent**.

C. Choose **Multiple Graphs**, and then select **On the same graph**.

Bar Chart

FILE
Marital_Status

A. (Replicating Figure 2.2) From the menu, choose **Graph > Bar Chart**. From **Bars Represent** select **Values from a Table**, and from **Two-way Table** select **Cluster**.

B. In the *Bar Chart—Two-Way Table—Cluster* dialog box, select 1960 and 2010 as **Graph variables**. Select Marital Status as **Row labels**. Under **Table Arrangement**, choose **Rows are outermost categories and columns are innermost**.

Histogram

From Raw Data:

FILE *MV_Houses*

A. (Replicating Figure 2.5) From the menu, choose **Graph > Histogram > Simple**. Click **OK**.

B. Select House Price as **Graph Variables**. Click **OK**.

C. Double-click *x*-axis and select **Edit Scale**. Under **Major Tick Positions**, choose **Position of Ticks** and enter 300 400 500 600 700 800. Under **Scale Range**, deselect **Auto** for *Minimum* and enter 300. Then deselect **Auto** for *Maximum* and enter 800. Select the **Binning** tab. Under **Interval Type**, select **Cutpoint**. Under **Interval Definition**, select **Midpoint/Cutpoint Definitions** and enter 300 400 500 600 700 800.

From a Frequency Distribution:

FILE *MV_Frequency*

A. (Replicating Figure 2.5) From the menu, choose **Graph > Bar Chart**. From **Bars Represent** select **A function of a variable**, and from **One Y** select **Simple**. Click **OK**.

B. Under **Function** select **Sum**. Under **Graph variables** select **Frequency**, and under **Categorical variable** select **Class (in $1,000s)**.

C. Double-click *x*-axis. Under **Space Between Scale Categories**, deselect **Gap between Cluster** and enter 0.

Polygon

FILE *Polygon*

A. (Replicating Figure 2.8) From the menu, choose **Graph > Scatterplot > With Connect Line**.

B. Under **Y variables** select y, and under **X variables** select x.

Ogive

FILE *Ogive*

A. (Replicating Figure 2.9) From the menu, choose **Graph > Scatterplot > With Connect Line**.

B. Under **Y variables** select y, and under **X variables** select x.

Scatterplot

FILE *Edu_Inc*

A. (Replicating Figure 2.14) From the menu, choose **Graph > Scatterplot > Simple**.

B. Under **Y variables** select Income, and under **X variables** select Education.

SPSS

Pie Chart

FILE *Marital_Status*

A. (Replicating Figure 2.1) From the menu, choose **Graphs > Legacy Dialogs > Pie**. Under **Data in Chart Are**, select **Values of individual cases**. Click **Define**.

B. Under **Slices Represent**, select 1960. Under **Slices Labels**, select **Variable**, then select Marital Status.

C. Double-click on the graph to open **Chart Editor**, and then choose **Elements > Show Data Labels**. In the **Properties** dialog box, under **Displayed** select Percent and Marital Status.

Bar Chart

FILE *Marital_Status*

A. (Replicating Figure 2.2) From the menu, choose **Graphs > Legacy Dialogs > Bar**. Choose **Clustered**. Under **Data in Chart Are**, select **Values of individual cases**. Click **Define**.

B. Under **Bars Represent**, select 1960 and 2010. Under **Category Labels**, select **Variable**, then select Marital Status.

Histogram

FILE *MV_Houses*

A. (Replicating Figure 2.5) From the menu, choose **Graphs > Legacy Dialogs > Histogram**. Under **Variable**, select HousePrice.

B. In the Output window, double-click on Frequency (*y*-axis title), choose the **Scale** tab, and under **Range**, enter 0 as **Minimum**, 15 as **Maximum**, and 5 as **Major Increment**. Then click **Apply**.

C. Double-click on the bars. Choose the **Binning** tab, and under **X Axis**, select **Custom** and **Interval width**, and enter 100 for the interval width. Then click **Apply**.

Polygon

FILE *Polygon*

A. (Replicating Figure 2.8) From the menu, choose **Graphs > Legacy Dialogs > Scatter/Dot**. Choose **Simple Scatter** and then click **Define**.

B. Under **Y Axis** select y, and under **X Axis** select x. Click **OK**.

C. In the Output window, double-click on the graph to open the **Chart Editor**, then from the menu choose **Elements > Interpolation Line**.

Ogive

FILE *Ogive*

A. (Replicating Figure 2.9) From the menu, choose **Graphs > Legacy Dialogs > Scatter/Dot**. Choose **Simple Scatter**. Then click **Define**.

B. Under **Y Axis**, select y, and under **X Axis**, select x. Click **OK**.

C. In the Output window, double-click on the graph to open the **Chart Editor**, then from the menu choose **Elements > Interpolation Line**. Then click **Apply**. From the menu choose **Edit > Select X Axis**. Choose the **Scale** tab, and under **Range**, enter 300 as **Minimum**, 800 as **Maximum**, 100 as **Major Increment**, and 300 as **Origin**. Then click **Apply**.

Scatterplot

FILE *Edu_Inc*

A. (Replicating Figure 2.14) From the menu, choose **Graphs > Legacy Dialogs > Scatter/Dot**. Choose **Simple Scatter**. Then click **Define**.

B. Under **Y Axis**, select Income, and under **X Axis**, select Education. Click **OK**.

C. In the Output window, double-click on the graph to open the **Chart Editor**. From the menu, choose **Edit > Select Y Axis**. Under **Range**, enter 0 as **Minimum**, 120 as **Maximum**, and 20 as **Major Increment**. Then click **Apply**. From the menu choose **Edit > Select X Axis**. Choose the **Scale** tab, and under **Range**, enter 0 as **Minimum**, 10 as **Maximum**, and 2 as **Major Increment**. Then click **Apply**.

JMP

Pie Chart

FILE *Marital_Status*

(Replicating Figure 2.1) From the menu, choose **Graph > Chart**. Under **Select Columns**, select Marital Status, and then under **Cast Selected Columns into Roles**, select **Categories, X, Levels**. Under **Select Columns**, select 1960 and 2010, and then select **Statistics > % of Total**. Under **Options**, choose **Pie Chart**. In order to add percentages to the pie chart, click the red arrow next to **Chart > Label Options > Label by Percent of Total Values**.

Bar Chart

FILE *Marital_Status*

(Replicating Figure 2.2) From the menu, choose **Graph > Chart**. Under **Select Columns**, select **Marital Status**, and then under **Cast Selected Columns into Roles**, select **Categories, X, Levels**. Under Select Columns, select 1960 and 2010, and select **Statistics > Data**. Under **Options**, select **Overlay** and **Bar Chart**.

Histogram

FILE *MV_Houses*

A. (Replicating Figure 2.5) From the menu, choose **Analyze > Distribution**. Under **Select Columns**, select House Price, then under **Cast Selected Columns into Roles**, select **Y, columns**.

B. Right-click on the *y*-axis and select **Axis Settings**. For **Minimum**, enter 300; for **Maximum** enter 800; and for **Increment**, enter 100. (JMP automatically produces a histogram in a vertical layout. In order to change the layout to horizontal, click the red arrow next to **House Price > Display Options > Horizontal Layout**.)

Polygon

FILE *Polygon*

A. (Replicating Figure 2.8) From the menu, choose **Graph > Overlay Plot**. Under **Select Columns**, select x, and then under **Cast Selected Columns into Roles**, select **X**. Under **Select Columns**, select y, and then under **Cast Selected Columns into Roles**, select **Y**.

B. Click on the red triangle next to the title **Overlay Plot**. Select **Y Options > Connect Points**.

Ogive

FILE *Ogive*

A. (Replicating Figure 2.9) From the menu, choose **Graph > Overlay Plot**. Under **Select Columns**, select x, and then under **Cast Selected Columns into Roles**, select **X**. Under **Select Columns**, select y, and then under **Cast Selected Columns into Roles**, select **Y**.

B. Click on the red triangle next to the title **Overlay Plot.** Select **Y Options > Connect Points**.

Scatterplot

FILE *Edu_Inc*

(Replicating Figure 2.14) From the menu, choose **Graph > Overlay Plot**. Under **Select Columns**, select Education, and then under **Cast Selected Columns into Roles**, select **X**. Under **Select Columns**, select Income, and then under **Cast Selected Columns into Roles**, select **Y**.

R

Pie Chart

FILE *Marital_Status*

(Replicating Figure 2.1) Use the **pie** function. For options within the function, use *labels* to indicate the names for each category and *main* to designate a title. Enter:

```
> pie(Marital_Status$`1960', labels = Marital_Status$`Marital Status',
  main = "Marital Status, 1960")
```

Histogram

FILE *MV_Houses*

(Replicating Figure 2.5) Use the **hist** function. For options within the function, use *breaks* to denote the number of distinct intervals, *main* to designate a title, and *xlab* to label the *x*-axis. Enter:

```
> hist(MV_Houses$`House Price', breaks = 5, main = "Histogram",
  xlab = "House Prices (in $1,000s)")
```

Polygon

A. (Replicating Figure 2.8) Use the **plot** function. For options within the function, use *ylab* and *xlab* to label the *y*-axis and the *x*-axis, respectively. Enter:

FILE
Polygon

```
> plot(Polygon$`y' ~  Polygon$`x', ylab="Relative Frequency",
  xlab="House Prices (in $1,000s)")
```

B. Add lines to the scatterplot using the **lines** function. Enter:

```
> lines(Polygon$`y' ~  Polygon$`x')
```

Ogive

A. (Replicating Figure 2.9) Refer to the instructions for the polygon for specifics about the **plot** and the **line** functions. Enter:

FILE
Ogive

```
> plot(Ogive$`y' ~  Ogive$`x', ylab="Cumulative Relative Frequency",
  xlab="House Prices (in $1,000s)")
> lines(Ogive$`y' ~  Ogive$`x')
```

Scatterplot

(Replicating Figure 2.14) Refer to the instructions for the polygon for specifics about the **plot** function. Enter:

FILE
Edu_Inc

```
> plot(Edu_Inc$`Income' ~  Edu_Inc$`Education', ylab = "Income",
  xlab = "Education")
```

3 Numerical Descriptive Measures

Learning Objectives

After reading this chapter you should be able to:

LO **3.1** Calculate and interpret measures of central location.

LO **3.2** Interpret a percentile and a boxplot.

LO **3.3** Calculate and interpret measures of dispersion.

LO **3.4** Explain mean-variance analysis and the Sharpe ratio.

LO **3.5** Apply Chebyshev's theorem, the empirical rule, and *z*-scores.

LO **3.6** Calculate summary measures for grouped data.

LO **3.7** Calculate and interpret measures of association.

In Chapter 2, we used tables and graphs in order to extract meaningful information from data. In this chapter, we focus on numerical descriptive measures. These measures provide precise, objectively determined values that are easy to calculate, interpret, and compare with one another. We first calculate several measures of central location, which attempt to find a typical or central value for the data. In addition to analyzing the center, we examine how the data vary around the center. Measures of dispersion gauge the underlying variability of the data. We use measures of central location and dispersion to introduce some popular applications, including the Sharpe ratio, Chebyshev's theorem, the empirical rule, and the *z*-score. Finally, we discuss measures of association that examine the linear relationship between two variables. These measures assess whether two variables have a positive linear relationship, a negative linear relationship, or no linear relationship.



Introductory Case

Investment Decision

Jacqueline Brennan works as a financial advisor at a large investment firm. She meets with an inexperienced investor who has some questions regarding two approaches to mutual fund investing: growth investing versus value investing. The investor has heard that growth funds invest in companies whose stock prices are expected to grow at a faster rate, relative to the overall stock market, and value funds invest in companies whose stock prices are below their true worth. The investor has also heard that the main component of investment return is through capital appreciation in growth funds and through dividend income in value funds. The investor shows Jacqueline the annual return data for Vanguard's Growth Index mutual fund (henceforth, Growth) and Vanguard's Value Index mutual fund (henceforth, Value). Table 3.1 shows the annual return data for these two mutual funds for the years 2007–2016.

TABLE 3.1 Returns (in percent) for the Growth and the Value Funds

Year	Growth	Value	Year	Growth	Value
2007	12.56	0.09	2012	16.89	15.00
2008	−38.32	−35.97	2013	32.16	32.85
2009	36.29	19.58	2014	13.47	13.05
2010	16.96	14.28	2015	3.17	−1.03
2011	1.71	1.00	2016	5.99	16.75

Source: finance.yahoo.com, data retrieved February 17, 2017.

FILE
Growth_Value

In addition to clarifying the style differences in growth investing versus value investing, Jacqueline will use the above sample information to

1. Calculate and interpret the typical return for these two mutual funds.
2. Calculate and interpret the investment risk for these two mutual funds.
3. Determine which mutual fund provides the greater return relative to risk.

A synopsis of this case is provided at the end of Section 3.4.

LO **3.1**

Calculate and interpret measures of central location.

3.1 MEASURES OF CENTRAL LOCATION

The term *central location* relates to the way quantitative data tend to cluster around some middle or central value. Measures of central location attempt to find a typical or central value that describes the data. Examples include finding a typical value that describes the return on an investment, the number of defects in a production process, the salary of a business graduate, the rental price in a neighborhood, the number of customers at a local convenience store, and so on.

The Mean

The **arithmetic mean** is the primary measure of central location. Generally, we refer to the arithmetic mean as simply the **mean** or the **average**. In order to calculate the mean of a data set, we simply add up the observations and divide by the number of observations in the population or sample.

EXAMPLE 3.1

Let's use the data in Table 3.1 in the introductory case to calculate and interpret the mean return for the Growth mutual fund and the mean return for the Value mutual fund.

SOLUTION: Let's start with the mean return for the Growth mutual fund. We first add all the returns and then divide by the number of returns as follows:

$$\text{Growth mutual fund mean return} = \frac{12.56 + (-38.32) + \cdots + 5.99}{10} = \frac{100.88}{10} = 10.09\%.$$

Similarly, we calculate the mean return for the Value mutual fund as:

$$\text{Value mutual fund mean return} = \frac{0.09 + (-35.97) + \cdots + 16.75}{10} = \frac{75.60}{10} = 7.56\%.$$

Thus, over the 10-year period 2007–2016, the mean return for the Growth mutual fund was greater than the mean return for the Value mutual fund, or equivalently, $10.09\% > 7.56\%$. These means represent typical annual returns resulting from one-year investments. We will see throughout this chapter, however, that we would be ill-advised to invest in a mutual fund solely on the basis of its average return.

All of us have calculated a mean before. What might be new for some of us is the notation used to express the mean as a formula. For instance, when calculating the mean return for the Growth mutual fund, we let $x_1 = 12.56$, $x_2 = -38.32$, and so on, and let n represent the number of observations in the sample. So our calculation for the mean can be written as

$$\text{Mean} = \frac{x_1 + x_2 + \cdots + x_{10}}{n}.$$

The mean of the sample is referred to as $\bar{x}$ (pronounced x-bar). Also, we can denote the numerator of this formula using summation notation, which yields the following

compact formula for the **sample mean**: $\bar{x} = \frac{\Sigma x_i}{n}$. We should also point out that if we had all the return data for this mutual fund, instead of just the data for the past 10 years, then we would have been able to calculate the **population mean** μ as $\mu = \frac{\Sigma x_i}{N}$, where μ is the Greek letter mu (pronounced as "mew") and N is the number of observations in the population.

MEASURE OF CENTRAL LOCATION: THE MEAN

For sample values $x_1, x_2, \ldots, x_n$, the sample mean $\bar{x}$ is computed as

$$\bar{x} = \frac{\Sigma x_i}{n}.$$

For population values $x_1, x_2, \ldots, x_N$, the population mean μ is computed as

$$\mu = \frac{\Sigma x_i}{N}.$$

The calculation method is identical for the sample mean and the population mean except that the sample mean uses n observations and the population mean uses N observations, where $n < N$. We refer to the population mean as a **parameter** and the sample mean as a **statistic**. Since the population mean is generally unknown, we often use the sample mean to estimate the population mean.

The mean is used extensively in statistics. However, it can give a misleading description of the center of the distribution in the presence of extremely small or large values, also referred to as **outliers**.

The mean is the most commonly used measure of central location. However, one weakness of this measure is that it is unduly influenced by outliers.

Example 3.2 highlights the main weakness of the mean.

EXAMPLE 3.2

Seven people work at Acetech, a small technology firm in Seattle. Their salaries (in $) over the past year are listed in Table 3.2. Compute the mean salary for this firm and discuss whether it accurately indicates a typical value.

TABLE 3.2 Salaries of Employees at Acetech

Title	Salary
Administrative Assistant	40,000
Research Assistant	40,000
Computer Programmer	65,000
Senior Research Associate	90,000
Senior Sales Associate	145,000
Chief Financial Officer	150,000
President (and owner)	550,000

SOLUTION: Since the salaries of all employees of Acetech are included in Table 3.2, we calculate the population mean salary as

$$\mu = \frac{\Sigma x_i}{N} = \frac{40{,}000 + 40{,}000 + \cdots + 550{,}000}{7} = 154{,}286.$$

It is true that the mean salary for this firm is $154,286, but this value does not reflect the typical salary at this firm. In fact, six of the seven employees earn less than $154,286. This example highlights the main weakness of the mean — that is, it is sensitive to outliers.

The Median

Since the mean can be affected by outliers, we often also calculate the **median** as a measure of central location. The median is the middle value of a data set. It divides the data in half; an equal number of observations lie above and below the median. Many government publications and other data sources publish both the mean and the median in order to accurately portray a data set's typical value. If the values of the mean and the median differ significantly, then it is likely that the data set contains outliers. For instance, in 2015 the U.S. Census Bureau determined that the median income for American households was $52,353, whereas the mean income was $71,932. It is well documented that a small number of households in the United States have income that is considerably higher than the typical American household income. As a result, these top-earning households influence the mean by pushing its value significantly above the value of the median.

MEASURE OF CENTRAL LOCATION: THE MEDIAN

The median is the middle value of a data set. The data are arranged in ascending order (smallest to largest) and the median is calculated as

- The middle value if the number of observations is odd, or
- The average of the two middle values if the number of observations is even.

The median is especially useful when outliers are present.

EXAMPLE 3.3

Use the data in Table 3.2 to calculate the median salary of employees at Acetech.

SOLUTION: In Table 3.2, the data are already arranged in ascending order. We reproduce the salaries along with their relative positions.

Position:	1	2	3	4	5	6	7
Value:	40,000	40,000	65,000	90,000	145,000	150,000	550,000

Given seven salaries, the median occupies the 4th position. Thus, the median is $90,000. Three salaries are less than $90,000 and three salaries are greater than $90,000. As compared to the mean income of $154,286, the median in this case better reflects the typical salary.

EXAMPLE 3.4

Use the data in Table 3.1 in the introductory case to calculate and interpret the median returns for the Growth and the Value mutual funds.

SOLUTION: Let's start with the median return for the Growth mutual fund. We first arrange the data in ascending order:

Position:	1	2	3	4	5	6	7	8	9	10
Value:	–38.32	1.71	3.17	5.99	12.56	13.47	16.89	16.96	32.16	36.29

Given 10 observations, the median is the average of the values in the 5th and 6th positions. These values are 12.56 and 13.47, so the median is $\frac{12.56 + 13.47}{2} = 13.02$. Over the period 2007–2016, the Growth mutual fund had a median return of 13.02%, which indicates that 5 years had returns less than 13.02% and 5 years had returns greater than 13.02%. For the Growth mutual fund, the median return of 13.02% differs from the mean return of 10.09% by approximately 3 percentage points. For the Value mutual fund, a similar comparison of these two measures of central location reveals a bigger gap. The median return of 13.67% (calculations not shown) is more than 6 percentage points greater than the mean return of 7.56%. This result is not surprising since a casual inspection of the data reveals that the relative magnitude of very small values is larger for the Value mutual fund. In order to give a more transparent description of a data's center, it is wise to report both the mean and the median.

The Mode

The **mode** of a data set is the value that occurs most frequently. A data set can have more than one mode, or even no mode. For instance, if we try to calculate the mode return for either the Growth mutual fund or the Value mutual fund in Table 3.1, we see that no value in either mutual fund occurs more than once. Thus, there is no mode for either mutual fund. If a data set has one mode, then we say it is unimodal. If two or more modes exist, then the data set is multimodal; it is common to call it bimodal in the case of two modes. Generally, the mode's usefulness as a measure of central location tends to diminish with data sets that have more than three modes.

MEASURE OF CENTRAL LOCATION: THE MODE

The mode is the most frequently occurring value in a data set. A data set may have no mode or more than one mode.

EXAMPLE 3.5

Use the data in Table 3.2 to calculate the modal salary for employees at Acetech.

SOLUTION: The salary $40,000 is earned by two employees. Every other salary occurs just once. So $40,000 is the modal salary. Just because a value occurs with the most frequency does not guarantee that it best reflects the center of the data. It is true that the modal salary at Acetech is $40,000, but most employees earn considerably more than this amount.

In the preceding examples, we used measures of central location to describe quantitative data. However, in many instances we want to summarize qualitative data, where the mode is the only meaningful measure of central location.

EXAMPLE 3.6

Kenneth Forbes is a manager at the University of Wisconsin campus bookstore. There has been a recent surge in the sale of women's sweatshirts, which are available in three sizes: Small (S), Medium (M), and Large (L). Kenneth notes that the campus bookstore sold 10 sweatshirts over the weekend in the following sizes:



©fizkes/Shutterstock

S	L	L	M	S	L	M	L	L	M

Comment on the data set and use the appropriate measure of central location that best reflects the typical size of a sweatshirt.

SOLUTION: This data set is an example of qualitative data. Here, the mode is the only relevant measure of central location. The modal size is L since it appears 5 times, as compared to S and M, which appear 2 and 3 times, respectively. Often, when examining issues relating to the demand for a product, such as replenishing stock, the mode tends to be the most relevant measure of central location.

The Weighted Mean

So far we have focused on applications where each observation in the data contributed equally to the mean. The **weighted mean** is relevant when some observations contribute more than others. For example, a student is often evaluated on the basis of the weighted mean since the score on the final exam is typically worth more than the score on the midterm.

MEASURE OF CENTRAL LOCATION: THE WEIGHTED MEAN

Let $w_1, w_2, \ldots, w_n$ denote the weights of the sample observations $x_1, x_2, \ldots, x_n$ such that $w_1 + w_2 + \cdots + w_n = 1$. The weighted mean for the sample is computed as

$$\bar{x} = \Sigma w_i x_i.$$

The weighted mean for the population is computed similarly.

EXAMPLE 3.7

A student scores 60 on Exam 1, 70 on Exam 2, and 80 on Exam 3. What is the student's average score for the course if Exams 1, 2, and 3 are worth 25%, 25%, and 50% of the grade, respectively?

SOLUTION: We define the weights as $w_1 = 0.25$, $w_2 = 0.25$, and $w_3 = 0.50$. We compute the average score as $\bar{x} = \Sigma w_i x_i = 0.25(60) + 0.25(70) + 0.50(80) = 72.50$. Note that the unweighted mean is only 70 because it does not incorporate the higher weight given to the score on Exam 3.

Using Excel to Calculate Measures of Central Location

In general, Excel offers a couple of ways to calculate most of the descriptive measures that we discuss in this chapter.

Using Excel's Function Option

Excel provides built-in formulas for virtually every summary measure that we may need. To illustrate, we follow these steps to calculate the mean for the Growth mutual fund.

FILE
Growth_Value

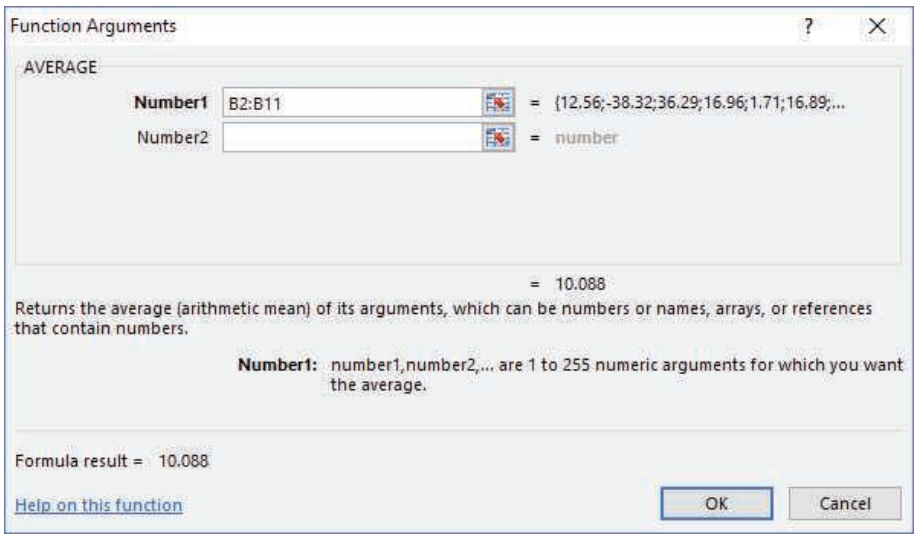
A. Open the ***Growth_Value*** data (Table 3.1).

B. From the menu choose **Formulas > Insert Function**. In the *Insert Function* dialog box, choose **Statistical** under *Select a Category.* Here you will see a list of all the relevant summary measures that Excel calculates.

C. Since we want to calculate the mean return for the Growth mutual fund, under *Select a Function,* choose **AVERAGE**. Click **OK**.

D. See Figure 3.1. In the *AVERAGE* dialog box, click on the box to the right of *Number 1* and then select the Growth data. Click **OK**. You should see the value 10.088, which, when rounded to two decimal places, equals the value that we calculated manually. In order to calculate the median and the mode, we repeat these steps, but we choose MEDIAN and MODE as the functions instead of AVERAGE.

FIGURE 3.1 Excel's AVERAGE dialog box



Source: Microsoft Excel

Once you get familiar with Excel's function names, an easier way to perform these calculations is to select an empty cell in the spreadsheet and input "=Function Name(array)", where you replace Function Name with Excel's syntax for that particular function and select the relevant data for the array or input the cell designations. For example, if we want to calculate the mean return for the Growth mutual fund, and we know that the data occupy cells B2 through B11 on the spreadsheet, we input "=AVERAGE(B2:B11)". After choosing <Enter>, Excel returns the function result in the cell. When introducing new functions later in this chapter and other chapters, we will follow this format. Table 3.3 shows various descriptive measures and corresponding function names in Excel. We will refer back to Table 3.3 on numerous occasions in this chapter.

TABLE 3.3 Descriptive Measures and Corresponding Function Names in Excel

Descriptive Measure	Excel's Function Name
Location	
Mean	=AVERAGE(array)
Median	=MEDIAN(array)
Mode	=MODE(array)
Minimum	=MIN(array)
Maximum	=MAX(array)
Dispersion	
Range	=MAX(array)-MIN(array)
Mean Absolute Deviation	=AVEDEV(array)
Sample Variance	=VAR.S(array)
Sample Standard Deviation	=STDEV.S(array)
Population Variance	=VAR.P(array)
Population Standard Deviation	=STDEV.P(array)
Association	
Sample Covariance	=COVARIANCE.S(array1,array2)
Population Covariance	=COVARIANCE.P(array1,array2)
Correlation	=CORREL(array1,array2)

Using Excel's Data Analysis Toolpak Option

Another way to obtain values for the mean, the median, and the mode is to use Excel's Data Analysis Toolpak option. One advantage of this option is that it provides numerous summary measures using a single command. Again, we illustrate this option using the data from the introductory case.

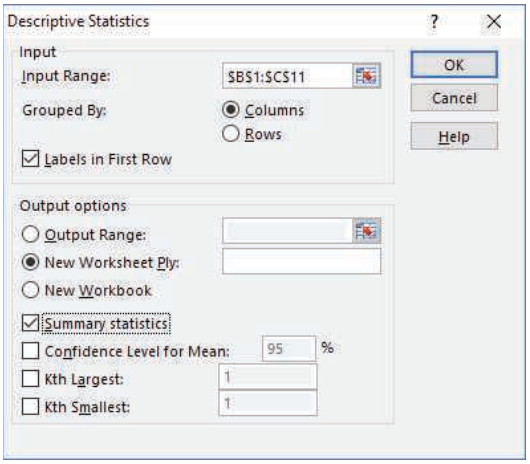
FILE
Growth_Value

A. Open the ***Growth_Value*** data (Table 3.1).

B. From the menu, choose **Data > Data Analysis > Descriptive Statistics > OK**. (Note: As mentioned in Chapter 2, if you do not see **Data Analysis** under **Data**, you must *Add-in* the Analysis Toolpak option.)

C. See Figure 3.2. In the *Descriptive Statistics* dialog box, click on the box next to *Input Range,* then select the Growth and Value data. If you included the fund names

FIGURE 3.2 Excel's Descriptive Statistics dialog box



Source: Microsoft Excel

when you highlighted the data, make sure you click on the option next to *Labels in First Row*. Click the box in front of *Summary Statistics*. Then click **OK**.

D. Table 3.4 presents the Excel output. If the output is difficult to read, highlight the output and choose **Home > Format > Column > Autofit Selection**. As noted earlier, Excel provides numerous summary measures; we have put the measures of central location in boldface. (Measures of dispersion are also in boldface; we analyze these measures in more detail shortly.) Note that Excel reports the mode as #N/A, which means no value is available; this is consistent with our finding that no value in the data appeared more than once.

TABLE 3.4 Excel's Output Using Data Analysis Toolpak Option

Growth		Value	
Mean	**10.088**	**Mean**	**7.56**
Standard Error	6.46609616	Standard Error	5.837290277
Median	**13.015**	**Median**	**13.665**
Mode	**#N/A**	**Mode**	**#N/A**
Standard Deviation	**20.44759144**	**Standard Deviation**	**18.45913264**
Sample Variance	**418.1039956**	**Sample Variance**	**340.7395778**
Kurtosis	3.416405191	Kurtosis	3.262709799
Skewness	−1.380719627	Skewness	−1.418700256
Range	74.61	Range	68.82
Minimum	−38.32	Minimum	−35.97
Maximum	36.29	Maximum	32.85
Sum	100.88	Sum	75.6
Count	10	Count	10

Note on Symmetry

In Chapter 2, we used histograms to discuss **symmetry** and **skewness**. Recall that the distribution is symmetric if one side of the histogram is a mirror image of the other side. For a symmetric and unimodal distribution, the mean, the median, and the mode are equal. In business applications, it is common to encounter data that are skewed. The mean is usually greater than the median when the data are positively skewed and less than the median when the data are negatively skewed. We would also like to comment on the numerical measure of skewness that Excel reports in Table 3.4, even though we will not discuss its calculation. A skewness coefficient of zero indicates that the values are evenly distributed on both sides of the mean. A positive skewness coefficient implies that extreme values are concentrated in the right tail of the distribution, pulling the mean up, relative to the median, and the bulk of values lie to the left of the mean. Similarly, a negative skewness coefficient implies that extreme values are concentrated in the left tail of the distribution, pulling the mean down, relative to the median, and the bulk of values lie to the right of the mean. For both mutual funds, we see that the returns are negatively skewed.

EXERCISES 3.1

Mechanics

1. Given the following observations from a sample, calculate the mean, the median, and the mode.

8	10	9	12	12

2. Given the following observations from a sample, calculate the mean, the median, and the mode.

−4	0	−6	1	−3	−4

3. Given the following observations from a population, calculate the mean, the median, and the mode.

150	257	55	110	110	43	201	125	55

4. Given the following observations from a population, calculate the mean, the median, and the mode.

20	15	25	20	10	15	25	20	15

5. FILE ***Excel_1.*** Given the accompanying data, use Excel's function options to find the mean and the median.

6. FILE ***Excel_2.*** Given the accompanying data, use Excel's function options to find the mean and the median.

Applications

7. At a small firm in Boston, seven employees were asked to report their one-way commute time (in minutes) into the city. Their responses were as follows.

20	35	90	45	40	35	50

   a. How long was the shortest commute? The longest commute?
   b. Calculate the mean, the median, and the mode.

8. In order to get an idea on current buying trends, a real estate agent collects data on 10 recent house sales in the area. Specifically, she notes the number of bedrooms in each house as follows:

3	4	3	3	5	2	4	4	5	3

   a. Calculate the mean, the median, and the mode.
   b. Which measure of central location best reflects the typical value with respect to the number of bedrooms in recent house sales?

9. The following table shows the 10 highest-paid chief executive officers of the last decade.

Name	Firm	Compensation (in $ millions)
Lawrence Ellison	Oracle	1,835.7
Barry Diller	IAC, Expedia	1,142.9
Ray Irani	Occidental Petroleum	857.1
Steve Jobs	Apple	748.8
Richard Fairbank	Capital One	568.5
Angelo Mozilo	Countrywide	528.6
Eugene Isenberg	Nabors Industries	518.0
Terry Semel	Yahoo	489.6
Henry Silverman	Cendant	481.2
William McGuire	UnitedHealth Group	469.3

Source: *The Wall Street Journal,* July 27, 2010.

   a. Calculate the mean compensation for the 10 highest-paid chief executive officers.
   b. Does the mean accurately reflect the center of the data? Explain.

10. An investor bought common stock of Microsoft Corporation on three occasions at the following prices.

Date	Price Per Share	Number of Shares
January 2009	19.58	70
July 2009	24.06	80
December 2009	29.54	50

Calculate the average price per share at which the investor bought these shares.

11. You score 90 on the midterm, 60 on the final, and 80 on the class project. What is your average score if the midterm is worth 30%, the final is worth 50%, and the class project is worth 20%?

12. An investor bought common stock of Apple Inc. on three occasions at the following prices.

Date	Price Per Share
January 2016	94.81
July 2016	102.67
December 2016	115.32

   a. What is the average price per share if the investor had bought 100 shares in January, 60 in July, and 40 in December?
   b. What is the average price per share if the investor had bought 40 shares in January, 60 in July, and 100 in December?

13. FILE ***ERA.*** One important statistic in baseball is a pitcher's earned run average, or ERA. This number represents the average number of earned runs given up by the pitcher per nine innings. The following table lists a portion of the ERAs for pitchers playing for the New York Yankees (NYY) and the Baltimore Orioles (BO) as of July 22, 2010.

Player NYY	ERA NYY	Player BO	ERA BO
Sabathia	3.13	Guthrie	4.58
Pettitte	2.88	Millwood	5.77
⋮	⋮	⋮	⋮

Source: www.mlb.com.

   a. Calculate the mean and the median ERAs for the New York Yankees.
   b. Calculate the mean and the median ERAs for the Baltimore Orioles.
   c. Based solely on your calculations above, which team is likely to have the better winning record? Explain.

14. FILE ***Largest_Corporations.*** The following table shows Fortune 500's rankings of America's 10 largest corporations for 2010. Next to each corporation is its market capitalization

(in $ billions as of March 26, 2010) and its total return (in %) to investors for the year 2009.

Company	Mkt. Cap.	Total Return
Walmart	209	–2.7
Exxon Mobil	314	–12.6
Chevron	149	8.1
General Electric	196	–0.4
Bank of America	180	7.3
ConocoPhillips	78	2.9
AT&T	155	4.8
Ford Motor	47	336.7
JP Morgan Chase	188	19.9
Hewlett-Packard	125	43.1

Source: money.cnn.com, data retrieved May 3, 2010.

a. Calculate the mean and the median for market capitalization.
b. Calculate the mean and the median for total return.
c. For each variable (market capitalization and total return), comment on which measure better reflects central location.

15. FILE ***MV_Houses.*** The following table shows a portion of the sale price (in $1,000s) for 36 homes sold in Mission Viejo, CA, during June 2010.

Number	Price
1	430
2	520
⋮	⋮
36	430

Calculate the mean, the median, and the mode.

16. FILE ***Gas_Prices_2012.*** The accompanying table shows a portion of the average price of gas (in $ per gallon) for the 50 states in the United States.

State	Price
Alabama	4.36
Alaska	3.79
⋮	⋮
Wyoming	3.63

Source: AAA.com, data retrieved April 16, 2012.

Find the mean, the median, and the mode for the price per gallon in the U.S.

17. FILE ***Life_Expectancy.*** The following table lists a portion of U.S. life expectancy (in years) for the 50 states.

Rank	State	Life Expectancy
1	Hawaii	81.5
2	Alaska	80.9
⋮	⋮	⋮
50	Mississippi	74.8

Source: en.wikipedia.org/wiki/List_of_U.S._states_by_life_expectancy, data retrieved April 25, 2012.

Find the mean, the median, and the mode of life expectancy.

3.2 PERCENTILES AND BOXPLOTS

LO 3.2

Interpret a percentile and a boxplot.

As discussed earlier, the median is a measure of central location that divides the data in half; that is, half of the data points fall below the median and half fall above the median. The median is also called the 50th percentile. In many instances, we are interested in a **percentile** other than the 50th percentile. Here we discuss calculating and interpreting percentiles. Generally, percentiles are calculated for large data sets; for ease of exposition, we show their use with a small data set. In addition, we construct a boxplot, which is, more or less, a visual representation of particular percentiles. It also helps us identify outliers and skewness in the data.

Percentiles provide detailed information about how data are spread over the interval from the smallest value to the largest value. You have probably been exposed to percentiles. For example, the SAT is the most widely used test in the undergraduate admissions process. Scores on the math portion of the SAT range from 200 to 800. Suppose you obtained a raw score of 650 on this section of the test. It may not be readily apparent how you did relative to other students that took the same test. However, if you know that the raw score corresponds to the 75th percentile, then you know that approximately 75% of

students had scores lower than your score and approximately 25% of students had scores higher than your score.

PERCENTILES

In general, the pth percentile divides a data set into two parts:

- Approximately p percent of the observations have values less than the pth percentile.
- Approximately $(100 - p)$ percent of the observations have values greater than the pth percentile.

Calculating the pth Percentile

A. First arrange the data in ascending order (smallest to largest).

B. Locate the approximate position of the percentile by calculating L_p:

$$L_p = (n + 1)\frac{p}{100},$$

where L_p indicates the location of the desired pth percentile and n is the sample size; for the population percentile, replace n with N. For example, we set $p = 50$ for the median because it is the 50th percentile.

C. Once you find the value for L_p, observe whether or not L_p is an integer:

- If L_p is an integer, then L_p denotes the location of the pth percentile. For instance, if L_{20} is equal to 2, then the 20th percentile is equal to the second observation in the ordered data set.
- If L_p is not an integer, we need to interpolate between two observations to approximate the desired percentile. So if L_{20} is equal to 2.25, then we need to interpolate 25% of the distance between the second and third observations in order to find the 20th percentile.

EXAMPLE 3.8

Consider the information presented in the introductory case of this chapter. Calculate and interpret the 25th and the 75th percentiles for the Growth mutual fund.

SOLUTION: The first step is to arrange the data in ascending order:

Position:	1	2	3	4	5	6	7	8	9	10
Value:	−38.32	1.71	3.17	5.99	12.56	13.47	16.89	16.96	32.16	36.29

For the 25th percentile: $L_{25} = (n + 1)\,\frac{p}{100} = (10 + 1)\frac{25}{100} = 2.75$. So, the 25th percentile is located 75% of the distance between the second and third observations; it is calculated as

$$1.71 + 0.75(3.17 - 1.71) = 1.71 + 1.10 = 2.81.$$

Thus, approximately 25% of the returns were less than 2.81%, and approximately 75% of the returns were greater than 2.81%.

For the 75th percentile: $L_{75} = (n+1)\frac{p}{100} = (10+1)\frac{75}{100} = 8.25$. So, the 75th percentile is located 25% of the distance between the eighth and ninth observations; it is calculated as

$$16.96 + .25(32.16 - 16.96) = 16.96 + 3.80 = 20.76.$$

Thus, approximately 75% of the returns were less than 20.76%, and approximately 25% of the returns were greater than 20.76%.

Earlier, we calculated the median or the 50th percentile for the Growth mutual fund and obtained a value of 13.02%. When we calculate the 25th, the 50th, and the 75th percentiles for a data set, we have effectively divided the data into four equal parts, or quarters. Thus, the 25th percentile is also referred to as the first quartile (Q1), the 50th percentile is referred to as the second quartile (Q2), and the 75th percentile is referred to as the third quartile (Q3).

Note on Calculating Percentiles

As mentioned in the introduction, the use of percentiles is most meaningful for large data sets. The use of a small data set in this section is simply for expositional purposes. In addition, software packages, like Excel, use different algorithms to calculate percentiles. Differences in values tend to be more dramatic with small sample sizes. With larger sample sizes, the differences, if any, tend to be negligible.

Constructing and Interpreting a Boxplot

The minimum value (Min), the quartiles (Q1, Q2, and Q3), and the maximum value (Max) are often referred to as the five-number summary of a data set. Table 3.5 shows the five-number summary for the Growth mutual fund. A **boxplot**, also referred to as a box-and-whisker plot, is a convenient way to graphically display the five-number summary of a data set.

TABLE 3.5 Summary Values for the Growth Mutual Fund

Min	Q1	Q2	Q3	Max
−38.32	2.81	13.02	20.76	36.29

Boxplots are particularly useful when comparing similar information gathered at another place or time. They are also an effective tool for identifying outliers and skewness. In Section 3.1, we discussed that the mean is unduly influenced by outliers. Sometimes outliers may indicate bad data due to incorrectly recorded observations or incorrectly included observations in the data set. In such cases, the relevant observations should be corrected or simply deleted from the data set. Alternatively, outliers may just be due to random variations, in which case the relevant observations should remain in the data set. In any event, it is important to be able to identify potential outliers so that one can take corrective actions, if needed.

In order to construct a boxplot, we follow these steps.

A. Plot the five-number summary values in ascending order on the horizontal axis.

B. Draw a box encompassing the first and third quartiles.

C. Draw a dashed vertical line in the box at the median.

D. To determine if a given observation is an outlier, first calculate the difference between Q3 and Q1. This difference is called the **interquartile range** or IQR. Therefore, the length of the box is equal to the IQR and the span of the box contains the middle half of the data. Draw a line ("whisker") that extends from

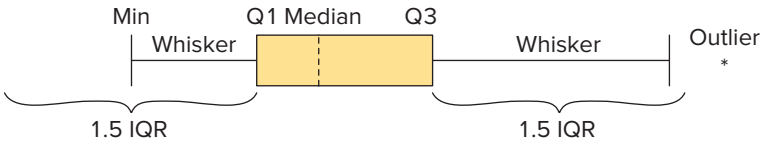
Q1 to the minimum data value that is not farther than 1.5 × IQR from Q1. Similarly, draw a line that extends from Q3 to the maximum data value that is not farther than 1.5 × IQR from Q3.

E. Use an asterisk (or comparable symbol) to indicate points that are farther than 1.5 × IQR from the box. These points are considered outliers.

Consider the boxplot in Figure 3.3 for illustration. In the figure, the left whisker extends from Q1 to Min since Min is not farther than 1.5 × IQR from Q1. The right whisker, on the other hand, does not extend from Q3 to Max since there is an observation that is farther than 1.5 × IQR from Q3. The asterisk on the right indicates that this observation is considered an outlier.

Boxplots are also used to informally gauge the shape of the distribution. Symmetry is implied if the median is in the center of the box and the left and right whiskers are equidistant from their respective quartiles. If the median is left of center and the right whisker is longer than the left whisker, then the distribution is positively skewed. Similarly, if the median is right of center and the left whisker is longer than the right whisker, then the distribution is negatively skewed. If outliers exist, we need to include them when comparing the lengths of the left and right whiskers. From Figure 3.3, we note that the median is located to the left of center and that an outlier exists on the right side. Here the right whisker is longer than the left whisker, and, if the outlier is included, then the right whisker becomes even longer. This indicates that the underlying distribution is positively skewed.

FIGURE 3.3 A sample boxplot

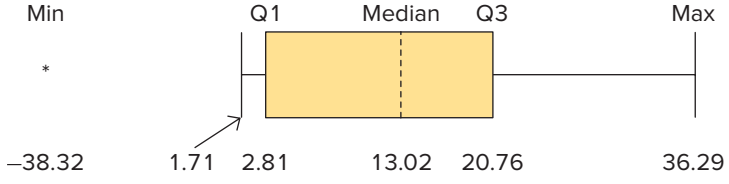


EXAMPLE 3.9

Use the information in Table 3.5 to construct and interpret the boxplot for the Growth mutual fund.

SOLUTION: Based on the information in Table 3.5, we calculate the IQR as the difference between Q3 and Q1, or IQR = 20.76 − 2.81 = 17.95. We then calculate 1.5 × IQR = 1.5 × 17.95 = 26.93. The distance between Q1 and the smallest value, 2.81 − (−38.32) = 41.13, exceeds the limit of 26.93, thus the value −38.32 is considered an outlier and will be designated as such. The next smallest value in the data set is 1.71. The distance between Q1 and this value, 2.81 − 1.71 = 1.10, is well within the limit of 26.93, so the left whisker will extend up to the point 1.71. The distance between the largest value and Q3, 36.29 − 20.76 = 15.53, is also within the limit of 26.93, so the right whisker will extend to the maximum value of 36.29. See Figure 3.4.

FIGURE 3.4 Boxplot for the Growth mutual fund



From this boxplot we can quickly grasp several points concerning the distribution of returns for the Growth mutual fund. First, returns range from −38.32%

to 36.29%, with about half being less than 13.02% and half being greater than 13.02%. We make two further observations: (1) the median is off-center within the box, being located to the right of center, and (2) the outlier on the left side implies that if the left whisker were to continue to this outlier, the left whisker would be longer than the right whisker. These two observations suggest the distribution is negatively skewed.

EXERCISES 3.2

Mechanics

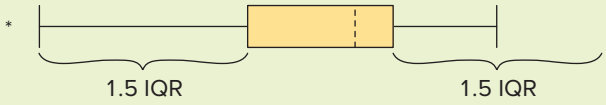
18. Calculate the 20th, 50th, and 80th percentiles for the following data set:

120	215	187	343	268	196	312

19. Calculate the 20th, 40th, and 70th percentiles for the following data set:

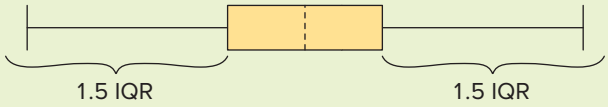
−300	−257	−325	−234	−297	−362	−255

20. Consider the following boxplot.



a. Does the boxplot indicate possible outliers in the data?
b. Comment on the skewness of the underlying distribution.

21. Consider the following boxplot.



a. Does the boxplot indicate possible outliers in the data?
b. Comment on the skewness of the underlying distribution.

22. Consider the following five-point summary that was obtained from a data set with 200 observations.

Min	Q1	Median	Q3	Max
34	54	66	78	98

a. Interpret Q1 and Q3.
b. Calculate the interquartile range. Determine whether any outliers exist.
c. Is the distribution symmetric? If not, comment on its skewness.

23. Consider the following five-point summary that was obtained from a data set with 500 observations.

Min	Q1	Median	Q3	Max
125	200	300	550	1300

a. Interpret Q1 and Q3.
b. Calculate the interquartile range. Determine whether any outliers exist.
c. Is the distribution symmetric? If not, comment on its skewness.

Applications

24. Consider the return data (in percent) for the Value mutual fund in Table 3.1.
a. Calculate and interpret the 25th, 50th, and 75th percentiles.
b. Calculate the interquartile range. Are there any outliers?
c. Is the distribution symmetric? If not, comment on its skewness.

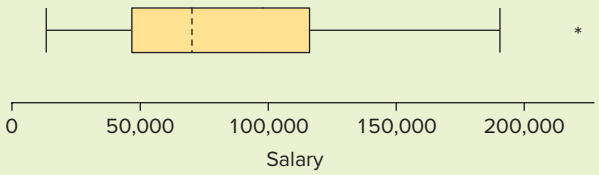
25. Scores on the final in a statistics class are as follows.

75	25	75	52	80	85	80	99	90	60
86	92	40	74	72	55	87	70	85	70

a. Calculate and interpret the 25th, 50th, and 75th percentiles.
b. Calculate the interquartile range. Are there any outliers?
c. Is the distribution symmetric? If not, comment on its skewness.

26. The following five-point summary and boxplot represent the salaries (in $) of 100 employees at a large firm.

Min	Q1	Median	Q3	Max
13,305	46,702	70,267	116,288	221,086



a. Interpret Q1 and Q3.
b. Do outliers exist in the data? Explain.
c. Is the distribution symmetric? If not, comment on its skewness.

27. FILE ***PE_Ratio.*** A price-earnings ratio or P/E ratio is calculated as a firm's share price compared to the income or profit earned by the firm per share. Generally, a high P/E ratio suggests that investors are expecting higher earnings growth in the future compared to companies with a lower P/E ratio. The accompanying table shows a portion of companies that comprise the Dow Jones Industrial Average (DJIA) and their P/E ratios as of May 17, 2012 (at the time data were retrieved, the P/E ratio for Bank of America was not available).

Company	P/E Ratio
3M (MMM)	14
Alcoa (AA)	24
⋮	⋮
Walt Disney (DIS)	14

a. Calculate and interpret the 25th, 50th, and 75th percentiles.

b. Construct a boxplot. Are there any outliers? Is the distribution symmetric? If not, comment on its skewness.

28. FILE ***Census.*** The accompanying table shows a portion of median household income (Income in $) and median house value (House Value in $) for the 50 states in 2010.

State	Income	House Value
Alabama	42081	117600
Alaska	66521	229100
⋮	⋮	⋮
Wyoming	53802	174000

Source: 2010 U.S. Census.

a. Construct a boxplot for household income and use it to identify outliers, if any, and comment on skewness.

b. Construct a boxplot for median house value and use it to identify outliers, if any, and comment on skewness.

LO **3.3**

Calculate and interpret measures of dispersion.

3.3 MEASURES OF DISPERSION

In Section 3.1, we focused on measures of central location in an attempt to find a typical or central value that describes the data. It is also important to analyze how the data vary around the center. Recall that over the 10-year period 2007–2016, the average returns for the Growth and Value mutual funds were 10.09% and 7.56%, respectively. As an investor, you might ask why anyone would put money in the Value mutual fund when, on average, this fund has a lower return. The answer to this question will become readily apparent once we analyze measures of variability or dispersion.

Table 3.6 shows each fund's minimum and maximum returns, as well as each fund's average return, over this time period. Note that the minimum and the maximum values for the Growth mutual fund are more extreme compared to the Value mutual fund; that is, the minimum value is smaller ($-38.32\% < -35.97\%$) and the maximum value is larger ($36.29\% > 32.85\%$). This indicates that returns for the Growth mutual fund may be more dispersed from the mean. The comparison of the funds illustrates that the average is not sufficient when summarizing a data set; that is, it fails to describe the underlying variability of the data.

TABLE 3.6 Select Measures for the Growth and Value Mutual Funds, 2007–2016

	Minimum Return	Average Return	Maximum Return
Growth	−38.32%	10.09%	36.29%
Value	−35.97%	7.56%	32.85%

We now discuss several measures of dispersion that gauge the variability of a data set. Each measure is a numerical value that equals zero if all data values are identical, and increases as data values become more varied.

Range

The **range** is the simplest measure of dispersion; it is the difference between the maximum value and the minimum value in a data set.

MEASURE OF DISPERSION: THE RANGE

The range is calculated by taking the difference between the maximum value (Max) and minimum value (Min) in a data set:

$$\text{Range} = \text{Max} - \text{Min}$$

EXAMPLE 3.10

Use the data in Table 3.6 to calculate the range for the Growth and the Value mutual funds.

SOLUTION:

$$\text{Growth} : 36.29\% - (-38.32\%) = 74.61\%$$
$$\text{Value} : 32.85\% - (-35.97\%) = 68.82\%$$

The Growth mutual fund has the higher value for the range, indicating that it has more dispersion with respect to its minimum and maximum values.

The range is not considered a good measure of dispersion because it focuses solely on the extreme values and ignores every other observation in the data set. While the interquartile range, IQR = Q3 − Q1, discussed in Section 3.2, does not depend on extreme values, this measure still does not incorporate all the data.

The Mean Absolute Deviation

A good measure of dispersion should consider differences of all observations from the mean. If we simply average all differences from the mean, the positives and the negatives will cancel out, even though they both contribute to dispersion, and the resulting average will equal zero. The **mean absolute deviation** (MAD) is an average of the absolute differences between the observations and the mean.

MEASURE OF DISPERSION: THE MEAN ABSOLUTE DEVIATION (MAD)

For sample values, $x_1, x_2, \ldots, x_n$, the sample MAD is computed as

$$\text{Sample MAD} = \frac{\Sigma|x_i - \bar{x}|}{n}.$$

For population values, $x_1, x_2, \ldots, x_N$, the population MAD is computed as

$$\text{Population MAD} = \frac{\Sigma|x_i - \mu|}{N}.$$

EXAMPLE 3.11

Use the data in Table 3.1 to calculate MAD for the Growth and the Value mutual funds.

SOLUTION: We first compute MAD for the Growth mutual fund. The second column in Table 3.7 shows differences from the sample mean, $\bar{x} = 10.09$. As mentioned earlier, the sum of these differences equals zero (or a number very close to zero due to rounding). The third column shows the absolute value of each deviation from the mean. Summing these values yields the numerator for the MAD formula.

TABLE 3.7 MAD Calculations for the Growth Mutual Fund

x_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x} \rvert$
12.56	12.56 – 10.09 = 2.47	2.47
–38.32	–38.32 – 10.09 = –48.41	48.41
⋮	⋮	⋮
5.99	5.99 – 10.09 = –4.10	4.10
	Total = 0 (subject to rounding)	Total = 135.60

For the Growth mutual fund: MAD $= \frac{\Sigma|x_i - \bar{x}|}{n} = \frac{135.60}{10} = 13.56$.

Similar calculations for the Value mutual fund yield: MAD $= \frac{\Sigma|x_i - \bar{x}|}{n} = \frac{132.30}{10} = 13.23$.

The Value mutual fund has a slightly smaller value for MAD than the Growth mutual fund, again indicating a less dispersed data set.

The Variance and the Standard Deviation

The **variance** and the **standard deviation** are the two most widely used measures of dispersion. Instead of calculating the average of the absolute differences from the mean, as in MAD, we calculate the average of the squared differences from the mean. The squaring of differences from the mean emphasizes larger differences more than smaller ones; MAD weighs large and small differences equally.

The variance is defined as the average of the squared differences between the observations and the mean. The formula for the variance differs depending on whether we have a sample or a population. Also, whatever the units of measurement of the original data, the variance has squared units. In order to return to the original units of measurement, we take the positive square root of the variance, which gives us the standard deviation.

MEASURES OF DISPERSION: THE VARIANCE AND THE STANDARD DEVIATION

For sample values $x_1, x_2, \ldots, x_n$, the sample variance s^2 and the sample standard deviation s are computed as

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} \qquad \text{and} \qquad s = \sqrt{s^2}.$$

For population values $x_1, x_2, \ldots, x_N$, the population variance σ^2 (the Greek letter sigma, squared) and the population standard deviation σ are computed as

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} \qquad \text{and} \qquad \sigma = \sqrt{\sigma^2}.$$

Note: The sample variance uses $n - 1$ rather than n in the denominator to ensure that the sample variance is an unbiased estimator for the population variance, a topic discussed in Chapter 8.

EXAMPLE 3.12

Use the data in Table 3.1 to calculate the sample variance and the sample standard deviation for the Growth and the Value mutual funds. Express the answers in the correct units of measurement.

SOLUTION: We will show the calculations for the Growth mutual fund, which has a mean return of 10.09%. The second column in Table 3.8 shows each return less the mean. The third column shows the square of each deviation from the mean. Summing these values yields the numerator for the sample variance formula.

TABLE 3.8 Sample Variance Calculation for the Growth Fund

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
12.56	12.56 – 10.09 = 2.47	$(2.47)^2 = 6.10$
–38.32	–38.32 – 10.09 = –48.41	$(-48.41)^2 = 2343.53$
⋮	⋮	⋮
5.99	5.99 – 10.09 = –4.10	$(-4.10)^2 = 16.81$
	Total = 0 (subject to rounding)	Total = 3762.94

For the Growth mutual fund: $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{3{,}762.94}{10-1} = 418.10(\%)^2$. Note that the units of measurement are squared. The sample standard deviation is $s = \sqrt{418.10} = 20.45(\%)$.

Similar calculations for the Value mutual fund yield

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{3{,}066.66}{10-1} = 340.74(\%)^2 \text{ and } s = \sqrt{340.74} = 18.46(\%).$$

Based on all measures of dispersion discussed thus far, we can conclude that the Value mutual fund is less dispersed than the Growth mutual fund. With financial data, standard deviation tends to be the most common measure of risk. Therefore, the investment risk of the Value mutual fund is lower than that of the Growth mutual fund over this 10-year period.

The Coefficient of Variation

In some instances, analysis entails comparing the variability of two or more data sets that have different means or units of measurement. The **coefficient of variation** (**CV**) serves as a relative measure of dispersion and adjusts for differences in the magnitudes of the means. Calculated by dividing a data set's standard deviation by its mean, CV is a unitless measure that allows for direct comparisons of mean-adjusted dispersion across different data sets.

MEASURE OF DISPERSION: THE COEFFICIENT OF VARIATION (CV)

The coefficient of variation (CV) for a data set is calculated by dividing the standard deviation by the mean. For a sample, it is calculated as $s/\bar{x}$. For a population, it is calculated as σ/μ.

EXAMPLE 3.13

Calculate and interpret the coefficient of variation for the Growth and the Value mutual funds.

SOLUTION: We use the sample means and the sample standard deviations computed earlier.

$$\text{Growth : CV} = \frac{s}{\bar{x}} = \frac{20.45\%}{10.09\%} = 2.03.$$

$$\text{Value : CV} = \frac{s}{\bar{x}} = \frac{18.46\%}{7.56\%} = 2.44.$$

Even though all other measures of variability show that returns for the Growth mutual fund are more dispersed than returns for the Value mutual fund, the coefficient of variation indicates that returns for the Value mutual fund have more relative dispersion.

Using Excel to Calculate Measures of Dispersion

Table 3.3 in Section 3.1 shows the function names for various measures of dispersion in Excel. To illustrate, we provide instructions to find the sample standard deviation for the Growth fund.

Using Excel's Function Option

FILE
Growth_Value

We open the ***Growth_Value*** data file and insert "=STDEV.S(B2:B11)". Excel returns a value of 20.45, which matches the value that we calculated by hand.

Using Excel's Data Analysis Toolpak Option

In Section 3.1, we also discussed using Excel's Data Analysis Toolpak option for calculating summary measures. For measures of dispersion, Excel treats the data as a sample and calculates the range, the sample variance, and the sample standard deviation. These values for the Growth and the Value mutual funds are shown in boldface in Table 3.4.

EXERCISES 3.3

Mechanics

29. Consider the following population data:

34	42	12	10	22

a. Calculate the range.
b. Calculate MAD.
c. Calculate the population variance.
d. Calculate the population standard deviation.

30. Consider the following population data:

0	−4	2	−8	10

a. Calculate the range.
b. Calculate MAD.
c. Calculate the population variance.
d. Calculate the population standard deviation.

31. Consider the following sample data:

40	48	32	52	38	42

a. Calculate the range.
b. Calculate MAD.
c. Calculate the sample variance.
d. Calculate the sample standard deviation.

32. Consider the following sample data:

−10	12	−8	−2	−6	8

a. Calculate the range.

b. Calculate MAD.
c. Calculate the sample variance and the sample standard deviation.

33. FILE ***Excel_1.*** Given the accompanying data, use Excel's function options to find MAD, the sample variance, and the sample standard deviation.

34. FILE ***Excel_2.*** Given the accompanying data, use Excel's function options to find MAD, the sample variance, and the sample standard deviation.

Applications

35. The Department of Transportation (DOT) fields thousands of complaints about airlines each year. The DOT categorizes and tallies complaints, and then periodically publishes rankings of airline performance. The following table presents the 2006 results for the 10 largest U.S. airlines.

Airline	Complaints*	Airline	Complaints*
Southwest	1.82	Northwest	8.84
JetBlue	3.98	Delta	10.35
Alaska	5.24	American	10.87
AirTran	6.24	US	13.59
Continental	8.83	United	13.60

Source: Department of Transportation; *per million passengers.

a. Which airline fielded the least amount of complaints? Which airline fielded the most? Calculate the range.
b. Calculate the mean and the median number of complaints for this sample.
c. Calculate the variance and the standard deviation.

36. The monthly closing stock prices (rounded to the nearest dollar) for Starbucks Corp. and Panera Bread Co. for the first six months of 2016 are reported in the following table.

Month	Starbucks Corp.	Panera Bread Co.
January 2016	61	194
February 2016	58	207
March 2016	60	205
April 2016	55	215
May 2016	57	219
June 2016	58	212

Source: finance.yahoo.com.

a. Calculate the sample variance and the sample standard deviation for each firm's stock price.
b. Which firm's stock price had greater variability as measured by the standard deviation?
c. Which firm's stock price had the greater relative dispersion?

37. FILE ***AnnArbor_Rental.*** Real estate investment in college towns continues to promise good returns (*The Wall Street Journal,* September 24, 2010). Marcela Treisman works for an investment firm in Michigan. Her assignment is to analyze the rental market in Ann Arbor, which is home to the University of Michigan. She gathers data on monthly rent for 2011 along with the square footage of 40 homes. A portion of the data is shown in the accompanying table.

Monthly Rent	Square Footage
645	500
675	648
⋮	⋮
2,400	2,700

Source: www.zillow.com.

a. Calculate the mean and the standard deviation for monthly rent.
b. Calculate the mean and the standard deviation for square footage.
c. Which sample data exhibit greater relative dispersion?

38. FILE ***Largest_Corporations.*** The accompanying data file shows the Fortune 500 rankings of America's largest corporations for 2010. Next to each corporation are its market capitalization (in $ billions as of March 26, 2010) and its total return (in %) to investors for the year 2009.
a. Calculate the coefficient of variation for market capitalization.
b. Calculate the coefficient of variation for total return.
c. Which sample data exhibit greater relative dispersion?

39. FILE ***Census.*** The accompanying data file shows, among other variables, median household income and median house value for the 50 states.
a. Compute and discuss the range of household income and house value.
b. Compute the sample MAD and the sample standard deviation of household income and house value.
c. Discuss why we cannot directly compare the sample MAD and the standard deviations of the two data sets.

3.4 MEAN-VARIANCE ANALYSIS AND THE SHARPE RATIO

LO 3.4
Explain mean-variance analysis and the Sharpe ratio.

In the introduction to Section 3.3, we asked why any rational investor would invest in the Value mutual fund over the Growth mutual fund since the average return for the Value mutual fund over the 2007–2016 period was 7.56%, whereas the average return for the

Growth mutual fund was 10.09%. It turns out that, in general, investments with higher returns also carry higher risk. Investments include financial assets such as stocks, bonds, and mutual funds. The average return represents an investor's reward, whereas variance, or equivalently standard deviation, corresponds to risk.

According to mean-variance analysis, we can measure performance of any risky asset solely on the basis of the average and the variance of its returns.

MEAN-VARIANCE ANALYSIS

Mean-variance analysis postulates that the performance of an asset is measured by its rate of return, and this rate of return is evaluated in terms of its reward (mean) and risk (variance). In general, investments with higher average returns are also associated with higher risk.

Consider Table 3.9, which summarizes the mean and the variance for the Growth and the Value mutual funds. It is true that an investment in the Growth mutual fund rather than the Value mutual fund provided an investor with a higher reward over this 10-year period, as measured by the mean return. However, this same investor encountered more risk, as measured by the variance.

TABLE 3.9 Mean-Variance Analysis for the Two Mutual Funds, 2007–2016

Fund	Mean Return %	Variance $\%^2$
Growth	10.09	418.10
Value	7.56	340.74

A discussion of mean-variance analysis seems almost incomplete without mention of the **Sharpe ratio**. Nobel Laureate William Sharpe developed what he originally referred to as the "reward-to-variability" ratio. However, academics and finance professionals prefer to call it the "Sharpe ratio." The Sharpe ratio is used to characterize how well the return of an asset compensates for the risk that the investor takes. Investors are often advised to pick investments that have high Sharpe ratios.

The Sharpe ratio is defined with the reward specified in terms of the population mean and the variability specified in terms of the population standard deviation. However, we often compute the Sharpe ratio in terms of the sample mean and the sample standard deviation, where the return is usually expressed as a percent and not a decimal.

THE SHARPE RATIO

The Sharpe ratio measures the extra reward per unit of risk. The Sharpe ratio for an investment I is computed as

$$\frac{\bar{x}_I - \bar{R}_f}{s_I},$$

where $\bar{x}_I$ is the mean return for the investment, $\bar{R}_f$ is the mean return for a risk-free asset such as a Treasury bill (T-bill), and s_I is the standard deviation for the investment.

The numerator of the Sharpe ratio measures the extra reward that investors receive for the added risk taken—this difference is often called excess return. The higher the Sharpe ratio, the better the investment compensates its investors for risk.

EXAMPLE 3.14

Calculate and interpret the Sharpe ratios for the Growth and the Value mutual funds given that the return on a 1-year T-bill is 2%.

SOLUTION: Since the return on a 1-year T-bill is 2%, $\bar{R}_f = 2$. Plugging in the values of the relevant means and standard deviations into the Sharpe ratio yields

$$\text{Sharpe ratio for the Growth mutual fund: } \frac{\bar{x}_I - \bar{R}_f}{s_I} = \frac{10.09 - 2}{20.45} = 0.40.$$

$$\text{Sharpe ratio for the Value mutual fund: } \frac{\bar{x}_I - \bar{R}_f}{s_I} = \frac{7.56 - 2}{18.46} = 0.30.$$

We had earlier shown that the Growth mutual fund had a higher return, which is good, along with a higher variance, which is bad. We can use the Sharpe ratio to make a valid comparison between the funds. The Growth mutual fund provides a higher Sharpe ratio than the Value mutual fund ($0.40 > 0.30$); therefore, the Growth mutual fund offered more reward per unit of risk compared to the Value mutual fund.

SYNOPSIS OF INTRODUCTORY CASE



©Ingram Publishing/Getty Images

Growth and value are two fundamental styles in stock and mutual fund investing. Proponents of growth investing believe that companies that are growing faster than their peers are trendsetters and will be able to maintain their superior growth. By investing in the stocks of these companies, they expect their investment to grow at a rate faster than the overall stock market. By comparison, value investors focus on the stocks of companies that are trading at a discount relative to the overall market or a specific sector. Investors of value stocks believe that these stocks are undervalued and that their price will increase once their true value is recognized by other investors. The debate between growth and value investing is age-old, and which style dominates depends on the sample period used for the analysis.

An analysis of annual return data for Vanguard's Growth Index mutual fund (Growth) and Vanguard's Value Index mutual fund (Value) for the years 2007 through 2016 provides important information for an investor trying to determine whether to invest in a growth mutual fund, a value mutual fund, or both types of mutual funds. Over this period, the mean return for the Growth fund of 10.09% is greater than the mean return for the Value fund of 7.56%. While the mean return typically represents the reward of investing, it does not incorporate the risk of investing.

Standard deviation tends to be the most common measure of risk with financial data. Since the standard deviation for the Growth fund (20.45%) is greater than the standard deviation for the Value fund (18.46%), the Growth fund is likelier to have returns farther above and below its mean. Finally, given a risk-free rate of 2%, the Sharpe ratio for the Growth fund is 0.40 compared to that for the Value fund of 0.30, indicating that the Growth fund provides more reward per unit of risk. Assuming that the behavior of these returns will continue, the investor will favor investing in Growth over Value. A commonly used disclaimer, however, states that past performance is no guarantee of future results. Since the two styles often complement each other, it might be advisable for the investor to add diversity to his portfolio by using them together.

EXERCISES 3.4

Mechanics

40. Consider the following data for two investments, A and B:

Investment A:	$\bar{x} = 10\%$ and $s = 5\%$
Investment B:	$\bar{x} = 15\%$ and $s = 10\%$

a. Which investment provides the higher return? Which investment provides less risk? Explain.

b. Given a risk-free rate of 1.4%, calculate the Sharpe ratio for each investment. Which investment provides the higher reward per unit of risk? Explain.

41. Consider the following data for two investments, A and B:

Investment A:	$\bar{x} = 8\%$ and $s = 5\%$
Investment B:	$\bar{x} = 10\%$ and $s = 7\%$

a. Which investment provides the higher return? Which investment provides less risk? Explain.

b. Given a risk-free rate of 2%, calculate the Sharpe ratio for each investment. Which investment provides the higher reward per unit of risk? Explain.

42. Consider the following returns for two investments, A and B, over the past four years:

Investment 1:	2%	8%	−4%	6%
Investment 2:	6%	12%	−8%	10%

a. Which investment provides the higher return?

b. Which investment provides less risk?

c. Given a risk-free rate of 1.2%, calculate the Sharpe ratio for each investment. Which investment has performed better? Explain.

Applications

43. Consider the following summary measures for the annual returns for Vanguard's Energy Fund and Vanguard's Health Care Fund from 2005 through 2017.

Energy: $\bar{x} = 9.62\%$ and $s = 23.58\%$
Health Care: $\bar{x} = 12.38\%$ and $s = 15.45\%$

a. Which fund had the higher average return?

b. Which fund was riskier over this time period? Given your answer in part (a), is this result surprising? Explain.

c. Given a risk-free rate of 3%, which fund has the higher Sharpe ratio? What does this ratio imply?

44. FILE ***Fidelity_Country.*** The accompanying table shows a portion of the annual returns (in %) for the Fidelity Latin America Fund and the Fidelity Canada Fund from 2000 through 2017.

Year	Latin America	Canada
2000	−17.46	12.28
2001	−6.04	−9.61
⋮	⋮	⋮
2017	30.48	14.43

Source: finance.yahoo.com.

a. Which fund had the higher average return?

b. Which fund was riskier over this time period?

c. Given a risk-free rate of 3%, which fund has the higher Sharpe ratio? What does this ratio imply?

45. FILE ***Fidelity_Select.*** The accompanying table shows a portion of the annual return (in %) for the Fidelity Select Technology Fund and Fidelity Select Energy Fund from 2000 through 2016.

Year	Technology	Energy
2000	−32.30	31.77
2001	−31.70	−11.97
⋮	⋮	⋮
2016	11.94	33.84

Source: finance.yahoo.com.

a. Compare the sample means and the sample standard deviations of the two funds.

b. Use a risk-free rate of 2% to compare the Sharpe ratios of the two funds.

LO **3.5**

Apply Chebyshev's theorem, the empirical rule, and *z*-scores.

3.5 ANALYSIS OF RELATIVE LOCATION

The mean and the standard deviation are the most extensively used measures of central location and dispersion, respectively. Unlike the mean, it is not easy to interpret the standard deviation intuitively. All we can say is that a low value for the standard deviation indicates that the data points are close to the mean, while a high value for the standard deviation indicates that the data are spread out. In this section, we will use Chebyshev's theorem and the empirical rule to make precise statements regarding the percentage of data values that fall within a specified number of standard deviations from the mean. We will also use the mean and the standard deviation to compute *z*-scores; *z*-scores measure the relative location of a value within a data set, and they are also used to detect outliers.

Chebyshev's Theorem

As we will see in more detail in later chapters, it is important to be able to use the standard deviation to make statements about the proportion of observations that fall within certain intervals. Fortunately, a Russian mathematician named Pavroty Chebyshev (1821–1894) found bounds for the proportion of the data that lie within a specified number of standard deviations from the mean.

CHEBYSHEV'S THEOREM

For any data set, the proportion of observations that lie within k standard deviations from the mean is at least $1 - 1/k^2$, where k is any number greater than 1.

This theorem holds both for a sample and for a population. For example, it implies that at least 0.75, or 75%, of the observations fall within $k = 2$ standard deviations from the mean. Similarly, at least 0.89, or 89%, of the observations fall within $k = 3$ standard deviations from the mean.

EXAMPLE 3.15

A large lecture class has 280 students. The professor has announced that the mean score on an exam is 74 with a standard deviation of 8. At least how many students scored within 58 and 90?

SOLUTION: The score 58 is two standard deviations below the mean ($\bar{x} - 2s = 74 - (2 \times 8) = 58$), while the score 90 is two standard deviations above the mean ($\bar{x} + 2s = 74 + (2 \times 8) = 90$). Using Chebyshev's theorem and $k = 2$, we have $1 - 1/2^2 = 0.75$. In other words, Chebyshev's theorem asserts that at least 75% of the scores will fall within 58 and 90. Therefore, at least 75% of 280 students, or $0.75(280) = 210$ students, scored within 58 and 90.

The main advantage of Chebyshev's theorem is that it applies to all data sets, regardless of the shape of the distribution. However, it results in conservative bounds for the percentage of observations falling in a particular interval. The actual percentage of observations lying in the interval may in fact be much larger.

The Empirical Rule

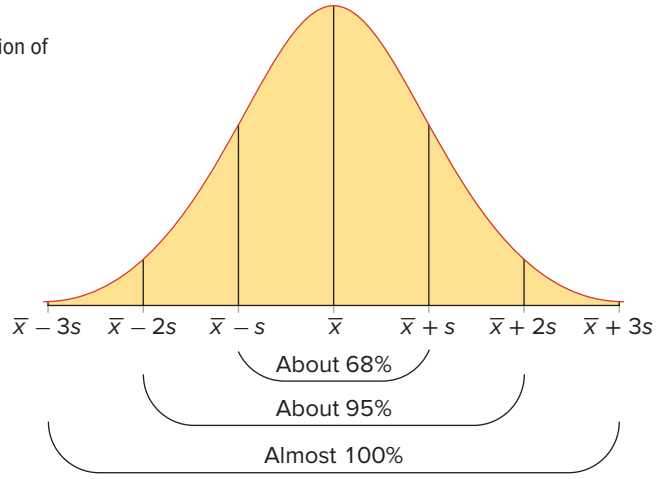
If we know that the data are drawn from a relatively symmetric and bell-shaped distribution—perhaps by a visual inspection of its histogram—then we can make more precise statements about the percentage of observations that fall within certain intervals. Symmetry and bell-shape are characteristics of the normal distribution, a topic that we discuss in Chapter 6. The normal distribution is often used as an approximation for many real-world applications. The **empirical rule** is illustrated in Figure 3.5. It provides the approximate percentage of observations that fall within 1, 2, or 3 standard deviations from the mean.

THE EMPIRICAL RULE

Given a sample mean $\bar{x}$, a sample standard deviation s, and a relatively symmetric and bell-shaped distribution:

- Approximately 68% of all observations fall in the interval $\bar{x} \pm s$,
- Approximately 95% of all observations fall in the interval $\bar{x} \pm 2s$, and
- Almost all observations fall in the interval $\bar{x} \pm 3s$.

FIGURE 3.5
Graphical description of the empirical rule



EXAMPLE 3.16

Let's revisit Example 3.15 regarding a large lecture class with 280 students with a mean score of 74 and a standard deviation of 8. Assume that the distribution is symmetric and bell-shaped.

a. Approximately how many students scored within 58 and 90?

b. Approximately how many students scored more than 90?

SOLUTION:

a. As shown in Example 3.15, the score 58 is two standard deviations below the mean, while the score 90 is two standard deviations above the mean. The empirical rule states that approximately 95% of the observations fall within two standard deviations of the mean. Therefore, about 95% of 280 students, or $0.95(280) = 266$ students, scored within 58 and 90.

b. We know that the score 90 is two standard deviations above the mean. Since approximately 95% of the observations fall within two standard deviations of the mean, we can infer that 5% of the observations fall outside the interval. Therefore, given the symmetry of the distribution, about half of 5%, or 2.5%, of 280 students scored above 90. Equivalently, about 7 students (0.025×280) scored above 90 on the exam. If the professor uses a cutoff score above 90 for an A, then only seven students in the class are expected to get an A.

The main difference between Chebyshev's theorem and the empirical rule is that Chebyshev's theorem applies to all data sets, whereas the empirical rule is appropriate when the distribution is symmetric and bell-shaped. In the preceding two examples, while Chebyshev's theorem asserts that at least 75% of the students scored between 58 and 90, we are able to make a more precise statement with the empirical rule that suggests that about 95% of the students scored between 58 and 90. It is preferable to use the empirical rule if the histogram or other visual and numerical measures suggest a symmetric and bell-shaped distribution.

z-Scores

It is often instructive to use the mean and the standard deviation to find the relative location of values within a data set. Suppose a student gets a score of 90 on her accounting exam and 90 on her marketing exam. While the student's scores are identical in both classes, her relative position in these classes may be quite different. What if the mean score was different in the classes? Even with the same mean scores, what if the standard

deviation was different in the classes? Both the mean and the standard deviation are needed to find the relative position of this student in both classes.

We use the ***z*-score** to find the relative position of a sample value within the data set by dividing the deviation of the sample value from the mean by the standard deviation.

z-SCORE

A z-score is computed as

$$z = \frac{x - \bar{x}}{s},$$

where x is a sample value and $\bar{x}$ and s are the sample mean and the sample standard deviation, respectively.

A z-score is a unitless measure since its numerator and denominator have the same units, which cancel out with each other. It measures the distance of a given sample value from the mean in terms of standard deviations. For example, a z-score of 2 implies that the given sample value is 2 standard deviations above the mean. Similarly, a z-score of -1.5 implies that the given sample value is 1.5 standard deviations below the mean. Converting sample data into z-scores is also called **standardizing** the data.

EXAMPLE 3.17

The mean and the standard deviation of scores on an accounting exam are 74 and 8, respectively. The mean and standard deviation of scores on a marketing exam are 78 and 10, respectively. Find the z-scores for a student who scores 90 in both classes.

SOLUTION: The z-score in the accounting class is $z = \frac{90 - 74}{8} = 2$. Similarly, the z-score in the marketing class is $z = \frac{90 - 78}{10} = 1.2$. Therefore, the student has fared relatively better in accounting since she is two standard deviations above the mean, as compared to marketing where she is only 1.2 standard deviations above the mean.

In Section 3.2, we used boxplots as an effective tool to identify outliers. If the data are relatively symmetric and bell-shaped, we can also use z-scores to detect outliers. Since almost all observations fall within three standard deviations of the mean, it is common to treat an observation as an outlier if its z-score is more than 3 or less than -3. Such observations must be reviewed to determine if they should remain in the data set.

EXAMPLE 3.18

Consider the information presented in the introductory case of this chapter. Use z-scores to determine if there are outliers in the Growth mutual fund data. Is this result consistent with the boxplot constructed in Figure 3.4?

SOLUTION: The smallest and the largest observations in the data set are -38.32 and 36.29, respectively. The z-score for the smallest observation is $z = \frac{-38.32 - 10.09}{20.45} = -2.37$ and the z-score for the largest observation is $z = \frac{36.29 - 10.09}{20.45} = 1.28$. Since the absolute value of both z-scores is less than 3, it would suggest that there are no outliers in the Growth mutual fund data. However, the boxplot in Figure 3.4 in Section 3.2 indicates that there is an outlier. How do we resolve this apparent inconsistency? Remember that z-scores are reliable indicators of outliers when the distribution is relatively bell-shaped and symmetric. Since the Growth mutual fund data are shown to be negatively skewed, we are better served identifying outliers in this case with a boxplot.

EXERCISES 3.5

Mechanics

46. A data set has a mean of 80 and a standard deviation of 5.
    a. Using Chebyshev's theorem, what percentage of the observations fall between 70 and 90?
    b. Using Chebyshev's theorem, what percentage of the observations fall between 65 and 95?
47. A data set has a mean of 1,500 and a standard deviation of 100.
    a. Using Chebyshev's theorem, what percentage of the observations fall between 1,300 and 1,700?
    b. Using Chebyshev's theorem, what percentage of the observations fall between 1,100 and 1,900?
48. A data set has a mean of 500 and a standard deviation of 25.
    a. Using Chebyshev's theorem, find the interval that encompasses at least 75% of the data.
    b. Using Chebyshev's theorem, find the interval that encompasses at least 89% of the data.
49. Data are drawn from a bell-shaped distribution with a mean of 20 and a standard deviation of 2.
    a. Approximately what percentage of the observations fall between 18 and 22?
    b. Approximately what percentage of the observations fall between 16 and 24?
    c. Approximately what percentage of the observations are less than 16?
50. Consider a bell-shaped distribution with a mean of 750 and a standard deviation of 50. There are 500 observations in the data set.
    a. Approximately what percentage of the observations are less than 700?
    b. Approximately how many observations are less than 700?
51. Data are drawn from a bell-shaped distribution with a mean of 25 and a standard deviation of 4. There are 1,000 observations in the data set.
    a. Approximately what percentage of the observations are less than 33?
    b. Approximately how many observations are less than 33?
52. Data are drawn from a bell-shaped distribution with a mean of 5 and a standard deviation of 2.5.
    a. Approximately what percentage of the observations are positive?
    b. Approximately what percentage of the observations are not positive?
53. Data are drawn from a bell-shaped distribution with a mean of 50 and a standard deviation of 12. There are 250 observations in the data set. Approximately how many observations are more than 74?
54. Consider a sample with six observations of 6, 9, 12, 10, 9, and 8. Compute the *z*-score for each observation.
55. Consider a sample with 10 observations of −3, 8, 4, 2, −4, 15, 6, 0, −4, and 5. Use *z*-scores to determine if there are any outliers in the data; assume a bell-shaped distribution.

Applications

56. A sample of the salaries of assistant professors on the business faculty at a local university revealed a mean income of $72,000 with a standard deviation of $3,000.
    a. Using Chebyshev's theorem, what percentage of the faculty earns at least $66,000 but no more than $78,000?
    b. Using Chebyshev's theorem, what percentage of the faculty earns at least $63,000 but no more than $81,000?
57. The historical returns on a portfolio had an average return of 8% and a standard deviation of 12%. Assume that returns on this portfolio follow a bell-shaped distribution.
    a. Approximately what percentage of returns were greater than 20%?
    b. Approximately what percentage of returns were below −16%?
58. It is often assumed that IQ scores follow a bell-shaped distribution with a mean of 100 and a standard deviation of 16.
    a. Approximately what percentage of scores are between 84 and 116?
    b. Approximately what percentage of scores are less than 68?
    c. Approximately what percentage of scores are more than 116?
59. An investment strategy has an expected return of 8% and a standard deviation of 6%. Assume investment returns are bell-shaped.
    a. How likely is it to earn a return between 2% and 14%?
    b. How likely is it to earn a return greater than 14%?
    c. How likely is it to earn a return below −4%?
60. On average, an American professional football game lasts about three hours, even though the ball is actually in play only 11 minutes (www.sportsgrid.com, January 14, 2014). Let the standard deviation be 0.4 hour.
    a. Use Chebyshev's theorem to approximate the proportion of games that last between 2.2 hours and 3.8 hours.
    b. Assume a bell-shaped distribution to approximate the proportion of games that last between 2.2 hours and 3.8 hours.
61. FILE ***Census.*** The accompanying data file shows, among other variables, median household income and median house value for the 50 states in 2010. Assume that income and house value data are bell-shaped.
    a. Use *z*-scores to determine if there are any outliers in the household income data.
    b. Use *z*-scores to determine if there are any outliers in the house value data.

62. FILE ***Fidelity_Select.*** The accompanying data file shows the annual return (in percent) for the Fidelity Select Technology Fund and the Fidelity Select Energy Fund from 2000 through 2016. Assume that the return data are bell-shaped.
    a. Use *z*-scores to determine if there are any outliers in the technology return data.
    b. Use *z*-scores to determine if there are any outliers in the energy return data.

3.6 SUMMARIZING GROUPED DATA

LO **3.6**

Calculate summary measures for grouped data.

The mean and the variance are the most widely used descriptive measures in statistics. However, the formulas in Sections 3.1 and 3.3 apply to ungrouped or raw data. In many instances, we access data that are in the form of a frequency distribution or grouped data. This is especially true of secondary data, such as data we obtain from government publications. When data are grouped or aggregated, the formulas for the mean and the variance must be modified.

THE MEAN AND THE VARIANCE FOR A FREQUENCY DISTRIBUTION

- Sample: $\bar{x} = \dfrac{\Sigma m_i f_i}{n}$ and $s^2 = \dfrac{\Sigma (m_i - \bar{x})^2 f_i}{n-1}$
- Population: $\mu = \dfrac{\Sigma m_i f_i}{N}$ and $\sigma^2 = \dfrac{\Sigma (m_i - \mu)^2 f_i}{N}$,

where m_i and f_i are the midpoint and the frequency of the ith class, respectively. The standard deviation is the positive square root of the variance.

We note that by aggregating, some of the data information is lost. Therefore, unlike in the case of raw data, we can only compute approximate values for the summary measures with grouped data.

EXAMPLE 3.19

Recall the frequency distribution of house prices (in $1,000s) that we constructed in Chapter 2.

Class	Frequency
300 up to 400	4
400 up to 500	11
500 up to 600	14
600 up to 700	5
700 up to 800	2

a. Calculate the average house price.

b. Calculate the sample variance and the sample standard deviation.

SOLUTION: Table 3.10 shows the frequency f_i and the midpoint m_i for each class in the second and third columns, respectively.

TABLE 3.10 The Sample Mean and the Sample Variance Calculation for Grouped Data

Class	f_i	m_i	$m_i f_i$	$(m_i - \bar{x})^2 f_i$
300 up to 400	4	350	1400	$(350 - 522)^2 \times 4 = 118336$
400 up to 500	11	450	4950	$(450 - 522)^2 \times 11 = 57024$
500 up to 600	14	550	7700	$(550 - 522)^2 \times 14 = 10976$
600 up to 700	5	650	3250	$(650 - 522)^2 \times 5 = 81920$
700 up to 800	2	750	1500	$(750 - 522)^2 \times 2 = 103968$
Total	36		Total = 18800	Total = 372224

a. For the mean, we multiply each class's midpoint by its respective frequency, as shown in the fourth column of Table 3.10. Finally, we sum the values in the fourth column and divide by the sample size. Or,

$$\bar{x} = \frac{\Sigma m_i f_i}{n} = \frac{18{,}800}{36} = 522.$$ The average house price is thus \$522,000.

b. For the sample variance, we first calculate the sum of the weighted squared differences from the mean. The last column in Table 3.10 shows the appropriate calculations for each class. Summing the values in the last column yields the numerator for the variance formula. We calculate the variance as

$$s^2 = \frac{\Sigma (m_i - \bar{x})^2 f_i}{n - 1} = \frac{372{,}224}{36 - 1} = 10{,}635.$$

The standard deviation is simply the positive square root of the sample variance, or $s = \sqrt{10{,}635} = 103.13$. The standard deviation for house price is thus \$103,130.

Many times, the data from secondary sources are distributed in the form of a relative frequency distribution rather than a frequency distribution. In order to use the formulas for the mean and variance for grouped data, first convert the relative frequency distribution into a frequency distribution, as discussed in Section 2.2 of Chapter 2.

EXERCISES 3.6

Mechanics

63. Consider the following frequency distribution.

Class	Frequency
2 up to 4	20
4 up to 6	60
6 up to 8	80
8 up to 10	20

a. Calculate the population mean.
b. Calculate the population variance and the population standard deviation.

64. Consider the following frequency distribution.

Class	Frequency
50 up to 60	10
60 up to 70	15
70 up to 80	8
80 up to 100	2

a. Calculate the sample mean.
b. Calculate the sample variance and the sample standard deviation.

65. The following relative frequency distribution was constructed from a population of 200. Calculate the population mean, the population variance, and the population standard deviation.

Class	Relative Frequency
−20 up to −10	0.35
−10 up to 0	0.20
0 up to 10	0.40
10 up to 20	0.05

66. The following relative frequency distribution was constructed from a sample of 50. Calculate the sample mean, the sample variance, and the sample standard deviation.

Class	Relative Frequency
0 up to 2	0.34
2 up to 4	0.20
4 up to 6	0.40
6 up to 8	0.06

Applications

67. Fifty cities provided information on vacancy rates (in %) for local apartments in the following frequency distribution.

Vacancy Rate	Frequency
0 up to 3	5
3 up to 6	5
6 up to 9	10
9 up to 12	20
12 up to 15	10

a. Calculate the average vacancy rate.
b. Calculate the variance and the standard deviation for this sample.

68. A local hospital provided the following frequency distribution summarizing the weights of babies (in pounds) delivered over the month of January.

Weight	Number of Babies
2 up to 4	3
4 up to 6	8
6 up to 8	25
8 up to 10	30
10 up to 12	4

a. Calculate the mean weight.
b. Calculate the variance and the standard deviation for this sample.

69. A researcher conducts a mileage economy test involving 80 cars. The accompanying frequency distribution summarizes the results concerning miles per gallon (MPG).

MPG	Frequency
15 up to 20	15
20 up to 25	30
25 up to 30	15
30 up to 35	10
35 up to 40	7
40 up to 45	3

a. Calculate the mean mpg.
b. Calculate the variance and the standard deviation.

70. The Boston Security Analysts Society, Inc. (BSAS) is a nonprofit association that serves as a forum for the exchange of ideas for the investment community. Suppose the ages of its members are based on the following frequency distribution.

Age	Frequency
21–31	11
32–42	44
43–53	26
54–64	7

a. Calculate the mean age.
b. Calculate the sample variance and the sample standard deviation.

71. The National Sporting Goods Association (NSGA) conducted a survey of the ages of people who purchased athletic footwear in 2009. The ages are summarized in the following percent frequency distribution.

Age of Purchaser	Percent
Under 14 years old	19
14 to 17 years old	6
18 to 24 years old	10
25 to 34 years old	13
35 to 44 years old	14
45 to 64 years old	25
65 years old and over	13

Suppose the survey was based on 100 individuals. Calculate the average age of this distribution. Calculate the sample standard deviation. Use 10 as the midpoint of the first class and 75 as the midpoint of the last class.

LO **3.7**

Calculate and interpret measures of association.

3.7 MEASURES OF ASSOCIATION

In Chapter 2, we introduced the idea of a scatterplot to visually assess whether two variables had some type of linear relationship. In this section, we present two numerical measures of association that quantify the direction and strength of the linear relationship between two variables, x and y. It is important to point out that these measures are not appropriate when the underlying relationship between the variables is nonlinear.

A numerical measure that reveals the direction of the linear relationship between two variables is called the **covariance**. We use s_{xy} to refer to a sample covariance, and σ_{xy} to refer to a population covariance.

MEASURE OF ASSOCIATION: THE COVARIANCE

The covariance shows the direction of the linear relationship between two variables. For values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the sample covariance s_{xy} is computed as

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1}.$$

For values $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, the population covariance σ_{xy} is computed as

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N}.$$

Note: As in the case of the sample variance, the sample covariance uses $n - 1$ rather than n in the denominator.

The covariance can assume a negative value, a positive value, or a value of zero.

- A negative value for covariance indicates a negative linear relationship between the two variables; on average, if x is above its mean, then y tends to be below its mean, and vice versa.
- A positive value for covariance indicates a positive linear relationship between the two variables; on average, if x is above its mean, then y tends to be above its mean, and vice versa.
- The covariance is zero if y and x have no linear relationship.

The covariance is difficult to interpret because it is sensitive to the units of measurement. That is, the covariance between two variables might be 100 and the covariance between another two variables might be 100,000, yet all we can conclude is that both sets of variables are positively related. We cannot comment on the strength of the relationships. An easier measure to interpret is the **correlation coefficient**; it describes both the direction and the strength of the linear relationship between x and y. We use r_{xy} to refer to a sample correlation coefficient and ρ_{xy} (the Greek letter rho) to refer to a population correlation coefficient.

MEASURE OF ASSOCIATION: THE CORRELATION COEFFICIENT

The correlation coefficient shows the direction and the strength of the linear relationship between two variables. The sample correlation coefficient is computed as $r_{xy} = \dfrac{s_{xy}}{s_x s_y}$, and the population correlation coefficient is computed as $\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$.

The correlation coefficient is unit-free since the units in the numerator cancel with those in the denominator. The value of the correlation coefficient falls between -1 and 1. A perfect positive relationship exists if it equals 1, and a perfect negative relationship

exists if it equals −1. Other values for the correlation coefficient must be interpreted with reference to −1, 0, or 1. For instance, a correlation coefficient equal to −0.80 indicates a strong negative relationship, whereas a correlation coefficient equal to 0.12 indicates a weak positive relationship.

EXAMPLE 3.20

Use the data in Table 3.1 to calculate and interpret the covariance and the correlation coefficient for the Growth (x) and the Value (y) mutual funds. Recall that $\bar{x} = 10.09$, $s_x = 20.45$, $\bar{y} = 7.56$, and $s_y = 18.46$.

SOLUTION: As a first step, Figure 3.6 shows a scatterplot of the return data for the Growth and Value mutual funds; scatterplots were introduced in Section 2.4. It appears that there is a positive linear relationship between the two mutual fund returns.

FIGURE 3.6 Scatterplot of return data for the Growth and the Value mutual funds

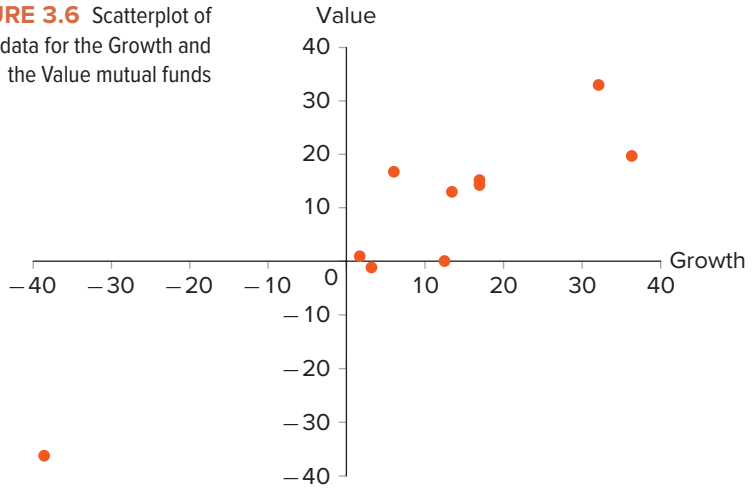


Table 3.11 shows the return data for each fund in the first two columns. The third column shows the product of differences from the mean.

TABLE 3.11 Covariance Calculation for the Growth and Value Mutual Funds

x_i	y_i	$(x_i - \bar{x})(y_i - \bar{y})$
12.56	0.09	(12.56 − 10.09)(0.09 − 7.56) = −18.45
−38.32	−35.97	(−38.32 − 10.09)(−35.97 − 7.56) = 2107.29
⋮	⋮	⋮
5.99	16.75	(5.99 − 10.09)(16.75 − 7.56) = −37.68
		Total = 3153.96

Summing the values in the third column yields the numerator for the covariance formula. Thus, we calculate the covariance as

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{3{,}153.96}{10 - 1} = 350.44.$$

The covariance of 350.44 indicates that the variables have a positive linear relationship. In other words, on average, when one fund's return is above its mean, the other fund's return is above its mean, and vice versa. The covariance is used to compute the correlation coefficient as

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{350.44}{(20.45)(18.46)} = 0.93.$$

The correlation coefficient of 0.93 indicates a strong positive linear relationship. In order to diversify the risk in an investor's portfolio, an investor is often advised to invest in assets (such as stocks, bonds, and mutual funds) whose returns are not strongly correlated. If asset returns do not have a strong positive correlation, then if one investment does poorly, the other may still do well.

Using Excel to Calculate Measures of Association

FILE
Growth_Value

Table 3.3 in Section 3.1 provides Excel function names for finding the covariance and the correlation coefficient. To illustrate, we provide instructions for finding the correlation coefficient between the returns for the Growth and Value mutual funds.

Open the ***Growth_Value*** data file. Note that the data for the Growth mutual fund are in cells B2 through B11 (array1) and the data for the Value mutual fund are in cells C2 through C11 (array2). We enter "=CORREL(B2:B11, C2:C11)", and Excel returns 0.93; this is the value that we calculated manually.

EXERCISES 3.7

Mechanics

72. Consider the following sample data:

x	−2	0	3	4	7
y	−2	−3	−8	−9	−10

a. Calculate the covariance.
b. Calculate and interpret the correlation coefficient.

73. Consider the following sample data:

x	12	18	20	22	25
y	15	20	25	22	27

a. Calculate the covariance.
b. Calculate and interpret the correlation coefficient.

74. FILE ***Excel_3.*** Given the accompanying data, use Excel's function options to find the sample covariance and the sample correlation coefficient.

75. FILE ***Excel_4.*** Given the accompanying data, use Excel's function options to find the sample covariance and the sample correlation coefficient.

Applications

76. In an attempt to determine whether a linear relationship exists between the price of a home (*x* in $1,000s) and the number of days it takes to sell the home (*y*), a real estate agent collected the following data from recent sales in his city.

x	265	225	160	325	430	515	180	423
y	136	125	120	140	145	121	122	145

a. Calculate the covariance. What kind of linear relationship exists?
b. Calculate and interpret the correlation coefficient.

77. The following table shows the annual returns (in %) for T. Rowe Price's Value and International Stock funds for the time period 2005–2009.

Year	Value Fund	International Fund
2005	6.30	16.27
2006	19.75	19.26
2007	0.75	13.43
2008	−39.76	−48.02
2009	37.15	52.20

a. Calculate and interpret the covariance between the returns.
b. Calculate and interpret the correlation coefficient.

78. A social scientist wants to analyze the relationship between educational attainment and salary. He interviews eight people. The accompanying table shows each person's years of higher education (Education in years) and corresponding salary (Salary in $1,000s).

Education	3	4	6	2	5	4	8	0
Salary	40	53	60	35	55	50	80	35

a. Calculate the covariance. What kind of linear relationship exists?
b. Calculate and interpret the correlation coefficient.

79. The director of graduate admissions at a local university is analyzing the relationship between scores on the Graduate Record Examination (GRE) and subsequent performance in graduate school, as measured by a student's grade point average (GPA). She uses a sample of 10 students who graduated within the past five years.

GRE	GPA
1,500	3.4
1,400	3.5
1,000	3.0
1,050	2.9
1,100	3.0
1,250	3.3
800	2.7
850	2.8
950	3.2
1,350	3.3

a. Calculate and interpret the covariance.
b. Calculate and interpret the correlation coefficient. Does an applicant's GRE score seem to be a good indicator of subsequent performance in graduate school?

80. FILE ***Census.*** Access the data accompanying this exercise.
a. Calculate and interpret the correlation coefficient for household income and house value.
b. Calculate and interpret the correlation coefficient for household income and the percentage of the residents who are foreign born.
c. Calculate and interpret the correlation coefficient for household income and the percentage of the residents who are without a high school diploma.

81. FILE ***Happiness_Age.*** Many attempts have been made to relate happiness with various factors. One such study relates happiness with age and finds that holding everything else constant, people are least happy when they are in their mid-40s (*The Economist,* December 16, 2010). Data are collected on a respondent's age and his/her perception of well-being on a scale from 0 to 100; a portion of the data is presented in the accompanying table.

Age	Happiness
49	62
51	66
⋮	⋮
69	72

a. Calculate and interpret the correlation coefficient between age and happiness.
b. Construct a scatterplot to point out a flaw with the above correlation analysis.

WRITING WITH STATISTICS

Many environmental groups and politicians are suggesting a return to the federal 55-mile-per-hour (mph) speed limit on America's highways. They argue that not only will a lower national speed limit reduce greenhouse emissions, it will also increase traffic safety.

Cameron Grinnell believes that a lower speed limit will not increase traffic safety. He believes that traffic safety is based on the variability of the speeds with which people are driving, rather than the average speed. The person who drives 20 mph below the pace of traffic is often as much a safety menace as the speeder. Cameron gathers the speeds of 40 cars from a highway with a speed limit of 55 mph (Highway 1) and the speeds of 40 cars from a highway with a speed limit of 65 mph (Highway 2). A portion of the data is shown in Table 3.12.



©Mike Watson Images/moodboard/Getty Images Plus/Getty Images

TABLE 3.12 Speed of Cars from Highway 1 and Highway 2

FILE ***Highway_Speeds***

Highway 1	Highway 2
60	70
55	65
⋮	⋮
52	65

Cameron would like to use the above sample information to

1. Compute and interpret the typical speed on these highways.
2. Compute and interpret the variability of speed on these highways.
3. Discuss if the reduction in the speed limit to 55 mph would increase safety on the highways.

Sample Report—Analyzing Speed Limits

Recently, many concerned citizens have lobbied for a return to the federal 55-mile-per-hour (mph) speed limit on America's highways. The reduction may lower gas emissions and save consumers on gasoline costs, but whether it will increase traffic safety is not clear. Many researchers believe that traffic safety is based on the variability of the speed rather than the average speed with which people are driving—the more variability in speed, the more dangerous the roads. Is there less variability in speed on a highway with a 55-mph speed limit as opposed to a 65-mph speed limit?

To compare average speeds, as well as the variability of speeds on highways, the speeds of 40 cars were recorded on a highway with a 55-mph speed limit (Highway 1) and on a highway with a 65-mph speed limit (Highway 2). Table 3.A shows the most relevant descriptive measures for the analysis.

TABLE 3.A Summary Measures for Highway 1 and Highway 2

	Highway 1	Highway 2
Mean	57	66
Median	56	66
Mode	50	70
Minimum	45	60
Maximum	74	70
Standard deviation	7.0	3.0
Coefficient of variation	0.12	0.05
Number of cars	40	40

The average speed of a car on Highway 1 was 57 mph, as opposed to 66 mph on Highway 2. On Highway 1, half of the 40 cars drove faster than 56 mph and half drove slower than 56 mph, as measured by the median; the median for Highway 2 was 66 mph. The mode shows that the most common speeds on Highway 1 and Highway 2 were 50 mph and 70 mph, respectively. Based on each measure of central location, Highway 2 experiences higher speeds as compared to Highway 1.

While measures of central location typically represent where the data cluster, these measures do not relay information about the variability in the data. The range of speeds is 29 mph for Highway 1 as compared to a range of just 10 mph for Highway 2. Generally, standard deviation is a more credible measure of dispersion, since range is based entirely on the minimum and the maximum values. The standard deviation for Highway 1 is substantially greater than the standard deviation for Highway 2 (7.0 mph > 3.0 mph). Therefore, the speeds on Highway 1 are more variable than the speeds on Highway 2. Even adjusting for differences in the magnitudes of the means by calculating the coefficient of variation, the speeds on Highway 1 are still more dispersed than on Highway 2 (0.12 > 0.05).

On average, it is true that the speeds on Highway 2 are higher than the speeds on Highway 1; however, the variability of speeds is greater on Highway 1. If traffic safety improves when the variability of speeds declines, then the data suggest that a return to a federal 55-mph speed limit may not enhance the well-being of highway travelers.

CONCEPTUAL REVIEW

LO 3.1 Calculate and interpret measures of central location.

The mean (average) is the most widely used measure of central location. The **sample mean** and the **population mean** are computed as $\bar{x} = \frac{\Sigma x_i}{n}$ and $\mu = \frac{\Sigma x_i}{N}$, respectively. One weakness of the mean is that it is unduly influenced by **outliers**—extremely small or large values.

The **median** is the middle value of a data set and is especially useful when outliers are present. We arrange the data in ascending order (smallest to largest) and find the median as the middle value if the number of observations is odd, or the average of the two middle values if the number of observations is even.

The **mode** is the value in the data set that occurs with the most frequency. A data set may have no mode or more than one mode. If the data are qualitative, then the mode is the only meaningful measure of central location.

LO 3.2 Interpret a percentile and a boxplot.

Percentiles provide detailed information about how the data are spread over the interval from the smallest value to the largest value. In general, the pth percentile divides the data set into two parts, where approximately p percent of the values are less than the pth percentile and the rest are greater than the pth percentile. The 25th percentile is also referred to as the first quartile (Q1), the 50th percentile is referred to as the second quartile (Q2), and the 75th percentile is referred to as the third quartile (Q3).

A **boxplot** displays the five-number summary (the minimum value, Q1, Q2, Q3, and the maximum value) for the data set. Boxplots are particularly useful when comparing similar information gathered at another place or time. They are also used as an effective tool for identifying outliers and skewness.

LO 3.3 Calculate and interpret measures of dispersion.

The **range** is the difference between the maximum and the minimum values in a data set.

The **mean absolute deviation** (MAD) is an average of the absolute differences between the observations and the mean of a data set. The sample MAD and the population MAD are computed as $\text{MAD} = \frac{\Sigma|x_i - \bar{x}|}{n}$ and $\text{MAD} = \frac{\Sigma|x_i - \mu|}{N}$, respectively.

The **variance** and the **standard deviation**, which are based on squared differences from the mean, are the two most widely used measures of dispersion. The sample variance s^2 and the sample standard deviation s are computed as $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$ and $s = \sqrt{s^2}$, respectively. The population variance σ^2 and the population standard deviation σ are computed as $\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$ and $\sigma = \sqrt{\sigma^2}$, respectively. Whatever the units of measurement of the original data, the variance has squared units. By calculating the standard deviation, we return to the original units of measurement.

The **coefficient of variation CV** is a relative measure of dispersion. The CV allows comparisons of variability between data sets with different means or different units of measurement. The sample CV and the population CV are computed as $\text{CV} = s/\bar{x}$ and $\text{CV} = \sigma/\mu$, respectively.

LO 3.4 Explain mean-variance analysis and the Sharpe ratio.

Mean-variance analysis postulates that we measure the performance of an asset by its rate of return and evaluate this rate of return in terms of its reward (mean) and risk (variance). In general, investments with higher average returns are also associated with higher risk.

The **Sharpe ratio** measures extra reward per unit of risk. The Sharpe ratio for an investment I is computed as $\frac{\bar{x}_I - \bar{R}_f}{s_I}$, where $\bar{R}_f$ denotes the mean return on a risk-free asset. The higher the Sharpe ratio, the better the investment compensates its investors for risk.

LO 3.5 Apply Chebyshev's theorem, the empirical rule, and *z*-scores.

Chebyshev's theorem dictates that for any data set, the proportion of observations that lie within k standard deviations from the mean will be at least $1 - 1/k^2$, where k is any number greater than 1.

Given a sample mean $\bar{x}$, a sample standard deviation s, and a bell-shaped distribution, the **empirical rule** dictates that

- Approximately 68% of all observations fall in the interval $\bar{x} \pm s$,
- Approximately 95% of all observations fall in the interval $\bar{x} \pm 2s$, and
- Almost all observations fall in the interval $\bar{x} \pm 3s$.

A ***z*-score,** calculated as $(x - \bar{x})/s$, measures the relative location of the sample value x. For a relatively symmetric and bell-shaped distribution, it is also used to detect outliers.

LO 3.6 Calculate summary measures for grouped data.

When analyzing grouped data, the formulas for the mean and the variance are modified as follows:

- The sample mean and the population mean are computed as $\bar{x} = \frac{\Sigma m_i f_i}{n}$ and $\mu = \frac{\Sigma m_i f_i}{N}$, respectively.
- The sample variance and the population variance are computed as $s^2 = \frac{\Sigma (m_i - \bar{x})^2 f_i}{n-1}$ and $\sigma^2 = \frac{\Sigma (m_i - \mu)^2 f_i}{N}$, respectively. As always, the standard deviation is calculated as the positive square root of the variance.

LO 3.7 Calculate and interpret measures of association.

The **covariance** and the **correlation coefficient** are measures of association that assess the direction and strength of a linear relationship between two variables, x and y. The sample covariance s_{xy} and the population covariance σ_{xy} are computed as $s_{xy} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n-1}$ and $\sigma_{xy} = \frac{\Sigma (x_i - \mu_x)(y_i - \mu_y)}{N}$, respectively. The sample correlation coefficient r_{xy} and the population correlation coefficient ρ_{xy} are computed as $r_{xy} = \frac{s_{xy}}{s_x s_y}$ and $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$, respectively.

ADDITIONAL EXERCISES AND CASE STUDIES

Exercises

82. The following table lists the sales (in $ millions) of the top Italian restaurant chains in 2009.

Restaurant	Sales
Olive Garden	3,300
Carrabba's Italian Grill	629
Romano's Macaroni Grill	583
Maggiano's	366
Carino's Italian Grill	356
Buca di Beppo	220
Bertucci's	210

Source: *The Boston Globe*, July 31, 2010.

Calculate the mean, the median, and the mode. Which measure of central tendency best reflects typical sales? Explain.

83. The following table shows the annual returns (in percent) for Fidelity's Electronic and Utilities funds.

Year	Electronic	Utilities
2005	13.23	9.36
2006	1.97	32.33
2007	2.77	21.03
2008	−50.00	−35.21
2009	81.65	14.71

Source: www.finance.yahoo.com.

a. Calculate the sample mean, the sample variance, and the sample standard deviation for each fund.
b. Which fund had the higher average return?
c. Which fund was riskier over this time period? Use both the standard deviation and the coefficient of variation in your explanation.
d. Given a risk-free rate of 4%, which fund has the higher Sharpe ratio? What does this ratio imply?

84. The manager at a water park constructed the following frequency distribution to summarize attendance for 60 days in July and August.

Attendance	Frequency
1,000 up to 1,250	5
1,250 up to 1,500	6
1,500 up to 1,750	10
1,750 up to 2,000	20
2,000 up to 2,250	15
2,250 up to 2,500	4

a. Calculate the mean attendance.
b. Calculate the variance and the standard deviation.

85. Monthly stock prices (in $) for two competing firms are as follows.

Month	Firm A	Firm B
January	28	21
February	31	24
March	32	24
April	35	27
May	34	25
June	28	20

a. Calculate the sample mean, the sample variance, and the sample standard deviation for each firm's stock price.
b. Which firm had the higher average stock price over the time period?
c. Which firm's stock price had greater variability as measured by the standard deviation? Which firm's stock price had the greater relative dispersion?

86. The National Sporting Goods Association (NSGA) conducted a survey of the ages of individuals that purchased skateboarding footwear. The ages of this survey are summarized in the following percent frequency distribution.

Age of User	Percent
Under 14 years old	35
14 to 17 years old	41
18 to 24 years old	15
25 to 34 years old	4
35 to 44 years old	4
45 to 64 years old	1

Suppose the survey was based on a sample of 200 individuals. Calculate the mean and the standard deviation of the age of individuals that purchased skateboarding shoes. Use 10 as the midpoint of the first class.

87. A manager of a local retail store analyzes the relationship between advertising (in $100s) and sales (in $1,000s) by reviewing the store's data for the previous six months.

Advertising	Sales
20	15
25	18
30	20
22	16
27	19
26	20

a. Calculate the mean of advertising and the mean of sales.
b. Calculate the standard deviation of advertising and the standard deviation of sales.
c. Calculate and interpret the covariance between advertising and sales.
d. Calculate and interpret the correlation coefficient.

88. The following table shows the annual returns (in %) for two of Putnam's mutual funds: the Voyager Growth Fund and the George Putnam Balanced Fund.

Year	Growth	Balanced
2011	−17.76	2.71
2012	14.39	12.36
2013	43.93	17.76
2014	9.55	10.48
2015	4.25	−1.11

Source: www.finance.yahoo.com.

a. Calculate and interpret the covariance.
b. Calculate and interpret the correlation coefficient.

89. FILE ***Debt_Payments.*** An economist wishes to summarize sample data from 26 metropolitan areas in the United States. The following table lists a portion of each area's 2010–2011 median income (Income in $1,000s) as well as the monthly unemployment rate (in %) and average consumer debt (in $) for August 2010.

Metropolitan Area	Income	Unemployment	Debt
Washington, D.C.	103.50	6.3	1285
Seattle	81.70	8.5	1135
⋮	⋮	⋮	⋮
Pittsburgh	63.00	8.3	763

Source: eFannieMae.com reports 2010–2011 area median incomes; www.bls.gov gives monthly unemployment rates for August 2010; Experian.com collected average monthly consumer debt payments in August 2010 and published the data in November 2010.

Compute summary measures for income, the monthly unemployment rate, and average consumer debt. Interpret these summary measures.

90. FILE ***Car_Theft.*** The accompanying table shows a portion of the number of cases of car thefts for the 50 states during 2010.

State	Car Theft
Alabama	658
Alaska	280
⋮	⋮
Wyoming	84

Source: www.fbi.gov.

a. Calculate the mean, the median, and the mode for the number of car thefts.
b. Use z-scores to determine if there are any outliers in the data. Are you surprised by the result?

91. FILE ***Quarterback_Salaries.*** American football is the highest paying sport on a per-game basis. Given that the quarterback is considered the most important player on an NFL team, he is typically well-compensated. Consider a portion of the following quarterback salary data (in $ millions) in 2009.

Name	Salary
Philip Rivers	25.5566
Jay Cutler	22.0441
⋮	⋮
Tony Romo	0.6260

Source: www.nfl.com.

a. Compute and interpret the mean and the median salary for a quarterback.
b. Compute and interpret the range and the standard deviation for quarterback salaries.

92. FILE ***Gambling.*** The accompanying table shows a portion of the number of cases of crime related to gambling (Gambling) and offenses against family and children (Abuse) for the 50 states in the United States during 2010.

State	Gambling	Abuse
Alabama	47	1022
Alaska	10	315
⋮	⋮	⋮
Wyoming	0	194

Source: www.fbi.gov.

a. Construct a boxplot for gambling and use it to identify outliers, if any.
b. Construct a boxplot for abuse and use it to identify outliers, if any.
c. Calculate and interpret the sample correlation coefficient between gambling and abuse.

93. FILE ***Gas_Prices_2012.*** The accompanying table shows a portion of the average price of gas (in $ per gallon) for the 50 states during April 2012.

State	Price
Alabama	4.36
Alaska	3.79
⋮	⋮
Wyoming	3.63

Source: AAA.com, data retrieved April 16, 2012.

a. Construct a boxplot for the gasoline price and use it to identify outliers, if any.
b. Confirm your analysis by using z-scores to determine if there are any outliers in the gasoline price.

CASE STUDIES

CASE STUDY 3.1 An article in *The Wall Street Journal* (July 11, 2008) outlined a number of reasons as to why the 16 teams in Major League Baseball's National League (NL) are inferior to the 14 teams in the American League (AL). One reason for the imbalance pointed to the disparity in opening-day payrolls: the average AL payroll is greater than the NL average. A portion of the data showing opening-day payroll (in $) for each team is shown in the accompanying table.

Data for Case Study 3.1 Major League Baseball's Opening-Day Payrolls, 2010

FILE
MLB_Salaries

American League	Payroll	National League	Payroll
New York Yankees	206333389	Chicago Cubs	146609000
Boston Red Sox	162447333	Philadelphia Phillies	141928379
⋮	⋮	⋮	⋮

Source: www.bizofbaseball.com.

In a report, use the sample information to

1. Discuss the mean and the median of AL and NL opening-day salaries and comment on skewness.
2. Compare the range and the standard deviation of AL and NL opening-day salaries.
3. Use these summary measures to comment on the findings in *The Wall Street Journal.*

CASE STUDY 3.2 Five years after graduating from college, Lucia Li feels that she is finally ready to invest some of her earnings. She has eliminated her credit card debt and has established an emergency fund. Her parents have been pleased with the performance of their mutual fund investments with Janus Capital Group. She has narrowed her search down to two mutual funds:

The Janus Balanced Fund (JANBX): This "core" fund consists of stocks and bonds and its goal is diversification. It has historically produced solid long-term returns through different market cycles.

The Janus Overseas Fund (JAOSX): This fund invests in overseas companies based on their individual merits instead of their geography or industry sector.

The following table reports a portion of the annual returns (in percent) for these two funds from 2000–2016.

Data for Case Study 3.2 Returns (in percent) for Janus Funds

FILE
Janus_Funds

Year	Balanced	Overseas
2000	−2.16	−18.57
2001	−5.04	−23.11
⋮	⋮	⋮
2016	4.51	−6.91

Source: finance.yahoo.com, data retrieved March 1, 2017.

In a report, use the sample information to

1. Calculate measures of central location to describe the similarities and the differences in these two funds' returns.
2. Calculate measures of dispersion to assess the risk of each fund.
3. Calculate and interpret measures of correlation between the two funds.

CASE STUDY 3.3 Due to a crisis in subprime lending, obtaining a mortgage has become difficult even for people with solid credit. In a report by the Associated Press (August 25, 2007), sales of existing homes fell for a 5th consecutive month, while home prices dropped for a record 12th month in July 2007. Mayan Horowitz, a research analyst for QuantExperts, wishes to study how the mortgage crunch has impacted the once-booming market of Florida. He collects data on the sale prices (in $1,000s) of 25 single-family homes in Fort Myers, Florida, in January 2007 and collects another sample in July 2007. For a valid comparison, he samples only three-bedroom homes, each with 1,500 square feet or less of space on a lot size of 10,000 square feet or less. A portion of the data is shown in the accompanying table.

Data for Case Study 3.3 Home Prices (in $1,000s) in January 2007 and July 2007

FILE *FortMyers_Sales*

Number	January	July
1	100	136
2	190	235
⋮	⋮	⋮
25	200	180

Source: www.zillow.com.

In a report, use the sample information to

1. Compare the mean, the median, and the mode in each of the two sample periods.
2. Compare the standard deviation and the coefficient of variation in each of the two sample periods.
3. Discuss significant changes in the housing market in Fort Myers over the 6-month period.

APPENDIX 3.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, JMP, and R. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands. When importing data into R, use the menu-driven option: File > Import Dataset > From Excel.

Minitab

Calculating Summary Measures

FILE *Growth_Value*

A. (Replicating Table 3.4) From the menu, choose **Stat > Basic Statistics > Display Descriptive Statistics**. Then, under **Variables**, select Growth and Value. Click **Statistics**.

B. Choose the summary measures that you wish to calculate, such as Mean, Standard deviation, etc.

Constructing a Boxplot

FILE *Growth_Value*

A. (Replicating Figure 3.4) From the menu, choose **Graph > Boxplot > One Y > Simple**.

B. Under **Graph variables**, select Growth. Click on **Data View**. Choose **Interquartile range box**, **Outlier symbols**, **Individual symbols**, and **Median connect line**.

C. Click on **Scale** and select the **Transpose value and category scales** box.

Calculating the Covariance and the Correlation Coefficient

FILE *Growth_Value*

(Replicating Example 3.20) From the menu, choose **Stat > Basic Statistics > Covariance** (choose **Correlation** to calculate the correlation coefficient). Under **Variables**, select Growth and Value.

SPSS

Calculating Summary Measures

A. (Replicating Table 3.4) From the menu, choose **Analyze > Descriptive Statistics > Descriptives**.

FILE *Growth_Value*

B. Under **Variables**, select Growth and Value. Choose **Options**. Select the summary measures that you wish to calculate, such as Mean, Std. deviation, etc.

Calculating the Covariance and the Correlation Coefficient

A. (Replicating Example 3.20) From the menu, choose **Analyze > Correlate > Bivariate**.

FILE *Growth_Value*

B. Under **Variables**, select Growth and Value. Under **Correlation Coefficients**, select **Pearson**. Choose **Options**. Under **Statistics**, select **Cross-product deviations and covariances**.

JMP

Calculating Summary Measures and Constructing a Boxplot

(Replicating Table 3.4 and Figure 3.4) From the menu, choose **Analyze > Distribution**. Under **Select Columns**, select Growth and Value, and under **Cast Selected Columns into Roles**, choose **Y**, **Columns**.

FILE *Growth_Value*

Calculating the Covariance and the Correlation Coefficient

A. (Replicating Example 3.20) From the menu, choose **Analyze > Multivariate Methods > Multivariate**. Under **Select Columns**, select Growth and Value, and under **Cast Selected Columns into Roles**, select **Y**, **Columns**.

FILE *Growth_Value*

B. Click the red triangle beside **Multivariate**. Select **Covariance Matrix**.

R

Calculating Summary Measures

A. (Replicating Table 3.4) Use the **mean** function to find the mean for a specified variable. In order to find the mean for the Growth mutual fund, enter:

FILE *Growth_Value*

```
> mean(Growth_Value$`Growth')
```

The median, the sample variance, and the sample standard deviation for a variable can be found using the functions **median**, **var**, and **sd**, respectively.

B. Use the **summary** function to find the minimum, first quartile, median, mean, third quartile, and maximum values for each variable in a data frame. Enter:

```
> summary(Growth_Value)
```

Constructing a Boxplot

(Replicating Figure 3.4) Use the **boxplot** function. For options within the function, use *xlab* to label the *x*-axis, *names* to label the variable, and *horizontal* to construct a horizontal boxplot, rather than a vertical boxplot. Enter:

FILE *Growth_Value*

```
> boxplot(Growth_Value$`Growth', xlab = "Annual Returns, 2007-2016
  (in percent)", names ="Growth", horizontal = TRUE)
```

Calculating the Correlation Coefficient

(Replicating Example 3.20) Use the **cor** function to find all pairwise correlations in a data frame. Enter:

FILE *Growth_Value*

```
> cor(Growth_Value)
```

4 Introduction to Probability

Learning Objectives

After reading this chapter you should be able to:

LO **4.1** Describe fundamental probability concepts.

LO **4.2** Apply the rules of probability.

LO **4.3** Distinguish between independent and dependent events.

LO **4.4** Calculate and interpret probabilities from a contingency table.

LO **4.5** Apply the total probability rule.

LO **4.6** Apply Bayes' theorem.

Every day we make choices about issues in the presence of uncertainty. Uncertainty describes a situation where a variety of events are possible. Usually, we either implicitly or explicitly assign probabilities to these events and plan or act accordingly. For instance, we read the paper, watch the news, or check the Internet to determine the likelihood of rain and whether we should carry an umbrella. Retailers strengthen their sales force before the end-of-year holiday season in anticipation of an increase in shoppers. The Federal Reserve cuts interest rates when it believes the economy is at risk for weak growth and raises interest rates when it feels that inflation is the greater risk. By figuring out the chances of various events, we are better prepared to make the more desirable choices. This chapter presents the essential probability tools needed to frame and address many real-world issues involving uncertainty. Probability theory turns out to be the very foundation for statistical inference, and numerous concepts introduced in this chapter are essential for understanding later chapters.


©Fab Fernandez/Image Source/Getty Images

Introductory Case

Sportswear Brands

Annabel Gonzalez is chief retail analyst at Longmeadow Consultants, a marketing firm. One aspect of her job is to track sports-apparel sales and uncover any particular trends that may be unfolding in the industry. Recently, she has been following Under Armour, Inc., the pioneer in the compression-gear market. Compression garments are meant to keep moisture away from a wearer's body during athletic activities in warm and cool weather. Under Armour has been in a fierce competition with Nike and Adidas for market share for a decade; it is also concerned about a new line of sports apparel called Second Skins that may take away customers from the compression garment giants because of its more desirable pricing (www.barrons.com, April 26, 2017).

As part of her analysis, Annabel would first like to examine whether the age of the customer matters when buying compression clothing. Her initial feeling is that the Under Armour brand attracts a younger customer, whereas the more established companies, Nike and Adidas, draw an older clientele. She believes this information is relevant to advertisers and retailers in the sporting goods industry, as well as to some in the financial community. She collects data on 600 recent purchases in the compression-gear market. She cross-classifies the data by age group and brand name, as shown in Table 4.1.

TABLE 4.1 Purchases of Compression Garments Based on Age and Brand Name

	Brand Name		
Age Group	Under Armour	Nike	Adidas
Under 35 years	174	132	90
35 years and older	54	72	78

Annabel wants to use the sample information to

1. Calculate and interpret relevant probabilities concerning brand name and age.
2. Determine whether the appeal of the Under Armour brand is mostly to younger customers.

A synopsis of this case is provided at the end of Section 4.3.

LO **4.1**

Describe fundamental probability concepts.

4.1 FUNDAMENTAL PROBABILITY CONCEPTS

Since many choices we make involve some degree of uncertainty, we are better prepared for the eventual outcome if we can use probabilities to describe which events are likely and which are unlikely. A **probability** is defined as follows.

A probability is a numerical value that measures the likelihood that an event occurs. This value is between zero and one, where a value of zero indicates an *impossible* event and a value of one indicates a *definite* event.

In order to define an event and assign the appropriate probability to it, it is useful to first establish some terminology and impose some structure on the situation.

An **experiment** is a process that leads to one of several possible outcomes. The diversity of the outcomes of an experiment is due to the uncertainty of the real world. When you purchase a new computer, there is no guarantee as to how long it will last before any repair work is needed. It may need repair in the first year, in the second year, or after two years. You can think of this as an experiment because the actual outcome will be determined only over time. Other examples of an experiment include whether a roll of a fair die will result in a value of 1, 2, 3, 4, 5, or 6; whether the toss of a coin results in heads or tails; whether a project is finished early, on time, or late; whether the economy will improve, stay the same, or deteriorate; and whether a ball game will end in a win, loss, or tie.

A **sample space**, denoted by S, of an experiment contains all possible outcomes of the experiment. For example, suppose the sample space representing the letter grade in a course is given by $S = \{\text{A, B, C, D, F}\}$. The sample space for an experiment need not be unique. For example, in the above experiment, we can also define the sample space with just P (pass) and F (fail) outcomes; that is, $S = \{\text{P, F}\}$. Note that if the teacher also gives out an I (incomplete) grade, then neither of the sample spaces defined above are valid because they do not contain all possible outcomes of the experiment.

An experiment is a process that leads to one of several possible outcomes. A sample space, denoted S, of an experiment contains all possible outcomes of the experiment.

EXAMPLE 4.1

A snowboarder competing in the Winter Olympic Games is trying to assess her probability of earning a medal in her event, the ladies' halfpipe. Construct the appropriate sample space.

SOLUTION: The athlete's attempt to predict her chances of earning a medal is an experiment because, until the Winter Games occur, the outcome is unknown. We formalize an experiment by constructing its sample space. The athlete's competition has four possible outcomes: gold medal, silver medal, bronze medal, and no medal. We formally write the sample space as $S = \{\text{gold, silver, bronze, no medal}\}$.

Events

An **event** is a subset of the sample space. A simple event consists of just one of the possible outcomes of an experiment. Getting an A in a course is an example of a simple event. An event may also contain several outcomes of an experiment. For example, we can define an event as getting a passing grade in a course; this event is formed by the subset of outcomes A, B, C, and D.

An event is any subset of outcomes of the experiment. It is called a simple event if it contains a single outcome.

Let us define two events from Example 4.1, where one event represents "earning a medal" and the other denotes "failing to earn a medal." These events are **exhaustive** because they include all outcomes in the sample space. In the earlier grade-distribution example, the events of getting grades A and B are not exhaustive events because they do not include many feasible grades in the sample space. However, the events P and F, defined as "pass" and "fail," respectively, are exhaustive.

Another important probability concept concerns **mutually exclusive** events. For two mutually exclusive events, the occurrence of one event precludes the occurrence of the other. Suppose we define the two events "at least earning a silver medal" (outcomes of gold and silver) and "at most earning a silver medal" (outcomes of silver, bronze, no medal). These two events are exhaustive because no outcome of the experiment is omitted. However, in this case, the events are not mutually exclusive because the outcome "silver" appears in both events. Going back to the grade-distribution example, while the events of getting grades A and B are not exhaustive, they are mutually exclusive, since you cannot possibly get an A as well as a B in the same course. However, getting grades P and F are mutually exclusive and exhaustive. Similarly, the events defined as "at least earning a silver medal" and "at most earning a bronze medal" are mutually exclusive and exhaustive.

Events are exhaustive if all possible outcomes of an experiment belong to the events. Events are mutually exclusive if they do not share any common outcome of an experiment.

For any experiment, we can define events based on one or more outcomes of the experiment and also combine events to form new events. The **union** of two events, denoted $A \cup B$, is the event consisting of all outcomes in A or B. A useful way to illustrate these concepts is through the use of a Venn diagram, named after the British mathematician John Venn (1834–1923). Figure 4.1 shows a Venn diagram where the rectangle represents the sample space S and the two circles represent events A and B. The union $A \cup B$ is the portion in the Venn diagram that is included in either A or B.

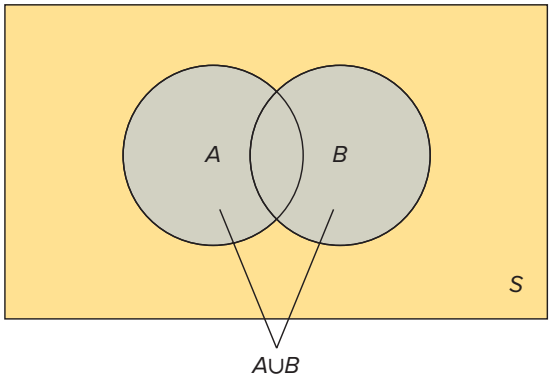
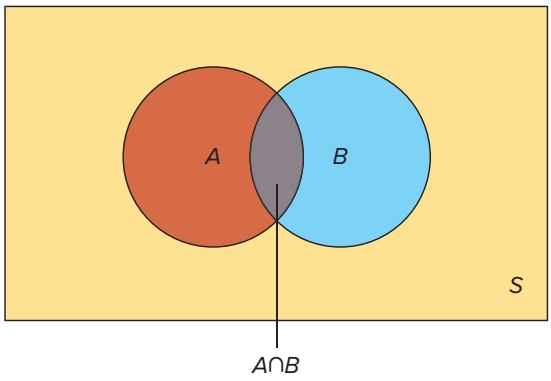


FIGURE 4.1
The union of two events, $A \cup B$

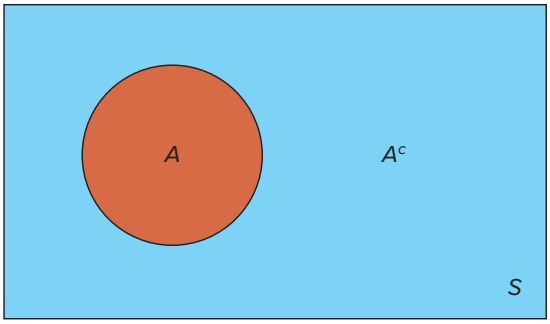
The **intersection** of two events, denoted $A \cap B$, is the event consisting of all outcomes in A and B. Figure 4.2 depicts the intersection of two events A and B. The intersection $A \cap B$ is the portion in the Venn diagram that is included in both A and B.

FIGURE 4.2
The intersection of two events, $A \cap B$



The **complement** of event A, denoted A^c, is the event consisting of all outcomes in the sample space S that are not in A. In Figure 4.3, A^c is everything in S that is not included in A.

FIGURE 4.3
The complement of an event, A^c



COMBINING EVENTS

- The union of two events, denoted $A \cup B$, is the event consisting of all outcomes in A or B.
- The intersection of two events, denoted $A \cap B$, is the event consisting of all outcomes in A and B.
- The complement of event A, denoted A^c, is the event consisting of all outcomes in the sample space S that are not in A.

EXAMPLE 4.2

Recall that the snowboarder's sample space from Example 4.1 is defined as $S = \{\text{gold, silver, bronze, no medal}\}$. Now suppose the snowboarder defines the following three events:

- $A = \{\text{gold, silver, bronze}\}$; that is, event A denotes earning a medal;
- $B = \{\text{silver, bronze, no medal}\}$; that is, event B denotes earning at most a silver medal; and
- $C = \{\text{no medal}\}$; that is, event C denotes failing to earn a medal.

a. Find $A \cup B$ and $B \cup C$.

b. Find $A \cap B$ and $A \cap C$.

c. Find B^c.

SOLUTION:

a. The union of A and B denotes all outcomes common to A or B; here, the event $A \cup B = \{\text{gold, silver, bronze, no medal}\}$. Note that there is no double counting of the outcomes "silver" or "bronze" in $A \cup B$. Similarly, we have the event $B \cup C = \{\text{silver, bronze, no medal}\}$.

b. The intersection of A and B denotes all outcomes common to A and B; here, the event $A \cap B = \{\text{silver, bronze}\}$. The event $A \cap C = \emptyset$, where $\emptyset$ denotes the null (empty) set; no common outcomes appear in both A and C.

c. The complement of B denotes all outcomes in S that are not in B; here, the event $B^c = \{\text{gold}\}$.

Assigning Probabilities

Now that we have described a valid sample space and the various ways in which we can define events from that sample space, we are ready to assign probabilities. When we arrive at a probability, we generally are able to categorize the probability as a subjective probability, an empirical probability, or a classical probability. Regardless of the method used, there are two defining properties of probability.

THE TWO DEFINING PROPERTIES OF PROBABILITY

1. The probability of any event A is a value between 0 and 1; that is, $0 \leq P(A) \leq 1$.
2. The sum of the probabilities of any list of mutually exclusive and exhaustive events equals 1.

Suppose the snowboarder from Example 4.1 believes that there is a 10% chance that she will earn a gold medal, a 15% chance that she will earn a silver medal, a 20% chance that she will earn a bronze medal, and a 55% chance that she will fail to earn a medal. She has assigned a **subjective probability** to each of the simple events. She made a personal assessment of these probabilities without referencing any data.

The snowboarder believes that the most likely outcome is failing to earn a medal since she gives that outcome the greatest chance of occurring at 55%. When formally writing out the probability that an event occurs, we generally construct a probability statement. Here, the probability statement might take the form: $P(\{\text{no medal}\}) = 0.55$, where $P(\text{"event"})$ represents the probability that a given event occurs. Table 4.2 summarizes these events and their respective subjective probabilities. Note that here the events are mutually exclusive and exhaustive.

TABLE 4.2 Snowboarder's Subjective Probabilities

Event	Probability
Gold	0.10
Silver	0.15
Bronze	0.20
No medal	0.55

Reading from the table we can readily see, for instance, that she assesses that there is a 15% chance that she will earn a silver medal, or $P(\{\text{silver}\}) = 0.15$. We should note that all the probabilities are between the values of zero and one, and they add up to one, thus meeting the defining properties of probability.

Suppose the snowboarder wants to calculate the probability of earning a medal. In Example 4.2, we defined "earning a medal" as event A, so the probability statement takes the form $P(A)$. We calculate this probability by summing the probabilities of the outcomes in A, or equivalently,

$$P(A) = P(\{\text{gold}\}) + P(\{\text{silver}\}) + P(\{\text{bronze}\}) = 0.10 + 0.15 + 0.20 = 0.45.$$

EXAMPLE 4.3

Given the events in Example 4.2 and the probabilities in Table 4.2, calculate the following probabilities.

a. $P(B \cup C)$

b. $P(A \cap C)$

c. $P(B^c)$

SOLUTION:

a. The probability that event B or event C occurs is

$$\begin{aligned} P(B \cup C) &= P(\{\text{silver}\}) + P(\{\text{bronze}\}) + P(\{\text{no medal}\}) \\ &= 0.15 + 0.20 + 0.55 = 0.90. \end{aligned}$$

b. The probability that event A and event C occur is

$P(A \cap C) = 0$; recall that there are no common outcomes in A and C.

c. The probability that the complement of B occurs is

$$P(B^c) = P(\{\text{gold}\}) = 0.10.$$

In many instances, we calculate probabilities by referencing data based on the observed outcomes of an experiment. The **empirical probability** of an event is the observed relative frequency with which an event occurs. The experiment must be repeated a large number of times for empirical probabilities to be accurate.

EXAMPLE 4.4

The frequency distribution in Table 4.3 summarizes the ages of the richest 400 Americans. Suppose we randomly select one of these individuals.

a. What is the probability that the individual is at least 50 but less than 60 years old?

b. What is the probability that the individual is younger than 60 years old?

c. What is the probability that the individual is at least 80 years old?

TABLE 4.3 Frequency Distribution of Ages of 400 Richest Americans

Ages	Frequency
< 40	13
40 up to 50	24
50 up to 60	67
60 up to 70	113
70 up to 80	117
80 up to 90	55
≥ 90	11

Source: www.forbes.com, 2016 Ranking.

SOLUTION: In Table 4.3a, we first label each outcome with letter notation; for instance, the outcome "< 40" is denoted as event A. Next we calculate the relative frequency of each event and use the relative frequency to denote the probability of the event.

TABLE 4.3a Relative Frequency Distribution of Ages of 400 Richest Americans

Ages	Event	Frequency	Relative Frequency
< 40	A	13	13/400 = 0.0325
40 up to 50	B	24	0.0600
50 up to 60	C	67	0.1675
60 up to 70	D	113	0.2825
70 up to 80	E	117	0.2925
80 up to 90	F	55	0.1375
≥ 90	G	11	0.0275

a. The probability that an individual is at least 50 but less than 60 years old is

$$P(C) = \frac{67}{400} = 0.1675.$$

b. The probability that an individual is younger than 60 years old is

$$P(A \cup B \cup C) = \frac{13 + 24 + 67}{400} = 0.26.$$

c. The probability that an individual is at least 80 years old is

$$P(F \cup G) = \frac{55 + 11}{400} = 0.165.$$

In a more narrow range of well-defined problems, we can sometimes deduce probabilities by reasoning about the problem. The resulting probability is a **classical probability**. Classical probabilities are often used in games of chance. They are based on the assumption that all outcomes of an experiment are equally likely. Therefore, the classical probability of an event is computed as the number of outcomes belonging to the event divided by the total number of outcomes.

EXAMPLE 4.5

Suppose our experiment consists of rolling a six-sided die. Then we can define the appropriate sample space as $S = \{1, 2, 3, 4, 5, 6\}$.

a. What is the probability that we roll a 2?

b. What is the probability that we roll a 2 or 5?

c. What is the probability that we roll an even number?

SOLUTION: Here we recognize that each outcome is equally likely. So with 6 possible outcomes, each outcome has a 1/6 chance of occurring.

a. The probability that we roll a 2, $P(\{2\})$, is thus 1/6.

b. The probability that we roll a 2 or 5, $P(\{2\}) + P(\{5\})$, is 1/6 + 1/6 = 1/3.

c. The probability that we roll an even number, $P(\{2\}) + P(\{4\}) + P(\{6\})$, is 1/6 + 1/6 + 1/6 = 1/2.

CATEGORIZING PROBABILITIES

- A subjective probability is calculated by drawing on personal and subjective judgment.
- An empirical probability is calculated as a relative frequency of occurrence.
- A classical probability is based on logical analysis rather than on observation or personal judgment.

Since empirical and classical probabilities generally do not vary from person to person, they are often grouped as objective probabilities.

According to the famous **law of large numbers**, the empirical probability approaches the classical probability if the experiment is run a very large number of times. Consider, for example, flipping a fair coin 10 times. It is possible that heads may not show up exactly 5 times and, therefore, the relative frequency may not be 0.50. However, if we flip the fair coin a very large number of times, heads will show up approximately half of the time. This would make the empirical probability equal to the classical probability of 0.50.

EXERCISES 4.1

Mechanics

1. Determine whether the following probabilities are best categorized as subjective, empirical, or classical probabilities.
   a. Before flipping a fair coin, Sunil assesses that he has a 50% chance of obtaining tails.
   b. At the beginning of the semester, John believes he has a 90% chance of receiving straight A's.
   c. A political reporter announces that there is a 40% chance that the next person to come out of the conference room will be a Republican, since there are 60 Republicans and 90 Democrats in the room.
2. A sample space S yields three mutually exclusive and exhaustive events, A, B, and C, such that $P(A) = 0.25$ and $P(A \cup B) = 0.70$.
   a. Find $P(C)$.
   b. Find $P(B^C)$.
   c. Find $P(A \cap C)$.
3. A sample space S yields five equally likely events, A, B, C, D, and E.
   a. Find $P(D)$.
   b. Find $P(B^C)$.
   c. Find $P(A \cup C \cup E)$.
4. You roll a die with the sample space $S = \{1, 2, 3, 4, 5, 6\}$. You define A as $\{1, 2, 3\}$, B as $\{1, 2, 3, 5, 6\}$, C as $\{4, 6\}$, and D as $\{4, 5, 6\}$. Determine which of the following events are exhaustive and/or mutually exclusive.
   a. A and B
   b. A and C
   c. A and D
   d. B and C
5. A sample space, S, yields four simple events, A, B, C, and D, such that $P(A) = 0.35$, $P(B) = 0.10$, and $P(C) = 0.25$.
   a. Find $P(D)$.
   b. Find $P(C^c)$.
   c. Find $P(A \cup B)$.

Applications

6. Survey data, based on 65,000 mobile phone subscribers, shows that 44% of the subscribers use smartphones (*Forbes*, December 15, 2011). Based on this information, you infer that the probability that a mobile phone subscriber uses a smartphone is 0.44. Would you consider this probability estimate accurate? Would you label this probability as subjective, empirical, or classical?
7. Jane Peterson has taken Amtrak to travel from New York to Washington, DC, on six occasions, of which three times the train was late. Therefore, Jane tells her friends that the probability that this train will arrive on time is 0.50. Would you label this probability as empirical or classical? Why would this probability not be accurate?
8. Consider the following scenarios to determine if the mentioned combination of attributes represents a union or an intersection.
   a. A marketing firm is looking for a candidate with a business degree and at least five years of work experience.
   b. A family has decided to purchase a Toyota minivan or a Honda minivan.
9. Consider the following scenarios to determine if the mentioned combination of attributes represents a union or an intersection.
   a. There are two courses that seem interesting to you, and you would be happy if you can take at least one of them.
   b. There are two courses that seem interesting to you, and you would be happy if you can take both of them.

10. You apply for a position at two firms. Let event A represent the outcome of getting an offer from the first firm and event B represent the outcome of getting an offer from the second firm.
    a. Explain why events A and B are not exhaustive.
    b. Explain why events A and B are not mutually exclusive.
11. An alarming number of U.S. adults are either overweight or obese. The distinction between overweight and obese is made on the basis of body mass index (BMI), expressed as weight/height^2. An adult is considered overweight if the BMI is 25 or more but less than 30. An obese adult will have a BMI of 30 or greater. According to a January 2012 article in the *Journal of the American Medical Association*, 33.1% of the adult population in the United States is overweight and 35.7% is obese. Use this information to answer the following questions.
    a. What is the probability that a randomly selected adult is either overweight or obese?
    b. What is the probability that a randomly selected adult is neither overweight nor obese?
    c. Are the events "overweight" and "obese" exhaustive?
    d. Are the events "overweight" and "obese" mutually exclusive?
12. Many communities are finding it more and more difficult to fill municipal positions such as town administrators, finance directors, and treasurers. The following table shows the proportion of municipal managers by age group in the United States for the years 1971 and 2006.

Age	1971	2006
Under 30	0.26	0.01
30 to 40	0.45	0.12
41 to 50	0.21	0.28
51 to 60	0.05	0.48
Over 60	0.03	0.11

Source: *The International City-County Management Association.*

    a. In 1971, what was the probability that a municipal manager was 40 years old or younger? In 2006, what was the probability that a municipal manager was 40 years old or younger?
    b. In 1971, what was the probability that a municipal manager was 51 years old or older? In 2006, what was the probability that a municipal manager was 51 years old or older?
    c. What trends in ages can you detect from municipal managers in 1971 versus municipal managers in 2006?
13. At four community health centers on Cape Cod, Massachusetts, 15,164 patients were asked to respond to questions designed to detect depression (*The Boston Globe*, June 11, 2008). The survey produced the following results.

Diagnosis	Number
Mild	3,257
Moderate	1,546
Moderately Severe	975
Severe	773
No Depression	8,613

    a. What is the probability that a randomly selected patient suffered from mild depression?
    b. What is the probability that a randomly selected patient did not suffer from depression?
    c. What is the probability that a randomly selected patient suffered from moderately severe to severe depression?
    d. Given that the national figure for moderately severe to severe depression is approximately 6.7%, does it appear that there is a higher rate of depression in this summer resort community? Explain.

4.2 RULES OF PROBABILITY

In the previous section, we discussed how the probability of an event is assigned. Here we present various rules used to combine probabilities of events.

The Complement Rule

LO **4.2**

Apply the rules of probability.

The **complement rule** follows from one of the defining properties of probability: the sum of probabilities assigned to simple events in a sample space must equal one. Note that since S is a collection of all possible outcomes of the experiment (nothing else can happen), $P(S) = 1$. Let's revisit the sample space that we constructed when we rolled a six-sided die: $S = \{1, 2, 3, 4, 5, 6\}$. Suppose event A is defined as an even-numbered outcome or $A = \{2, 4, 6\}$. We then know that the complement of A, A^c, is the set consisting of $\{1, 3, 5\}$. Moreover, we can deduce that $P(A) = 1/2$ and $P(A^c) = 1/2$, so $P(A) + P(A^c) = 1$. Rearranging this equation, we obtain the complement rule: $P(A^c) = 1 - P(A)$.

THE COMPLEMENT RULE

The complement rule states that the probability of the complement of an event, $P(A^c)$, is equal to one minus the probability of the event; that is, $P(A^c) = 1 - P(A)$.

The complement rule is quite straightforward, but it is widely used and powerful.

EXAMPLE 4.6

According to the 2010 U.S. Census, 37% of women ages 25 to 34 have earned at least a college degree, as compared with 30% of men in the same age group.

a. What is the probability that a randomly selected woman between the ages of 25 to 34 does not have a college degree?

b. What is the probability that a randomly selected man between the ages of 25 to 34 does not have a college degree?

SOLUTION:

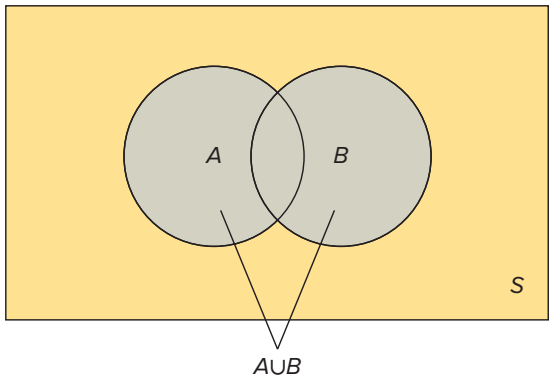
a. Let's define A as the event that a randomly selected woman between the ages of 25 and 34 has a college degree; thus, $P(A) = 0.37$. In this problem, we are interested in the complement of A. So $P(A^c) = 1 - P(A) = 1 - 0.37 = 0.63$.

b. Similarly, we define B as the event that a randomly selected man between the ages of 25 to 34 has a college degree, so $P(B) = 0.30$. Thus, $P(B^c) = 1 - P(B) = 1 - 0.30 = 0.70$.

The Addition Rule

The **addition rule** allows us to find the probability of the union of two events. Suppose we want to find the probability that either A occurs or B occurs, so in probability terms, $P(A \cup B)$. We reproduce the Venn diagram, used earlier in Figure 4.1, to help in exposition. Figure 4.4 shows a sample space S with the two events A and B. Recall that the union, $A \cup B$, is the portion in the Venn diagram that is included in either A or B. The intersection, $A \cap B$, is the portion in the Venn diagram that is included in both A and B.

FIGURE 4.4
Finding the probability of the union of two events, $P(A \cup B)$



If we try to obtain $P(A \cup B)$ by simply summing $P(A)$ with $P(B)$, then we overstate the probability because we double-count the probability of the intersection of A and B, $P(A \cap B)$. When implementing the addition rule, we sum $P(A)$ and $P(B)$ and then subtract $P(A \cap B)$ from this sum.

THE ADDITION RULE

The addition rule states that the probability that A or B occurs, or that at least one of these events occurs, is equal to the probability that A occurs, plus the probability that B occurs, minus the probability that both A and B occur. Equivalently,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

EXAMPLE 4.7

Anthony feels that he has a 75% chance of getting an A in Statistics and a 55% chance of getting an A in Managerial Economics. He also believes he has a 40% chance of getting an A in both classes.

a. What is the probability that he gets an A in at least one of these courses?

b. What is the probability that he does not get an A in either of these courses?

SOLUTION:

a. Let $P(A_S)$ correspond to the probability of getting an A in Statistics and $P(A_M)$ correspond to the probability of getting an A in Managerial Economics. Thus, $P(A_S) = 0.75$ and $P(A_M) = 0.55$. In addition, there is a 40% chance that Anthony gets an A in both classes; that is, $P(A_S \cap A_M) = 0.40$. In order to find the probability that he receives an A in at least one of these courses, we calculate

$$P(A_S \cup A_M) = P(A_S) + P(A_M) - P(A_S \cap A_M) = 0.75 + 0.55 - 0.40 = 0.90.$$

b. The probability that he does not receive an A in either of these two courses is actually the complement of the union of the two events; that is, $P((A_S \cup A_M)^c)$. We calculated the union in part a, so using the complement rule we have

$$P((A_S \cup A_M)^c) = 1 - P(A_S \cup A_M) = 1 - 0.90 = 0.10.$$

An alternative expression that correctly captures the required probability is $P(A_S^c \cap A_M^c)$, which is the probability that he does not get an A in Statistics and he does not get an A in Managerial Economics. A common mistake is to calculate the probability as $1 - P(A_S \cap A_M) = 1 - 0.40 = 0.60$, which simply indicates that there is a 60% chance that Anthony will not get an A in both courses. This is clearly not the required probability that Anthony does not get an A in either course.

The Addition Rule for Mutually Exclusive Events

As mentioned earlier, mutually exclusive events do not share any outcome of an experiment. Figure 4.5 shows the Venn diagram for two mutually exclusive events; note that the circles do not intersect.

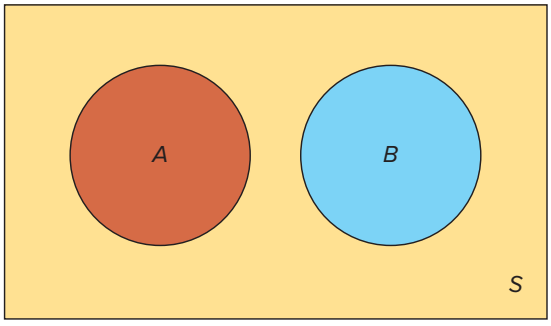


FIGURE 4.5
Mutually exclusive events

For mutually exclusive events A and B, the probability of their intersection is zero; that is, $P(A \cap B) = 0$. We need not concern ourselves with double-counting, and, therefore, the probability of the union is simply the sum of the two probabilities.

THE ADDITION RULE FOR MUTUALLY EXCLUSIVE EVENTS

If A and B are mutually exclusive events, then $P(A \cap B) = 0$ and, therefore, the addition rule simplifies to $P(A \cup B) = P(A) + P(B)$.

EXAMPLE 4.8

Samantha Greene, a college senior, contemplates her future immediately after graduation. She thinks there is a 25% chance that she will join the Peace Corps and teach English in Madagascar for the next few years. Alternatively, she believes there is a 35% chance that she will enroll in a full-time law school program in the United States.

a. What is the probability that she joins the Peace Corps or enrolls in law school?

b. What is the probability that she does not choose either of these options?

SOLUTION:

a. We can write the probability that Samantha joins the Peace Corps as $P(A) = 0.25$ and the probability that she enrolls in law school as $P(B) = 0.35$. Immediately after college, Samantha cannot choose both of these options. This implies that these events are mutually exclusive, so $P(A \cap B) = 0$. Thus, when solving for the probability that Samantha joins the Peace Corps or enrolls in law school, $P(A \cup B)$, we can simply sum $P(A)$ and $P(B)$: $P(A \cup B) = P(A) + P(B) = 0.25 + 0.35 = 0.60$.

b. In order to find the probability that she does not choose either of these options, we need to recognize that this probability is the complement of the union of the two events; that is, $P((A \cup B)^c)$. Therefore, using the complement rule, we have

$$P((A \cup B)^c) = 1 - P(A \cup B) = 1 - 0.60 = 0.40.$$

Conditional Probability

In business applications, the probability of interest is often a conditional probability. Examples include the probability that the housing market will improve conditional on the Federal Reserve taking remedial actions; the probability of making a six-figure salary conditional on getting an MBA; the probability that a company's stock price will go up conditional on higher-than-expected profits; and the probability that sales will improve conditional on the firm launching a new marketing campaign.

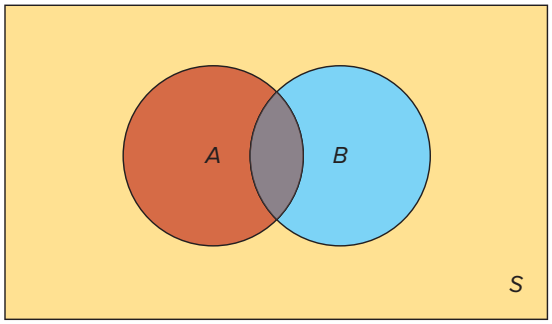
Let's use an example to illustrate the concept of conditional probability. Suppose the probability that a recent business college graduate finds a suitable job is 0.80. The probability of finding a suitable job is 0.90 if the recent business college graduate has prior work experience. This type of probability is called a **conditional probability**, where the probability of an event is conditional on the occurrence of another event. If A represents "finding a job" and B represents "prior work experience," then $P(A) = 0.80$ and the conditional probability is denoted as $P(A \mid B) = 0.90$. The vertical mark $\mid$ means "given that," and the conditional probability is typically read as "the probability of A given B." In this example, the probability of finding a suitable job increases from 0.80 to 0.90 when

conditioned on prior work experience. In general, the conditional probability, $P(A \mid B)$, is greater than the **unconditional probability**, $P(A)$, if B exerts a positive influence on A. Similarly, $P(A \mid B)$ is less than $P(A)$ when B exerts a negative influence on A. Finally, if B exerts no influence on A, then $P(A \mid B)$ equals $P(A)$. It is common to refer to "unconditional probability" simply as "probability."

As we will see later, it is important that we write the event that has already occurred after the vertical mark, since in most instances $P(A \mid B) \neq P(B \mid A)$. In the previous example $P(B \mid A)$ would represent the probability of prior work experience conditional on having found a job.

We again rely on the Venn diagram in Figure 4.6 to explain the conditional probability.

FIGURE 4.6
Finding the conditional probability, $P(A \mid B)$



Since $P(A \mid B)$ represents the probability of A conditional on B, the original sample space S reduces to B. The conditional probability $P(A \mid B)$ is based on the portion of A that is included in B. It is derived as the ratio of the probability of the intersection of A and B to the probability of B.

CONDITIONAL PROBABILITY

Given two events A and B, each with a positive probability of occurring, the probability that A occurs given that B has occurred (A conditioned on B) is equal to $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$. Similarly, the probability that B occurs given that A has occurred (B conditioned on A) is equal to $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$.

EXAMPLE 4.9

Economic globalization is defined as the integration of national economies into the international economy through trade, foreign direct investment, capital flows, migration, and the spread of technology. Although globalization is generally viewed favorably, it also increases the vulnerability of a country to economic conditions of the other country. An economist predicts a 60% chance that country A will perform poorly and a 25% chance that country B will perform poorly. There is also a 16% chance that both countries will perform poorly.

a. What is the probability that country A performs poorly given that country B performs poorly?

b. What is the probability that country B performs poorly given that country A performs poorly?

c. Interpret your findings.

SOLUTION: We first write down the available information in probability terms. Defining A as "country A performing poorly" and B as "country B performing poorly," we have the following information: $P(A) = 0.60$, $P(B) = 0.25$, and $P(A \cap B) = 0.16$.

a. $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.16}{0.25} = 0.64$

b. $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{0.16}{0.60} = 0.27$

c. It appears that globalization has definitely made these countries vulnerable to the economic woes of the other country. The probability that country A performs poorly increases from 60% to 64% when country B has performed poorly. Similarly, the probability that country B performs poorly increases from 25% to 27% when conditioned on country A performing poorly.

LO **4.3**

Distinguish between independent and dependent events.

Independent and Dependent Events

Of particular interest to researchers is whether or not two events influence one another. Two events are **independent** if the occurrence of one event does not affect the probability of the occurrence of the other event. Let's revisit the earlier example where the probability of finding a job is 0.80 and the probability of finding a job given prior work experience is 0.90. Prior work experience exerts a positive influence on finding a job because the conditional probability, $P(A \mid B) = 0.90$, exceeds the probability, $P(A) = 0.80$. Now consider the probability of finding a job given that your neighbor has bought a red car. Obviously, your neighbor's decision to buy a red car has no influence on your probability of finding a job, which remains at 0.80.

Events are considered **dependent** if the occurrence of one is related to the probability of the occurrence of the other. We determine the independence of two events by comparing the conditional probability of one event, for instance $P(A \mid B)$, to the probability, $P(A)$. If these two probabilities are the same, we say that the two events, A and B, are independent; if the probabilities differ, the two events are dependent.

INDEPENDENT VERSUS DEPENDENT EVENTS

Two events, A and B, are independent if $P(A \mid B) = P(A)$ or, equivalently, $P(B \mid A) = P(B)$. Otherwise, the events are dependent.

EXAMPLE 4.10

Suppose that for a given year there is a 2% chance that your desktop computer will crash and a 6% chance that your laptop computer will crash. Moreover, there is a 0.12% chance that both computers will crash. Is the reliability of the two computers independent of each other?

SOLUTION: Let event D represent the outcome that your desktop crashes and event L represent the outcome that your laptop crashes. Therefore, $P(D) = 0.02$, $P(L) = 0.06$, and $P(D \cap L) = 0.0012$. The reliability of the two computers is independent because

$$P(D \mid L) = \frac{P(D \cap L)}{P(L)} = \frac{0.0012}{0.06} = 0.02 = P(D).$$

In other words, if your laptop crashes, it does not alter the probability that your desktop also crashes. Equivalently,

$$P(L \mid D) = \frac{P(D \cap L)}{P(D)} = \frac{0.0012}{0.02} = 0.06 = P(L).$$

The Multiplication Rule

In some situations, we are interested in finding the probability that two events, A and B, both occur; that is, $P(A \cap B)$. In order to obtain this probability, we can rearrange the formula for conditional probability to derive $P(A \cap B)$. For instance, from $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$, we can easily derive $P(A \cap B) = P(A \mid B)P(B)$. Similarly, from $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$, we derive $P(A \cap B) = P(B \mid A)P(A)$. Since we calculate the product of two probabilities to find $P(A \cap B)$, we refer to it as the **multiplication rule** for probabilities.

THE MULTIPLICATION RULE

The multiplication rule states that the probability that A and B both occur is equal to the probability that A occurs given that B has occurred, times the probability that B occurs; that is, $P(A \cap B) = P(A \mid B)P(B)$. Equivalently, we can also arrive at this probability as $P(A \cap B) = P(B \mid A)P(A)$.

EXAMPLE 4.11

A stockbroker knows from past experience that the probability that a client owns stocks is 0.60 and the probability that a client owns bonds is 0.50. The probability that a client owns bonds if he/she already owns stocks is 0.55.

a. What is the probability that a client owns both of these securities?

b. Given that a client owns bonds, what is the probability that he/she owns stocks?

SOLUTION:

a. Let S correspond to the event that a client owns stocks and B correspond to the event that a client owns bonds. Thus, the probability that a client owns stocks is $P(S) = 0.60$ and the probability that a client owns bonds is $P(B) = 0.50$. The conditional probability that a client owns bonds given that he/she owns stocks is $P(B \mid S) = 0.55$. We calculate the probability that a client owns both of these securities as $P(S \cap B) = P(B \mid S)P(S) = 0.55 \times 0.60 = 0.33$.

b. We need to calculate the conditional probability that a client owns stocks given that he/she owns bonds, or $P(S \mid B)$. Using the formula for conditional probability and the answer from part a, we find $P(S \mid B) = \frac{P(S \cap B)}{P(B)} = \frac{0.33}{0.50} = 0.06$.

The Multiplication Rule for Independent Events

We know that two events, A and B, are independent if $P(A \mid B) = P(A)$. With independent events, the multiplication rule $P(A \cap B) = P(A \mid B)P(B)$ simplifies to $P(A \cap B) = P(A)P(B)$. We can also use this rule to determine whether or not two events are

independent. That is, two events are independent if the probability $P(A \cap B)$ equals the product of their probabilities, $P(A)P(B)$. In Example 4.10, we were given the probabilities $P(D) = 0.02$, $P(L) = 0.06$, and $P(D \cap L) = 0.0012$. Consistent with the earlier result, events D and L are independent because $P(D \cap L) = 0.0012$ equals $P(D)P(L) = 0.02 \times 0.06 = 0.0012$.

THE MULTIPLICATION RULE FOR INDEPENDENT EVENTS

If A and B are independent events, then the probability that A and B both occur equals the product of the probability of A and the probability of B; that is, $P(A \cap B) = P(A)P(B)$.

EXAMPLE 4.12

The probability of passing the Level 1 CFA (Chartered Financial Analyst) exam is 0.50 for John Campbell and 0.80 for Linda Lee. The prospect of John's passing the exam is completely unrelated to Linda's success on the exam.

a. What is the probability that both John and Linda pass the exam?

b. What is the probability that at least one of them passes the exam?

SOLUTION: We can write the probabilities that John passes the exam and that Linda passes the exam as $P(J) = 0.50$ and $P(L) = 0.80$, respectively.

a. Since we are told that John's chances of passing the exam are not influenced by Linda's success at the exam, we can conclude that these events are independent, so $P(J) = P(J \mid L) = 0.50$ and $P(L) = P(L \mid J) = 0.80$. Thus, when solving for the probability that both John and Linda pass the exam, we calculate the product of the probabilities: $P(J \cap L) = P(J)\ P(L) = 0.50 \times 0.80 = 0.40$.

b. We calculate the probability that at least one of them passes the exam as $P(J \cup L) = P(J) + P(L) - P(J \cap L) = 0.50 + 0.80 - 0.40 = 0.90$.

EXERCISES 4.2

Mechanics

14. Let $P(A) = 0.65$, $P(B) = 0.30$, and $P(A|B) = 0.45$.
    a. Calculate $P(A \cap B)$.
    b. Calculate $P(A \cup B)$.
    c. Calculate $P(B|A)$.
15. Let $P(A) = 0.55$, $P(B) = 0.30$, and $P(A \cap B) = 0.10$.
    a. Calculate $P(A|B)$.
    b. Calculate $P(A \cup B)$.
    c. Calculate $P((A \cup B)^C)$.
16. Let A and B be mutually exclusive events with $P(A) = 0.25$ and $P(B) = 0.30$.
    a. Calculate $P(A \cap B)$.
    b. Calculate $P(A \cup B)$.
    c. Calculate $P(A|B)$.
17. Let A and B be independent events with $P(A) = 0.40$ and $P(B) = 0.50$.
    a. Calculate $P(A \cap B)$.
    b. Calculate $P((A \cup B)^C)$.
    c. Calculate $P(A|B)$.
18. Let $P(A) = 0.65$, $P(B) = 0.30$, and $P(A|B) = 0.45$.
    a. Are A and B independent events? Explain.
    b. Are A and B mutually exclusive events? Explain.
    c. What is the probability that neither A nor B takes place?
19. Let $P(A) = 0.15$, $P(B) = 0.10$, and $P(A \cap B) = 0.05$.
    a. Are A and B independent events? Explain.
    b. Are A and B mutually exclusive events? Explain.
    c. What is the probability that neither A nor B takes place?

20. Consider the following probabilities: $P(A) = 0.25$, $P(B^c) = 0.40$, and $P(A \cap B) = 0.08$. Find:
    a. $P(B)$
    b. $P(A|B)$
    c. $P(B|A)$
21. Consider the following probabilities: $P(A^c) = 0.30$, $P(B) = 0.60$, and $P(A \cap B^c) = 0.24$. Find:
    a. $P(A|B^c)$
    b. $P(B^c|A)$
    c. Are A and B independent events? Explain.
22. Consider the following probabilities: $P(A) = 0.40$, $P(B) = 0.50$, and $P(A^c \cap B^c) = 0.24$. Find:
    a. $P(A^c|B^c)$
    b. $P(A^c \cup B^c)$
    c. $P(A \cup B)$

Applications

23. Survey data, based on 65,000 mobile phone subscribers, show that 44% of the subscribers use smartphones (*Forbes*, December 15, 2011). Moreover, 51% of smartphone users are women.
    a. Find the probability that a mobile phone subscriber is a woman who uses a smartphone.
    b. Find the probability that a mobile phone subscriber is a man who uses a smartphone.
24. Only 20% of students in a college ever go to their professor during office hours. Of those who go, 30% seek minor clarification and 70% seek major clarification.
    a. What is the probability that a student goes to the professor during her office hours for a minor clarification?
    b. What is the probability that a student goes to the professor during her office hours for a major clarification?
25. The probabilities that stock A will rise in price is 0.40 and that stock B will rise in price is 0.60. Further, if stock B rises in price, the probability that stock A will also rise in price is 0.50.
    a. What is the probability that at least one of the stocks will rise in price?
    b. Are events A and B mutually exclusive? Explain.
    c. Are events A and B independent? Explain.
26. Despite government bailouts and stimulus money, unemployment in the United States had not decreased significantly as economists had expected (*U.S. News & World Report*, July 2, 2010). Many analysts predicted only an 18% chance of a reduction in U.S. unemployment. However, if Europe slipped back into a recession, the probability of a reduction in U.S. unemployment would drop to 0.06.
    a. What is the probability that there is not a reduction in U.S. unemployment?
    b. Assume there is an 8% chance that Europe slips back into a recession. What is the probability that there is not a reduction in U.S. unemployment and that Europe slips into a recession?
27. Dr. Miriam Johnson has been teaching accounting for over 20 years. From her experience, she knows that 60% of her students do homework regularly. Moreover, 95% of the students who do their homework regularly pass the course. She also knows that 85% of her students pass the course.
    a. What is the probability that a student will do homework regularly and also pass the course?
    b. What is the probability that a student will neither do homework regularly nor will pass the course?
    c. Are the events "pass the course" and "do homework regularly" mutually exclusive? Explain.
    d. Are the events "pass the course" and "do homework regularly" independent? Explain.
28. Records show that 5% of all college students are foreign students who also smoke. It is also known that 50% of all foreign college students smoke. What percent of the students at this university are foreign?
29. An analyst estimates that the probability of default on a seven-year AA-rated bond is 0.06, while that on a seven-year A-rated bond is 0.13. The probability that they will both default is 0.04.
    a. What is the probability that at least one of the bonds defaults?
    b. What is the probability that neither the seven-year AA-rated bond nor the seven-year A-rated bond defaults?
    c. Given that the seven-year AA-rated bond defaults, what is the probability that the seven-year A-rated bond also defaults?
30. Mike Danes has been delayed in going to the annual sales event at one of his favorite apparel stores. His friend has just texted him that there are only 20 shirts left, of which 8 are in size M, 10 in size L, and 2 in size XL. Also 9 of the shirts are white, 5 are blue, and the remaining are of mixed colors. Mike is interested in getting a white or a blue shirt in size L. Define the events A = Getting a white or a blue shirt and B = Getting a shirt in size L.
    a. Find $P(A)$, $P(A^c)$, and $P(B)$.
    b. Are the events A and B mutually exclusive? Explain.
    c. Would you describe Mike's preference by the events $A \cup B$ or $A \cap B$?
31. In general, shopping online is supposed to be more convenient than going to stores. However, according to a Harris Interactive poll, 87% of people have experienced problems with an online transaction (*The Wall Street Journal*, October 2, 2007). Forty-two percent of people who

experienced a problem abandoned the transaction or switched to a competitor's website. Fifty-three percent of people who experienced problems contacted customer-service representatives.

a. What proportion of people did not experience problems with an online transaction?
b. What proportion of people experienced problems with an online transaction and abandoned the transaction or switched to a competitor's website?
c. What proportion of people experienced problems with an online transaction and contacted customer-service representatives?

32. A manufacturing firm just received a shipment of 20 assembly parts, of slightly varied sizes, from a vendor. The manager knows that there are only 15 parts in the shipment that would be suitable. He examines these parts one at a time.
    a. Find the probability that the first part is suitable.
    b. If the first part is suitable, find the probability that the second part is also suitable.
    c. If the first part is suitable, find the probability that the second part is not suitable.
33. Apple products have become a household name in America, with 51% of all households owning at least one Apple product (*CNN*, March 19, 2012). In the Midwest, the likelihood of owning an Apple product is 61% for households with kids and 48% for households without kids. Suppose there are 1,200 households in a representative community, of which 820 are with kids and the rest are without kids.
    a. Are the events "household with kids" and "household without kids" mutually exclusive and exhaustive? Explain.
    b. What is the probability that a household is without kids?
    c. What is the probability that a household is with kids and owns an Apple product?
    d. What is the probability that a household is without kids and does not own an Apple product?
34. Despite the repeated effort by the government to reform how Wall Street pays its executives, some of the nation's biggest banks are continuing to pay out bonuses nearly as large as those in the best years before the Great Recession (*The Washington Post*, January 15, 2010). It is known that 10 out of 15 members of the board of directors of a company were in favor of the bonus. Suppose two members were randomly selected by the media.
    a. What is the probability that both of them were in favor of the bonus?
    b. What is the probability that neither of them was in favor of the bonus?
35. Christine Wong has asked Dave and Mike to help her move into a new apartment on Sunday morning. She has asked them both, in case one of them does not show up. From past experience, Christine knows that there is a 40% chance that Dave will not show up and a 30% chance that Mike will not show up. Dave and Mike do not know each other and their decisions can be assumed to be independent.
    a. What is the probability that both Dave and Mike will show up?
    b. What is the probability that at least one of them will show up?
    c. What is the probability that neither Dave nor Mike will show up?
36. According to the Census's Population Survey, the percentage of children with two parents at home is the highest for Asians and lowest for blacks (*USA TODAY*, February 26, 2009). It is reported that 85% of Asian, 78% of white, 70% of Hispanic, and 38% of black children have two parents at home. Suppose there are 500 students in a representative school, of which 280 are white, 50 are Asian, 100 are Hispanic, and 70 are black.
    a. Are the events "Asians" and "black" mutually exclusive and exhaustive? Explain.
    b. What is the probability that a child is not white?
    c. What is the probability that a child is white and has both parents at home?
    d. What is the probability that a child is Asian and does not have both parents at home?
37. A study shows that unemployment does not impact white-collar and blue-collar workers equally (*Newsweek*, April 20, 2009). According to the Bureau of Labor Statistics report, while the national unemployment rate is 8.5%, it is only 4.3% for those with a college degree. It is fair to assume that 27% of people in the labor force are college educated. You have just heard that another worker in a large firm has been laid off. What is the probability that the worker is college educated?
38. According to a survey by two United Nations agencies and a nongovernmental organization, two in every three women in the Indian capital of New Delhi are likely to face some form of sexual harassment in a year (*BBC World News*, July 9, 2010). The study also reports that women who use public transportation are especially vulnerable. Suppose the corresponding probability of harassment for women who use public transportation is 0.82. It is also known that 28% of women use public transportation.
    a. What is the probability that a woman takes public transportation and also faces sexual harassment?
    b. If a woman is sexually harassed, what is the probability that she had taken public transportation?
39. According to results from the Spine Patient Outcomes Research Trial, or SPORT, surgery for a painful, common back condition resulted in significantly reduced back pain and better physical function than treatment with drugs and physical therapy (*The Wall Street Journal*, February 21, 2008). SPORT followed 803 patients, of whom 398 ended up getting surgery. After two years, of those who had surgery, 63% said they had a major

improvement in their condition, compared with 29% among those who received nonsurgical treatment.

a. What is the probability that a patient had surgery? What is the probability that a patient did not have surgery?
b. What is the probability that a patient had surgery and experienced a major improvement in his or her condition?
c. What is the probability that a patient received nonsurgical treatment and experienced a major improvement in his or her condition?

40. A study challenges the media narrative that foreclosures are dangerously widespread (*The New York Times*, March 2, 2009). According to this study, 62% of all foreclosures were centered in only four states, namely, Arizona, California, Florida, and Nevada. The national average rate of foreclosures in 2008 was 0.79%. What percent of the homes in the United States were foreclosed in 2008 and also centered in Arizona, California, Florida, or Nevada?

4.3 CONTINGENCY TABLES AND PROBABILITIES

LO 4.4

Calculate and interpret probabilities from a contingency table.

In Chapter 2 we learned that it is useful to construct a frequency distribution when organizing qualitative data. While it is true that a frequency distribution is an effective tool for summarizing one variable, it is often the case that we want to examine or compare two qualitative variables. On these occasions, a **contingency table** proves very useful. Contingency tables are widely used in marketing and biomedical research, as well as in the social sciences.

A CONTINGENCY TABLE

A contingency table generally shows frequencies for two qualitative (categorical) variables, x and y, where each cell represents a mutually exclusive combination of the pair of x and y values.

Table 4.4, first presented in the introductory case study of this chapter, is an example of a contingency table where the qualitative variables of interest, x and y, are Age Group and Brand Name, respectively. Age Group has two possible categories: (1) under 35 years and (2) 35 years and older; Brand Name, has three possible categories: (1) Under Armour, (2) Nike, and (3) Adidas.

TABLE 4.4 Purchases of Compression Garments Based on Age and Brand Name

	Brand Name		
Age Group	**Under Armour**	**Nike**	**Adidas**
Under 35 years	174	132	90
35 years and older	54	72	78

Each cell in Table 4.4 represents a frequency; for example, there are 174 customers under the age of 35 who purchase an Under Armour product, whereas there are 54 customers at least 35 years old who purchase an Under Armour product. Recall that we estimate an empirical probability by calculating the relative frequency of the occurrence of the event. To make calculating these probabilities less cumbersome, it is often useful to denote each event with letter notation and calculate totals for each column and row as shown in Table 4.4a.

TABLE 4.4a A Contingency Table Labeled Using Event Notation

	Brand Name			
Age Group	B_1	B_2	B_3	**Total**
A	174	132	90	396
A^c	54	72	78	204
Total	228	204	168	600

Thus, let events A and A^c correspond to "under 35 years" and "35 years and older," respectively; similarly, let events B_1, B_2, and B_3 correspond to "Under Armour," "Nike," and "Adidas," respectively. In addition, after calculating row totals, it is now easier to recognize that 396 of the customers are under 35 years old and 204 of the customers are at least 35 years old. Similarly, column totals indicate that 228 customers purchase Under Armour, 204 purchase Nike, and 168 purchase Adidas. Finally, the frequency corresponding to the cell in the last column and the last row is 600. This value represents the total number of customers in the sample. We arrive at this value by either summing the values in the last column (396 + 204) or summing the values in the last row (228 + 204 + 168).

The following example illustrates how to calculate probabilities when the data are presented in the form of a contingency table.

EXAMPLE 4.13

Using the information in Table 4.4a, answer the following questions.

a. What is the probability that a randomly selected customer is younger than 35 years old?

b. What is the probability that a randomly selected customer purchases an Under Armour garment?

c. What is the probability that a customer is younger than 35 years old and purchases an Under Armour garment?

d. What is the probability that a customer is either younger than 35 years old or purchases an Under Armour garment?

e. What is the probability that a customer is under 35 years of age, given that the customer purchases an Under Armour garment?

SOLUTION:

a. $P(A) = \frac{396}{600} = 0.66$; there is a 66% chance that a randomly selected customer is less than 35 years old.

b. $P(B_1) = \frac{228}{600} = 0.38$; there is a 38% chance that a randomly selected customer purchases an Under Armour garment.

c. $P(A \cap B_1) = \frac{174}{600} = 0.29$; there is a 29% chance that a randomly selected customer is younger than 35 years old and purchases an Under Armour garment.

d. $P(A \cup B_1) = \frac{174 + 132 + 90 + 54}{600} = \frac{450}{600} = 0.75$; there is a 75% chance that a randomly selected customer is either younger than 35 years old or purchases an Under Armour garment. Alternatively, we can use the addition rule to solve this problem as $P(A \cup B_1) = P(A) + P(B_1) - P(A \cap B_1) = 0.66 + 0.38 - 0.29 = 0.75$.

e. We wish to calculate the conditional probability, $P(A \mid B_1)$. When the information is in the form of a contingency table, calculating a conditional probability is rather straightforward. We are given the information that the customer purchases an Under Armour garment, so the sample size shrinks from 600 customers to 228 customers. We can ignore all customers that make Nike or Adidas purchases, or all outcomes in events B_2 and B_3. Thus, of the 228 customers who make an Under Armour purchase, 174 of them are under 35 years of age. Therefore, the probability that a customer is under 35 years of age given that the customer makes an Under Armour purchase is calculated as $P(A \mid B_1) = \frac{174}{228} = 0.76$. Alternatively, we can use the conditional probability formula to solve the problem as $P(A \mid B_1) = \frac{P(A \cap B_1)}{P(B_1)} = \frac{174/600}{228/600} = \frac{174}{228} = 0.76$.

Arguably, a more convenient way of calculating relevant probabilities is to convert the contingency table to a **joint probability table**. The frequency in each cell is divided by the number of outcomes in the sample space, which in this example is 600 customers. Table 4.4b shows the results.

TABLE 4.4b Converting a Contingency Table to a Joint Probability Table

	Brand Name			
Age Group	B_1	B_2	B_3	Total
A	0.29	0.22	0.15	0.66
A^c	0.09	0.12	0.13	0.34
Total	0.38	0.34	0.28	1.00

The values in the interior of the table represent the probabilities of the intersection of two events, also referred to as **joint probabilities**. For instance, the probability that a randomly selected person is under 35 years of age and makes an Under Armour purchase, denoted $P(A \cap B_1)$, is 0.29. Similarly, we can readily read from this table that 12% of the customers purchase a Nike garment and are at least 35 years old, or $P(A^c \cap B_2) = 0.12$.

The values in the margins of Table 4.4b represent unconditional probabilities. These probabilities are also referred to as **marginal probabilities**. For example, the probability that a randomly selected customer is under 35 years of age, $P(A)$, is simply 0.66. Also, the probability of purchasing a Nike garment, $P(B_2)$, is 0.34.

Note that the conditional probability is basically the ratio of a joint probability to an unconditional probability. Since $P(A \mid B_1) = \frac{P(A \cap B_1)}{P(B_1)}$, the numerator is the joint probability, $P(A \cap B_1)$, and the denominator is the unconditional probability, $P(B_1)$. Let's refer back to the probability that we calculated earlier; that is, the probability that a customer is under 35 years of age, given that the customer purchases an Under Armour product. This conditional probability is easily computed as $P(A \mid B_1) = \frac{P(A \cap B_1)}{P(B_1)} = \frac{0.29}{0.38} = 0.76$.

EXAMPLE 4.14

Given the information in Table 4.4b, what is the probability that a customer purchases an Under Armour product, given that the customer is under 35 years of age?

SOLUTION: Now we are solving for $P(B_1 \mid A)$. So

$$P(B_1 \mid A) = \frac{P(A \cap B_1)}{P(A)} = \frac{0.29}{0.66} = 0.44.$$

Note that $P(B_1 \mid A) = 0.44 \neq P(A \mid B_1) = 0.76$.

EXAMPLE 4.15

Determine whether the events "under 35 years old" and "Under Armour" are independent.

SOLUTION: In order to determine whether two events are independent, we compare an event's conditional probability to its unconditional probability; that is, events A and B are independent if $P(A \mid B) = P(A)$. In the Under Armour example, we have already found that $P(A \mid B_1) = 0.76$. In other words, there is a 76% chance that a customer is under 35 years old given that the customer purchases an Under Armour product. We compare this conditional probability to its unconditional probability, $P(A) = 0.66$. Since these probabilities differ, the events "under 35 years old" and "Under Armour" are not independent events. We could have compared $P(B_1 \mid A)$ to $P(B_1)$ and found

that $0.44 \neq 0.38$, which leads us to the same conclusion that the events are dependent. As discussed in the preceding section, an alternative approach is to compare the joint probability with the product of the two unconditional probabilities. Events are independent if $P(A \cap B_1) = P(A)P(B_1)$. In this example, $P(A \cap B_1) = 0.29$ does not equal $P(A)P(B_1) = 0.66 \times 0.38 = 0.25$, so the two events are not independent.

A Note on Independence

It is important to note that the conclusions about independence, such as the one made in Example 4.15, are informal since they are based on empirical probabilities computed from given sample information. In the preceding example, these probabilities will change if a different sample of 600 customers is used. Formal tests of independence are discussed in Chapter 11.

SYNOPSIS OF INTRODUCTORY CASE



©Digital Vision/Photodisc/Getty Images

After careful analysis of the contingency table representing customer purchases of compression garments based on age and brand name, several interesting remarks can be made. From a sample of 600 customers, it appears that the majority of the customers who purchase these products tend to be younger: 66% of the customers were younger than 35 years old, whereas 34% were at least 35 years old. It is true that more customers chose to purchase Under Armour garments (with 38% of purchases) as compared to Nike or Adidas garments (with 34% and 28% of purchases, respectively). However, given that Under Armour was the pioneer in the compression-gear market, this company should be concerned with the competition posed by Nike and Adidas. Further inspection of the contingency table reveals that if a customer was under 35 years old, the chance of the customer purchasing an Under Armour garment rises to about 44%. This result indicates that the age of a customer seems to influence the brand name purchased. In other words, 38% of the customers choose to buy Under Armour products, but as soon as the attention is confined to those customers who are under 35 years old, the likelihood of a purchase from Under Armour rises to about 44%. The information that the Under Armour brand appeals to younger customers is relevant not only to Under Armour and how the firm may focus its advertising efforts, but also to competitors and retailers in the compression garment market.

EXERCISES 4.3

Mechanics

41. Consider the following contingency table.

	B	B^c
A	26	34
A^c	14	26

a. Convert the contingency table into a joint probability table.
b. What is the probability that A occurs?
c. What is the probability that A and B occur?
d. Given that B has occurred, what is the probability that A occurs?
e. Given that A^c has occurred, what is the probability that B occurs?
f. Are A and B mutually exclusive events? Explain.
g. Are A and B independent events? Explain.

42. Consider the following joint probability table.

	B_1	B_2	B_3	B_4
A	0.09	0.22	0.15	0.20
A^c	0.03	0.10	0.09	0.12

a. What is the probability that A occurs?
b. What is the probability that B_2 occurs?

c. What is the probability that A^c and B_4 occur?
d. What is the probability that A or B_3 occurs?
e. Given that B_2 has occurred, what is the probability that A occurs?
f. Given that A has occurred, what is the probability that B_4 occurs?

Applications

43. According to an online survey by Harris Interactive for job site CareerBuilder.com, more than half of IT (information technology) workers say they have fallen asleep at work (*InformationWeek*, September 27, 2007). Sixty-four percent of government workers admitted to falling asleep on the job. Consider the following contingency table that is representative of the survey results.

	Job Category	
Slept on the Job?	IT Professional	Government Professional
Yes	155	256
No	145	144

a. Convert the contingency table into a joint probability table.
b. What is the probability that a randomly selected worker is an IT professional?
c. What is the probability that a randomly selected worker slept on the job?
d. If a randomly selected worker slept on the job, what is the probability that he/she is an IT professional?
e. If a randomly selected worker is a government professional, what is the probability that he/she slept on the job?
f. Are the events "IT Professional" and "Slept on the Job" independent? Explain using probabilities.

44. A report suggests that business majors spend the least amount of time on course work than all other college students (*The New York Times*, November 17, 2011). A provost of a university decides to conduct a survey where students are asked if they study hard, defined by spending at least 20 hours per week on course work. Of 120 business majors included in the survey, 20 said that they studied hard, as compared to 48 out of 150 nonbusiness majors who said that they studied hard.
a. Construct a contingency table that shows the frequencies for the qualitative variables Major (business or nonbusiness) and Study Hard (yes or no).
b. Find the probability that a business major spends less than 20 hours per week on course work.
c. What is the probability that a student studies hard?
d. If a student spends at least 20 hours on course work, what is the probability that he/she is a business major? What is the corresponding probability that he/she is a nonbusiness major?

45. A poll asked 16- to 21-year-olds whether or not they are likely to serve in the U.S. military. The following joint probability table, cross-classified by gender and race, reports the proportion of those polled who responded that they are likely or very likely to serve in the active-duty military.

	Race		
Gender	Hispanic	Black	White
Male	0.335	0.205	0.165
Female	0.145	0.105	0.045

Source: Defense Human Resources Activity telephone poll of 3,228 Americans conducted October through December 2005.

a. What is the probability that a randomly selected respondent is female?
b. What is the probability that a randomly selected respondent is Hispanic?
c. Given that a respondent is female, what is the probability that she is Hispanic?
d. Given that a respondent is white, what is the probability that the respondent is male?
e. Are the events "Male" and "White" independent? Explain using probabilities.

46. According to a Michigan State University researcher, Americans are becoming increasingly polarized on issues pertaining to the environment (msutoday.msu.edu, April 19, 2011). It is reported that 70% of Democrats see signs of global warming as compared to only 30% of Republicans who feel the same. Suppose the survey was based on 400 Democrats and 400 Republicans.
a. Construct a contingency table that shows frequencies for the qualitative variables Political Affiliation (Democrat or Republican) and Global Warming (yes or no).
b. Find the probability that a Republican sees signs of global warming.
c. Find the probability that a person does not see signs of global warming.
d. If a person sees signs of global warming, what is the probability that this person is a Democrat?

47. Merck & Co. conducted a study to test the promise of its experimental AIDS vaccine (*The Boston Globe*, September 22, 2007). Volunteers in the study were all free of the human immunodeficiency virus (HIV), which causes AIDS, at the start of the study, but all were at high risk for getting the virus. Volunteers were given either the vaccine or a dummy shot; 24 of 741 volunteers who got the vaccine became infected with HIV, whereas 21 of 762 volunteers who got the dummy shot became infected with HIV. The following table summarizes the results of the study.

	Vaccinated	Dummy Shot
Infected	24	21
Not Infected	717	741

a. Convert the contingency table into a joint probability table.
b. What is the probability that a randomly selected volunteer got vaccinated?
c. What is the probability that a randomly selected volunteer became infected with the HIV virus?
d. If the randomly selected volunteer was vaccinated, what is the probability that he/she got infected?
e. Are the events "Vaccinated" and "Infected" independent? Explain using probabilities. Given your answer, is it surprising that Merck & Co. ended enrollment and vaccination of volunteers in the study? Explain.

48. More and more households are struggling to pay utility bills given high heating costs (*The Wall Street Journal*, February 14, 2008). Particularly hard hit are households with homes heated with propane or heating oil. Many of these households are spending twice as much to stay warm this winter compared to those who heat with natural gas or electricity. A representative sample of 500 households was taken to investigate if the type of heating influences whether or not a household is delinquent in paying its utility bill. The following table reports the results.

Delinquent in Payment?	Type of Heating			
	Natural Gas	Electricity	Heating Oil	Propane
Yes	50	20	15	10
No	240	130	20	15

a. What is the probability that a randomly selected household uses heating oil?
b. What is the probability that a randomly selected household is delinquent in paying its utility bill?
c. What is the probability that a randomly selected household uses heating oil and is delinquent in paying its utility bill?
d. Given that a household uses heating oil, what is the probability that it is delinquent in paying its utility bill?
e. Given that a household is delinquent in paying its utility bill, what is the probability that the household uses electricity?
f. Are the events "Heating Oil" and "Delinquent in Payment" independent? Explain using probabilities.

49. The research team at a leading perfume company is trying to test the market for its newly introduced perfume. In particular the team wishes to look for gender and international differences in the preference for this perfume. They sample 2,500 people internationally and each person in the sample is asked to try the new perfume and list his/her preference. The following table reports the results.

Preference	Gender	America	Europe	Asia
Like it	Men	210	150	120
	Women	370	310	180
Don't like it	Men	290	150	80
	Women	330	190	120

a. What is the probability that a randomly selected man likes the perfume?
b. What is the probability that a randomly selected Asian likes the perfume?
c. What is the probability that a randomly selected European woman does not like the perfume?
d. What is the probability that a randomly selected American man does not like the perfume?
e. Are the events "Men" and "Like Perfume" independent in (i) America, (ii) Europe, and (iii) Asia? Explain using probabilities.
f. Internationally, are the events "Men" and "Like Perfume" independent? Explain using probabilities.

4.4 THE TOTAL PROBABILITY RULE AND BAYES' THEOREM

In this section, we present two important rules in probability theory: the total probability rule and Bayes' theorem. The **total probability rule** is a useful tool for breaking the computation of a probability into distinct cases. **Bayes' theorem** uses this rule to update the probability of an event that has been affected by a new piece of evidence.

LO **4.5**

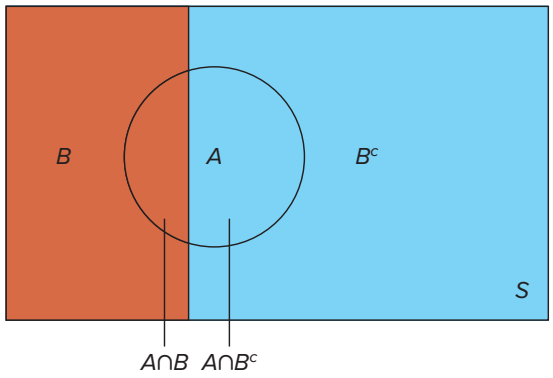
Apply the total probability rule.

The Total Probability Rule

Sometimes the probability of an event is not readily apparent from the given information. The total probability rule expresses the probability of an event in terms of joint or conditional probabilities. Let $P(A)$ denote the probability of an event of interest. We can express $P(A)$ as the sum of probabilities of the intersections of A with some mutually exclusive and exhaustive events corresponding to an experiment. For instance, consider

event B and its complement B^c. Figure 4.7 shows the sample space partitioned entirely into these two mutually exclusive and exhaustive events. The circle, representing event A, consists entirely of its intersections with B and B^c. According to the total probability rule, $P(A)$ equals the sum of $P(A \cap B)$ and $P(A \cap B^c)$.

FIGURE 4.7
The total probability rule: $P(A) = P(A \cap B) + P(A \cap B^c)$



Oftentimes the joint probabilities needed to compute the total probability are not explicitly specified. Therefore, we use the multiplication rule to derive these probabilities from the conditional probabilities as $P(A \cap B) = P(A \mid B)P(B)$ and $P(A \cap B^c) = P(A \mid B^c)P(B^c)$.

THE TOTAL PROBABILITY RULE CONDITIONAL ON TWO EVENTS

The total probability rule expresses the probability of an event, A, in terms of probabilities of the intersection of A with any mutually exclusive and exhaustive events. The total probability rule based on two events, B and B^c, is

$$P(A) = P(A \cap B) + P(A \cap B^c),$$

or equivalently,

$$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c).$$

An intuitive way to express the total probability rule is with the help of a **probability tree**. Whenever an experiment can be broken down into stages, with a different aspect of the result observed at each stage, we can use a probability tree to represent the various possible sequences of observations. We also use an alternative tabular method for computing the probability $P(A)$. The following example illustrates the mechanics of a probability tree and the tabular method.

EXAMPLE 4.16

Even though a certain statistics professor does not require attendance as part of a student's overall grade, she has noticed that those who regularly attend class have a higher tendency to get a final grade of A. The professor calculates that there is an 80% chance that a student attends class regularly. Moreover, given that a student attends class regularly, there is a 35% chance that the student receives an A grade; however, if a student does not attend class regularly, there is only a 5% chance of an A grade. Use this information to answer the following questions.

a. What is the probability that a student does not attend class regularly?

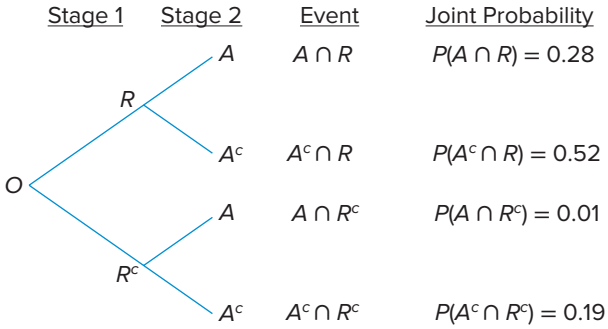
b. What is the probability that a student attends class regularly and receives an A grade?

c. What is the probability that a student does not attend class regularly and receives an A grade?

d. What is the probability that a student receives an A grade?

SOLUTION: We first let A correspond to the event that a student receives an A grade and R correspond to the event that a student attends class regularly. From the preceding information, we then have the following probabilities: $P(R) = 0.80$, $P(A \mid R) = 0.35$, and $P(A \mid R^c) = 0.05$. Figure 4.8 shows a probability tree that consists of nodes (junctions) and branches (lines) where the initial node O is called the origin. The branches emanating from O represent the possible outcomes that may occur at the first stage. Thus, at stage 1 we have events R and R^c originating from O. These events become the nodes at the second stage. The sum of the probabilities coming from any particular node is equal to one.

FIGURE 4.8 Probability tree for class attendance and final grade in statistics



a. Using the complement rule, if we know that there is an 80% chance that a student attends class regularly, $P(R) = 0.80$, then the probability that a student does not attend class regularly is found as $P(R^c) = 1 - P(R) = 1 - 0.80 = 0.20$.

In order to arrive at a subsequent stage, and deduce the corresponding probabilities, we use the information obtained from the previous stage. For instance, given that a student attends class regularly, there is a 35% chance that the student receives an A grade; that is, $P(A \mid R) = 0.35$. Given that a student regularly attends class, the likelihood of not receiving an A grade is 65% because $P(A^c \mid R) = 1 - P(A \mid R) = 0.65$. Similarly, given $P(A \mid R^c) = 0.05$, we compute $P(A^c \mid R^c) = 1 - P(A \mid R^c) = 1 - 0.05 = 0.95$. Any path through branches of the tree from the origin to a terminal node defines the intersection of the earlier two events. Thus, following the top branches, we arrive at the event $A \cap R$, meaning that a student attends class regularly and receives an A grade. The probability of this event is the product of the probabilities attached to the branches forming that path; here we are simply applying the multiplication rule. Now we are prepared to answer parts b and c.

b. Multiplying the probabilities attached to the top branches, we obtain $P(A \cap R) = P(A \mid R)P(R) = 0.35 \times 0.80 = 0.28$; there is a 28% chance that a student attends class regularly and receives an A grade.

c. In order to find the probability that a student does not attend class regularly and receives an A grade, we compute $P(A \cap R^c) = P(A \mid R^c)$ $P(R^c) = 0.05 \times 0.20 = 0.01$.

d. The probability that a student receives an A grade, $P(A)$, is not explicitly given in Example 4.16. However, we can sum the relevant joint probabilities in parts b and c to obtain this probability:

$$P(A) = P(A \cap R) + P(A \cap R^c) = 0.28 + 0.01 = 0.29.$$

An alternative method uses a tabular representation of probabilities. Table 4.5 contains all relevant probabilities that are directly or indirectly specified in Example 4.16.

TABLE 4.5 Tabular Method for Computing $P(A)$

Unconditional Probability	Conditional Probability	Joint Probability
$P(R) = 0.80$	$P(A \mid R) = 0.35$	$P(A \cap R) = P(A \mid R)\, P(R) = 0.28$
$P(R^c) = 0.20$	$P(A \mid R^c) = 0.05$	$P(A \cap R^c) = P(A \mid R^c)\, P(R^c) = 0.01$
$P(R) + P(R^c) = 1$		$P(A) = P(A \cap R) + P(A \cap R^c) = 0.29$

As we saw earlier, each joint probability is computed as a product of its conditional probability and the corresponding unconditional probability; that is, $P(A \cap R) = P(A \mid R)P(R) = 0.35 \times 0.80 = 0.28$. Similarly, $P(A \cap R^c) = P(A \mid R^c)P(R^c) = 0.05 \times 0.20 = 0.01$. Therefore, $P(A) = P(A \cap R) + P(A \cap R^c) = 0.29$.

Bayes' Theorem

LO **4.6**

Apply Bayes' theorem.

The total probability rule is also needed to derive Bayes' theorem, developed by the Reverend Thomas Bayes (1702–1761). Bayes' theorem is a procedure for updating probabilities based on new information. The original probability is an unconditional probability called a **prior probability**, in the sense that it reflects only what we know now before the arrival of any new information. On the basis of new information, we update the prior probability to arrive at a conditional probability called a **posterior probability**.

Suppose we know that 99% of the individuals who take a lie detector test tell the truth. Therefore, the prior probability of telling the truth is 0.99. Suppose an individual takes the lie detector test and the results indicate that the individual lied. Bayes' theorem updates a prior probability to compute a posterior probability, which in the above example is essentially a conditional probability based on the information that the lie detector has detected a lie.

Let $P(B)$ denote the prior probability and $P(B \mid A)$ the posterior probability. Note that the posterior probability is conditional on event A, representing new information. Recall the conditional probability formula from Section 4.2:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}.$$

In some instances, we may have to evaluate $P(B \mid A)$, but we do not have explicit information on $P(A \cap B)$ or $P(A)$. However, given information on $P(B)$, $P(A \mid B)$, and $P(A \mid B^c)$, we can use the total probability rule and the multiplication rule to find $P(B \mid A)$ as follows:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^c)} = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)}.$$

BAYES' THEOREM

The posterior probability $P(B \mid A)$ can be found using the information on the prior probability $P(B)$ along with the conditional probabilities $P(A \mid B)$ and $P(A \mid B^c)$ as

$$P(B \mid A) = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^c)} = \frac{P(A \mid B)\, P(B)}{P(A \mid B)\, P(B) + P(A \mid B^c)\, P(B^c)}.$$

In the above formula, we have used Bayes' theorem to update the prior probability $P(B)$ to the posterior probability $P(B \mid A)$. Equivalently, we can use Bayes' theorem to update the prior probability $P(A)$ to derive the posterior probability $P(A \mid B)$ by interchanging the events A and B in the above formula.

EXAMPLE 4.17

In a lie-detector test, an individual is asked to answer a series of questions while connected to a polygraph (lie detector). This instrument measures and records several physiological responses of the individual on the basis that false answers will produce distinctive measurements. Assume that 99% of the individuals who go in for a polygraph test tell the truth. These tests are considered to be 95% reliable. In other words, there is a 95% chance that the test will detect a lie if an individual actually lies. Let there also be a 0.5% chance that the test erroneously detects a lie even when the individual is telling the truth. An individual has just taken a polygraph test and the test has detected a lie. What is the probability that the individual was actually telling the truth?

SOLUTION: First we define some events and their associated probabilities. Let D and T correspond to the events that the polygraph detects a lie and that an individual is telling the truth, respectively. We are given that $P(T) = 0.99$, implying that $P(T^c) = 1 - 0.99 = 0.01$. In addition, we formulate $P(D \mid T^c) = 0.95$ and $P(D \mid T) = 0.005$. We need to find $P(T \mid D)$ when we are not explicitly given $P(D \cap T)$ and $P(D)$. We can use Bayes' theorem to find

$$P(T \mid D) = \frac{P(D \cap T)}{P(D)} = \frac{P(D \cap T)}{P(D \cap T) + P(D \cap T^c)} = \frac{P(D \mid T)P(T)}{P(D \mid T)P(T) + P(D \mid T^c)P(T^c)}.$$

Although we can use this formula to solve the problem directly, it is often easier to solve it systematically with the help of the following table.

TABLE 4.6 Computing Posterior Probabilities for Example 4.17

Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
$P(T) = 0.99$	$P(D \mid T) = 0.005$	$P(D \cap T) = 0.00495$	$P(T \mid D) = 0.34256$
$P(T^c) = 0.01$	$P(D \mid T^c) = 0.95$	$P(D \cap T^c) = 0.00950$	$P(T^c \mid D) = 0.65744$
$P(T) + P(T^c) = 1$		$P(D) = 0.01445$	$P(T \mid D) + P(T^c \mid D) = 1$

The first column presents prior probabilities and the second column shows related conditional probabilities. We first compute the denominator of Bayes' theorem by using the total probability rule, $P(D) = P(D \cap T) + P(D \cap T^c)$. Joint probabilities are calculated as products of conditional probabilities with their corresponding prior probabilities. For instance, in Table 4.6, in order to obtain $P(D \cap T)$, we multiply $P(D \mid T)$ with $P(T)$, which yields $P(D \cap T) = 0.005 \times 0.99 = 0.00495$. Similarly, we find $P(D \cap T^c) = 0.95 \times 0.01 = 0.00950$. Thus, according to the total probability rule, $P(D) = 0.00495 + 0.00950 = 0.01445$. Finally, $P(T \mid D) = \frac{P(D \cap T)}{P(D \cap T) + P(D \cap T^c)} = \frac{0.00495}{0.01445} = 0.34256$. The prior probability of an individual telling the truth is 0.99. However, given the new information that the polygraph detected the individual telling a lie, the posterior probability of this individual telling the truth is now revised downward to 0.34256.

So far we have used the total probability rule as well as Bayes' theorem based on two mutually exclusive and exhaustive events, namely, B and B^c. We can easily extend the analysis to include n mutually exclusive and exhaustive events, $B_1, B_2, \ldots, B_n$.

EXTENSIONS OF THE TOTAL PROBABILITY RULE AND BAYES' THEOREM

If $B_1, B_2, \ldots B_n$ represent n mutually exclusive and exhaustive events, then the total probability rule extends to

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n),$$

or equivalently,

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n).$$

Similarly, Bayes' theorem, for any $i = 1, 2, \ldots, n$, extends to

$$P(B_i \mid A) = \frac{P(A \cap B_i)}{P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)},$$

or equivalently,

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)\,P(B_1) + P(A \mid B_2)\,P(B_2) + \cdots + P(A \mid B_n)P(B_n)}.$$

EXAMPLE 4.18

Scott Myers is a security analyst for a telecommunications firm called Webtalk. Although he is optimistic about the firm's future, he is concerned that its stock price will be considerably affected by the condition of credit flow in the economy. He believes that the probability is 0.20 that credit flow will improve significantly, 0.50 that it will improve only marginally, and 0.30 that it will not improve at all. He also estimates that the probability that the stock price of Webtalk will go up is 0.90 with significant improvement in credit flow in the economy, 0.40 with marginal improvement in credit flow in the economy, and 0.10 with no improvement in credit flow in the economy.

a. Based on Scott's estimates, what is the probability that the stock price of Webtalk goes up?

b. If we know that the stock price of Webtalk has gone up, what is the probability that credit flow in the economy has improved significantly?

SOLUTION: As always, we first define the relevant events and their associated probabilities. Let S, M, and N denote significant, marginal, and no improvement in credit flow, respectively. Then $P(S) = 0.20$, $P(M) = 0.50$, and $P(N) = 0.30$. In addition, if we allow G to denote an increase in stock price, we formulate $P(G \mid S) = 0.90$, $P(G \mid M) = 0.40$, and $P(G \mid N) = 0.10$. We need to calculate $P(G)$ in part a and $P(S \mid G)$ in part b. Table 4.7 aids in assigning probabilities.

TABLE 4.7 Computing Posterior Probabilities for Example 4.18

Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
$P(S) = 0.20$	$P(G \mid S) = 0.90$	$P(G \cap S) = 0.18$	$P(S \mid G) = 0.4390$
$P(M) = 0.50$	$P(G \mid M) = 0.40$	$P(G \cap M) = 0.20$	$P(M \mid G) = 0.4878$
$P(N) = 0.30$	$P(G \mid N) = 0.10$	$P(G \cap N) = 0.03$	$P(N \mid G) = 0.0732$
$P(S) + P(M) + P(N) = 1$		$P(G) = 0.41$	$P(S \mid G) + P(M \mid G) + P(N \mid G) = 1$

a. In order to calculate $P(G)$, we use the total probability rule, $P(G) = P(G \cap S) + P(G \cap M) + P(G \cap N)$. The joint probabilities are calculated as products of conditional probabilities with their corresponding prior probabilities. For instance, in Table 4.7, $P(G \cap S) = P(G \mid S)P(S) = 0.90 \times 0.20 = 0.18$.

Therefore, the probability that the stock price of Webtalk goes up equals $P(G) = 0.18 + 0.20 + 0.03 = 0.41$.

b. According to Bayes' theorem, $P(S \mid G) = \frac{P(G \cap S)}{P(G)} = \frac{P(G \cap S)}{P(G \cap S) + P(G \cap M) + P(G \cap N)}$. From Table 4.7, we find $P(S \mid G) = \frac{0.18}{0.41} = 0.4390$. Note that the prior probability of a significant improvement in credit flow is revised upward from 0.20 to a posterior probability of 0.4390.

EXERCISES 4.4

Mechanics

50. Let $P(B) = 0.60$, $P(A \mid B) = 0.80$, and $P(A \mid B^c) = 0.10$. Calculate the following probabilities:
    a. $P(B^c)$
    b. $P(A \cap B)$ and $P(A \cap B^c)$
    c. $P(A)$
    d. $P(B \mid A)$

51. Let $P(A) = 0.70$, $P(B \mid A) = 0.55$, and $P(B \mid A^c) = 0.10$. Use a probability tree to calculate the following probabilities:
    a. $P(A^c)$
    b. $P(A \cap B)$ *and* $P(A^c \cap B)$
    c. $P(B)$
    d. $P(A \mid B)$

52. Complete the following probability table.

Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
$P(A) = 0.30$	$P(B \mid A) = 0.25$	$P(A \cap B) =$	$P(A \mid B) =$
$P(A^c) =$	$P(B \mid A^c) = 0.80$	$P(A^c \cap B) =$	$P(A^c \mid B) =$
Total =		$P(B) =$	Total =

53. Complete the following probability table.

Prior Probability	Conditional Probability	Joint Probability	Posterior Probability
$P(B) = 0.85$	$P(A \mid B) = 0.05$	$P(A \cap B) =$	$P(B \mid A) =$
$P(B^c) =$	$P(A \mid B^c) = 0.80$	$P(A \cap B^c) =$	$P(B^c \mid A) =$
Total =		$P(A) =$	Total =

54. Let a sample space be partitioned into three mutually exclusive and exhaustive events, B_1, B_2, and B_3. Complete the following probability table.

Prior Probabilities	Conditional Probabilities	Joint Probabilities	Posterior Probabilities
$P(B_1) = 0.10$	$P(A \mid B_1) = 0.40$	$P(A \cap B_1) =$	$P(B_1 \mid A) =$
$P(B_2) =$	$P(A \mid B_2) = 0.60$	$P(A \cap B_2) =$	$P(B_2 \mid A) =$
$P(B_3) = 0.30$	$P(A \mid B_3) = 0.80$	$P(A \cap B_3) =$	$P(B_3 \mid A) =$
Total =		$P(A) =$	Total =

Applications

55. Christine has always been weak in mathematics. Based on her performance prior to the final exam in Calculus, there is a 40% chance that she will fail the course if she does not have a tutor. With a tutor, her probability of failing decreases to 10%. There is only a 50% chance that she will find a tutor at such short notice.
    a. What is the probability that Christine fails the course?
    b. Christine ends up failing the course. What is the probability that she had found a tutor?

56. An analyst expects that 20% of all publicly traded companies will experience a decline in earnings next year. The analyst has developed a ratio to help forecast this decline. If the company is headed for a decline, there is a 70% chance that this ratio will be negative. If the company is not headed for a decline, there is a 15% chance that the ratio will be negative. The analyst randomly selects a company and its ratio is negative. What is the posterior probability that the company will experience a decline?

57. The State Police are trying to crack down on speeding on a particular portion of the Massachusetts Turnpike. To aid in this pursuit, they have purchased a new radar gun that promises greater consistency and reliability. Specifically, the gun advertises ± one-mile-per-hour accuracy 98% of the time; that is, there is a 0.98 probability that the gun will detect a speeder, if the driver is actually speeding. Assume there is a 1% chance that the gun erroneously detects a speeder even when the driver is below the speed limit. Suppose that 95% of the drivers drive below the speed limit on this stretch of the Massachusetts Turnpike.
    a. What is the probability that the gun detects speeding and the driver was speeding?
    b. What is the probability that the gun detects speeding and the driver was not speeding?
    c. Suppose the police stop a driver because the gun detects speeding. What is the probability that the driver was actually driving below the speed limit?

58. According to a study, cell phones are the main medium for teenagers to stay connected with friends and family (*CNN*, March 19, 2012). It is estimated that 90% of older teens and 60% of younger teens own a cell phone. Suppose 70% of all teens are older teens.
    a. What is the implied probability that a teen owns a cell phone?
    b. Given that a teen owns a cell phone, what is the probability that he/she is an older teen?

c. Given that the teen owns a cell phone, what is the probability that he/she is a younger teen?

59. According to data from the *National Health and Nutrition Examination Survey*, 33% of white, 49.6% of black, 43% of Hispanic, and 8.9% of Asian women are obese. In a representative town, 48% of women are white, 19% are black, 26% are Hispanic, and the remaining 7% are Asian.
    a. Find the probability that a randomly selected woman in this town is obese.
    b. Given that a woman is obese, what is the probability that she is white?
    c. Given that a woman is obese, what is the probability that she is black?
    d. Given that a woman is obese, what is the probability that she is Asian?

60. A crucial game of the Los Angeles Lakers basketball team depends on the health of their key player. According to his doctor's report, there is a 40% chance that he will be fully fit to play, a 30% chance that he will be somewhat fit to play, and a 30% chance that he will not be able to play at all. The coach has estimated the chances of winning at 80% if the player is fully fit, 60% if he is somewhat fit, and 40% if he is unable to play.
    a. What is the probability that the Lakers will win the game?
    b. You have just heard that the Lakers won the game. What is the probability that the key player had been fully fit to play in the game?

61. An analyst thinks that next year there is a 20% chance that the world economy will be good, a 50% chance that it will be neutral, and a 30% chance that it will be poor. She also predicts probabilities that the performance of a start-up firm, Creative Ideas, will be good, neutral, or poor for each of the economic states of the world economy. The following table presents probabilities for three states of the world economy and the corresponding conditional probabilities for Creative Ideas.

State of the World Economy	Probability of Economic State	Performance of Creative Ideas	Conditional Probability of Creative Ideas
Good	0.20	Good	0.60
		Neutral	0.30
		Poor	0.10
Neutral	0.50	Good	0.40
		Neutral	0.30
		Poor	0.30
Poor	0.30	Good	0.20
		Neutral	0.30
		Poor	0.50

    a. What is the probability that the performance of the world economy will be neutral and that of Creative Ideas will be poor?
    b. What is the probability that the performance of Creative Ideas will be poor?
    c. The performance of Creative Ideas was poor. What is the probability that the performance of the world economy had also been poor?

WRITING WITH STATISTICS

Support for marijuana legalization in the United States has grown remarkably over the past few decades. In 1969, when the question was first presented, only 12% of Americans were in favor of its legalization. This support had increased to over 25% by the late 1970s. While support was stagnant from 1981 to 1997, the turn of the century brought a renewed interest in its legalization, with the percentage of Americans in favor exceeding 30% by 2000 and 40% by 2009.

Alexis Lewis works for a drug policy institute that focuses on science, health, and human rights. She is analyzing the demographic breakdown of marijuana supporters. Using 2016 results from a Pew Research Center survey, she has found that support for marijuana legalization varies considerably depending on a person's age group. Alexis compiles information on support based on age group as shown in Table 4.8.



©Seastock/Shutterstock

TABLE 4.8 Percentage Support for Legalizing Marijuana by Age Group

Age Group	Support
Millennial (18–35)	71%
Generation X (36–51)	57%
Baby Boomer (52–70)	56%
Silent (71 and older)	33%

Source: Pew Research Center, survey conducted August 23 – September 2, 2016.

Alexis finds that another important factor determining the fate of marijuana legalization concerns each age group's ability to sway the vote. For adults eligible to vote as of 2016, she breaks down each age group's voting power. The Millennial, Generation X, Baby Boomer, and Silent generations account for 31%, 25%, 31%, and 13% of the voting population, respectively.

Alexis wants to use this information to

1. Calculate and interpret relevant conditional, unconditional, and joint probabilities.
2. Calculate and interpret the probability of all Americans who support the legalization of marijuana.

Sample Report—Linking Support for Legalizing Marijuana with Age Group

Driven by growing public support, the legalization of marijuana in America has been moving at a breakneck speed in recent years. As of 2016, marijuana is now legal in some form in 28 states and in Washington, DC. Even recreational marijuana is gaining support, becoming legal in Alaska, California, Colorado, Maine, Massachusetts, Nevada, Oregon, Washington, and Washington, DC. Changing demographics can help explain how the tide has turned in marijuana's favor, especially since Millennials (those between the ages of 18 and 35) are on the verge of becoming the nation's largest living generation.

A 2016 Pew Research Study provides interesting data regarding support for marijuana legalization. Two factors seem to drive support for the issue: generation (or age group) and the relative size of a generation's voting bloc. For ease of interpretation, let M, G, B, and S denote "Millennial," "Generation X," "Baby Boomer," and "Silent" generations, respectively. Based on data from the study, the following probability statements can be formulated with respect to the relative size of each generation's voting bloc: $P(M) = 0.31$, $P(G) = 0.25$, $P(B) = 0.31$, $P(S) = 0.13$. In other words, Millennials and Baby Boomers have the most voting power, each comprising 31% of the voting population; the Generation X and Silent generations represent 25% and 13% of the voting population, respectively.

Now let L denote "support for legalizing marijuana." Again, based on data from the study, conditional probabilities can be specified as $P(L \mid M) = 0.71$, $P(L \mid G) = 0.57$, $P(L \mid B) = 0.56$, and $P(L \mid S) = 0.33$. Therefore, the probability that a randomly selected adult supports legal marijuana and is in the Millennial generation is determined as $P(L \cap M) = 0.71 \times 0.31 = 0.2201$. Similarly, $P(L \cap G) = 0.1425$, $P(L \cap B) = 0.1736$, and $P(L \cap S) = 0.0429$. By combining all generations, we deduce the total probability of support for legalizing marijuana as $P(L) = 0.2201 + 0.1425 + 0.1736 + 0.0429 = 0.5791$; in 2016, a staggering 58% of Americans support the legalization of marijuana. Table 4.A is the joint probability table that summarizes unconditional and joint probabilities.

TABLE 4.A Joint Probability Table for the Support for Legalizing Marijuana by Age Group

	Legalizing Marijuana		
Age Group	Support	Do not Support	Total
Millennial (18–35)	0.2201	0.0899	0.31
Generation X (36–51)	0.1425	0.1075	0.25
Baby Boomer (52–70)	0.1736	0.1364	0.31
Silent (71 and older)	0.0429	0.0871	0.13
Total	0.5791	0.4209	1.00

To put it in perspective, suppose that there are 1,000 randomly selected adult attendees at a conference. The results imply that there would be about 310 Millennial, 250 Generation X, 310 Baby Boomer, and 130 Silent attendees. Further, the supporters of marijuana legalization would include about 220 Millennial, 143 Generation X, 174 Baby Boomer, and 43 Silent attendees.

Millennials, with roughly 31% of the overall electorate, are now as large a political force as Baby Boomers. In general, Millennials tend to be liberal on social issues such as gay rights, immigration, and marijuana. This shift in population has not gone unnoticed by political parties, which all hope to court the more than 75 million of these eligible young voters.

CONCEPTUAL REVIEW

LO 4.1 Describe fundamental probability concepts.

In order to assign the appropriate probability to an uncertain event, it is useful to establish some terminology. An **experiment** is a process that leads to one of several possible outcomes. A **sample space**, denoted S, of an experiment contains all possible outcomes of the experiment. An **event** is any subset of outcomes of an experiment, and is called a simple event if it contains a single outcome. Events are **exhaustive** if all possible outcomes of an experiment belong to the events. Events are **mutually exclusive** if they do not share any common outcome of an experiment.

A **probability** is a numerical value that measures the likelihood that an event occurs. It assumes a value between zero and one, where a value zero indicates an impossible event and a value one indicates a definite event. The two defining properties of a probability are (1) the probability of any event A is a value between 0 and 1, $0 \le P(A) \le 1$, and (2) the sum of the probabilities of any list of mutually exclusive and exhaustive events equals 1.

A **subjective** probability is calculated by drawing on personal and subjective judgment. An **empirical probability** is calculated as a relative frequency of occurrence. A **classical probability** is based on logical analysis rather than on observation or personal judgment.

LO 4.2 Apply the rules of probability.

Rules of probability allow us to calculate the probabilities of more complex events. The **complement rule** states that the probability of the complement of an event can be found by subtracting the probability of the event from one: $P(A^c) = 1 - P(A)$.

The probability that at least one of two events occurs is calculated by using the **addition rule**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Since $P(A \cap B) = 0$ for mutually exclusive events, the addition rule then simplifies in these instances to $P(A \cup B) = P(A) + P(B)$.

The probability of event A, denoted $P(A)$, is an **unconditional probability**. It is the probability that A occurs without any additional information. The probability that A occurs given that B has already occurred, denoted $P(A|B)$, is a **conditional probability**. A conditional probability is computed as $P(A|B) = \frac{P(A \cap B)}{P(B)}$. We rearrange the conditional probability formula to arrive at the **multiplication rule**. When using this rule, we find the probability that two events, A and B, both occur; that is, $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$.

LO 4.3 Distinguish between independent and dependent events.

Two events, A and B, are **independent** if $P(A|B) = P(A)$, or if $P(B|A) = P(B)$. Otherwise, the events are **dependent**. For independent events, the multiplication rule simplifies to $P(A \cap B) = P(A)P(B)$.

LO 4.4 **Calculate and interpret probabilities from a contingency table.**

A **contingency table** generally shows frequencies for two qualitative (categorical) variables, x and y, where each cell represents a mutually exclusive combination of x-y values. Empirical probabilities are easily calculated as the relative frequency of the occurrence of the event.

LO 4.5 **Apply the total probability rule.**

The **total probability rule** expresses the probability of an event A in terms of probabilities of the intersection of A with two mutually exclusive and exhaustive events, B and B^c:

$$P(A) = P(A \cap B) + P(A \cap B^c), \text{ or equivalently,}$$
$$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c)$$

We can extend this rule where the sample space is partitioned into n mutually exclusive and exhaustive events, $B_1, B_2, \ldots, B_n$. The total probability rule is

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n), \text{ or equivalently,}$$
$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n).$$

LO 4.6 **Apply Bayes' theorem.**

Bayes' theorem provides a procedure for updating probabilities based on new information. Let $P(B)$ be the prior probability and $P(B \mid A)$ be the posterior probability based on new information provided by A. Then,

$$P(B \mid A) = \frac{P(A \cap B)}{P(A \cap B) + P(A \cap B^c)} = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)}.$$

For the extended case, Bayes' theorem, for any $i = 1, 2, \ldots, n$, is

$$P(B_i \mid A) = \frac{P(A \cap B_i)}{P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)}, \text{ or equivalently,}$$
$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n)}.$$

ADDITIONAL EXERCISES AND CASE STUDIES

Exercises

62. According to a global survey of 4,400 parents of children between the ages of 14 to 17, 44% of parents spy on their teen's Facebook account (www.msnbc.com, April 25, 2012). Assume that American parents account for 10% of all parents of teens with Facebook accounts, of which 60% spy on their teen's Facebook account. Suppose a parent is randomly selected, and the following events are defined: A = selecting an American parent and B = selecting a spying parent.
    a. Based on the above information, what are the probabilities that can be established?
    b. Are the events A and B mutually exclusive and/or exhaustive? Explain.
    c. Are the events A and B independent? Explain.
    d. What is the probability of selecting an American parent given that she/he is a spying parent?

63. High blood pressure is common in adults who are overweight and are black American. According to the American Heart Association, 47% of black men and 43% of black women have high blood pressure. Suppose 84% of black men with high blood pressure are overweight and 92% of black women with high blood pressure are overweight.
    a. Find the percentage of black men who have high blood pressure and are overweight.
    b. Find the percentage of black women who have high blood pressure and are overweight.

64. According to eMarketer estimates, 88.3% of 12- to 17-year-olds had a mobile phone in 2016. Among those with mobile phones, 84.0% had smartphones. Calculate the percentage of 12- to 17-year-olds who had smartphones.

65. Henry Chow is a stockbroker working for Merrill Lynch. He knows from past experience that there is a 70% chance that his new client will want to include U.S. equity in her portfolio and a 50% chance that she will want to include foreign equity. There is also a 40% chance that she will want to include both U.S. equity and foreign equity in her portfolio.
    a. What is the probability that the client will want to include U.S. equity if she already has foreign equity in her portfolio?
    b. What is the probability that the client decides to include neither U.S. equity nor foreign equity in her portfolio?

66. The Easy Credit Company reports the following table representing a breakdown of customers according to the amount they owe and whether a cash advance has been made. An auditor randomly selects one of the accounts.

	Cash Advance?	
Amounts owed by customers	Yes	No
$0 – 199.99	245	2,890
$200 – 399.99	380	1,700
$400 – 599.99	500	1,425
$600 – 799.99	415	940
$800 – 999.99	260	480
$1,000 or more	290	475
Total Customers	2,090	7,910

    a. What is the probability that a customer received a cash advance?
    b. What is the probability that a customer owed less than $200 and received a cash advance?
    c. What is the probability that a customer owed less than $200 or received a cash advance?
    d. Given that a customer received a cash advance, what is the probability that the customer owed $1,000 or more?
    e. Given that a customer owed $1,000 or more, what is the probability that the customer received a cash advance?
    f. Are the events "receiving a cash advance" and "owing $1,000 or more" mutually exclusive? Explain using probabilities.
    g. Are the events "receiving a cash advance" and "owing $1,000 or more" independent? Explain using probabilities.

67. The following frequency distribution shows the ages of India's 40 richest individuals. One of these individuals is selected at random.

Ages	Frequency
30 up to 40	3
40 up to 50	8
50 up to 60	15
60 up to 70	9
70 up to 80	5

Source: www.forbes.com.

    a. What is the probability that the individual is between 50 and 60 years of age?
    b. What is the probability that the individual is younger than 50 years of age?
    c. What is the probability that the individual is at least 60 years of age?

68. How much you smile in your younger days can predict your later success in marriage (www.msnbc.com, April 16, 2009). The analysis is based on the success rate in marriage of people over age 65 and their smiles when they were only 10 years old. Researchers found that only 11% of the biggest smilers had been divorced, while 31% of the biggest frowners had experienced a broken marriage.
    a. Suppose it is known that 2% of the people are the biggest smilers at age 10 and divorced in later years. What percent of people are the biggest smilers?
    b. If 25% of people are considered to be the biggest frowners, calculate the probability that a person is a biggest frowner at age 10 and divorced later in life.

69. Anthony Papantonis, owner of Nauset Construction, is bidding on two projects, A and B. The probability that he wins project A is 0.40 and the probability that he wins project B is 0.25. Winning Project A and winning Project B are independent events.
    a. What is the probability that he wins project A or project B?
    b. What is the probability that he does not win either project?

70. Wooden boxes are commonly used for the packaging and transportation of mangoes. A convenience store in Morganville, New Jersey, regularly buys mangoes from a wholesale dealer. For every shipment, the manager randomly inspects two mangoes from a box containing 20 mangoes for damages due to transportation. Suppose the chosen box contains exactly 3 damaged mangoes.
    a. Find the probability that the first mango is not damaged.

b. Find the probability that neither of the mangoes is damaged.
c. Find the probability that both mangoes are damaged.

71. A study shows that unemployment does not impact males and females in the same way (*Newsweek*, April 20, 2009). According to a Bureau of Labor Statistics report, 8.5% of those who are eligible to work are unemployed. The unemployment rate is 8.8% for eligible men and only 7.0% for eligible women. Suppose 52% of the eligible workforce in the U.S. consists of men.
    a. You have just heard that another worker in a large firm has been laid off. What is the probability that this worker is a man?
    b. You have just heard that another worker in a large firm has been laid off. What is the probability that this worker is a woman?

72. According to the CGMA Economic Index, which measures executive sentiment across the world, 18% of all respondents expressed optimism about the global economy (www.aicpa.org, March 29, 2012). Moreover, 22% of the respondents from the United States and 9% from Asia felt optimistic about the global economy.
    a. What is the probability that an Asian respondent is not optimistic about the global economy?
    b. If 28% of all respondents are from the United States, what is the probability that a respondent is from the United States and is optimistic about the global economy?
    c. Suppose 22% of all respondents are from Asia. If a respondent feels optimistic about the global economy, what is the probability that the respondent is from Asia?

73. A professor of management has heard that eight students in his class of 40 have landed an internship for the summer. Suppose he runs into two of his students in the corridor.
    a. Find the probability that neither of these students has landed an internship.
    b. Find the probability that both of these students have landed an internship.

74. It has generally been believed that it is not feasible for men and women to be just friends (*The New York Times*, April 12, 2012). Others argue that this belief may not be true anymore since gone are the days when men worked and women stayed at home and the only way they could get together was for romance. In a survey, 186 heterosexual college students were asked if it was feasible for men and women to be just friends. Thirty-two percent of females and 57% of males reported that it was not feasible for men and women to be just friends. Suppose the study consisted of 100 female and 86 male students.
    a. Construct a contingency table that shows frequencies for the qualitative variables Gender (men or women) and Feasible (yes or no).
    b. Find the probability that a student believes that men and women can be friends.
    c. If a student believes that men and women can be friends, what is the probability that this student is a male? Find the corresponding probability that this student is a female.

75. At a local bar in a small Midwestern town, beer and wine are the only two alcoholic options. The manager noted that of all male customers who visited over the weekend, 150 ordered beer, 40 ordered wine, and 20 asked for soft drinks. Of female customers, 38 ordered beer, 20 ordered wine, and 12 asked for soft drinks.
    a. Construct a contingency table that shows frequencies for the qualitative variables Gender (male or female) and Drink Choice (beer, wine, or soft drink).
    b. Find the probability that a customer orders wine.
    c. What is the probability that a male customer orders wine?
    d. Are the events "Wine" and "Male" independent? Explain using probabilities.

76. A study in the *Journal of the American Medical Association* (February 20, 2008) found that patients who go into cardiac arrest while in the hospital are more likely to die if it happens after 11 pm. The study investigated 58,593 cardiac arrests that occurred during the day or evening. Of those, 11,604 survived to leave the hospital. There were 28,155 cardiac arrests during the shift that began at 11 pm, commonly referred to as the graveyard shift. Of those, 4,139 survived for discharge. The following contingency table summarizes the results of the study.

	Survived for Discharge	Did not Survive for Discharge	Total
Day or Evening Shift	11,604	46,989	58,593
Graveyard Shift	4,139	24,016	28,155
Total	15,743	71,005	86,748

    a. What is the probability that a randomly selected patient experienced cardiac arrest during the graveyard shift?
    b. What is the probability that a randomly selected patient survived for discharge?
    c. Given that a randomly selected patient experienced cardiac arrest during the graveyard shift, what is the probability the patient survived for discharge?

d. Given that a randomly selected patient survived for discharge, what is the probability the patient experienced cardiac arrest during the graveyard shift?

e. Are the events "Survived for Discharge" and "Graveyard Shift" independent? Explain using probabilities. Given your answer, what type of recommendations might you give to hospitals?

77. It has been reported that women end up unhappier than men later in life, even though they start out happier (*Yahoo News*, August 1, 2008). Early in life, women are more likely to fulfill their family life and financial aspirations, leading to greater overall happiness. However, men report a higher satisfaction with their financial situation and family life, and are thus happier than women, in later life. Suppose the results of the survey of 300 men and 300 women are presented in the following table.

Response to the question "Are you satisfied with your financial and family life?"

	Age		
Response by Women	20 to 35	35 to 50	Over 50
Yes	73	36	32
No	67	54	38

	Age		
Response by Men	20 to 35	35 to 50	Over 50
Yes	58	34	38
No	92	46	32

a. What is the probability that a randomly selected woman is satisfied with her financial and family life?

b. What is the probability that a randomly selected man is satisfied with his financial and family life?

c. For women, are the events "Yes" and "20 to 35" independent? Explain using probabilities.

d. For men, are the events "Yes" and "20 to 35" independent? Explain using probabilities.

78. An analyst predicts that there is a 40% chance that the U.S. economy will perform well. If the U.S. economy performs well, then there is an 80% chance that Asian countries will also perform well. On the other hand, if the U.S. economy performs poorly, the probability of Asian countries performing well goes down to 0.30.

a. What is the probability that both the U.S. economy and the Asian countries will perform well?

b. What is the probability that the Asian countries will perform well?

c. What is the probability that the U.S. economy will perform well, given that the Asian countries perform well?

79. Apparently, depression significantly increases the risk of developing dementia later in life (*BBC News*, July 6, 2010). In a study, it was reported that 22% of those who had depression went on to develop dementia, compared to only 17% of those who did not have depression. Suppose 10% of all people suffer from depression.

a. What is the probability of a person developing dementia?

b. If a person has developed dementia, what is the probability that the person suffered from depression earlier in life?

80. According to data from the *National Health and Nutrition Examination Survey*, 36.5% of adult women and 26.6% of adult men are at a healthy weight. Suppose 50.52% of the adult population consists of women.

a. What proportion of adults is at a healthy weight?

b. If an adult is at a healthy weight, what is the probability that the adult is a woman?

c. If an adult is at a healthy weight, what is the probability that the adult is a man?

81. Suppose that 60% of students do homework regularly. It is also known that 80% of students who had been doing homework regularly end up doing well in the course (get a grade of A or B). Only 20% of students who had not been doing homework regularly end up doing well in the course.

a. What is the probability that a student does well in the course?

b. Given that a student did well in the course, what is the probability that the student had been doing homework regularly?

82. There is a growing public support for marijuana law reform, with polls showing more than half the country is in favor of some form of marijuana legalization. However, opinions on marijuana are divided starkly along political party lines. The results of the Pew Research Center survey conducted in 2016 are shown in the following table. In addition, assume that 27% of Americans identify as Republicans, 30% as Democrats, and 43% as independents.

Political Party	Support
Republican	41%
Democrat	66%
Independent	63%

a. Calculate the probability that a randomly selected American adult supports marijuana legalization and is a Republican.

b. Calculate the probability that a randomly selected American adult supports marijuana legalization and is a Democrat.

c. Calculate the probability that a randomly selected American adult supports marijuana legalization and is an independent.
d. What percentage of American adults support marijuana legalization?
e. If a randomly selected American adult supports marijuana legalization, what is the probability that this adult is a Republican?

83. A 2015 national survey by the Washington Post-Kaiser Family Foundation finds that there is a big gender divide between Americans when identifying as feminist or strong feminist. The results of the survey are shown in the following table. In addition, per the 2010 U.S. Census Current Population Survey, 50.8% of the American population is female and 49.2% is male.

Gender	Feminist or Strong Feminist
Female	66%
Male	41%

a. Calculate the probability that a randomly selected American adult is a female who also identifies as feminist or strong feminist.
b. Calculate the probability that a randomly selected American adult is a male who also identifies as feminist or strong feminist.
c. What percentage of American adults identify as feminist or strong feminist?
d. If a randomly selected American adult identifies as feminist or strong feminist, what is the probability that this adult is a female?

84. According to the Census's Population Survey, the percentage of children with two parents at home is the highest for Asians and lowest for blacks (*USA Today*, February 26, 2009). It is reported that 85% of Asian children have two parents at home versus 78% of white, 70% of Hispanic, and 38% of black. Suppose there are 500 students in a representative school of which 280 are white, 50 are Asian, 100 are Hispanic, and 70 are black.
a. What is the probability that a child has both parents at home?
b. If both parents are at home, what is the probability the child is Asian?
c. If both parents are at home, what is the probability the child is black?

CASE STUDIES

CASE STUDY 4.1 Ever since the introduction of New Coke failed miserably in the 1980s, most food and beverage companies have been cautious about changing the taste or formula of their signature offerings. In an attempt to attract more business, Starbucks recently introduced a new milder brew, Pike Place Roast, as its main drip coffee at the majority of its locations nationwide. The idea was to offer a more approachable cup of coffee with a smoother finish. However, the strategy also downplayed the company's more established robust roasts; initially, the milder brew was the only option for customers after noon. Suppose on a recent afternoon, 100 customers were asked whether or not they would return in the near future for another cup of Pike Place Roast. The following contingency table (cross-classified by type of customer and whether or not the customer will return) lists the results:

Data for Case Study 4.1

Return in Near Future?	Customer Type	
	First-time Customer	Established Customer
Yes	35	10
No	5	50

In a report, use the sample information to

1. Calculate and interpret unconditional probabilities.
2. Calculate the probability that a customer will return given that the customer is an established customer.
3. Determine whether the events "Return in Near Future" and "Customer Type" are independent. Shortly after the introduction of Pike Place Roast, Starbucks decided to offer its bolder brew again in the afternoon at many of its locations. Do your results support Starbucks' decision? Explain.

CASE STUDY 4.2 It is common to ignore the thyroid gland of women during pregnancy (*The New York Times*, April 13, 2009). This gland makes hormones that govern metabolism, helping to regulate body weight, heart rate, and a host of other factors. If the thyroid malfunctions, it can produce too little or too much of these hormones. Hypothyroidism, caused by an untreated underactive thyroid in pregnant women, carries the risk of impaired intelligence in the child. According to one research study, 62 out of 25,216 pregnant women were identified with hypothyroidism. Nineteen percent of the children born to women with an untreated underactive thyroid had an I.Q. of 85 or lower, compared with only 5% of those whose mothers had a healthy thyroid. It was also reported that if mothers have their hypothyroidism treated, their children's intelligence would not be impaired.

In a report, use the sample information to

1. Find the likelihood that a woman suffers from hypothyroidism during pregnancy and later has a child with an I.Q. of 85 or lower.
2. Determine the number of children in a sample of 100,000 that are likely to have an I.Q. of 85 or lower if the thyroid gland of pregnant women is ignored.
3. Compare and comment on your answer to part b with the corresponding number if all pregnant women are tested and treated for hypothyroidism.

CASE STUDY 4.3 Enacted in 1998, the Children's Online Privacy Protection Act requires firms to obtain parental consent before tracking the information and the online movement of children; however, the act applies to those children ages 12 and under. Teenagers are often oblivious to the consequences of sharing their lives online. Data reapers create huge libraries of digital profiles and sell these profiles to advertisers, who use it to detect trends and micro-target their ads back to teens. For example, a teen searching online for ways to lose weight could become enticed by an ad for dietary supplements, fed into his/her network by tracking cookies. As a preliminary step in gauging the magnitude of teen usage of social networking sites, an economist surveys 200 teen girls and 200 teen boys. Of teen girls, 166 use social networking sites; of teen boys, 156 use social networking sites.

In a report, use the sample information to

1. Construct a contingency table that shows frequencies for the qualitative variables Gender (male or female) and Use of Social Networking Sites (Yes or No).
2. Determine the probability that a teen uses social networking sites.
3. Determine the probability that a teen girl uses a social networking site.
4. A bill before Congress would like to extend the Children's Online Privacy Protection Act to apply to 15-year-olds. In addition, the bill would also ban Internet companies from sending targeted advertising to children under 16 and give these children and their parents the ability to delete their digital footprint and profile with an "eraser button" (*The Boston Globe*, May 20, 2012). Given the probabilities that you calculated with respect to teen usage of social networking sites, do you think that this legislation is necessary? Explain.

CASE STUDY 4.4 In 2008, it appeared that rising gas prices had made Californians less resistant to offshore drilling. A Field Poll survey showed that a higher proportion of Californians supported the idea of drilling for oil or natural gas along the state's coast than in 2005 (*The Wall Street Journal*, July 17, 2008). Assume that random drilling for oil only succeeds 5% of the time.

An oil company has just announced that it has discovered new technology for detecting oil. The technology is 80% reliable. That is, if there is oil, the technology will signal "oil" 80% of the time. Let there also be a 1% chance that the technology erroneously detects oil, when in fact no oil exists.

In a report, use the above information to

1. Prepare a table that shows the relevant probabilities.
2. Find the probability that, on a recent expedition, oil actually existed but the technology detected "no oil" in the area.

5 Discrete Probability Distributions

Learning Objectives

After reading this chapter you should be able to:

LO **5.1** Describe a discrete random variable and its probability distribution.

LO **5.2** Calculate and interpret summary measures for a discrete random variable.

LO **5.3** Calculate and interpret probabilities for a binomial random variable.

LO **5.4** Calculate and interpret probabilities for a Poisson random variable.

LO **5.5** Calculate and interpret probabilities for a hypergeometric random variable.

In this chapter, we extend our discussion about probability by introducing the concept of a random variable. A random variable summarizes the results of an experiment in terms of numerical values. It can be classified as discrete or continuous depending on the range of values that it assumes. A discrete random variable assumes a countable number of distinct values, whereas a continuous random variable is characterized by uncountable values. In this chapter, we focus on a discrete random variable and its associated probability distribution. Examples of discrete random variables include the number of credit cards carried by consumers, the number of foreclosures in a sample of 100 households, and the number of cars lined up at a toll booth. We calculate summary measures for a discrete random variable, including its mean, variance, and standard deviation. Finally, we discuss three widely used discrete probability distributions: the binomial, the Poisson, and the hypergeometric distributions.



©Richard Gardner/REX/Shutterstock

Introductory Case

Available Staff for Probable Customers

In addition to its previous plan to shut 100 stores, Starbucks announced plans in 2008 to close 500 more U.S. locations (*The Wall Street Journal,* July 9, 2008). Executives claimed that a weak economy and higher gas and food prices led to a drop in domestic store traffic. Others speculate that Starbucks' rapid expansion produced a saturated market. The locations that will close are not profitable, are not expected to be profitable, or are located near an existing company-operated Starbucks.

Anne Jones, a manager at a local Starbucks, has been reassured by headquarters that her store will remain open. She is concerned about how other nearby closings might affect business at her store. Anne knows that a typical Starbucks customer visits the chain between 15 and 18 times a month, making it among the nation's most frequented retailers. She believes that her loyal Starbucks customers, along with displaced customers, will average 18 visits to the store over a 30-day month. To decide staffing needs, Anne knows that she needs a solid understanding about the probability distribution of customer arrivals. If too many employees are ready to serve customers, some employees will be idle, which is costly to the store. However, if not enough employees are available to meet demand, this could result in losing angry customers who choose not to wait for service.

Anne wants to use the above information to

1. Calculate the expected number of visits from a typical Starbucks customer in a specified time period.
2. Calculate the probability that a typical Starbucks customer visits the chain a certain number of times in a specified time period.

A synopsis of this case is provided in Section 5.4.

LO 5.1

Describe a discrete random variable and its probability distribution.

5.1 RANDOM VARIABLES AND DISCRETE PROBABILITY DISTRIBUTIONS

We often have to make important decisions in the face of uncertainty. For example, a car dealership has to determine the number of cars to hold on its lot when the actual demand for cars is unknown. Similarly, an investor has to select a portfolio when the actual outcomes of investment returns are not known. This uncertainty is captured by what we call a **random variable**. A random variable summarizes outcomes of an experiment with numerical values.

We generally use the letter X to denote a random variable. A **discrete random variable** assumes a countable number of distinct values such as x_1, x_2, x_3, and so on. A **continuous random variable**, on the other hand, is characterized by uncountable values. In other words, a continuous random variable can take on any value within an interval.

DISCRETE VERSUS CONTINUOUS RANDOM VARIABLES

A random variable is a function that assigns numerical values to the outcomes of an experiment. A discrete random variable assumes a countable number of distinct values. A continuous random variable, on the other hand, is characterized by uncountable values in an interval.

Recall from Chapter 4, the sample space S is a set of all outcomes of an experiment. Whenever some numerical values are assigned to these outcomes, a random variable X can be defined. Consider the following experiments, and some examples of discrete random variables (with their possible values shown) that are associated with the experiments:

Experiment 1. Rolling a six-sided die; $S = \{1, 2, 3, 4, 5, 6\}$.

Let X = Win \$10 if odd number, lose \$10 if even number; possible values of $X = \{-10, 10\}$

Experiment 2. Two shirts are selected from the production line and each is either defective (D) or nondefective (N); $S = \{(D, D), (D, N), (N, D), (N, N)\}$.

Let X = the number of defective shirts; possible values of $X = \{0, 1, 2\}$

Experiment 3. Reviewing multiple mortgage applications and, for each client, deciding whether the client gets approved (A) or denied (D); S = the set of all possible infinite sequences whose elements are A or D.

Let X = the number of approvals; possible values of $X = \{0, 1, 2, 3, \ldots\}$

The random variables defined for Experiments 1 and 2 have a finite and countable number of distinct values, whereas the random variable defined for Experiment 3 has an infinite but countable number of distinct values.

Sometimes, we can define a random variable *directly* by identifying its values with some numerical outcomes. For example, we may be interested in the number of students who get financial aid out of the 100 students who applied. Then the set of possible values of the random variable, equivalent to the sample space, is $\{0, 1, \ldots, 100\}$. In a similar way, we can define a discrete random variable by the infinite number of distinct values that it may take. For example, consider the number of cars that cross the Brooklyn Bridge between 9:00 am and 10:00 am on a Monday morning. Here the discrete random variable takes an infinite but countable number of distinct values from $\{0, 1, 2, \ldots\}$. Note that we cannot specify an upper bound on the observed number of cars.

Although, we explore discrete random variables in this chapter, random variables can also be continuous. For example, the time taken by a student to complete a 60-minute exam may assume any value between 0 and 60 minutes. Thus, the set of such values is uncountable; that is, it is impossible to put all real numbers from the interval [0, 60] in a sequence. Here, the random variable is continuous because the outcomes are uncountable. Some students may think that time in this example is countable in seconds; however, this is not the case once we consider fractions of a second. We will discuss the details of continuous random variables in the next chapter.

The Discrete Probability Distribution

Every random variable is associated with a **probability distribution** that describes it completely. It is common to define discrete random variables in terms of their **probability mass function** and continuous random variables in terms of their **probability density function**. Discrete and continuous random variables can also be defined in terms of their **cumulative distribution function**, or equivalently, $P(X \leq x)$.

The probability mass function for a discrete random variable X is a list of the values of X with the associated probabilities; that is, the list of all possible pairs $(x, P(X = x))$. The cumulative distribution function of X is defined as $P(X \leq x)$.

For convenience, we will use terms like "probability distribution" and "distribution" for the probability mass function. Similarly, we will use "cumulative probability distribution" for the cumulative distribution function.

We can view a discrete probability distribution in several ways, including tabular, algebraic, and graphical forms. Example 5.1 shows one of two tabular forms. In general, we can construct a table in two different ways. The first approach directly specifies the probability that the random variable assumes a specific value.

EXAMPLE 5.1

Consider an experiment of rolling a fair six-sided die, with the random variable X defined as the number rolled. Present the probability distribution in a tabular form.

SOLUTION: A probability distribution for rolling a six-sided die is shown in Table 5.1.

TABLE 5.1 Probability Distribution for Example 5.1

x	1	2	3	4	5	6
$P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6

From Table 5.1, we can deduce, for instance, that $P(X = 5)$ equals 1/6. For that matter, the probability that X assumes any of the six possible values is 1/6.

The probability distribution defined in Example 5.1 illustrates two components of all discrete probability distributions.

TWO KEY PROPERTIES OF DISCRETE PROBABILITY DISTRIBUTIONS

- The probability that a discrete random variable X assumes a particular value x falls between 0 and 1, or equivalently, $0 \leq P(X = x) \leq 1$.
- The sum of the probabilities equals 1. In other words, $\Sigma P(X = x_i) = 1$, where the sum extends over all values x of X.

The second tabular view of a probability distribution is based on the cumulative probability distribution. The cumulative probability distribution is convenient when we are interested in finding the probability that the random variable assumes a range of values rather than a specific value. For the random variable defined in Example 5.1, the cumulative probability distribution is shown in Table 5.2.

TABLE 5.2 Cumulative Probability Distribution for Example 5.1

x	1	2	3	4	5	6
$P(X \leq x)$	1/6	2/6	3/6	4/6	5/6	6/6

If we are interested in finding the probability of rolling a four or less, $P(X \leq 4)$, we see from the cumulative probability distribution that this probability is 4/6. With the earlier probability representation, we would add up the probabilities to compute $P(X \leq 4)$ as

$$P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = 1/6 + 1/6 + 1/6 + 1/6 = 4/6.$$

At the same time, we can use the cumulative probability distribution to find the probability that the random variable assumes a specific value. For example, $P(X = 3)$ can be found as $P(X \leq 3) - P(X \leq 2) = 3/6 - 2/6 = 1/6$.

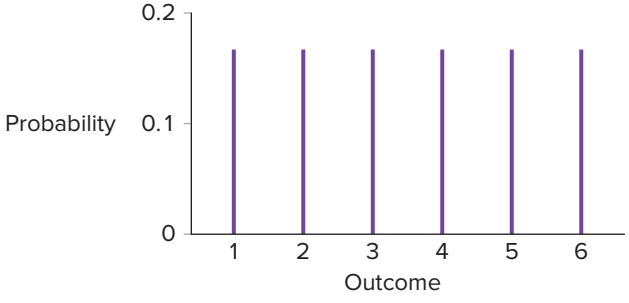
In many instances, we can express a probability distribution by applying an algebraic formula. A formula representation of the probability distribution for the random variable defined in Example 5.1 is

$$P(X = x) = \begin{cases} 1/6 & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, from the formula we can ascertain that $P(X = 5) = 1/6$ and $P(X = 7) = 0$.

In order to graphically depict a probability distribution, we place all values x of X on the horizontal axis and the associated probabilities $P(X = x)$ on the vertical axis. We then draw a line segment that emerges from each x and ends where its height equals $P(X = x)$. Figure 5.1 graphically illustrates the probability distribution for the random variable defined in Example 5.1 with probabilities equal to $1/6 = 0.1667$.

FIGURE 5.1 Probability distribution when rolling a six-sided die



The probability distribution in Figure 5.1 is an example of a **discrete uniform distribution**, which has the following characteristics:

- The distribution has a finite number of specified values.
- Each value is equally likely.
- The distribution is symmetric.

EXAMPLE 5.2

The number of homes that a realtor sells over a one-month period has the probability distribution shown in Table 5.3.

TABLE 5.3 Probability Distribution for the Number of Houses Sold

Number of Houses Sold	Probability
0	0.30
1	0.50
2	0.15
3	0.05

a. Is this a valid probability distribution?

b. What is the probability that the realtor does not sell any houses in a one-month period?

c. What is the probability that the realtor sells at most one house in a one-month period?

d. What is the probability that the realtor sells at least two houses in a one-month period?

e. Graphically depict the probability distribution and comment on its symmetry/skewness.

SOLUTION:

a. We first note that the random variable X denotes the number of houses that the realtor sells over a one-month period, and the possible values of X are 0, 1, 2, or 3. The probability distribution is valid because it satisfies the following two conditions: (1) all probabilities fall between 0 and 1, and (2) the probabilities sum to 1 $(0.30 + 0.50 + 0.15 + 0.05 = 1)$.

b. In order to find the probability that the realtor does not sell any houses in a one-month period, we find $P(X = 0) = 0.30$.

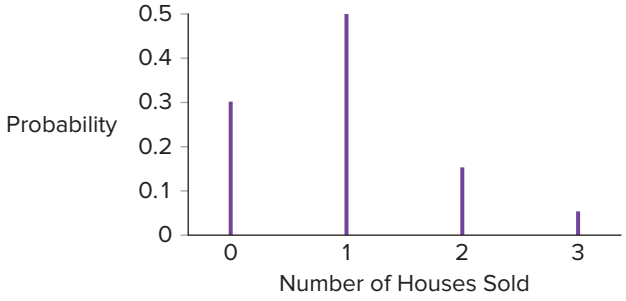
c. We find the probability that a realtor sells at most one house as $P(X \leq 1) = P(X = 0) + P(X = 1) = 0.30 + 0.50 = 0.80$.

d. We find the probability that the realtor sells at least two houses as $P(X \geq 2) = P(X = 2) + P(X = 3) = 0.15 + 0.05 = 0.20$.

Note that since the sum of the probabilities over all values of X equals 1, we can also find the above probability as $P(X \geq 2) = 1 - P(X \leq 1) = 1 - 0.80 = 0.20$.

e. The graph in Figure 5.2 shows that the distribution is not symmetric; rather, it is positively skewed. There are small chances of selling two or three houses in a one-month period. The most likely outcome by far is selling one house over a one-month period, with a probability of 0.50.

FIGURE 5.2 Probability distribution for the number of houses sold



EXERCISES 5.1

Mechanics

1. Consider the following discrete probability distribution.

x	15	22	34	40
$P(X = x)$	0.14	0.40	0.26	0.20

   a. Is this a valid probability distribution? Explain.
   b. Graphically depict this probability distribution.
   c. What is the probability that the random variable X is less than 40?
   d. What is the probability that the random variable X is between 10 and 30?
   e. What is the probability that the random variable X is greater than 20?

2. Consider the following discrete probability distribution.

x	−25	−15	10	20
$P(X = x)$	0.35	0.10		0.10

   a. Complete the probability distribution.
   b. Graphically depict the probability distribution and comment on the symmetry of the distribution.
   c. What is the probability that the random variable X is negative?
   d. What is the probability that the random variable X is greater than −20?
   e. What is the probability that the random variable X is less than 20?

3. Consider the following cumulative probability distribution.

x	0	1	2	3	4	5
$P(X \le x)$	0.15	0.35	0.52	0.78	0.84	1

   a. Calculate $P(X \le 3)$.
   b. Calculate $P(X = 3)$.
   c. Calculate $P(2 \le X \le 4)$.

4. Consider the following cumulative probability distribution.

x	−25	0	25	50
$P(X \le x)$	0.25	0.50	0.75	1

   a. Calculate $P(X \le 0)$.
   b. Calculate $P(X = 50)$.
   c. Is this a discrete uniform distribution? Explain.

Applications

5. Identify the possible values of the following random variables. Which of the random variables are discrete?
   a. The numerical grade a student receives in a course.
   b. The grade point average of a student.
   c. The salary of an employee, defined in figures (four-figure, five-figure, etc.).
   d. The salary of an employee defined in dollars.

6. Identify the possible values of the following random variables. Which of the random variables are discrete?
   a. The advertised size of a round Domino's pizza.
   b. The actual size of a round Domino's pizza.
   c. The number of daily visitors to Yosemite National Park.
   d. The age of a visitor to Yosemite National Park.

7. India is the second most populous country in the world, with a population of over 1 billion people. Although the government has offered various incentives for population control, some argue that the birth rate, especially in rural India, is still too high to be sustainable. A demographer assumes the following probability distribution for the household size in India.

Household Size	Probability
1	0.05
2	0.09
3	0.12
4	0.24
5	0.25
6	0.12
7	0.07
8	0.06

a. What is the probability that there are less than 5 members in a household in India?
b. What is the probability that there are 5 or more members in a household in India?
c. What is the probability that the number of members in a household in India is strictly between 3 and 6?
d. Graphically depict this probability distribution and comment on its symmetry.

8. A financial analyst creates the following probability distribution for the performance of an equity income mutual fund.

Performance	Numerical Score	Probability
Very poor	1	0.14
Poor	2	0.43
Neutral	3	0.22
Good	4	0.16
Very good	5	0.05

a. Comment on the optimism or pessimism depicted in the analyst's estimates.
b. Convert the above probability distribution to a cumulative probability distribution.
c. What is the probability that this mutual fund will do at least Good?

9. A basketball player is fouled while attempting to make a basket and receives two free throws. The opposing coach believes there is a 55% chance that the player will miss both shots, a 25% chance that he will make one of the shots, and a 20% chance that he will make both shots.
a. Construct the appropriate probability distribution.
b. What is the probability that he makes no more than one of the shots?
c. What is the probability that he makes at least one of the shots?

10. After Donald Trump won the election, the consumer confidence index rose to 93.8, a six-month high (www.bloomberg.com, November 23, 2016). Given new economic data, an analyst believes that there is a 75% chance that the index will fall below 90 and only a 5% chance that it will rise above 95. The analyst defines the confidence score as 1 if the index is below 90, 2 if it is between 90 and 95, and 3 if it is above 95.
a. According to the analyst, what is the probability that the confidence score is 2?
b. According to the analyst, what is the probability that the confidence score is not 1?

11. Professor Sanchez has been teaching Principles of Economics for over 25 years. He uses the following scale for grading.

Grade	Numerical Score	Probability
A	4	0.10
B	3	0.30
C	2	0.40
D	1	0.10
F	0	0.10

a. Depict the above probability distribution graphically. Comment on whether or not the probability distribution is symmetric.
b. Convert the above probability distribution to a cumulative probability distribution.
c. What is the probability of earning at least a B in Professor Sanchez's course?
d. What is the probability of passing Professor Sanchez's course?

12. Jane Wormley is a professor of management at a university. She expects to be able to use her grant money to fund up to two students for research assistance. While she realizes that there is a 5% chance that she may not be able to fund either student, there is an 80% chance that she will be able to fund two students.
a. What is the probability that Jane will fund one student?
b. Construct a cumulative probability distribution of the random variable defined as the number of students that Jane will be able to fund.

13. Fifty percent of the customers who go to Sears Auto Center for tires buy four tires and 30% buy two tires. Moreover, 18% buy fewer than two tires, with 5% buying none.
a. What is the probability that a customer buys three tires?
b. Construct a cumulative probability distribution for the number of tires bought.

5.2 EXPECTED VALUE, VARIANCE, AND STANDARD DEVIATION

LO **5.2**

Calculate and interpret summary measures for a discrete random variable.

The analysis of probability distributions is useful because it allows us to calculate various probabilities associated with the different values that the random variable assumes. In addition, it helps us calculate summary measures for a random variable. These summary measures include the mean, the variance, and the standard deviation.

Expected Value

One of the most important probabilistic concepts in statistics is that of the **expected value**, also referred to as the population mean. The expected value of the discrete random variable X, denoted by $E(X)$, or simply by μ, is a weighted average of all possible values of X. Before we present its formula, we would like to point out that the expected value of a random variable should not be confused with its most probable value. As we will see later, the expected value is, in general, not even one of the possible values of the random variable. We can think of the expected value as the long-run average value of the random variable over infinitely many independent repetitions of an experiment. Consider a simple experiment with a fair coin, where you win \$10 if it is heads and lose \$10 if it is tails. If you flip the coin many times, the expected gain is \$0, which is neither of the two possible values, namely \$10 or −\$10.

EXPECTED VALUE OF A DISCRETE RANDOM VARIABLE

For a discrete random variable X with values $x_1, x_2, x_3, \ldots$, which occur with probabilities $P(X = x_i)$, the expected value of X is calculated as

$$E(X) = \mu = \Sigma x_i P(X = x_i).$$

Variance and Standard Deviation

The mean μ of the random variable X provides us with a measure of the central location, but it does not give us information on how the various values are dispersed from μ. We again use the measures of variance and standard deviation to indicate whether the values of X are clustered about μ or widely scattered from μ.

VARIANCE AND STANDARD DEVIATION OF A DISCRETE RANDOM VARIABLE

For a discrete random variable X with values $x_1, x_2, x_3, \ldots$, which occur with probabilities $P(X = x_i)$, the variance of X is calculated as

$$Var(X) = \sigma^2 = \Sigma(x_i - \mu)^2 P(X = x_i).$$

The standard deviation of X is the positive square root of the variance of X or, equivalently, $SD(X) = \sigma = \sqrt{\sigma^2}$.

EXAMPLE 5.3

Brad Williams is the owner of a large car dealership in Chicago. Brad decides to construct an incentive compensation program that equitably and consistently compensates employees on the basis of their performance. He offers an annual bonus of \$10,000 for superior performance, \$6,000 for good performance, \$3,000 for fair performance, and \$0 for poor performance. Based on prior records, he expects an employee to perform at superior, good, fair, and poor performance levels with probabilities 0.15, 0.25, 0.40, and 0.20, respectively. Table 5.4 lists the bonus amount, performance type, and the corresponding probabilities.

TABLE 5.4 Probability Distribution for Compensation Program

Bonus (in \$1,000s)	Performance Type	Probability
10	Superior	0.15
6	Good	0.25
3	Fair	0.40
0	Poor	0.20

a. Calculate the expected value of the annual bonus amount.

b. Calculate the variance and the standard deviation of the annual bonus amount.

c. What is the total annual amount that Brad can expect to pay in bonuses if he has 25 employees?

SOLUTION:

a. Let the random variable X denote the bonus amount (in \$1,000s) for an employee. The first and second columns of Table 5.5 represent the probability distribution of X. The calculations for the mean are provided in the third column. We weigh each outcome by its respective probability, $x_i P(X = x_i)$, and then sum these weighted values. Thus, as shown at the bottom of the third column, $E(X) = \mu = \Sigma x_i P(X = x_i) = 4.2$, or \$4,200. Note that the expected value is not one of the possible values of X; that is, none of the employees will earn a bonus of \$4,200. This outcome reinforces the interpretation of expected value as a long-run average.

TABLE 5.5 Calculations for Example 5.3

Value, x_i	Probability, $P(X = x_i)$	Weighted Value, $x_i P(X = x_i)$	Weighted Squared Deviation, $(x_i - \mu)^2 P(X = x_i)$
10	0.15	$10 \times 0.15 = 1.5$	$(10 - 4.2)^2 \times 0.15 = 5.05$
6	0.25	$6 \times 0.25 = 1.5$	$(6 - 4.2)^2 \times 0.25 = 0.81$
3	0.40	$3 \times 0.40 = 1.2$	$(3 - 4.2)^2 \times 0.40 = 0.58$
0	0.20	$0 \times 0.20 = 0$	$(0 - 4.2)^2 \times 0.20 = 3.53$
		Total = 4.2	Total = 9.97

b. The last column of Table 5.5 shows the calculation for the variance. We first calculate each x_i's squared difference from the mean $(x_i - \mu)^2$, weigh each value by the appropriate probability, $(x_i - \mu)^2 P(X = x_i)$, and then sum these weighted squared differences. Thus, as shown at the bottom of the last column, $Var(X) = \sigma^2 = \Sigma(x_i - \mu)^2 P(X = x_i) = 9.97$, or 9.97 (in (\$1,000s)^2). The standard deviation is the positive square root of the variance, $SD(X) = \sigma = \sqrt{9.97} = 3.158$, or \$3,158.

c. In part a we found that the expected bonus for an employee is \$4,200. Since Brad has 25 employees, he can expect to pay \$4,200 × 25 = \$105,000 in bonuses.

Risk Neutrality and Risk Aversion

An important concept in economics, finance, and psychology relates to the behavior of consumers under uncertainty. Consumers are said to be **risk neutral** if they are indifferent to risk and care only about their expected gains. They are said to be **risk-averse** if they care about risk and, if confronted with two choices with the same expected gains, they prefer the one with lower risk. In other words, a risk-averse consumer will take a risk only if it entails a suitable compensation. Consider a seemingly fair gamble where you flip a coin and get \$10 if it is heads and lose \$10 if it is tails, resulting in an expected gain of zero ($10 \times 0.5 - 10 \times 0.5 = 0$). A risk-neutral consumer is indifferent about participating in this gamble. For a risk-averse consumer, the pain associated with losing \$10 is more than the pleasure of winning \$10. Therefore, the risk-averse consumer will not want to participate in this seemingly fair gamble because there is no reward to compensate for the risk. Example 5.4 expands on this type of consumer behavior.

A CONSUMER'S RISK PREFERENCE

A risk-neutral consumer completely ignores risk and makes his/her decisions solely on the basis of expected gains. A risk-averse consumer demands a positive expected gain as compensation for taking risk. This compensation increases with the level of risk taken and the degree of risk aversion.

EXAMPLE 5.4

You have a choice of receiving $1,000 in cash or receiving a beautiful painting from your grandmother. The actual value of the painting is uncertain. You are told that the painting has a 20% chance of being worth $2,000, a 50% chance of being worth $1,000, and a 30% chance of being worth $500. What should you do?

SOLUTION: Let the random variable X represent the worth of the painting. Table 5.6 shows the probability distribution of X.

TABLE 5.6 Probability Distribution for the Value of the Painting

x	$P(X = x)$
500	0.30
1,000	0.50
2,000	0.20

We calculate the expected value as

$$E(X) = \Sigma x_i P(X = x_i) = 500 \times 0.30 + 1{,}000 \times 0.50 + 2{,}000 \times 0.20$$
$$= \$1{,}050.$$

Since the expected value of the painting is more than $1,000, it may appear that the right choice is to pick the painting over $1,000 in cash. This choice, however, is based entirely on the expected value of the painting, paying no attention to risk. While the expected value of $1,050 is more than $1,000, the painting entails some risk. For instance, there is a 30% chance that it may be worth only $500. Therefore, a risk-neutral consumer will take the painting because its expected value exceeds the risk-free cash value of $1,000. This consumer is not concerned with risk. For a risk-averse consumer, however, the decision is not clear-cut. It depends on the risk involved in picking the painting and how much he/she wants to be compensated for this risk. One way to resolve this issue is to define the utility function of the consumer, which in essence conveys the degree of risk aversion. A risk-averse consumer will pick the risky prospect if the expected utility (not the expected money) of the risky prospect exceeds the utility of a risk-free alternative. Further details are beyond the scope of this text.

EXERCISES 5.2

Mechanics

14. Calculate the mean, the variance, and the standard deviation of the following discrete probability distribution.

x	5	10	15	20
$P(X = x)$	0.35	0.30	0.20	0.15

15. Calculate the mean, the variance, and the standard deviation of the following discrete probability distribution.

x	−23	−17	−9	−3
$P(X = x)$	0.50	0.25	0.15	0.10

Applications

16. The number of homes that a realtor sells over a one-month period has the following probability distribution.

Number of Houses Sold	Probability
0	0.30
1	0.50
2	0.15
3	0.05

    a. On average, how many houses is the realtor expected to sell over a one-month period?
    b. What is the standard deviation of this probability distribution?

17. A marketing firm is considering making up to three new hires. Given its specific needs, the firm feels that there is a 60% chance of hiring at least two candidates. There is only a 5% chance that it will not make any hires and a 10% chance that it will make all three hires.
    a. What is the probability that the firm will make at least one hire?
    b. Find the expected value and the standard deviation of the number of hires.

18. An analyst has developed the following probability distribution for the rate of return for a common stock.

Scenario	Probability	Rate of Return (in %)
1	0.30	−5
2	0.45	0
3	0.25	10

    a. Calculate the expected rate of return.
    b. Calculate the variance and the standard deviation of this probability distribution.

19. Organizers of an outdoor summer concert in Toronto are concerned about the weather conditions on the day of the concert. They will make a profit of $25,000 on a clear day and $10,000 on a cloudy day. They will take a loss of $5,000 if it rains. The weather channel has predicted a 60% chance of rain on the day of the concert. Calculate the expected profit from the concert if the likelihood is 10% that it will be sunny and 30% that it will be cloudy.

20. Mark Underwood is a professor of economics at Indiana University. He has been teaching Principles of Economics for over 25 years. Professor Underwood uses the following scale for grading.

Grade	Probability
A	0.10
B	0.30
C	0.40
D	0.10
F	0.10

    Calculate the expected numerical grade in Professor Underwood's class using 4.0 for A, 3.0 for B, etc.

21. The manager of a publishing company plans to give a $20,000 bonus to the top 15%, $10,000 to the next 30%, and $5,000 to the next 10% of sales representatives. If the publishing company has a total of 200 sales representatives, what is the expected bonus that the company will pay?

22. An appliance store sells additional warranties on its refrigerators. Twenty percent of the buyers buy the limited warranty for $100 and 5% buy the extended warranty for $200. What is the expected revenue for the store from the warranty if it sells 120 refrigerators?

23. You are considering buying insurance for your new laptop computer, which you have recently bought for $1,500. The insurance premium for three years is $80. Over the three-year period there is an 8% chance that your laptop computer will require work worth $400, a 3% chance that it will require work worth $800, and a 2% chance that it will completely break down with a scrap value of $100. Should you buy the insurance? (Assume risk neutrality.)

24. Four years ago, Victor purchased a very reliable automobile. His warranty has just expired, but the manufacturer has just offered him a 5-year, bumper-to-bumper warranty extension. The warranty costs $3,400. Victor constructs the following probability distribution with respect to anticipated costs if he chooses not to purchase the extended warranty.

Cost (in $)	Probability
1,000	0.25
2,000	0.45
5,000	0.20
10,000	0.10

    a. Calculate Victor's expected cost.
    b. Given your answer in part a, should Victor purchase the extended warranty? (Assume risk neutrality.) Explain.

25. An investor considers investing $10,000 in the stock market. He believes that the probability is 0.30 that the economy will improve, 0.40 that it will stay the same, and 0.30 that it will deteriorate. Further, if the economy improves, he expects his investment to grow to $15,000, but it can also go down to $8,000 if the economy deteriorates. If the economy stays the same, his investment will stay at $10,000.
    a. What is the expected value of his investment?
    b. Should he invest the $10,000 in the stock market if he is risk neutral?
    c. Is the decision clear-cut if he is risk averse? Explain.

26. You are considering two mutual funds as an investment. The possible returns for the funds are dependent on the state of the economy and are given in the accompanying table.

State of the Economy	Fund 1 (in %)	Fund 2 (in %)
Good	20	40
Fair	10	20
Poor	−10	−40

You believe that the likelihood is 20% that the economy will be good, 50% that it will be fair, and 30% that it will be poor.

a. Find the expected value and the standard deviation of returns for Fund 1.

b. Find the expected value and the standard deviation of returns for Fund 2.

c. Which fund will you pick if you are risk averse? Explain.

27. Investment advisors recommend risk reduction through international diversification. International investing allows you to take advantage of the potential for growth in foreign economies, particularly in emerging markets. Janice Wong is considering investment in either Europe or Asia. She has studied these markets and believes that both markets will be influenced by the U.S. economy, which has a 20% chance for being good, a 50% chance for being fair, and a 30% chance for being poor. Probability distributions of the returns for these markets are given in the accompanying table.

State of the U.S. Economy	Returns in Europe (in %)	Returns in Asia (in %)
Good	10	18
Fair	6	10
Poor	−6	−12

a. Find the expected value and the standard deviation of returns in Europe and Asia.

b. What will Janice pick as an investment if she is risk neutral?

c. Discuss Janice's decision if she is risk averse.

LO **5.3**

Calculate and interpret probabilities for a binomial random variable.

5.3 THE BINOMIAL DISTRIBUTION

Different types of experiments generate different probability distributions. In the next three sections, we discuss three special cases: the binomial, the Poisson, and the hypergeometric probability distributions. Here we focus on the binomial distribution. Before we can discuss the binomial distribution, we first must ensure that the experiment satisfies the conditions of a **Bernoulli process**, which is a particular type of experiment named after the person who first described it, the Swiss mathematician James Bernoulli (1654–1705).

A BERNOULLI PROCESS

A Bernoulli process consists of a series of n independent and identical trials of an experiment such that on each trial

- There are only two possible outcomes, conventionally labeled success and failure; and
- The probabilities of success and failure remain the same from trial to trial.

We use p to denote the probability of success, and therefore, $1 - p$ is the probability of failure.

A **binomial random variable** is defined as the number of successes achieved in the n trials of a Bernoulli process. The possible values of a binomial random variable include $0, 1, \ldots, n$. Many experiments fit the conditions of a Bernoulli process. For instance,

- A bank grants or denies a loan to a mortgage applicant.
- A consumer either uses or does not use a credit card.
- An employee travels or does not travel by public transportation.
- A life insurance policy holder dies or does not die.
- A drug is either effective or ineffective.
- A college graduate applies or does not apply to graduate school.

Our goal is to attach probabilities to various outcomes of a Bernoulli process. The result is a **binomial probability distribution**, or simply, a **binomial distribution**.

A binomial random variable X is defined as the number of successes achieved in the n trials of a Bernoulli process. The binomial distribution for X shows the probabilities associated with the possible values of X.

We will eventually arrive at a general formula that helps us derive a binomial distribution. First, however, we will use a specific example and construct a **probability tree** in order to illustrate the possible outcomes and their associated probabilities.

EXAMPLE 5.5

From past experience, a manager of an upscale shoe store knows that 85% of her customers will use a credit card when making purchases. Suppose three customers are in line to make a purchase.

a. Does this example satisfy the conditions of a Bernoulli process?

b. Construct a probability tree.

c. Using the probability tree, derive the binomial probability distribution.

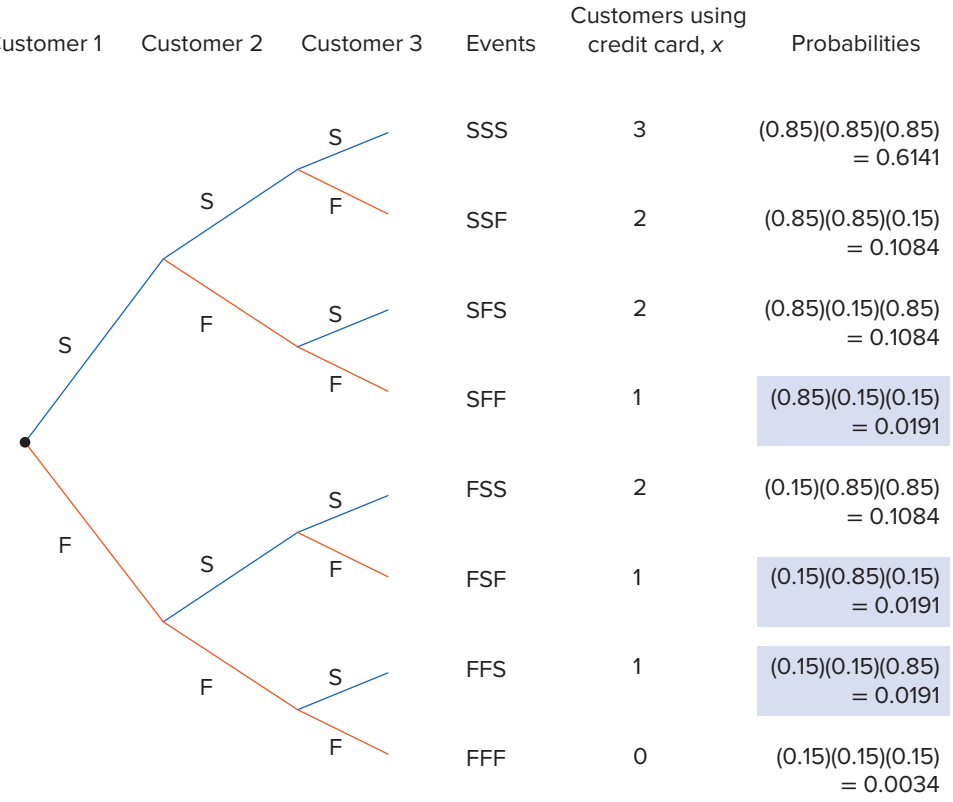
SOLUTION:

a. This example satisfies the conditions of a Bernoulli process because a customer either uses a credit card (labeled success), with an 85% likelihood, or does not use a credit card (labeled failure), with a 15% likelihood. Moreover, given a large number of customers, these probabilities of success and failure do not change from customer to customer.

b. Recall from Chapter 4 that we can use a probability tree whenever an experiment can be broken down into stages. Here we can view each stage as a trial. The probability tree for Example 5.5 is shown in Figure 5.3. We let S denote the outcome that a customer uses a credit card and F denote the outcome that a customer does not use a credit card. Starting from the unlabeled node on the left, customer 1 has an 85% chance of using a credit card and a 15% chance of not using one. The branches emanating from customer 1 denote conditional probabilities of customer 2 using a credit card, given whether or not customer 1 used a credit card. However, since we assume that the trials of a Bernoulli process are independent, the conditional probability is the same as the unconditional probability. In other words, customer 2 has the same 85% chance of using a credit card and a 15% chance of not using one regardless of what customer 1 uses. The same holds for the probabilities for customer 3. The fourth column shows that there are eight possible events at the end of the probability tree. We are able to obtain relevant probabilities by using the multiplication rule for independent events. For instance, following the top branches throughout the probability tree, we calculate the probability that all three customers use a credit card as $(0.85)(0.85)(0.85) = 0.6141$. The probabilities for the remaining events are found in a similar manner.

c. Since we are not interested in identifying the particular customer who uses a credit card, but rather the number of customers who use a credit card, we can combine events with the same number of successes, using the addition rule for mutually exclusive events. For instance, in order to find the probability that one customer uses a credit card, we add the probabilities that correspond to the outcome $x = 1$

FIGURE 5.3 Probability tree for Example 5.5



(see shaded areas in Figure 5.3): $0.0191 + 0.0191 + 0.0191 = 0.0573$. Similarly, we calculate the remaining probabilities corresponding to the other values of X and construct the probability distribution shown in Table 5.7.

TABLE 5.7 Binomial Probabilities for Example 5.5

x	$P(X = x)$
0	0.0034
1	0.0573
2	0.3252
3	0.6141
	Total = 1

Fortunately, we do not have to construct a probability tree each time we want to construct a binomial distribution. We can use the following formula for calculating probabilities associated with a binomial random variable.

THE BINOMIAL DISTRIBUTION

For a binomial random variable X, the probability of x successes in n Bernoulli trials is

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

for $x = 0, 1, 2, \ldots, n$. By definition, $0! = 1$.

The formula consists of two parts:

- The first term, $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, is referred to as the binomial coefficient. It tells us how many sequences with x successes and $n - x$ failures are possible in n trials. For instance, in order to calculate the number of sequences that contain exactly 1 credit card user in 3 trials, we substitute $x = 1$ and $n = 3$ into the formula and calculate $\binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{3!}{1!(3-1)!} = \frac{3 \times 2 \times 1}{(1) \times (2 \times 1)} = 3$. So there are three sequences having exactly 1 success—we can verify this result with Figure 5.3.
- The second part of the equation, $p^x(1-p)^{n-x}$, represents the probability of any particular sequence with x successes and $n - x$ failures. For example, we can obtain the probability of 1 success in 3 trials from row 4, row 6, or row 7 in the last column of Figure 5.3 (see shaded areas) as

$$\left.\begin{array}{l}\text{row 4: } 0.85 \times 0.15 \times 0.15\\ \text{row 6: } 0.15 \times 0.85 \times 0.15\\ \text{row 7: } 0.15 \times 0.15 \times 0.85\end{array}\right\} \quad \text{or} \quad (0.85)^1 \times (0.15)^2 = 0.0191.$$

  In other words, each sequence consisting of 1 success in 3 trials has a 1.91% chance of occurring.

  In order to obtain the overall probability of getting 1 success in 3 trials, we then multiply the binomial coefficient by the probability of obtaining the particular sequence, or here, $3 \times 0.0191 = 0.0573$. This is precisely the probability that we found for $P(X = 1)$ using the probability tree.

Moreover, we could use the formulas shown in Section 5.2 to calculate the expected value, the variance, and the standard deviation for any binomial random variable. Fortunately, for the binomial distribution, these formulas simplify to $E(X) = np$, $Var(X) = np(1 - p)$, and $SD(X) = \sqrt{np(1-p)}$. The simplified formula for the expected value is rather intuitive in that if we know the probability of success p of an experiment and we repeat the experiment n times, then on average, we expect np successes.

EXPECTED VALUE, VARIANCE, AND STANDARD DEVIATION FOR A BINOMIAL RANDOM VARIABLE

If X is a binomial random variable, then

$$E(X) = \mu = np,$$
$$Var(X) = \sigma^2 = np(1-p), \text{ and}$$
$$SD(X) = \sigma = \sqrt{np(1-p)}.$$

For instance, for the binomial probability distribution assumed in Example 5.5, we can derive the expected value with the earlier general formula as

$$E(X) = \Sigma x_i P(X = x_i) = (0 \times 0.0034) + (1 \times 0.0573) + (2 \times 0.3252) + (3 \times 0.6141) = 2.55.$$

However, an easier way is to use $E(X) = np$ and thus calculate the expected value as $3 \times 0.85 = 2.55$. Similarly, the variance and the standard deviation can be easily calculated as

$$Var(X) = \sigma^2 = np(1-p) = 3 \times 0.85 \times 0.15 = 0.38 \text{ and}$$
$$SD(X) = \sigma = \sqrt{np(1-p)} = \sqrt{0.38} = 0.62.$$

EXAMPLE 5.6

In the United States, about 30% of adults have four-year college degrees (*The Wall Street Journal*, April 26, 2012). Suppose five adults are randomly selected.

a. What is the probability that none of the adults has a college degree?

b. What is the probability that no more than two of the adults have a college degree?

c. What is the probability that at least two of the adults have a college degree?

d. Calculate the expected value, the variance, and the standard deviation of this binomial distribution.

e. Graphically depict the probability distribution and comment on its symmetry/skewness.

SOLUTION: First, this problem satisfies the conditions for a Bernoulli process with a random selection of five adults, $n = 5$. Here, an adult either has a college degree, with probability $p = 0.30$, or does not have a college degree, with probability $1 - p = 1 - 0.30 = 0.70$. Given a large number of adults, it fulfills the requirement that the probability that an adult has a college degree stays the same from adult to adult.

a. In order to find the probability that none of the adults has a college degree, we let $x = 0$ and find

$$\begin{aligned} P(X = 0) &= \frac{5!}{0!(5-0)!} \times (0.30)^0 \times (0.70)^{5-0} \\ &= \frac{5 \times 4 \times \cdots \times 1}{(1) \times (5 \times 4 \times \cdots \times 1)} \times 1 \times (0.70)^5 = 1 \times 1 \times 0.1681 \\ &= 0.1681. \end{aligned}$$

In other words, from a random sample of five adults, there is a 16.81% chance that none of the adults has a college degree.

b. We find the probability that no more than two adults have a college degree as

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2).$$

We have already found $P(X = 0)$ from part a. So we now compute $P(X = 1)$ and $P(X = 2)$:

$$P(X = 1) = \frac{5!}{1!(5-1)!} \times (0.30)^1 \times (0.70)^{5-1} = 0.3602$$

$$P(X = 2) = \frac{5!}{2!(5-2)!} \times (0.30)^2 \times (0.70)^{5-2} = 0.3087$$

Next we sum the three relevant probabilities and obtain $P(X \leq 2) = 0.1681 + 0.3602 + 0.3087 = 0.8370$. From a random sample of five adults, there is an 83.7% likelihood that no more than two of them will have a college degree.

c. We find the probability that at least two adults have a college degree as

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5).$$

We can solve this problem by calculating and then summing each of the four probabilities, from $P(X = 2)$ to $P(X = 5)$. A simpler method uses one of the key properties of a probability distribution, which states that the sum of the probabilities over all values of X equals 1. Therefore, $P(X \geq 2)$ can be written as

$1 - [P(X = 0) + P(X = 1)]$. We have already calculated $P(X = 0)$ and $P(X = 1)$ from parts a and b, so

$$P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)] = 1 - (0.1681 + 0.3602) = 0.4717.$$

From a random sample of five adults, there is a 47.17% likelihood that at least two adults will have a college degree.

d. We use the simplified formulas to calculate the mean, the variance, and the standard deviation as

$$E(X) = np = 5 \times 0.30 = 1.5 \text{ adults},$$
$$Var(X) = \sigma^2 = np(1 - p) = 5 \times 0.30 \times 0.70 = 1.05 \text{ (adults)}^2, \text{ and}$$
$$SD(X) = \sigma = \sqrt{np(1 - p)} = \sqrt{1.05} = 1.02 = 1.02 \text{ adults}.$$

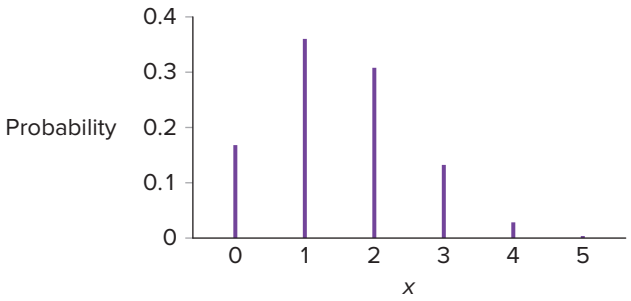
e. Before we graph this distribution, we first show the complete binomial distribution for Example 5.6 in Table 5.8.

TABLE 5.8 Binomial Distribution with $n = 5$ and $p = 0.30$

x	$P(X = x)$
0	0.1681
1	0.3602
2	0.3087
3	0.1323
4	0.0284
5	0.0024

This binomial distribution is graphically depicted in Figure 5.4. When randomly selecting five adults, the most likely outcome is that exactly one adult will have a college degree. The distribution is not symmetric; rather, it is positively skewed. In later chapters, we will learn that the binomial distribution is approximately symmetric when the sample size n is large.

FIGURE 5.4 Binomial distribution with $n = 5$ and $p = 0.30$



Using Excel to Obtain Binomial Probabilities

As you may have noticed, at times it is somewhat tedious and cumbersome to solve binomial distribution problems using the formulas. This issue becomes even more pronounced when we encounter large values for n and we wish to determine probabilities where X assumes a wide range of values. Table 5.9 shows Excel functions that we can use to solve problems associated with discrete probability distributions. Example 5.7

illustrates the use of these functions with respect to the binomial distribution. We will refer back to Table 5.9 in later sections of this chapter when we discuss the Poisson and hypergeometric distributions.

TABLE 5.9 Discrete Probability Distributions and Function Names in Excel

Distribution	Excel
Binomial	
$P(X = x)$	=BINOM.DIST(x, n, p, 0)
$P(X \leq x)$	=BINOM.DIST(x, n, p, 1)
Poisson	
$P(X = x)$	=POISSON.DIST(x, μ, 0)
$P(X \leq x)$	=POISSON.DIST(x, μ, 1)
Hypergeometric	
$P(X = x)$	=HYPGEOM.DIST(x, n, S, N, 0)
$P(X \leq x)$	=HYPGEOM.DIST(x, n, S, N, 1)

EXAMPLE 5.7

In the past decade, the use of technology has skyrocketed, with social media blooming into one of the most valuable methods of communication. People are turning to social media to stay in touch with friends and family members, connect with old friends, catch the news, look for employment, and be entertained. According to a 2016 Pew Research survey, 68% of all U.S. adults are Facebook users. Consider a sample of 100 randomly selected American adults.

a. What is the probability that exactly 70 American adults are Facebook users?

b. What is the probability that no more than 70 American adults are Facebook users?

c. What is the probability that at least 70 American adults are Facebook users?

SOLUTION: We let X denote the number of American adults who are Facebook users. We also know that $p = 0.68$ and $n = 100$.

We use Excel's **BINOM.DIST** function to calculate binomial probabilities. In order to find $P(X = x)$, we enter "=BINOM.DIST(x, n, p, 0)" where x is the number of successes, n is the number of trials, and p is the probability of success. If we enter a "1" for the last argument in the function, then Excel returns $P(X \leq x)$.

a. In order to find the probability that exactly 70 American adults are Facebook users, $P(X = 70)$, we enter "=BINOM.DIST(70, 100, 0.68, 0)" and Excel returns 0.0791.

b. In order to find the probability that no more than 70 American adults are Facebook users, $P(X \leq 70)$, we enter "=BINOM.DIST(70, 100, 0.68, 1)" and Excel returns 0.7007.

c. In order to find the probability that at least 70 American adults are Facebook users, $P(X \geq 70) = 1 - P(X \leq 69)$, we enter "=1−BINOM.DIST(69, 100, 0.68, 1)" and Excel returns 0.3784.

EXERCISES 5.3

Mechanics

28. Assume that X is a binomial random variable with $n = 5$ and $p = 0.35$. Calculate the following probabilities.
    a. $P(X = 0)$
    b. $P(X = 1)$
    c. $P(X \leq 1)$
29. Assume that X is a binomial random variable with $n = 6$ and $p = 0.68$. Calculate the following probabilities.
    a. $P(X = 5)$
    b. $P(X = 4)$
    c. $P(X \geq 4)$
30. Assume that X is a binomial random variable with $n = 8$ and $p = 0.32$. Calculate the following probabilities.
    a. $P(3 < X < 5)$
    b. $P(3 < X \leq 5)$
    c. $P(3 \leq X \leq 5)$
31. Let the probability of success on a Bernoulli trial be 0.30. In five Bernoulli trials, what is the probability that there will be (a) four failures, and (b) more than the expected number of failures?
32. Let X represent a binomial random variable with $n = 150$ and $p = 0.36$. Use Excel's function options to find the following probabilities.
    a. $P(X \leq 50)$
    b. $P(X = 40)$
    c. $P(X > 60)$
    d. $P(X \geq 55)$
33. Let X represent a binomial random variable with $n = 200$ and $p = 0.77$. Use Excel's function options to find the following probabilities.
    a. $P(X \leq 150)$
    b. $P(X > 160)$
    c. $P(155 \leq X \leq 165)$
    d. $P(X = 160)$

Applications

34. According to a survey by Transamerica Center for Health Studies, 15% of Americans still have no health insurance even after passage of the Affordable Care Act, better known as Obamacare (www.cbsnews.com, September 24, 2014). Suppose five individuals are randomly selected.
    a. What is the probability that all five have health insurance?
    b. What is the probability that no more than two have health insurance?
    c. What is the probability that at least four have health insurance?
    d. What is the expected number of individuals who have health insurance?
    e. Calculate the variance and the standard deviation for this probability distribution.
35. At a local community college, 40% of students who enter the college as freshmen go on to graduate. Ten freshmen are randomly selected.
    a. What is the probability that none of them graduates from the local community college?
    b. What is the probability that at most nine will graduate from the local community college?
    c. What is the expected number that will graduate?
36. In 2013, only 26% of Americans had confidence in U.S. banks, which is still far below the pre-recession level of 41% reported in June 2007 (www.gallup.com, June 26, 2014).
    a. What is the probability that fewer than half of four Americans in 2013 have confidence in U.S. banks?
    b. What would have been the corresponding probability in 2007?
37. Approximately 45% of Baby Boomers—those born between 1946 and 1964—are still in the workforce (www.pewresearch.org, May 11, 2015). Six Baby Boomers are selected at random.
    a. What is the probability that exactly one of the Baby Boomers is still in the workforce?
    b. What is the probability that at least five of the Baby Boomers are still in the workforce?
    c. What is the probability that less than two of the Baby Boomers are still in the workforce?
    d. What is the probability that more than the expected number of the Baby Boomers are still in the workforce?
38. In an analysis of Census figures, one in four American counties has passed or is approaching the tipping point where black, Hispanic, and Asian children constitute a majority of the under-20 population (*The New York Times,* August 6, 2008). Racial and ethnic minorities now account for 43% of Americans under 20.
    a. What is the expected number of whites in a random sample of 5,000 under-20 Americans? What is the corresponding standard deviation?
    b. What is the expected number of racial and ethnic minorities in a random sample of 5,000 under-20 Americans? What is the corresponding standard deviation?
    c. If you randomly sample six American counties, what is the probability that for the under-20 population, whites have a majority in all of the counties?
39. Sikhism, a religion founded in the 15th century in India, is going through turmoil due to a rapid decline in the number of Sikh youths who wear turbans (*The Washington Post,* March 29, 2009). The tedious task of combing and tying up long hair and a desire to assimilate has led to approximately 25% of Sikh youths giving up the turban.
    a. What is the probability that exactly two in a random sample of five Sikh youths wear a turban?
    b. What is the probability that two or more in a random sample of five Sikh youths wear a turban?

c. What is the probability that more than the expected number of Sikh youths wear a turban in a random sample of five Sikh youths?

d. What is the probability that more than the expected number of Sikh youths wear a turban in a random sample of 10 Sikh youths?

40. According to the U.S. Census, roughly half of all marriages in the United States end in divorce. Researchers from leading universities have shown that the emotions aroused by one person's divorce can transfer like a virus, making divorce contagious (*CNN,* June 10, 2010). A split-up between immediate friends increases a person's own chances of getting divorced from 36% to 63%, an increase of 75%.
    a. Compute the probability that more than half of four randomly selected marriages will end in divorce.
    b. Redo part a if it is known that the couple's immediate friends have split up.
    c. Redo part a if it is known that none of the couple's immediate friends has split up.

41. Sixty percent of a firm's employees are men. Suppose four of the firm's employees are randomly selected.
    a. What is more likely, finding three men and one woman or two men and two women?
    b. Do you obtain the same answer as in part a if 70% of the firm's employees had been men?

42. The principal of an architecture firm tells her client that there is at least a 50% chance of having an acceptable design by the end of the week. She knows that there is only a 25% chance that any one designer would be able to do so by the end of the week.
    a. Would she be correct in her statement to the client if she asks two of her designers to work on the design, independently?
    b. If not, what if she asks three of her designers to work on the design, independently?

43. Suppose 40% of recent college graduates plan on pursuing a graduate degree. Fifteen recent college graduates are randomly selected.
    a. What is the probability that no more than four of the college graduates plan to pursue a graduate degree?
    b. What is the probability that exactly seven of the college graduates plan to pursue a graduate degree?
    c. What is the probability that at least six but no more than nine of the college graduates plan to pursue a graduate degree?

44. At the University of Notre Dame Mendoza College of Business, 40% of the students seeking a master's degree specialize in finance (*Kiplinger's Personal Finance,* March 2009). Twenty master's degree students are randomly selected.
    a. What is the probability that exactly 10 of the students specialize in finance?
    b. What is the probability that no more than 10 of the students specialize in finance?
    c. What is the probability that at least 15 of the students specialize in finance?

45. The Washington, DC, region has one of the fastest-growing foreclosure rates in the nation, as 15,613 homes went into foreclosure during the one-year period ending in February 2008 (*The Washington Post,* June 19, 2008). Over the past year, the number of foreclosures per 10,000 homes is 131 for the Washington area, while it is 87 nationally. In other words, the foreclosure rate is 1.31% for the Washington, DC, area and 0.87% for the nation. Assume that the foreclosure rates remain stable.
    a. What is the probability that in a given year, fewer than 2 out of 100 houses in the Washington, DC, area will go up for foreclosure?
    b. What is the probability that in a given year, fewer than 2 out of 100 houses in the nation will go up for foreclosure?
    c. Comment on the above findings.

LO **5.4**

Calculate and interpret probabilities for a Poisson random variable.

5.4 THE POISSON DISTRIBUTION

Another important discrete probability distribution is the **Poisson distribution**, named after the French mathematician Simeon Poisson (1781–1849). It is particularly useful in problems that deal with finding the number of occurrences of a certain event over time or space, where space refers to area or region. For simplicity, we call these occurrences "successes." Before we can discuss the Poisson distribution, we first must ensure that our experiment satisfies the conditions of a **Poisson process**.

A POISSON PROCESS

An experiment satisfies a Poisson process if

- The number of successes within a specified time or space interval equals any integer between zero and infinity.
- The number of successes counted in nonoverlapping intervals are independent.
- The probability of success in any interval is the same for all intervals of equal size and is proportional to the size of the interval.

For a Poisson process, we define the number of successes achieved in a specified time or space interval as a **Poisson random variable**.

A Poisson random variable counts the number of occurrences (successes) of a certain event over a given interval of time or space.

Like the Bernoulli process, many experiments fit the conditions of a Poisson process. Consider the following examples of Poisson random variables categorized by those relating to time and those relating to space.

Examples of Poisson Random Variables with Respect to Time

- The number of cars that cross the Brooklyn Bridge between 9:00 am and 10:00 am on a Monday morning.
- The number of customers that use a McDonald's drive-thru in a day.
- The number of bankruptcies that are filed in a month.
- The number of homicides that occur in a year.

Examples of Poisson Random Variables with Respect to Space

- The number of defects in a 50-yard roll of fabric.
- The number of schools of fish in 100 square miles.
- The number of leaks in a specified stretch of a pipeline.
- The number of bacteria in a specified culture.

We use the following formula for calculating probabilities associated with a Poisson random variable.

THE POISSON DISTRIBUTION

For a Poisson random variable X, the probability of x successes over a given interval of time or space is

$$P(X = x) = \frac{e^{-\mu}\mu^{x}}{x!},$$

for $x = 0, 1, 2, \ldots$, where μ is the mean number of successes and $e \approx 2.718$ is the base of the natural logarithm.

As with the binomial random variable, we have simplified formulas to calculate the variance and the standard deviation of a Poisson random variable. An interesting fact is that the mean of the Poisson random variable is equal to the variance.

EXPECTED VALUE, VARIANCE, AND STANDARD DEVIATION FOR A POISSON RANDOM VARIABLE

If X is a Poisson random variable, then

$$E(X) = \mu,$$
$$Var(X) = \sigma^2 = \mu, \text{ and}$$
$$SD(X) = \sigma = \sqrt{\mu}.$$

EXAMPLE 5.8

We can now address questions first posed by Anne Jones in the introductory case of this chapter. Recall that Anne is concerned about staffing needs at the Starbucks that she manages. She has specific questions about the probability distribution of customer arrivals at her store. Anne believes that the typical Starbucks customer averages 18 visits to the store over a 30-day month. She has the following questions:

a. How many visits should Anne expect in a 5-day period from a typical Starbucks customer?

b. What is the probability that a customer visits the chain five times in a 5-day period?

c. What is the probability that a customer visits the chain no more than two times in a 5-day period?

d. What is the probability that a customer visits the chain at least three times in a 5-day period?

SOLUTION: In applications of the Poisson distribution, we first determine the mean number of successes in the relevant time or space interval. We use the Poisson process condition that the probability that success occurs in any interval is the same for all intervals of equal size and is proportional to the size of the interval. Here, the relevant mean will be based on the rate of 18 visits over a 30-day month.

a. Given the rate of 18 visits over a 30-day month, we can write the mean for the 30-day period as $\mu_{30} = 18$. For this problem, we compute the proportional mean for a 5-day period as $\mu_5 = 3$ because $\frac{18 \text{ visits}}{30 \text{ days}} = \frac{3 \text{ visits}}{5 \text{ days}}$.

In other words, on average, a typical Starbucks customer visits the store three times over a 5-day period.

b. In order to find the probability that a customer visits the chain five times in a 5-day period, we calculate

$$P(X = 5) = \frac{e^{-3}3^5}{5!} = \frac{(0.0498)(243)}{120} = 0.1008.$$

c. For the probability that a customer visits the chain no more than two times in a 5-day period, we find $P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2)$. We calculate the individual probabilities, and then find the sum:

$$P(X = 0) = \frac{e^{-3}3^0}{0!} = \frac{(0.0498)(1)}{1} = 0.0498,$$

$$P(X = 1) = \frac{e^{-3}3^1}{1!} = \frac{(0.0498)(3)}{1} = 0.1494, \quad \text{and}$$

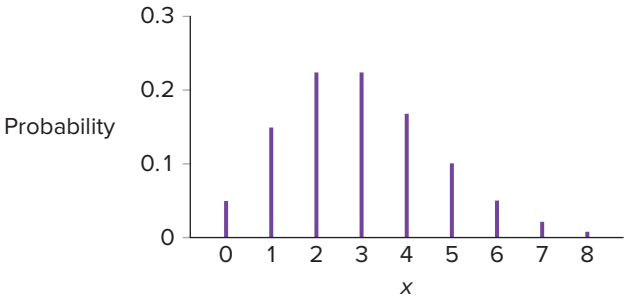
$$P(X = 2) = \frac{e^{-3}3^2}{2!} = \frac{(0.0498)(9)}{2} = 0.2241.$$

Thus, $P(X \le 2) = 0.0498 + 0.1494 + 0.2241 = 0.4233$. There is approximately a 42% chance that a customer visits the chain no more than two times in a 5-day period.

d. We write the probability that a customer visits the chain at least three times in a 5-day period as $P(X \ge 3)$. Initially, we might attempt to solve this problem by evaluating $P(X \ge 3) = P(X = 3) + P(X = 4) + P(X = 5) + \cdots$. However, given the infinite number of possible values, we cannot solve a Poisson problem this way. Here, we find $P(X \ge 3)$ as $1 - [P(X = 0) + P(X = 1) + P(X = 2)]$. Based on the probabilities in part c, we have $P(X \ge 3) = 1 - [0.0498 + 0.1494 + 0.2241] = 1 - 0.4233 = 0.5767$. Thus, there is about a 58% chance that a customer will frequent the chain at least three times in a 5-day period.

Figure 5.5 graphs the Poisson distribution $P(X = x)$ with $\mu = 3$, for x ranging from 0 to 8. The most likely outcomes are when x equals 2 and x equals 3, and the distribution is positively skewed. Remember that, theoretically, the values that the Poisson random variable assumes are infinitely countable, but the probabilities approach zero beyond those shown here.

FIGURE 5.5 Poisson distribution with $\mu = 3$



SYNOPSIS OF INTRODUCTORY CASE



©Monkey Business Images/Shutterstock

Anne Jones, the manager of a Starbucks store, is concerned about how other nearby store closings might affect foot traffic at her store. A solid understanding of the likelihood of customer arrivals is necessary before she can make further statistical inference. Historical data allow her to assume that a typical Starbucks customer averages 18 visits to a Starbucks store over a 30-day month. With this information and the knowledge that she can model customer arrivals using the Poisson distribution, she deduces that a typical customer averages three visits in a 5-day period. The likelihood that a typical customer frequents her store five times in a 5-day period is approximately 10%. Moreover, there is approximately a 42% chance that a typical customer goes to Starbucks no more than two times in a 5-day period, while the chances that this customer visits the chain at least three times is approximately 58%. These preliminary probabilities will prove vital as Anne plans her future staffing needs.

Using Excel to Obtain Poisson Probabilities

Like the binomial formula, the manual use of the Poisson formula can become quite cumbersome, especially when the values of x and μ become large. Excel again proves useful when calculating Poisson probabilities. Table 5.9 shows Excel functions that we can use to find Poisson probabilities. Example 5.9 illustrates the use of these functions.

EXAMPLE 5.9

Craft breweries that make beer in small batches are experiencing a spectacular growth in bars and liquor stores across the nation. The craft beer industry now boasts of 4,269 breweries, representing a 12% market share of the total beer market in the United States (*Fortune*, March 22, 2016). It has been estimated that 1.5 craft breweries open every day. Assume this number represents an average that remains constant over time.

a. What is the probability that no more than 10 craft breweries open every week?

b. What is the probability that exactly 10 craft breweries open every week?

SOLUTION: We let X denote the number of craft breweries that open every week and compute the weekly mean, $\mu = 1.5 \times 7 = 10.5$.

USING EXCEL

We use Excel's **POISSON.DIST** function to calculate Poisson probabilities. In order to find $P(X = x)$, we enter "=POISSON.DIST(x, μ, 0)" where x is the number of successes over some interval and μ is the mean over this interval. If we enter a "1" for the last argument in the function, then Excel returns $P(X \leq x)$.

a. In order to find the probability that no more than 10 craft breweries open every week, $P(X \leq 10)$, we enter "=POISSON.DIST(10, 10.5, 1)" and Excel returns 0.5207. There is a 52.07% chance that no more than 10 craft breweries open every week.

b. In order to find the probability that exactly 10 craft breweries open every week, $P(X = 10)$, we enter "=POISSON.DIST(10, 10.5, 0)" and Excel returns 0.1236. There is a 12.36% chance that 10 craft breweries open every week.

EXERCISES 5.4

Mechanics

46. Assume that X is a Poisson random variable with $\mu = 1.5$. Calculate the following probabilities.
    a. $P(X = 1)$
    b. $P(X = 2)$
    c. $P(X \geq 2)$

47. Assume that X is a Poisson random variable with $\mu = 4$. Calculate the following probabilities.
    a. $P(X = 4)$
    b. $P(X = 2)$
    c. $P(X \leq 1)$

48. Let the mean success rate of a Poisson process be 8 successes per hour.
    a. Find the expected number of successes in a half-hour period.
    b. Find the probability of at least two successes in a given half-hour period.
    c. Find the expected number of successes in a two-hour period.
    d. Find the probability of 10 successes in a given two-hour period.

49. Assume that X is a Poisson random variable with $\mu = 15$. Use Excel's function options to find the following probabilities.
    a. $P(X \leq 10)$
    b. $P(X = 13)$
    c. $P(X > 15)$
    d. $P(12 \leq X \leq 18)$

50. Assume that X is a Poisson random variable with $\mu = 20$. Use Excel's function options to find the following probabilities.
    a. $P(X < 14)$
    b. $P(X \geq 20)$
    c. $P(X = 25)$
    d. $P(18 \leq X \leq 23)$

Applications

51. Which of the following probabilities are likely to be found using a Poisson distribution?
    a. The probability that there will be six leaks in a specified stretch of a pipeline.
    b. The probability that at least 10 students in a class of 40 will land a job right after graduation.
    c. The probability that at least 50 families will visit Acadia National Park over the weekend.
    d. The probability that no customer will show up in the next five minutes.

52. Which of the following scenarios are likely to represent Poisson random variables?
    a. The number of violent crimes in New York over a six-week period.
    b. The number of customers of a bank manager who will default.
    c. The number of scratches on a 2-by-1-foot portion of a large wooden table.
    d. The number of patients of a doctor for whom the drug will be effective.

53. On average, there are 12 potholes per mile on a particular stretch of the state highway. Suppose the potholes are distributed evenly on the highway.
    a. Find the probability of finding fewer than two potholes in a quarter-mile stretch of the highway.
    b. Find the probability of finding more than one pothole in a quarter-mile stretch of the highway.

54. A tollbooth operator has observed that cars arrive randomly at an average rate of 360 cars per hour.
    a. Find the probability that two cars arrive during a specified one-minute period.
    b. Find the probability that at least two cars arrive during a specified one-minute period.
    c. Find the probability that 40 cars arrive between 10:00 am and 10:10 am.

55. A textile manufacturing process finds that on average, two flaws occur per every 50 yards of material produced.
    a. What is the probability of exactly two flaws in a 50-yard piece of material?
    b. What is the probability of no more than two flaws in a 50-yard piece of material?
    c. What is the probability of no flaws in a 25-yard piece of material?

56. Motorists arrive at a Gulf gas station at the rate of two per minute during morning hours.
    a. What is the probability that more than two motorists will arrive at the Gulf gas station during a one-minute interval in the morning?
    b. What is the probability that exactly six motorists will arrive at the Gulf gas station during a five-minute interval in the morning?
    c. How many motorists can an employee expect in her three-hour morning shift?

57. Airline travelers should be ready to be more flexible as airlines once again cancel thousands of flights this summer. The Coalition for Airline Passengers Rights, Health, and Safety averages 400 calls a day to help stranded travelers deal with airlines (seattlepi.com, July 10, 2008). Suppose the hotline is staffed for 16 hours a day.
    a. Calculate the average number of calls in a one-hour interval, 30-minute interval, and 15-minute interval.
    b. What is the probability of exactly six calls in a 15-minute interval?
    c. What is the probability of no calls in a 15-minute interval?
    d. What is the probability of at least two calls in a 15-minute interval?

58. On average, 400 people are struck by lightning in the United States each year (*The Boston Globe,* July 21, 2008).
    a. What is the probability that at most 425 people are struck by lightning in a year?
    b. What is the probability that at least 375 people are struck by lightning in a year?

59. According to a government report, the aging of the U.S. population is translating into many more visits to doctors' offices and hospitals (*USA Today,* August 7, 2008). It is estimated that an average person makes four visits a year to doctors' offices and hospitals.
    a. What are the mean and the standard deviation of an average person's number of visits to doctors' offices and hospitals in a month.
    b. What is the probability that an average person does not make any visits to doctors' offices and hospitals in a month?
    c. What is the probability that an average person makes at least one visit to doctors' offices and hospitals in a month?

60. Due to the advent of tablets, American adults are watching significantly less television than they did in previous decades. In 2016, Nielsen reported that American adults are watching an average of five hours and four minutes, or 304 minutes, of television per day.
    a. Find the probability that an average American adult watches more than 320 minutes of television per day.
    b. Find the probability that an average American adult watches more than 2,200 minutes of television per week.

61. In the fiscal year that ended September 30, 2008, there were 24,584 age-discrimination claims filed with the Equal Employment Opportunity Commission, an increase of 29% from the previous year (*The Wall Street Journal,* March 7–8, 2009). Assume there were 260 working days in the fiscal year for which a worker could file a claim.
    a. Calculate the average number of claims filed on a working day.
    b. What is the probability that exactly 100 claims were filed on a working day?
    c. What is the probability that no more than 100 claims were filed on a working day?

5.5 THE HYPERGEOMETRIC DISTRIBUTION

LO **5.5**

Calculate and interpret probabilities for a hypergeometric random variable.

In Section 5.3, we defined a binomial random variable X as the number of successes in the n trials of a Bernoulli process. The trials, according to a Bernoulli process, are independent and the probability of success does not change from trial to trial. The **hypergeometric distribution** is appropriate in applications where we cannot assume that the trials are independent.

Consider a box full of production items, of which 10% are known to be defective. Let success be labeled as the draw of a defective item. The probability of success may not be the same from trial to trial; it will depend on the size of the population and whether the sampling was done with or without replacement. Suppose the box consists of 20 items

of which 10%, or 2, are defective. The probability of success in the first draw is 0.10 (= 2/20). However, the probability of success in subsequent draws will depend on the outcome of the first draw. For example, if the first item was defective, the probability of success in the second draw will be 0.0526 (= 1/19), while if the first item was not defective, the probability of success in the second draw will be 0.1053 (= 2/19). Therefore, the binomial distribution is not appropriate because the trials are not independent and the probability of success changes from trial to trial.

In the preceding example, we assumed sampling without replacement; in other words, after an item is drawn, it is not put back in the box for subsequent draws. The binomial distribution would be appropriate if we sample with replacement since, in that case, for each draw there will be 20 items, of which 2 are defective, resulting in an unchanging probability of success. Moreover, the dependence of the trials can be ignored if the population size is very large relative to the sample size. For instance, if the box consists of 10,000, items, of which 10%, or 1,000, are defective, then the probability of success in the second draw will be either 999/9,999 or 1,000/9,999, which are both approximately equal to 0.10.

We use the hypergeometric distribution in place of the binomial distribution when we are sampling without replacement from a population whose size N is not significantly larger than the sample size n. The **hypergeometric random variable** is the number of successes achieved in the n trials of a two-outcome experiment, where the trials are not assumed to be independent.

THE HYPERGEOMETRIC DISTRIBUTION

For a hypergeometric random variable X, the probability of x successes in a random selection of n items is

$$P(X = x) = \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}},$$

for $x = 0, 1, 2, \ldots, n$ if $n \leq S$ or $x = 0, 1, 2, \ldots, S$ if $n > S$, where N denotes the number of items in the population of which S are successes.

The formula consists of three parts:

- The first term in the numerator, $\binom{S}{x} = \frac{S!}{x!(S-x)!}$, represents the number of ways x successes can be selected from S successes in the population.
- The second term in the numerator, $\binom{N-S}{n-x} = \frac{(N-S)!}{(n-x)!(N-S-n+x)!}$, represents the number of ways $(n - x)$ failures can be selected from $(N - S)$ failures in the population.
- The denominator, $\binom{N}{n} = \frac{N!}{n!(N-n)!}$, represents the number of ways a sample of size n can be selected from the population of size N.

As with the binomial and Poisson distributions, simplified formulas can be used to calculate the mean, the variance, and the standard deviation of a hypergeometric random variable.

EXPECTED VALUE, VARIANCE, AND STANDARD DEVIATION FOR A HYPERGEOMETRIC RANDOM VARIABLE

If X is a hypergeometric random variable, then

$$E(X) = \mu = n\left(\frac{S}{N}\right),$$

$$Var(X) = \sigma^2 = n\left(\frac{S}{N}\right)\left(1 - \frac{S}{N}\right)\left(\frac{N-n}{N-1}\right), \text{ and}$$

$$SD(X) = \sigma = \sqrt{n\left(\frac{S}{n}\right)\left(1 - \frac{S}{N}\right)\left(\frac{N-n}{N-1}\right)}.$$

EXAMPLE 5.10

Wooden boxes are commonly used for the packaging and transportation of mangoes. A convenience store in Morganville, New Jersey, regularly buys mangoes from a wholesale dealer. For every shipment, the manager randomly inspects five mangoes from a box containing 20 mangoes for damages due to transportation. Suppose the chosen box contains exactly two damaged mangoes.

a. What is the probability that one out of five mangoes used in the inspection is damaged?

b. If the manager decides to reject the shipment if one or more mangoes are damaged, what is the probability that the shipment will be rejected?

c. Calculate the expected value, the variance, and the standard deviation of the number of damaged mangoes used in the inspection.

SOLUTION: The hypergeometric distribution is appropriate because the probability of finding a damaged mango changes from draw to draw (sampling is without replacement and the population size N is not significantly more than the sample size n). We use the following values to solve the problems: $N = 20$, $n = 5$, $S = 2$.

a. The probability that one out of five mangoes is damaged is $P(X = 1)$. We calculate

$$P(X=1)=\frac{\binom{2}{1}\binom{20-2}{5-1}}{\binom{20}{5}}=\frac{\left(\frac{2!}{1!1!}\right)\left(\frac{18!}{4!14!}\right)}{\left(\frac{20!}{5!15!}\right)}=\frac{(2)(3{,}060)}{15{,}504}=0.3947.$$

Therefore, the likelihood that exactly one out of five mangoes is damaged is 39.47%.

b. In order to find the probability that one or more mangoes are damaged, we need to calculate $P(X \geq 1)$. We note that $P(X \geq 1) = 1 - P(X = 0)$ where

$$P(X=0)=\frac{\binom{2}{0}\binom{20-2}{5-0}}{\binom{20}{5}}=\frac{\left(\frac{2!}{0!2!}\right)\left(\frac{18!}{5!13!}\right)}{\left(\frac{20!}{5!15!}\right)}=\frac{(1)(8{,}568)}{15{,}504}=0.5526.$$

Therefore, the probability that the shipment will be rejected equals $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.5526 = 0.4474$.

c. We use the simplified formulas to obtain the mean, the variance, and the standard deviation as

$$E(X)=n\left(\frac{S}{N}\right)=5\left(\frac{2}{20}\right)=0.50,$$

$$Var(X)=n\left(\frac{S}{N}\right)\left(1-\frac{S}{N}\right)\left(\frac{N-n}{N-1}\right)=5\left(\frac{2}{20}\right)\left(1-\frac{2}{20}\right)\left(\frac{20-5}{20-1}\right)=0.3553,$$

$$SD(X)=\sqrt{0.3553}=0.5960.$$

Using Excel to Obtain Hypergeometric Probabilities

Since it is tedious to solve for hypergeometric probabilities by hand, we typically use the computer to aid in the calculations. Table 5.9 shows Excel functions that we can use to find hypergeometric probabilities. Example 5.11 illustrates the use of these functions.

EXAMPLE 5.11

Employment for management occupations is projected to grow 6% from 2014 to 2024, resulting in about 505,400 new jobs (*Bureau of Labor Statistics*, December 2015). Among 25 applicants for a management position, 15 have college degrees in business. Suppose four applicants are randomly chosen for interviews.

a. What is the probability that none of the applicants has a college degree in business?

b. What is the probability that no more than two of the applicants have college degrees in business?

SOLUTION: We let X denote the number of applicants with a college degree in business. We know that $n = 4$, $S = 15$, and $N = 25$.

USING EXCEL

We use Excel's **HYPGEOM.DIST** function to calculate hypergeometric probabilities. In order to find $P(X = x)$, we enter "=HYPGEOM.DIST(x, n, S, N, 0)" where x is the number of successes in the sample, n is the sample size, S is the number of successes in the population, and N is the population size. If we enter a "1" for the last argument in the function, then Excel returns $P(X \leq x)$.

a. In order to find the probability that none of the applicants has a college degree in business, $P(X = 0)$, we enter "=HYPGEOM.DIST(0, 4, 15, 25, 0)" and Excel returns 0.0166.

b. In order to find the probability that no more than two of the applicants have a college degree in business, $P(X \leq 2)$, we enter "=HYPGEOM.DIST(2, 4, 15, 25, 1)" and Excel returns 0.5324.

EXERCISES 5.5

Mechanics

62. Assume that X is a hypergeometric random variable with $N = 25$, $S = 3$, and $n = 4$. Calculate the following probabilities.
    a. $P(X = 0)$
    b. $P(X = 1)$
    c. $P(X \leq 1)$

63. Assume that X is a hypergeometric random variable with $N = 15$, $S = 4$, and $n = 3$. Calculate the following probabilities.
    a. $P(X = 1)$
    b. $P(X = 2)$
    c. $P(X \geq 2)$

64. Compute the probability of no successes in a random sample of three items obtained from a population of 12 items that contains two successes. What are the expected number and the standard deviation of the number of successes from the sample?

65. Assume that X is a hypergeometric random variable with $N = 50$, $S = 20$, and $n = 5$. Use Excel's function options to find the following probabilities.
    a. $P(X = 2)$
    b. $P(X \geq 2)$
    c. $P(X \leq 3)$

66. Compute the probability of at least eight successes in a random sample of 20 items obtained from a population of 100 items that contains 25 successes. What are the expected number and the standard deviation of the number of successes?

Applications

67. Suppose you have an urn of ten marbles, of which five are red and five are green. If you draw two marbles from this urn, what is the probability that both marbles are red? What is the probability that at least one of the marbles is red?

68. A professor of management has heard that eight students in his class of 40 have landed an internship for the summer. Suppose he runs into three of his students in the corridor.
    a. Find the probability that none of these students has landed an internship.
    b. Find the probability that at least one of these students has landed an internship.

69. Despite the repeated effort by the government to reform how Wall Street pays its executives, some of the nation's biggest banks are continuing to pay out bonuses nearly as large as those in the best years before the crisis (*The Washington Post,* January 15, 2010). It is known that 10 out of 15 members of the board of directors of a company were in favor of a bonus. Suppose three members were randomly selected by the media.
    a. What is the probability that all of them were in favor of a bonus?
    b. What is the probability that at least two members were in favor of a bonus?

70. Many programming teams work independently at a large software company. The management has been putting pressure on these teams to finish a project on time. The company currently has 18 large programming projects, of which only 12 are likely to finish on time. Suppose the manager decides to randomly supervise three such projects.
    a. What is the probability that all three projects finish on time?
    b. What is the probability that at least two projects finish on time?

71. David Barnes and his fiancée Valerie Shah are visiting Hawaii. There are 20 guests registered for orientation. It is announced that 12 randomly selected registered guests will receive a free lesson of Tahitian dance.
    a. What is the probability that both David and Valerie get picked for the Tahitian dance lesson?
    b. What is the probability that neither of them gets picked for the Tahitian dance lesson?

72. The National Science Foundation is fielding applications for grants to study climate change. Twenty universities apply for a grant, and only four of them will be awarded. If Syracuse University and Auburn University are among the 20 applicants, what is the probability that these two universities will receive a grant? Assume that the selection is made randomly.

73. A committee of 40 members consists of 24 men and 16 women. A subcommittee consisting of 10 randomly selected members will be formed.
    a. What are the expected number of men and women on the subcommittee?
    b. What is the probability that at least half of the members on the subcommittee will be women?

74. Powerball is a jackpot game with a grand prize starting at $20 million and often rolling over into the hundreds of millions. In 2006, the jackpot was $365 million. The winner may choose to receive the jackpot prize paid over 29 years or as a lump-sum payment. For $1 the player selects six numbers for the base game of Powerball. There are two independent stages of the game. Five balls are randomly drawn from 59 consecutively numbered white balls. Moreover, one ball, called the Powerball, is randomly drawn from 39 consecutively numbered red balls. To be a winner, the numbers selected by the player must match the numbers on the randomly drawn white balls as well as the Powerball.
    a. What is the probability that the player is able to match the numbers of two out of five randomly drawn white balls?
    b. What is the probability that the player is able to match the numbers of all five randomly drawn white balls?
    c. What is the probability that the player is able to match the Powerball for a randomly drawn red ball?
    d. What is the probability of winning the jackpot? [*Hint: Remember that the two stages of drawing white and red balls are independent.*]

WRITING WITH STATISTICS

Senior executives at Skyhigh Construction, Inc., participate in a pick-your-salary plan. They choose salaries in a range between $125,000 and $150,000. By choosing a lower salary, an executive has an opportunity to make a larger bonus. If Skyhigh does not generate an operating profit during the year, then no bonuses are paid. Skyhigh has just hired two new senior executives, Allen Grossman and Felicia Arroyo. Each must decide whether to choose *Option* 1: a base pay of $125,000 with a possibility of a large bonus or *Option* 2: a base pay of $150,000 with a possibility of a bonus, but the bonus would be one-half of the bonus under Option 1.



©Image Source/Getty Images

Grossman, 44 years old, is married with two young children. He bought his home at the height of the market and has a large monthly mortgage payment. Arroyo, 32 years old, just completed her MBA at a prestigious Ivy League university. She is single and has no student loans due to a timely inheritance upon entering graduate school. Arroyo just moved to the area so she has decided to rent an apartment for at least one year. Given their personal profiles, inherent perceptions of risk, and subjective views of the economy, Grossman and Arroyo construct their individual probability distributions with respect to bonus outcomes shown in Table 5.10.

TABLE 5.10 Grossman's and Arroyo's Probability Distributions

Bonus (in $)	Probability	
	Grossman	Arroyo
0	0.35	0.20
50,000	0.45	0.25
100,000	0.10	0.35
150,000	0.10	0.20

Jordan Lake, an independent human resources specialist, is asked to summarize the payment plans with respect to each executive's probability distribution.

Jordan would like to use the above probability distributions to

1. Compute expected values to evaluate payment plans for Grossman and Arroyo.
2. Help Grossman and Arroyo decide whether to choose Option 1 or Option 2 for his/her compensation package.

Sample Report—Comparison of Salary Plans

Skyhigh Construction, Inc., has just hired two new senior executives, Allen Grossman and Felicia Arroyo, to oversee planned expansion of operations. As senior executives, they participate in a pick-your-salary plan. Each executive is given two options for compensation:

Option 1: A base pay of $125,000 with a possibility of a large bonus.
Option 2: A base pay of $150,000 with a possibility of a bonus, but the bonus would be one-half of the bonus under Option 1.

Grossman and Arroyo understand that if the firm does not generate an operating profit in the fiscal year, then no bonuses are paid. Each executive has constructed a probability distribution given his/her personal background, underlying risk preferences, and subjective view of the economy.

Given the probability distributions and with the aid of expected values, the following analysis will attempt to choose the best option for each executive. Grossman, a married father with two young children, believes that Table 5.A best reflects his bonus payment expectations.

TABLE 5.A Calculating Grossman's Expected Bonus

Bonus (in $), x_i	Probability, $P(x_i)$	Weighted Value, $x_i P(x_i)$
0	0.35	$0 \times 0.35 = 0$
50,000	0.45	$50{,}000 \times 0.45 = 22{,}500$
100,000	0.10	$100{,}000 \times 0.10 = 10{,}000$
150,000	0.10	$150{,}000 \times 0.10 = 15{,}000$
		Total = 47,500

Expected bonus, $E(X)$, is calculated as a weighted average of all possible bonus values and is shown at the bottom of the third column of Table 5.A. Grossman's expected bonus is $47,500. Using this value for his bonus, his salary options are

Option 1: $125,000 + $47,500 = $172,500
Option 2: $150,000 + (1/2 × $47,500) = $173,750

Grossman should choose *Option 2* as his salary plan.

Arroyo is single with few financial constraints. Table 5.B shows the expected value of her bonus given her probability distribution.

TABLE 5.B Calculating Arroyo's Expected Bonus

Bonus (in $), x_i	Probability, $P(x_i)$	Weighted Value, $x_i P(x_i)$
0	0.20	$0 \times 0.20 = 0$
50,000	0.25	$50{,}000 \times 0.25 = 12{,}500$
100,000	0.35	$100{,}000 \times 0.35 = 35{,}000$
150,000	0.20	$150{,}000 \times 0.20 = 30{,}000$
		Total = 77,500

Arroyo's expected bonus amounts to $77,500. Thus, her salary options are

Option 1: $125,000 + $77,500 = $202,500
Option 2: $150,000 + (1/2 × $77,500) = $188,750

Arroyo should choose *Option 1* as her salary plan.

CONCEPTUAL REVIEW

LO 5.1 Describe a discrete random variable and its probability distribution.

A **random variable** summarizes outcomes of an experiment with numerical values. A **discrete random variable** assumes a countable number of distinct values, whereas a **continuous random variable** is characterized by uncountable values in an interval.

The **probability mass function** for a discrete random variable X is a list of the values of X with the associated probabilities; that is, the list of all possible pairs $(x, P(X = x))$. The **cumulative distribution function** of X is defined as $P(X \leq x)$.

LO 5.2 Calculate and interpret summary measures for a discrete random variable.

For a discrete random variable X with values $x_1, x_2, x_3, \ldots$, which occur with probabilities $P(X = x_i)$, the **expected value** of X is calculated as $E(X) = \mu = \Sigma x_i P(X = x_i)$. We interpret the expected value as the long-run average value of the random variable over infinitely many independent repetitions of an experiment. Measures of dispersion indicate whether the values of X are clustered about μ or widely scattered from μ. The variance of X is calculated as $Var(X) = \sigma^2 = \Sigma(x_i - \mu)^2 P(X = x_i)$. The standard deviation of X is $SD(X) = \sigma = \sqrt{\sigma^2}$.

In general, a **risk-averse consumer** expects a reward for taking risk. A risk-averse consumer may decline a risky prospect even if it offers a positive expected gain. A **risk-neutral consumer** completely ignores risk and always accepts a prospect that offers a positive expected gain.

LO 5.3 Calculate and interpret probabilities for a binomial random variable.

A **Bernoulli process** is a series of n independent and identical trials of an experiment such that on each trial there are only two possible outcomes, conventionally labeled "success" and "failure." The probabilities of success and failure, denoted p and $1 - p$, remain the same from trial to trial.

For a **binomial random variable** X, the probability of x successes in n Bernoulli trials is $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$ for $x = 0, 1, 2, \ldots, n$.

The expected value, the variance, and the standard deviation of a binomial random variable are $E(X) = np$, $Var(X) = \sigma^2 = np(1-p)$, and $SD(X) = \sigma = \sqrt{np(1-p)}$, respectively.

LO 5.4 **Calculate and interpret probabilities for a Poisson random variable.**

A **Poisson random variable** counts the number of occurrences of a certain event over a given interval of time or space. For simplicity, we call these occurrences "successes." For a Poisson random variable X, the probability of x successes over a given interval of time or space is $P(X = x) = \frac{e^{-\mu}\mu^x}{x!}$ for $x = 0, 1, 2, \ldots$, where μ is the mean number of successes and $e \approx 2.718$ is the base of the natural logarithm. The expected value, the variance, and the standard deviation of a Poisson distribution are $E(X) = \mu$, $Var(X) = \sigma^2 = \mu$, and $SD(X) = \sigma = \sqrt{\mu}$, respectively.

LO 5.5 **Calculate and interpret probabilities for a hypergeometric random variable.**

The hypergeometric distribution is appropriate in applications where the trials are not independent and the probability of success changes from trial to trial. We use it in place of the binomial distribution when we are sampling without replacement from a population whose size N is not significantly larger than the sample size n. For a **hypergeometric random variable** X, the probability of x successes in a random selection of n items is $P(X = x) = \frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}}$ for $x = 0, 1, 2, \ldots, n$ if $n \le S$ or $x = 0, 1, 2, \ldots, S$ if $n > S$, where N denotes the number of items in the population of which S are successes. The expected value, the variance, and the standard deviation of a hypergeometric distribution are $E(X) = n\left(\frac{S}{N}\right)$, $Var(X) = \sigma^2 = n\left(\frac{S}{N}\right)\left(1 - \frac{S}{N}\right)\left(\frac{N-n}{N-1}\right)$, and $SD(X) = \sigma = \sqrt{n\left(\frac{S}{N}\right)\left(1 - \frac{S}{N}\right)\left(\frac{N-n}{N-1}\right)}$, respectively.

ADDITIONAL EXERCISES AND CASE STUDIES

Exercises

75. An analyst developed the following probability distribution for the rate of return for a common stock.

Scenario	Probability	Rate of Return (in %)
1	0.25	−15
2	0.35	5
3	0.40	10

    a. Calculate the expected rate of return.
    b. Calculate the variance and the standard deviation of this probability distribution.

76. Facing the worst economic climate since the dot-com bust in the early 2000s, high-tech companies in the United States search for investment opportunities with cautious optimism (*USA TODAY*, February 17, 2009). Suppose the investment team at Microsoft is considering an innovative start-up project. According to its estimates, Microsoft can make a profit of $5 million if the project is very successful and $2 million if it is somewhat successful. It also stands to lose $4 million if the project fails. Calculate the expected profit or loss for Microsoft if the probabilities that the project is very successful and somewhat successful are 0.10 and 0.40, respectively, with the remaining amount being the failure probability.

77. A professor uses a relative scale for grading. She announces that 60% of the students will get at least a B, with 15% getting A's. Also, 5% will get a D and another 5% will get an F. Assume that no incompletes are given in the course. Let Score be defined by 4 for A, 3 for B, 2 for C, 1 for D, and 0 for F.
    a. Find the probability that a student gets a B.
    b. Find the probability that a student gets at least a C.
    c. Compute the expected value and the standard deviation of Score.

78. Fifty percent of the customers who go to Sears Auto Center for tires buy four tires and 30% buy two tires. Moreover, 18% buy fewer than two tires, with 5% buying none.

a. Find the expected value and the standard deviation of the number of tires a customer buys.
b. If Sears Auto Center makes a $15 profit on every tire it sells, what is its expected profit if it services 120 customers?

79. Forty-four percent of consumers with credit cards carry balances from month to month (bankrate.com, February 20, 2007). Four consumers with credit cards are randomly selected.
a. What is the probability that all four consumers carry a credit card balance?
b. What is the probability that fewer than two consumers carry a credit card balance?
c. Calculate the expected value, the variance, and the standard deviation for this distribution.

80. Rent-to-own (RTO) stores allow consumers immediate access to merchandise in exchange for a series of weekly or monthly payments. The agreement is for a fixed time period. At the same time, the customer has the flexibility to terminate the contract by returning the merchandise. Suppose an RTO store makes a $200 profit on appliances when the customer ends up owning the merchandise by making all payments. It makes a $20 profit when the customer returns the product and a loss of $600 when the customer defaults. Let the return and default probabilities be 0.60 and 0.05, respectively.
a. Construct a probability distribution for the profit per appliance.
b. What is the expected profit for a store that sells 200 rent-to-own contracts?

81. According to the Department of Transportation, 27% of domestic flights were delayed in 2007 (*Money*, May 2008). At New York's John F. Kennedy Airport, five flights are randomly selected.
a. What is the probability that all five flights are delayed?
b. What is the probability that all five are on time?

82. Apple products have become a household name in America, with 51% of all households owning at least one Apple product (*CNN*, March 19, 2012).
a. What is the probability that two in a random sample of four households own an Apple product?
b. What is the probability that all four in a random sample of four households own an Apple product?
c. In a random sample of 100 households, find the expected value and the standard deviation for the number of households that own an Apple product.

83. Twenty percent of U.S. mortgages are "underwater" (*The Boston Globe*, March 5, 2009). A mortgage is considered underwater if the value of the home is less than what is owed on the mortgage. Suppose 100 mortgage holders are randomly selected.
a. What is the probability that exactly 15 of the mortgages are underwater?
b. What is the probability that more than 20 of the mortgages are underwater?
c. What is the probability that at least 25 of the mortgages are underwater?

84. According to a survey by consulting firm Watson Wyatt, approximately 19% of employers have eliminated perks or plan to do so in the next year (*Kiplinger's Personal Finance*, February 2009). Suppose 30 employers are randomly selected.
a. What is the probability that exactly ten of the employers have eliminated or plan to eliminate perks?
b. What is the probability that at least ten employers, but no more than 20 employers, have eliminated or plan to eliminate perks?
c. What is the probability that at most eight employers have eliminated or plan to eliminate perks?

85. Studies have shown that bats can consume an average of ten mosquitoes per minute (berkshiremuseum.org).
a. Calculate the average number of mosquitoes that a bat consumes in a 30-second interval.
b. What is the probability that a bat consumes four mosquitoes in a 30-second interval?
c. What is the probability that a bat does not consume any mosquitoes in a 30-second interval?
d. What is the probability that a bat consumes at least one mosquito in a 30-second interval?

86. Despite the fact that home prices seem affordable and mortgage rates are at historic lows, real estate agents say they are showing more homes, but not selling more (*The Boston Globe*, March 7, 2009). A real estate company estimates that an average of five people show up at an open house to view a property. There is going to be an open house on Sunday.
a. What is the probability that at least five people will show up to view the property?
b. What is the probability that fewer than five people will show up to view the property?

87. The police have estimated that there are 12 major accidents per day on a particular 10-mile stretch of a national highway. Suppose the incidence of

accidents is evenly distributed on this 10-mile stretch of the highway.

a. Find the probability that there will be fewer than eight major accidents per day on this 10-mile stretch of the highway.
b. Find the probability that there will be more than two accidents per day on a 1-mile stretch of this highway.

88. Suppose you draw three cards, without replacement, from a deck of well-shuffled cards. Remember that each deck consists of 52 cards, with 13 each of spades, hearts, clubs, and diamonds.
    a. What is the probability that you draw all spades?
    b. What is the probability that you draw two or fewer spades?
    c. What is the probability that you draw all spades or hearts?
89. A professor has learned that three students in her class of 20 will cheat on the exam. She decides to focus her attention on four randomly chosen students during the exam.
    a. What is the probability that she finds at least one of the students cheating?
    b. What is the probability that she finds at least one of the students cheating if she focuses on six randomly chosen students?
90. Find the probability that an Internal Revenue Service (IRS) auditor will catch only 4 income tax returns with illegitimate deductions if he randomly selects 5 returns from among 20 returns, of which 10 contain illegitimate deductions.
91. A committee of 10 is to be chosen from 50 people, 25 of whom are Republicans and 25 Democrats. The committee is chosen at random.
    a. What is the probability that there will be five Republicans and five Democrats?
    b. What is the probability that a majority of the committee will be Republicans?
92. Many U.S. households still do not have Internet access. Suppose 20 out of 80 households in a small southern town do not have Internet access. A company that provides high-speed Internet has recently entered the market. As part of the marketing campaign, the company decides to randomly select ten households and offer them free laptops along with a brochure that describes their services. The aim is to build goodwill and, with a free laptop, tempt nonusers into getting Internet access.
    a. What is the probability that six laptop recipients do not have Internet access?
    b. What is the probability that at least five laptop recipients do not have Internet access?
    c. What is the probability that two or fewer laptop recipients do not have Internet access?
    d. What is the expected number of laptop recipients who do not have Internet access?

CASE STUDIES

CASE STUDY 5.1 An extended warranty is a prolonged warranty offered to consumers by the warranty administrator, the retailer, or the manufacturer. A report in *The New York Times* (November 23, 2009) suggests that 20.4% of laptops fail over three years. Roberto D'Angelo is interested in an extended warranty for his laptop. A good extended warranty is being offered at Compuvest.com for $74. It will cover any repair job that his laptop may need in the next three years. Based on his research, he determines that the likelihood of a repair job in the next three years is 13% for a minor repair, 8% for a major repair, and 3% for a catastrophic repair. The extended warranty will save him $80 for a minor repair, $320 for a major repair, and $500 for a catastrophic repair. These results are summarized in the following probability distribution.

Data for Case Study 5.1 Probability Distribution for Repair Cost

Type of Repair	Probability	Repair Cost (in $)
None	0.76	0
Minor	0.13	80
Major	0.08	320
Catastrophic	0.03	500

In a report, use the above information to

1. Calculate and interpret the expected value of the repair cost.
2. Analyze the expected gain or loss for a consumer who buys the extended warranty.
3. Determine what kind of a consumer (risk neutral, risk averse, or both) will buy this extended warranty.

CASE STUDY 5.2 According to figures released by the New York City government, smoking among New York City teenagers is on a decline, continuing a trend that began more than a decade ago (*The New York Times*, January 2, 2008). According to the New York City Youth Risk Behavior Survey, the teenage smoking rate dropped to 8.5% in 2007 from about 17.6% in 2001 and 23% in 1997. City officials attribute the lower smoking rate to factors including a cigarette tax increase, a ban on workplace smoking, and television and subway ads that graphically depict tobacco-related illnesses.

In a report, use the above information to

1. Calculate the probability that at least one in a group of 10 New York City teenagers smoked in 2007.
2. Calculate the probability that at least one in a group of 10 New York City teenagers smoked in 2001.
3. Calculate the probability that at least one in a group of 10 New York City teenagers smoked in 1997.
4. Comment on the smoking trend between 1997 and 2007.

CASE STUDY 5.3 Disturbing news regarding Scottish police concerns the number of crashes involving vehicles on operational duties (*BBC News*, March 10, 2008). Statistics showed that Scottish forces' vehicles had been involved in traffic accidents at the rate of 1,000 per year. The statistics included vehicles involved in 999 calls (the equivalent of 911 in the United States) and pursuits. Fire service and ambulance vehicles were not included in the figures.

In a report, use the above information to

1. Calculate and interpret the expected number of traffic accidents per day involving vehicles on operational duties.
2. Use this expected value to construct the probability distribution table that lists the probability of 0, 1, 2, . . ., 10 traffic accidents per day. Graph this distribution and summarize your findings.

APPENDIX 5.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, JMP, and R.

Minitab

The Binomial Distribution

A. (Replicating Example 5.7a) From the menu, choose **Calc > Probability Distributions >Binomial**.

B. Select **Probability** since we are finding $P(X = 70)$. (For cumulative probabilities, select **Cumulative probability**.) Enter 100 as the **Number of trials** and 0.68 as the **Event probability**. Select **Input constant** and enter the value 70.

The Poisson Distribution

A. (Replicating Example 5.9a) From the menu, choose **Calc > Probability Distributions > Poisson**.

B. Select **Cumulative probability** since we are finding $P(X \leq 10)$. (For calculating $P(X = x)$, select **Probability**.) Enter 10.5 for the **Mean**. Select **Input constant** and enter the value 10.

The Hypergeometric Distribution

A. (Replicating Example 5.11a) From the menu, choose **Calc > Probability Distributions > Hypergeometric**.

B. Select **Probability** since we are finding $P(X = 0)$. (For cumulative probabilities, select **Cumulative probability**.) Enter 25 for the **Population size (N)**, 15 for **Event count in population (M)**, and 4 for the **Sample size (n)**. Select **Input constant** and enter 0.

SPSS

Note: In order for the calculated probability to be seen on the spreadsheet, SPSS must first "view" data on the spreadsheet. For this purpose, enter a value of zero in the first cell of the first column.

The Binomial Distribution

A. (Replicating Example 5.7a) From the menu, choose **Transform > Compute Variable**.

B. Under **Target Variable**, type pdfbinomial. Under **Function group**, select **PDF & Noncentral PDF**, and under **Functions and Special Variables**, double-click on **Pdf.Binom**. (For cumulative probabilities, under **Function group** select **CDF & Noncentral CDF**, and under **Functions and Special Variables** double-click on **Cdf.Binom**.) In the **Numeric Expression** box, enter 70 for **quant**, 100 for **n**, and 0.68 for **prob**.

The Poisson Distribution

A. (Replicating Example 5.9a) From the menu, choose **Transform > Compute Variable**.

B. Under **Target Variable**, type cdfpoisson. Under **Function group**, select **CDF & Noncentral CDF**, and under **Functions and Special Variables**, double-click on **Pdf.Poisson**. (For calculating $P(X = x)$, under **Function group** select **PDF & Noncentral PDF**, and under **Functions and Special Variables**, double-click on **Pdf.Poisson**.) In the **Numeric Expression** box, enter 10 for **quant** and 10.5 for **Mean**.

The Hypergeometric Distribution

A. (Replicating Example 5.11a) From the menu, choose **Transform > Compute Variable**.

B. Under **Target Variable**, type pdfhyper. Under **Function group**, select **PDF & Noncentral PDF**, and under **Functions and Special Variables**, double-click on **Pdf.Hyper**. (For cumulative probabilities, under **Function group** select **CDF & Noncentral CDF**, and under **Functions and Special Variables** double-click on **Cdf.Hyper**.) In the **Numeric Expression** box, enter 0 for **quant**, 25 for **total**, 4 for **sample**, and 15 for **hits**.

JMP

Note: In order for the calculated probability to be seen on the spreadsheet, JMP must first "view" data on the spreadsheet. For this purpose, enter a value of zero in the first cell of the first column.

The Binomial Distribution

A. (Replicating Example 5.7a) Right-click at the top of the column in the spreadsheet view and select **Formula**. Under **Functions** (**grouped**), choose **Discrete Probability > Binomial Probability**. (For cumulative probabilities, select **Binomial Distribution**.)

B. Enter 0.68 for **p**, 100 for **n**, and 70 for **k**.

The Poisson Distribution

A. (Replicating Example 5.9a) Right-click at the top of the column in the spreadsheet view and select **Formula**. Under **Functions** (**grouped**), choose **Discrete Probability > Poisson Distribution**. (For calculating $P(X = x)$, select **Poisson Probability**.)

B. Enter 10.5 for **lambda** and 10 for **k**.

The Hypergeometric Distribution

A. (Replicating Example 5.11a) Right-click at the top of the column in the spreadsheet view and select **Formula**. Under **Functions** (**grouped**), choose **Discrete Probability > Hypergeometric Probability**. (For cumulative probabilities, select **Hypergeometric Distribution**.)

B. Enter 25 for **N**, 15 for **K**, 4 for **n**, and 0 for **x**.

R

The Binomial Distribution

(Replicating Example 5.7a) Use the **dbinom** and **pbinom** functions to calculate binomial probabilities. In order to find $P(X = x)$, enter "dbinom(x, n, p)", and to find $P(X \leq x)$, enter "pbinom(x, n, p)", where x is the number of successes, n is the number of trials, and p is the probability of success. Thus, to find $P(X = 70)$ with $n = 100$ and $p = 0.68$ enter:

```
> dbinom(70, 100, 0.68)
```

The Poisson Distribution

(Replicating Example 5.9a) Use the **dpois** and **ppois** functions to calculate Poisson probabilities. In order to find $P(X = x)$, enter "dpois(x, μ)", and to find $P(X \leq x)$, enter "ppois(x, μ)", where x is the number of successes over some interval and μ is the mean over this interval. Thus, to find $P(X \leq 10)$ with $\mu = 10.5$, enter:

```
> ppois(10, 10.5)
```

The Hypergeometric Distribution

(Replicating Example 5.11a) Use the **dhyper** and **phyper** functions to calculate hypergeometric probabilities. In order to find $P(X = x)$, enter "dhyper(x, S, $N - S$, n)", and to find $P(X \leq x)$, enter "phyper(x, S, $N - S$, n)", where x is the number of successes in the sample, S is the number of successes in the population, $N - S$ is the number of failures in the population, and n is the sample size. Thus, to find $P(X = 0)$ with $S = 15$, $N - S = 10$, and $n = 4$, enter:

```
> dhyper(0, 15, 10, 4)
```

6 Continuous Probability Distributions

Learning Objectives

After reading this chapter you should be able to:

LO **6.1** Describe a continuous random variable.

LO **6.2** Calculate and interpret probabilities for a random variable that follows the continuous uniform distribution.

LO **6.3** Explain the characteristics of the normal distribution.

LO **6.4** Calculate and interpret probabilities for a random variable that follows the normal distribution.

LO **6.5** Calculate and interpret probabilities for a random variable that follows the exponential distribution.

In Chapter 5, we classified a random variable as either discrete or continuous. A discrete random variable assumes a countable number of distinct values, such as the number of houses that a realtor sells in a month, the number of defective pieces in a sample of 20 machine parts, and the number of cars lined up at a toll booth. A continuous random variable, on the other hand, is characterized by uncountable values because it can take on any value within an interval. Examples of a continuous random variable include the investment return on a mutual fund, the waiting time at a toll booth, and the amount of soda in a cup. In all of these examples, it is impossible to list all possible values of the random variable. In this chapter, we focus on continuous random variables. Most of this chapter is devoted to the normal distribution, which is the most extensively used continuous probability distribution and is the cornerstone of statistical inference. Other important continuous distributions discussed are the continuous uniform and the exponential distributions.



Introductory Case

Demand for Salmon

Akiko Hamaguchi is the manager of a small sushi restaurant called Little Ginza in Phoenix, Arizona. As part of her job, Akiko has to purchase salmon every day for the restaurant. For the sake of freshness, it is important that she buys the right amount of salmon daily. Buying too much may result in wastage, and buying too little may disappoint some customers on high-demand days.

Akiko has estimated that the daily consumption of salmon is normally distributed with a mean of 12 pounds and a standard deviation of 3.2 pounds. She has always bought 20 pounds of salmon every day. Lately, she has been criticized by the owners because this amount of salmon was too often resulting in wastage. As part of cost cutting, Akiko is considering a new strategy. She will buy salmon that is sufficient to meet the daily demand of customers on 90% of the days.

Akiko wants to use the above information to

1. Calculate the probability that the demand for salmon at Little Ginza is above 20 pounds.
2. Calculate the probability that the demand for salmon at Little Ginza is below 15 pounds.
3. Determine the amount of salmon that should be bought daily so that the restaurant meets demand on 90% of the days.

A synopsis of this case is provided in Section 6.2.

LO **6.1**

Describe a continuous random variable.

6.1 CONTINUOUS RANDOM VARIABLES AND THE UNIFORM DISTRIBUTION

As discussed in Chapter 5, a discrete random variable X assumes a countable number of distinct values such as x_1, x_2, x_3, and so on. A continuous random variable, on the other hand, is characterized by uncountable values because it can take on any value within an interval. Unlike the case of a discrete random variable, we cannot describe the possible values of a continuous random variable X with a list x_1, x_2, . . . because the value $(x_1 + x_2)/2$, not in the list, might also be possible. Consider, for example, a continuous random variable defined by the amount of time a student takes to finish the exam. Here, it is impossible to put in a sequence all possible values of the random variable. Some students may think that time is countable in seconds; however, this may not be the case once we consider fractions of a second. Similarly, other continuous random variables, such as the investment return on a mutual fund and the amount of soda in a cup, are characterized by uncountable values.

For a discrete random variable, we can compute the probability that it assumes a particular value x, or written as a probability statement, $P(X = x)$. For instance, for a binomial random variable, we can calculate the probability of exactly one success in n trials; that is, $P(X = 1)$. We cannot make this calculation with a continuous random variable. The probability that a continuous random variable assumes a particular value x is zero; that is, $P(X = x) = 0$. This occurs because we cannot assign a nonzero probability to each of the uncountable values and still have the probabilities sum to one. Thus, for a continuous random variable, it is only meaningful to calculate the probability that the value of the random variable falls within some specified interval. Therefore, for a continuous random variable, $P(a \le X \le b) = P(a < X < b) = P(a \le X < b) = P(a < X \le b)$, since $P(X = a)$ and $P(X = b)$ are both zero.

In Chapter 5, we learned that a probability mass function for a discrete random variable X is a list of the values of X with the associated probabilities. For a continuous random variable, the counterpart to the probability mass function is called the probability density function, denoted by $f(x)$. As mentioned in Chapter 5, in this text we often use the term "probability distribution" to refer to both functions. The graph of $f(x)$ approximates the relative frequency polygon for the population. Unlike the probability mass function, $f(x)$ does not provide probabilities directly. The probability that the variable assumes a value within an interval, say $P(a \le X \le b)$, is defined as the area under $f(x)$ between points a and b. Moreover, the entire area under $f(x)$ over all values of x must equal one; this is equivalent to the fact that, for discrete random variables, the probabilities add up to one.

THE PROBABILITY DENSITY FUNCTION

The probability density function $f(x)$ for a continuous random variable X has the following properties:

- $f(x) \ge 0$ for all possible values x of X, and
- the area under $f(x)$ over all values x of X equals one.

As in the case for a discrete random variable, we can use the cumulative distribution function, denoted by $F(x)$, to compute probabilities for a continuous random variable. For a value x of the random variable X, $F(x) = P(X \le x)$ is simply the area under the probability density function up to the value x.

THE CUMULATIVE DISTRIBUTION FUNCTION

For any value x of the random variable X, the cumulative distribution function $F(x)$ is defined as

$$F(x) = P(X \leq x).$$

If you are familiar with calculus, then you will recognize that this cumulative probability is the integral of $f(u)$ for values less than or equal to x. Similarly, $P(a \leq X \leq b) = F(b) - F(a)$ is the integral of $f(u)$ between points a and b. Fortunately, we do not necessarily need the knowledge of integral calculus to compute probabilities for the continuous random variables discussed in this text.

The Continuous Uniform Distribution

LO 6.2

Calculate and interpret probabilities for a random variable that follows the continuous uniform distribution.

One of the simplest continuous probability distributions is called the **continuous uniform distribution**. This distribution is appropriate when the underlying random variable has an equally likely chance of assuming a value within a specified range. Examples of uniformly distributed random variables include the delivery time of an appliance, the scheduled flight time between cities, and the waiting time for a campus bus. Any specified range for each of the above random variables can be assumed to be equally probable.

Suppose you are informed that your new refrigerator will be delivered between 2:00 pm and 3:00 pm. Let the random variable X denote the delivery time of your refrigerator. This variable is bounded below by 2:00 pm and above by 3:00 pm for a total range of 60 minutes. It is reasonable to infer that the probability of delivery between 2:00 pm and 2:30 pm equals 0.50 (=30/60), as does the probability of delivery between 2:30 pm and 3:00 pm. Similarly, the probability of delivery in any 15-minute interval equals 0.25 (=15/60), and so on.

Figure 6.1 depicts the probability density function for a continuous uniform random variable. The values a and b on the horizontal axis represent its lower and upper limits, respectively. The continuous uniform distribution is symmetric around its mean μ, computed as $\frac{a+b}{2}$. In the refrigerator delivery example, the mean is $\mu = \frac{2+3}{2} = 2.5$, implying that you expect the delivery at 2:30 pm. The standard deviation σ of a continuous uniform variable equals $\sqrt{(b-a)^2/12}$.

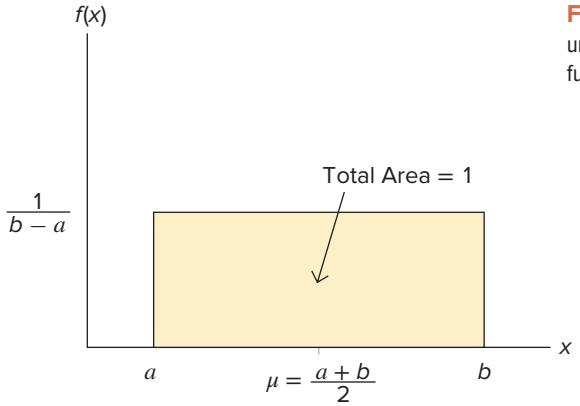


FIGURE 6.1 Continuous uniform probability density function

It is important to emphasize that the height of the probability density function does not directly represent a probability. As mentioned earlier, for all continuous random variables, it is the area under $f(x)$ that corresponds to probability. For the continuous uniform distribution, the probability is essentially the area of a rectangle, which is the base times

the height. Therefore, the probability is easily computed by multiplying the length of a specified interval (base) with $f(x) = \frac{1}{b-a}$ (height).

THE CONTINUOUS UNIFORM DISTRIBUTION

A random variable X follows the continuous uniform distribution if its probability density function is

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b, \text{ and} \\ 0 & \text{for } x < a \text{ or } x > b, \end{cases}$$

where a and b represent the lower and upper limits of values, respectively, that the random variable assumes.

The expected value and the standard deviation of X are computed as

$$E(X) = \mu = \frac{a+b}{2} \quad \text{and} \quad SD(X) = \sigma = \sqrt{(b-a)^2/12}.$$

EXAMPLE 6.1

A manager of a local drugstore is projecting next month's sales for a particular cosmetic line. She knows from historical data that sales follow a continuous uniform distribution with a lower limit of \$2,500 and an upper limit of \$5,000.

a. What are the mean and the standard deviation for this continuous uniform distribution?

b. What is the probability that sales exceed \$4,000?

c. What is the probability that sales are between \$3,200 and \$3,800?

SOLUTION:

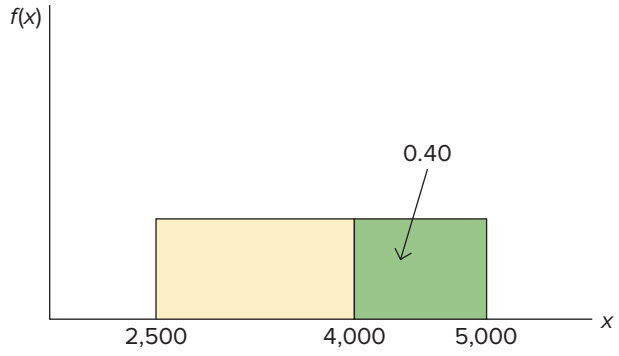
a. With a value for the lower limit of $a = 2{,}500$ and a value for the upper limit of $b = 5{,}000$, we calculate the mean and the standard deviation for this continuous uniform distribution as

$$\mu = \frac{a+b}{2} = \frac{2{,}500 + 5{,}000}{2} = 3{,}750, \text{ or } \$3{,}750, \text{ and}$$

$$\sigma = \sqrt{(b-a)^2/12} = \sqrt{(5{,}000 - 2{,}500)^2/12} = 721.69, \text{ or } \$721.69.$$

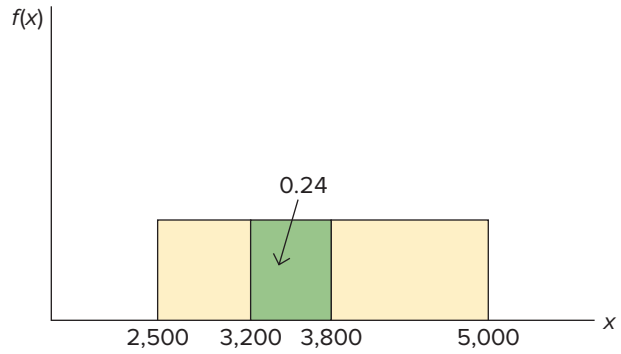
b. When solving for the probability that sales exceed \$4,000, we find $P(X > 4{,}000)$, which is the area between 4,000 and 5,000, as shown in Figure 6.2. The base of the rectangle equals $5{,}000 - 4{,}000 = 1{,}000$ and the height equals $\frac{1}{5{,}000 - 2{,}500} = 0.0004$. Thus, $P(X > 4{,}000) = 1{,}000 \times 0.0004 = 0.40$.

FIGURE 6.2 Area to the right of 4,000 (Example 6.1b)



c. When solving for the probability that sales are between \$3,200 and \$3,800, we find $P(3{,}200 \le X \le 3{,}800)$. Using the same methodology as in part b, we multiply the base times the height of the rectangle, as shown in Figure 6.3. Therefore, we obtain the probability as $(3{,}800 - 3{,}200) \times 0.0004 = 0.24$.

FIGURE 6.3 Area between 3,200 and 3,800 (Example 6.1c)



EXERCISES 6.1

Mechanics

1. The cumulative probabilities for a continuous random variable X are $P(X \le 10) = 0.42$ and $P(X \le 20) = 0.66$. Calculate the following probabilities.
   a. $P(X > 10)$
   b. $P(X > 20)$
   c. $P(10 < X < 20)$
2. For a continuous random variable X with an upper bound of 4, $P(0 \le X \le 2.5) = 0.54$ and $P(2.5 \le X \le 4) = 0.16$. Calculate the following probabilities.
   a. $P(X < 0)$
   b. $P(X > 2.5)$
   c. $P(0 \le X \le 4)$
3. For a continuous random variable X, $P(20 \le X \le 40) = 0.15$ and $P(X > 40) = 0.16$. Calculate the following probabilities.
   a. $P(X < 40)$
   b. $P(X < 20)$
   c. $P(X = 40)$
4. A random variable X follows the continuous uniform distribution with a lower bound of 5 and an upper bound of 35.
   a. What is the height of the density function $f(x)$?
   b. What are the mean and the standard deviation for the distribution?
   c. Calculate $P(X > 10)$.
5. A random variable X follows the continuous uniform distribution with a lower bound of -2 and an upper bound of 4.
   a. What is the height of the density function $f(x)$?
   b. What are the mean and the standard deviation for the distribution?
   c. Calculate $P(X \le -1)$.
6. A random variable X follows the continuous uniform distribution with a lower limit of 10 and an upper limit of 30.
   a. Calculate the mean and the standard deviation for the distribution.
   b. What is the probability that X is greater than 22?
   c. What is the probability that X is between 15 and 23?
7. A random variable X follows the continuous uniform distribution with a lower limit of 750 and an upper limit of 800.
   a. Calculate the mean and the standard deviation for the distribution.
   b. What is the probability that X is less than 770?

Applications

8. Suppose the average price of electricity for a New England customer follows the continuous uniform distribution with a lower bound of 12 cents per kilowatt-hour and an upper bound of 20 cents per kilowatt-hour.
   a. Calculate the average price of electricity for a New England customer.
   b. What is the probability that a New England customer pays less than 15.5 cents per kilowatt-hour?
   c. A local carnival is not able to operate its rides if the average price of electricity is more than 14 cents per kilowatt-hour. What is the probability that the carnival will need to close?
9. The arrival time of an elevator in a 12-story dormitory is equally likely at any time range during the next 4 minutes.
   a. Calculate the expected arrival time.
   b. What is the probability that an elevator arrives in less than 1½ minutes?
   c. What is the probability that the wait for an elevator is more than 1½ minutes?

10. The Netherlands is one of the world leaders in the production and sale of tulips. Suppose the heights of the tulips in the greenhouse of Rotterdam's Fantastic Flora follow a continuous uniform distribution with a lower bound of 7 inches and an upper bound of 16 inches. You have come to the greenhouse to select a bouquet of tulips, but only tulips with a height greater than 10 inches may be selected. What is the probability that a randomly selected tulip is tall enough to pick?
11. The scheduled arrival time for a daily flight from Boston to New York is 9:25 am. Historical data show that the arrival time follows the continuous uniform distribution with an early arrival time of 9:15 am and a late arrival time of 9:55 am.
    a. Calculate the mean and the standard deviation of the distribution.
    b. What is the probability that a flight arrives late (later than 9:25 am)?
12. You were informed at the nursery that your peach tree will definitely bloom sometime between March 18 and March 30. Assume that the bloom times follow a continuous uniform distribution between these specified dates.
    a. What is the probability that the tree does not bloom until March 25?
    b. What is the probability that the tree will bloom by March 20?
13. You have been informed that the assessor will visit your home sometime between 10:00 am and 12:00 pm. It is reasonable to assume that his visitation time is uniformly distributed over the specified two-hour interval. Suppose you have to run a quick errand at 10:00 am.
    a. If it takes 15 minutes to run the errand, what is the probability that you will be back before the assessor visits?
    b. If it takes 30 minutes to run the errand, what is the probability that you will be back before the assessor visits?

6.2 THE NORMAL DISTRIBUTION

The **normal probability distribution**, or simply the **normal distribution**, is the familiar **bell-shaped distribution**. It is also referred to as the Gaussian distribution.[1] The normal distribution is the most extensively used probability distribution in statistical work. One reason for this common use is that the normal distribution closely approximates the probability distribution for a wide range of random variables of interest. Examples of random variables that closely follow a normal distribution include

- Heights and weights of newborn babies.
- Scores on the SAT.
- Cumulative debt of college graduates.
- Advertising expenditure of firms.
- Rate of return on an investment.

Whenever possible, it is instructive to analyze the underlying data to determine if the normal distribution is appropriate for a given application. There are various ways to do this, including inspecting histograms (Chapter 2) and boxplots (Chapter 3) for symmetry and bell shape. In this chapter, we simply assume that the random variable in question is normally distributed and focus on finding probabilities associated with this type of random variable. The computation of these probabilities is easy and direct.

Another important function of the normal distribution is that it serves as the cornerstone of statistical inference. Recall from Chapter 1 that the study of statistics is divided into two branches: descriptive statistics and inferential statistics. Statistical inference is generally based on the assumption of the normal distribution and serves as the major topic in the remainder of this text.

[1]The discovery of the normal (Gaussian) distribution is often credited to Carl Friedrich Gauss (1777–1855), even though some attribute the credit to De Moivre (1667–1754), who had earlier discovered it in the context of simplifying the binomial distribution calculations.

Characteristics of the Normal Distribution

LO 6.3

Explain the characteristics of the normal distribution.

- The normal distribution is **bell-shaped** and **symmetric** around its mean; that is, one side of the mean is just the mirror image of the other side. The mean, the median, and the mode are all equal for a normally distributed random variable.
- The normal distribution is **completely described by two parameters**—the population mean μ and the population variance σ^2. The population mean describes the central location and the population variance describes the dispersion of the distribution.
- The normal distribution is **asymptotic** in the sense that the tails get closer and closer to the horizontal axis but never touch it. Thus, theoretically, a normal random variable can assume any value between minus infinity and plus infinity.

The probability density function for the normal distribution is defined as follows.

THE NORMAL DISTRIBUTION

A random variable X with mean μ and variance σ^2 follows the normal distribution if its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

where π equals approximately 3.14159 and $\exp(w) = e^w$ is the exponential function, where $e \approx 2.718$ is the base of the natural logarithm.

A graph depicting the normal probability density function is often referred to as the normal curve or the bell curve. The following example relates the normal curve to the location and the dispersion of the normally distributed random variable.

EXAMPLE 6.2

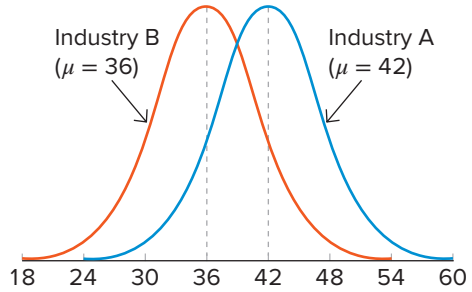
Suppose we know that the ages of employees in Industries A, B, and C are normally distributed. We are given the following information on the relevant parameters:

Industry A	Industry B	Industry C
$\mu = 42$ years	$\mu = 36$ years	$\mu = 42$ years
$\sigma = 5$ years	$\sigma = 5$ years	$\sigma = 8$ years

Graphically compare the ages of employees in Industry A with Industry B. Repeat the comparison for Industry A with Industry C.

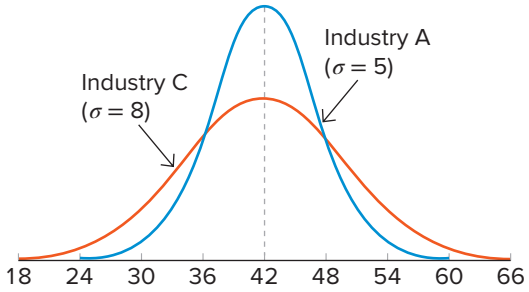
SOLUTION: Since the mean age of employees in Industry A is greater than that in Industry B, the normal curve for Industry A is located to the right of Industry B as shown in Figure 6.4. Both curves show equal dispersion from the mean, given that the standard deviations are the same.

FIGURE 6.4 Normal probability density function for two values of μ along with $\sigma = 5$



Since the mean age of employees in Industry A and Industry C is the same, the normal curves for each industry have the same center as shown in Figure 6.5. However, since the standard deviation for Industry A is less than that of Industry C, the normal curve for Industry A is less dispersed. Its peak is higher than that of Industry C, reflecting the fact that an employee's age is likelier to be closer to the mean age in Industry A. Figures 6.4 and 6.5 show that we can capture the entire distribution of any normally distributed random variable based on its mean and variance (or standard deviation).

FIGURE 6.5 Normal probability density function for two values of σ along with $\mu = 42$



We generally use the cumulative distribution function $F(x)$ to compute probabilities for a normally distributed random variable, where $F(x) = P(X \leq x)$ is simply the area under $f(x)$ up to the value x. As mentioned earlier, we do not necessarily need the knowledge of integral calculus to compute probabilities for the normal distribution. Instead, we rely on a table to find probabilities. We can also compute probabilities with certain calculators, Excel, and other statistical packages. The specifics of how to use the table are delineated next.

The Standard Normal Distribution

The **standard normal distribution** is a special case of the normal distribution with a mean equal to zero and a standard deviation (or variance) equal to one. Using the letter Z to denote a random variable with the standard normal distribution, we have $\mu = E(Z) = 0$ and $\sigma = SD(Z) = 1$. As usual, we use the lowercase letter z to denote the value that the standard normal variable Z may assume.

The value z is actually the z-score that we discussed in Chapter 3. It measures the number of standard deviations a given value is away from the mean. For example, a z-score of 2 implies that the given value is 2 standard deviations above the mean. Similarly, a z-score of -1.5 implies that the given value is 1.5 standard deviations below the mean. As mentioned in Chapter 3, converting values into z-scores is called standardizing the data.

We will first show how to compute probabilities related to the standard normal distribution. Later, we will show that any normal distribution is equivalent to the standard normal distribution when the unit of measurement is changed to measure standard deviations from the mean. Therefore, while most real-world normally distributed variables are not standard normal, we can always transform (standardize) them into standard normal to compute the relevant probabilities.

Virtually all introductory statistics texts include a **standard normal table**, also referred to as the ***z* table**, that provides areas (probabilities) under the z curve. However, the format of these tables is sometimes different. In this text, the z table provides cumulative probabilities $P(Z \leq z)$; this table appears on two pages in Appendix A and is labeled Table 1. The left-hand page provides cumulative probabilities for z values less than or equal to zero. The right-hand page shows cumulative probabilities for z values greater than or equal to zero. Given the symmetry of the normal distribution and the fact that the area under the entire curve is one, other probabilities can be easily computed.

THE STANDARD NORMAL DISTRIBUTION

The standard normal random variable Z is a normal random variable with $E(Z) = 0$ and $SD(Z) = 1$. The z table provides cumulative probabilities $P(Z \leq z)$ for positive and negative z values.

Figure 6.6 shows the standard normal probability density function (z distribution). Since the random variable Z is symmetric around its mean of zero, $P(Z < 0) = P(Z > 0) = 0.5$. As is the case with all continuous random variables, we can also write the probabilities as $P(Z \leq 0) = P(Z \geq 0) = 0.5$.

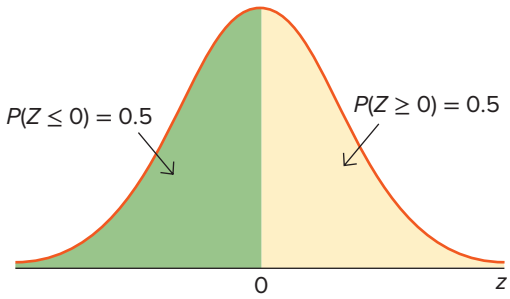


FIGURE 6.6
Standard normal probability density function

Finding a Probability for a Given *z* Value

As mentioned earlier, the z table provides cumulative probabilities $P(Z \leq z)$ for a given z. Consider, for example, a cumulative probability $P(Z \leq 1.52)$. Since $z = 1.52$ is positive, we can look up this probability from the right-hand page of the z table in Appendix A; Table 6.1 shows a portion of the table.

TABLE 6.1 Portion of the Right-Hand Page of the *z* Table

z	0.00	0.01	0.02
0.0	0.5000	0.5040	↓
0.1	0.5398	0.5438	↓
⋮	⋮	⋮	⋮
1.5	→	→	0.9357

The first column of the table, denoted as the z column, shows values of z up to the tenth decimal point, while the first row of the table, denoted as the z row, shows hundredths values. Thus, for $z = 1.52$, we match 1.5 on the z column with 0.02 on the z row to find a corresponding probability of 0.9357. The arrows in Table 6.1 indicate that $P(Z \leq 1.52) = 0.9357$.

In Figure 6.7, the cumulative probability corresponding to $z = 1.52$ is highlighted. Note that $P(Z \leq 1.52) = 0.9357$ represents the area under the z curve to the left of 1.52. Therefore, the area to the right of 1.52 can be computed as $P(Z > 1.52) = 1 - P(Z \leq 1.52) = 1 - 0.9357 = 0.0643$.

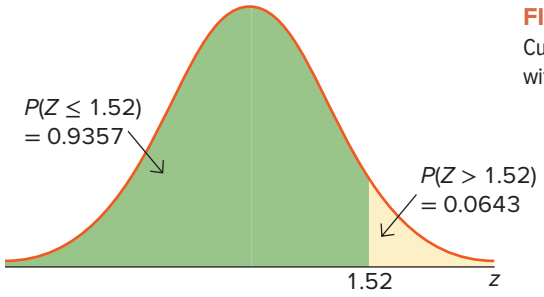


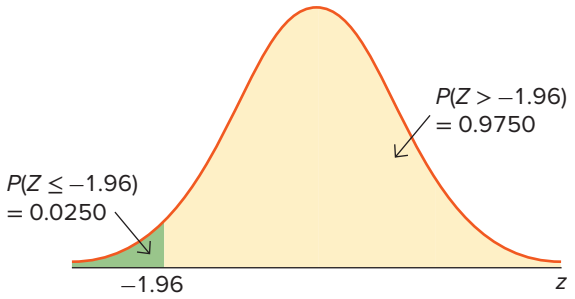
FIGURE 6.7
Cumulative probability with respect to $z = 1.52$

Suppose we want to find $P(Z \leq -1.96)$. Since z is a negative value, we can look up this probability from the left-hand page of the z table; Table 6.2 shows a portion of the table with arrows indicating that $P(Z \leq -1.96) = 0.0250$. Figure 6.8 highlights the corresponding probability. As before, the area to the right of -1.96 can be computed as $P(Z > -1.96) = 1 - P(Z \leq -1.96) = 1 - 0.0250 = 0.9750$.

TABLE 6.2 Portion of the Left-Hand Page of the z Table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
−3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	↓
−3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	↓
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−1.9	→	→	→	→	→	→	0.0250

FIGURE 6.8 Cumulative probability with respect to $z = -1.96$



EXAMPLE 6.3

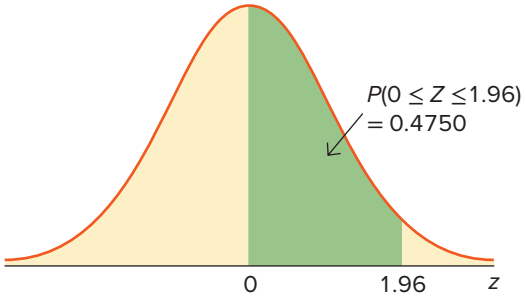
Find the following probabilities for the standard normal random variable Z.

a. $P(0 \leq Z \leq 1.96)$
b. $P(1.52 \leq Z \leq 1.96)$
c. $P(-1.52 \leq Z \leq 1.96)$
d. $P(Z > 4)$

SOLUTION: It always helps to start by highlighting the relevant probability in the z graph.

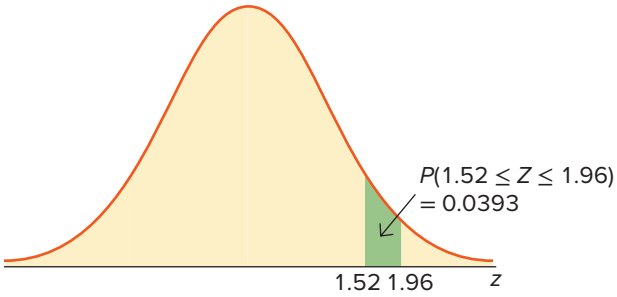
a. As shown in Figure 6.9, the area between 0 and 1.96 is equivalent to the area to the left of 1.96 minus the area to the left of 0. Therefore, $P(0 \leq Z \leq 1.96) = P(Z \leq 1.96) - P(Z < 0) = 0.9750 - 0.50 = 0.4750$.

FIGURE 6.9 Finding the probability between 0 and 1.96



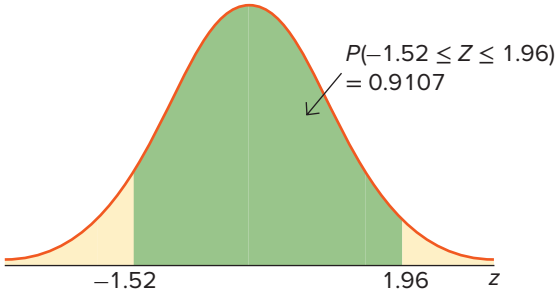
b. As in part a and shown in Figure 6.10, $P(1.52 \le Z \le 1.96) = P(Z \le 1.96) - P(Z < 1.52) = 0.9750 - 0.9357 = 0.0393$.

FIGURE 6.10 Finding the probability between 1.52 and 1.96



c. From Figure 6.11, $P(-1.52 \le Z \le 1.96) = P(Z \le 1.96) - P(Z < -1.52) = 0.9750 - 0.0643 = 0.9107$.

FIGURE 6.11 Finding the probability between −1.52 and 1.96



d. $P(Z > 4) = 1 - P(Z \le 4)$. However, the z table only goes up to 3.99 with $P(Z \le 3.99) = 1.0$ (approximately). In fact, for any z value greater than 3.99, it is acceptable to treat $P(Z \le z) = 1.0$. Therefore, $P(Z > 4) = 1 - P(Z \le 4) = 1 - 1 = 0$.

Finding a *z* Value for a Given Probability

So far we have computed probabilities for given z values. Now we will evaluate z values for given probabilities.

EXAMPLE 6.4

For the standard normal variable Z, find the z values that satisfy the following probability statements.

a. $P(Z \le z) = 0.6808$

b. $P(Z \le z) = 0.90$

c. $P(Z \le z) = 0.0643$

d. $P(Z > z) = 0.0212$

e. $P(-z \le Z \le z) = 0.95$

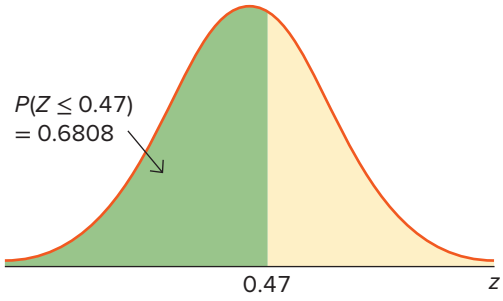
SOLUTION: As mentioned earlier, it helps to first highlight the relevant probability in the z graph. Recall, too, that the z table lists z values along with the corresponding cumulative probabilities. Noncumulative probabilities can be evaluated using symmetry.

a. Since the probability is already in a cumulative format—that is, $P(Z \leq z) = 0.6808$—we simply look up 0.6808 from the body of the table (right-hand side) to find the corresponding z value from the row/column of z. Table 6.3 shows the relevant portion of the z table, and Figure 6.12 depicts the corresponding area. Therefore, $z = 0.47$.

TABLE 6.3 Portion of the z Table for Example 6.4a

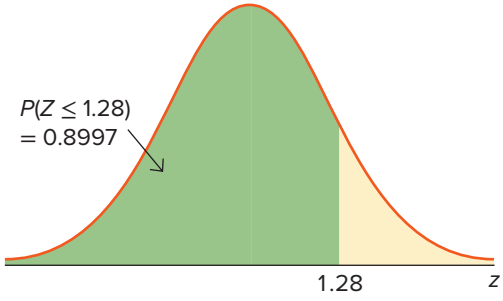
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	↑
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	↑
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.4	←	←	←	←	←	←	←	0.6808

FIGURE 6.12 Finding z given $P(Z \leq z) = 0.6808$



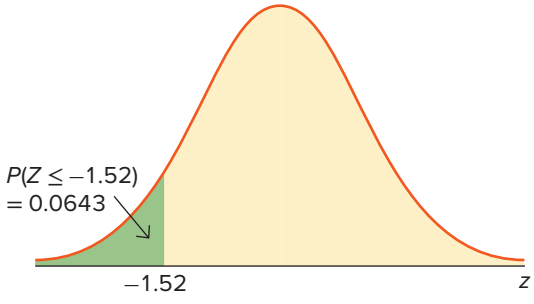
b. When deriving z for $P(Z \leq z) = 0.90$, we find that the z table (right-hand side) does not contain the cumulative probability 0.90. In such cases, we use the closest cumulative probability to solve the problem. Therefore, z is approximately equal to 1.28, which corresponds to a cumulative probability of 0.8997. Figure 6.13 shows this result graphically.

FIGURE 6.13 Finding z given $P(Z \leq z) = 0.90$



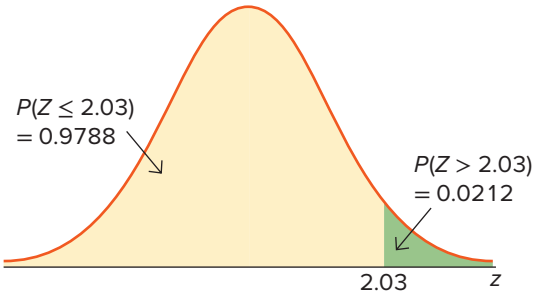
c. As shown in Figure 6.14, the z value that solves $P(Z \leq z) = 0.0643$ must be negative because the probability to its left is less than 0.50. We look up the cumulative probability 0.0643 in the table (left-hand side) to get $z = -1.52$.

FIGURE 6.14 Finding z given $P(Z \leq z) = 0.0643$



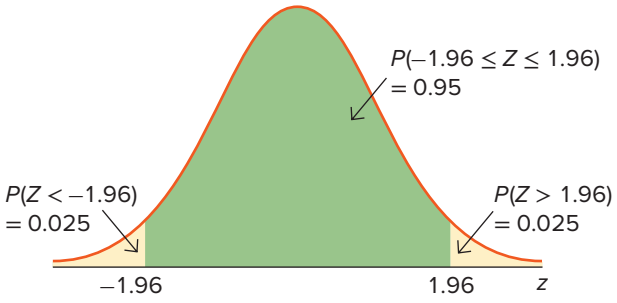
d. When deriving z for $P(Z > z) = 0.0212$, we have to find a z value such that the probability to the right of this value is 0.0212. Since the table states cumulative probabilities, we look up $P(Z \le z) = 1 - 0.0212 = 0.9788$ in the table (right-hand side) to get $z = 2.03$. Figure 6.15 shows the results.

FIGURE 6.15 Finding z given $P(Z > z) = 0.0212$



e. Since we know that the total area under the curve equals one, and we want to find $-z$ and z such that the area between the two values equals 0.95, we can conclude that the area in either tail is 0.025; that is, $P(Z < -z) = 0.025$ and $P(Z > z) = 0.025$. Figure 6.16 shows these results. We then use the cumulative probability, $P(Z \le z) = 0.95 + 0.025 = 0.975$, to find $z = 1.96$.

FIGURE 6.16 Finding z given $P(-z \le Z \le z) = 0.95$



The Transformation of Normal Random Variables

LO **6.4**

Calculate and interpret probabilities for a random variable that follows the normal distribution.

The importance of the standard normal distribution arises from the fact that any normal random variable can be transformed into the standard normal random variable to derive the relevant probabilities. In other words, any normally distributed random variable X with mean μ and standard deviation σ can be transformed (standardized) into the standard normal variable Z with mean zero and standard deviation one. We transform X into Z by subtracting from X its mean and dividing by its standard deviation; this is referred to as the **standard transformation**.

THE STANDARD TRANSFORMATION: CONVERTING X INTO Z

Any normally distributed random variable X with mean μ and standard deviation σ can be transformed into the standard normal random variable Z as

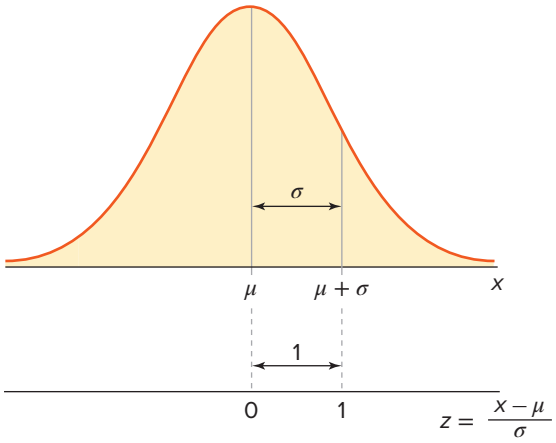
$$Z = \frac{X - \mu}{\sigma}.$$

Therefore, any value x has a corresponding value z given by

$$z = \frac{x - \mu}{\sigma}.$$

As illustrated in Figure 6.17, if the x value is at the mean—that is, $x = \mu$—then the corresponding z value is $z = \frac{\mu - \mu}{\sigma} = 0$. Similarly, if the x value is at one standard deviation above the mean—that is, $x = \mu + \sigma$—then the corresponding z value is $z = \frac{\mu + \sigma - \mu}{\sigma} = 1$. Therefore, by construction, $E(Z) = 0$ and $SD(Z) = 1$.

FIGURE 6.17
Transforming X with mean μ and standard deviation σ to Z



We are now in a position to solve any normal distribution problem by first transforming it to the z distribution.

EXAMPLE 6.5

Scores on a management aptitude exam are normally distributed with a mean of 72 and a standard deviation of 8.

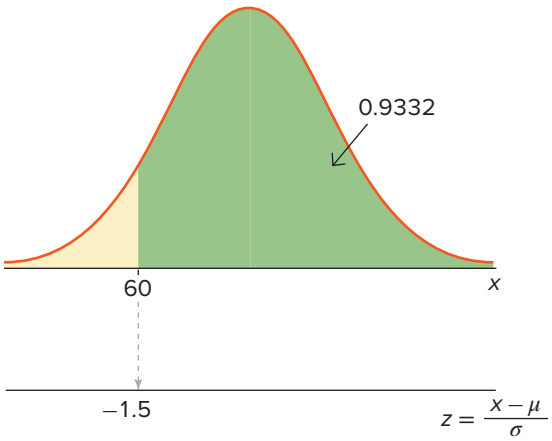
a. What is the probability that a randomly selected manager will score above 60?

b. What is the probability that a randomly selected manager will score between 68 and 84?

SOLUTION: Let X represent scores with $\mu = 72$ and $\sigma = 8$. We will use the standard transformation $z = \frac{x - \mu}{\sigma}$ to solve these problems.

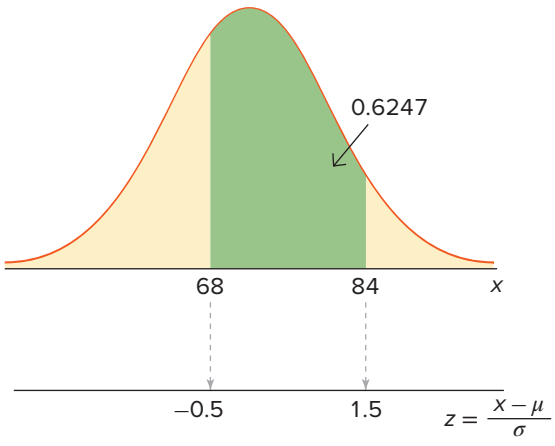
a. The probability that a manager scores above 60 is $P(X > 60)$. Figure 6.18 shows the probability as the shaded area to the right of 60. We derive $P(X > 60) = P\left(Z > \frac{60 - 72}{8}\right) = P(Z > -1.5)$. Since $P(Z > -1.5) = 1 - P(Z \leq -1.5)$, we look up -1.50 in the z table (left-hand side) to get this probability as $1 - 0.0668 = 0.9332$.

FIGURE 6.18 Finding $P(X > 60)$



b. When solving for the probability that a manager scores between 68 and 84, we find $P(68 \le X \le 84)$. The shaded area in Figure 6.19 shows this probability. We derive $P(68 \le X \le 84) = P\left(\frac{68-72}{8} \le Z \le \frac{84-72}{8}\right) = P(-0.5 \le Z \le 1.5)$. We compute this probability using the z table as $P(Z \le 1.5) - P(Z < -0.5) = 0.9332 - 0.3085 = 0.6247$.

FIGURE 6.19 Finding $P(68 \le X \le 84)$



So far we have used the standard transformation to compute probabilities for given x values. We can use the **inverse transformation**, $x = \mu + z\sigma$, to compute x values for given probabilities.

THE INVERSE TRANSFORMATION: CONVERTING Z INTO X

The standard normal variable Z can be transformed to the normally distributed random variable X with mean μ and standard deviation σ as $X = \mu + Z\sigma$.

Therefore, any value z has a corresponding value x given by $x = \mu + z\sigma$.

EXAMPLE 6.6

Scores on a management aptitude examination are normally distributed with a mean of 72 and a standard deviation of 8.

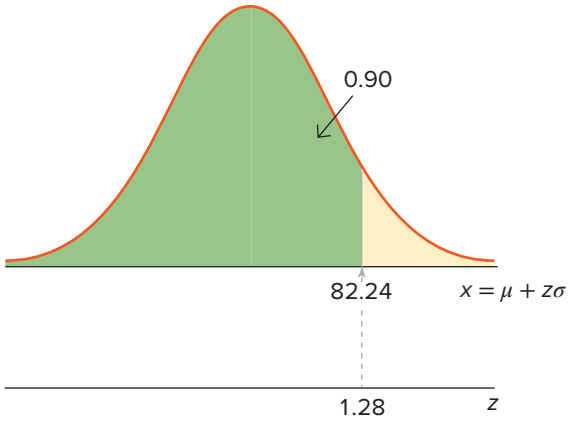
a. What is the lowest score that will place a manager in the top 10% (90th percentile) of the distribution?

b. What is the highest score that will place a manager in the bottom 25% (25th percentile) of the distribution?

SOLUTION: Let X represent scores on a management aptitude examination with $\mu = 72$ and $\sigma = 8$. We will use the inverse transformation $x = \mu + z\sigma$ to solve these problems.

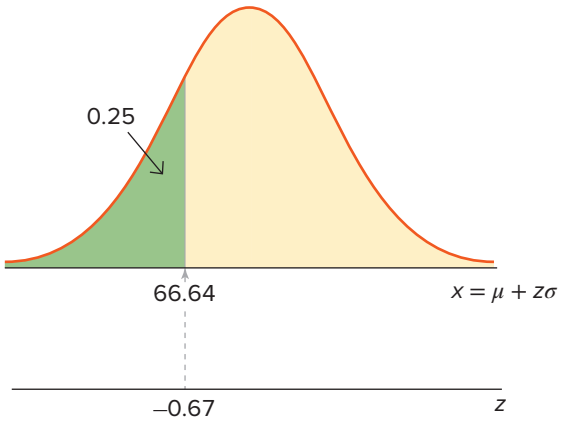
a. The 90th percentile is a numerical value x such that $P(X < x) = 0.90$. We look up 0.90 (or the closest value to 0.90) in the z table (right-hand side) to get $z = 1.28$ and use the inverse transformation to find $x = 72 + 1.28(8) = 82.24$. Therefore, a score of 82.24 or higher will place a manager in the top 10% of the distribution (see Figure 6.20).

FIGURE 6.20 Finding x given $P(X < x) = 0.90$



b. The 25th percentile is a numerical value x such that $P(X < x) = 0.25$. Using the z table (left-hand side), we find the corresponding z value that satisfies $P(Z < z) = 0.25$ as -0.67. We then solve $x = 72 - 0.67(8) = 66.64$. Therefore, a score of 66.64 or lower will place a manager in the bottom 25% of the distribution (see Figure 6.21).

FIGURE 6.21 Finding x given $P(X < x) = 0.25$



EXAMPLE 6.7

We can now answer the questions first posed by Akiko Hamaguchi in the introductory case of this chapter. Recall that Akiko would like to buy the right amount of salmon for daily consumption at Little Ginza. Akiko has estimated that the daily consumption of salmon is normally distributed with a mean of 12 pounds and a standard deviation of 3.2 pounds. She wants to answer the following questions:

a. What is the probability that the demand for salmon at Little Ginza is above 20 pounds?

b. What is the probability that the demand for salmon at Little Ginza is below 15 pounds?

c. How much salmon should be bought so that it meets customer demand on 90% of the days?

SOLUTION: Let X denote customer demand for salmon at the restaurant. We know that X is normally distributed with $\mu = 12$ and $\sigma = 3.2$.

a. When solving for the probability that the demand for salmon is more than 20 pounds, we find $P(X > 20) = P\left(Z > \frac{20 - 12}{3.2}\right) = P(Z > 2.50) = 1 - 0.9938 = 0.0062$.

b. When solving for the probability that the demand for salmon is less than 15 pounds, we find $P(X < 15) = P\left(Z < \frac{15 - 12}{3.2}\right) = P(Z < 0.94) = 0.8264$.

c. In order to compute the required amount of salmon that should be purchased to meet demand on 90% of the days, we solve for x in $P(X \leq x) = 0.90$. Since $P(X \leq x) = 0.90$ is equivalent to $P(Z \leq z) = 0.90$, we first derive $z = 1.28$. Given $x = \mu + z\sigma$, we find $x = 12 + 1.28(3.2) = 16.10$. Therefore, Akiko should buy 16.10 pounds of salmon daily to ensure that customer demand is met on 90% of the days.

SYNOPSIS OF INTRODUCTORY CASE

Akiko Hamaguchi is a manager at a small sushi restaurant called Little Ginza in Phoenix, Arizona. She is aware of the importance of purchasing the right amount of salmon daily. While purchasing too much salmon results in wastage, purchasing too little can disappoint customers who may choose not to frequent the restaurant in the future. In the past, she has always bought 20 pounds of salmon daily. A careful analysis of her purchasing habits and customer demand reveals that Akiko is buying too much salmon. The probability that the demand for salmon would exceed 20 pounds is very small at 0.0062. Even a purchase of 15 pounds satisfies customer demand on 82.64% of the days. In order to execute her new strategy of meeting daily demand of customers on 90% of the days, Akiko should purchase approximately 16 pounds of salmon daily.



©gkrphoto/Getty Images

A Note on the Normal Approximation of the Binomial Distribution

Recall from Chapter 5 that it is tedious to compute binomial probabilities with the formula when we encounter large values for n. As it turns out, with large values for n, the binomial distribution can be approximated by the normal distribution. Based on this normal distribution approximation, with mean $\mu = np$ and standard deviation $\sigma = \sqrt{npq}$, we can use the z table to compute relevant binomial probabilities. Some researchers believe that the discovery of the normal distribution in the 18th century was due to the need to simplify the binomial probability calculations. The popularity of this method, however, has been greatly reduced by the advent of computers. As we learned in Chapter 5, it is easy to compute exact binomial probabilities with Excel; thus, there is no reason to approximate. The normal distribution approximation, however, is extremely important when making an inference for the population proportion p, which is a key parameter of the binomial distribution. In later chapters, we will study the details of this approximation and how it is used for making inferences.

Using Excel for the Normal Distribution

Table 6.4 shows Excel functions that we can use to solve problems associated with continuous probability distributions. Example 6.8 illustrates the use of these functions with respect to the normal distribution. We will refer back to Table 6.4 with respect to the exponential distribution discussed in the next section.

TABLE 6.4 Continuous Probability Distributions and Function Names in Excel

Distribution	Excel Function
Standard Normal*	
$P(Z \leq z)$:	=NORM.S.DIST(z, 1)
Finding z:	=NORM.S.INV(*cumulprob*)
Normal	
$P(X \leq x)$:	=NORM.DIST(x, μ, σ, 1)
Finding x:	=NORM.INV(*cumulprob*, μ, σ)
Exponential	
$P(X \leq x)$:	=EXPON.DIST(x, λ, 1)
Finding x:	NA**

*Standard Normal functions are identical to Normal functions with $\mu = 0$ and $\sigma = 1$.

**NA denotes that this function is not readily available in Excel.

EXAMPLE 6.8

The Vanguard Balanced Index Fund seeks to maintain an allocation of 60% to stocks and 40% to bonds. With low fees and a consistent investment approach, this fund ranks fourth out of 792 funds that allocate 50% to 70% to stocks (*US News*, March 2017). Based on historical data, the expected return and standard deviation of this fund is estimated as 7.49% and 6.41%, respectively. Assume that the fund returns are stable and are normally distributed.

a. What is the probability that the fund will generate a return between 5% and 10%?

b. What is the lowest return of the fund that will place it in the top 10% (90th percentile) of the distribution?

SOLUTION: We let X denote the return on the Vanguard Balanced fund. We know that X is normally distributed with $\mu = 7.49$ and $\sigma = 6.41$.

We use Excel's **NORM.DIST** and **NORM.INV** functions to solve problems pertaining to the normal distribution. In order to find $P(X \leq x)$, we enter "=NORM.DIST(x, μ, σ, 1)" where x is the value for which we want to evaluate the cumulative probability, μ is the mean of the distribution, and σ is the standard deviation of the distribution. (If we enter "0" for the last argument in the function, then Excel returns the height of the normal distribution at the point x. This feature is useful if we want to plot the normal distribution.) If we want to find a particular x value for a given cumulative probability (*cumulprob*), then we enter "=NORM.INV(*cumulprob*, μ, σ)".

a. In order to find the probability of a return between 5% and 10%, $P(5 \leq X \leq 10)$, we enter "=NORM.DIST(10, 7.49, 6.41, 1) − NORM.DIST (5, 7.49, 6.41, 1)". Excel returns 0.3035.

b. In order to find the lowest return that will place it in the top 10% (90th percentile) of the distribution, $P(X > x) = 0.10$, we enter "=NORM.INV (0.90, 7.49, 6.41)". Excel returns 15.70.

EXERCISES 6.2

Mechanics

14. Find the following probabilities based on the standard normal variable Z.
    a. $P(Z > 1.32)$
    b. $P(Z \leq -1.32)$
    c. $P(1.32 \leq Z \leq 2.37)$
    d. $P(-1.32 \leq Z \leq 2.37)$
15. Find the following probabilities based on the standard normal variable Z.
    a. $P(Z > 0.74)$
    b. $P(Z \leq -1.92)$
    c. $P(0 \leq Z \leq 1.62)$
    d. $P(-0.90 \leq Z \leq 2.94)$
16. Find the following probabilities based on the standard normal variable Z.
    a. $P(-0.67 \leq Z \leq -0.23)$
    b. $P(0 \leq Z \leq 1.96)$
    c. $P(-1.28 \leq Z \leq 0)$
    d. $P(Z > 4.2)$
17. Find the following z values for the standard normal variable Z.
    a. $P(Z \leq z) = 0.9744$
    b. $P(Z > z) = 0.8389$
    c. $P(-z \leq Z \leq z) = 0.95$
    d. $P(0 \leq Z \leq z) = 0.3315$
18. Use Excel's function options to find the following z values for the standard normal variable Z.
    a. $P(Z \leq z) = 0.1020$
    b. $P(z \leq Z \leq 0) = 0.1772$
    c. $P(Z > z) = 0.9929$
    d. $P(0.40 \leq Z \leq z) = 0.3368$
19. Let X be normally distributed with mean $\mu = 10$ and standard deviation $\sigma = 6$.
    a. Find $P(X \leq 0)$.
    b. Find $P(X > 2)$.
    c. Find $P(4 \leq X \leq 10)$.
    d. Find $P(6 \leq X \leq 14)$.
20. Let X be normally distributed with mean $\mu = 10$ and standard deviation $\sigma = 4$.
    a. Find $P(X \leq 0)$.
    b. Find $P(X > 2)$.
    c. Find $P(4 \leq X \leq 10)$.
    d. Find $P(6 \leq X \leq 14)$.
21. Let X be normally distributed with mean $\mu = 120$ and standard deviation $\sigma = 20$.
    a. Find $P(X \leq 86)$.
    b. Find $P(80 \leq X \leq 100)$.
    c. Find x such that $P(X \leq x) = 0.40$.
    d. Find x such that $P(X > x) = 0.90$.
22. Let X be normally distributed with mean $\mu = 2.5$ and standard deviation $\sigma = 2$.
    a. Find $P(X > 7.6)$.
    b. Find $P(7.4 \leq X \leq 10.6)$.
    c. Find x such that $P(X > x) = 0.025$.
    d. Find x such that $P(x \leq X \leq 2.5) = 0.4943$.
23. Let X be normally distributed with mean $\mu = 2{,}500$ and standard deviation $\sigma = 800$.
    a. Find x such that $P(X \leq x) = 0.9382$.
    b. Find x such that $P(X > x) = 0.025$.
    c. Find x such that $P(2500 \leq X \leq x) = 0.1217$.
    d. Find x such that $P(X \leq x) = 0.4840$.
24. The random variable X is normally distributed. Also, it is known that $P(X > 150) = 0.10$.
    a. Find the population mean μ if the population standard deviation $\sigma = 15$.
    b. Find the population mean μ if the population standard deviation $\sigma = 25$.
    c. Find the population standard deviation σ if the population mean $\mu = 136$.
    d. Find the population standard deviation σ if the population mean $\mu = 128$.
25. Let X be normally distributed with $\mu = 254$ and $\sigma = 11$.
    a. Find $P(X \leq 266)$.
    b. Find $P(250 < X < 270)$.
    c. Find x such that $P(X \leq x) = 0.33$.
    d. Find x such that $P(X > x) = 0.33$.
26. Let X be normally distributed with $\mu = -15$ and $\sigma = 9$. Use Excel's function options for the following.
    a. Find $P(X > -12)$.
    b. Find $P(0 \leq X \leq 5)$.
    c. Find x such that $P(X \leq x) = 0.25$.
    d. Find x such that $P(X > x) = 0.25$.

Applications

27. The historical returns on a balanced portfolio have had an average return of 8% and a standard deviation of 12%. Assume that returns on this portfolio follow a normal distribution.
    a. What percentage of returns were greater than 20%?
    b. What percentage of returns were below −16%?
28. Assume that IQ scores follow a normal distribution with a mean of 100 and a standard deviation of 16.
    a. What is the probability that an individual scores between 84 and 116?

b. What is the probability that an individual scores less than 68?
c. What is the lowest score that will place an individual in the top 1% of IQ scores?

29. The average rent in a city is $1,500 per month with a standard deviation of $250. Assume rent follows the normal distribution.
    a. What percentage of rents are between $1,250 and $1,750?
    b. What percentage of rents are less than $1,250?
    c. What percentage of rents are greater than $2,000?
30. A professional basketball team averages 80 points per game with a standard deviation of 10 points. Assume points per game follow the normal distribution.
    a. What is the probability that a game's score is between 60 and 100 points?
    b. What is the probability that a game's score is more than 100 points? If there are 82 games in a regular season, in how many games will the team score more than 100 points?
31. The average high school teacher annual salary is $43,000 (Payscale.com, August 20, 2010). Let teacher salary be normally distributed with a standard deviation of $18,000.
    a. What percentage of high school teachers make between $40,000 and $50,000?
    b. What percentage of high school teachers make more than $80,000?
32. Americans are increasingly skimping on their sleep (*National Geographic News*, February 24, 2005). A health expert believes that American adults sleep an average of 6.2 hours on weekdays, with a standard deviation of 1.2 hours. Assume that sleep time on weekdays is normally distributed.
    a. What percentage of American adults sleep more than 8 hours on weekdays?
    b. What percentage of American adults sleep less than 6 hours on weekdays?
    c. What percentage of American adults sleep between 6 and 8 hours on weekdays?
33. The weight of turkeys is normally distributed with a mean of 22 pounds and a standard deviation of 5 pounds.
    a. Find the probability that a randomly selected turkey weighs between 20 and 26 pounds.
    b. Find the probability that a randomly selected turkey weighs less than 12 pounds.
34. Suppose that the miles-per-gallon (mpg) rating of passenger cars is a normally distributed random variable with a mean and a standard deviation of 33.8 mpg and 3.5 mpg, respectively.
    a. What is the probability that a randomly selected passenger car gets at least 40 mpg?
    b. What is the probability that a randomly selected passenger car gets between 30 and 35 mpg?
    c. An automobile manufacturer wants to build a new passenger car with an mpg rating that improves upon 99% of existing cars. What is the minimum mpg that would achieve this goal?
35. According to a company's website, the top 25% of the candidates who take the entrance test will be called for an interview. You have just been called for an interview. The reported mean and standard deviation of the test scores are 68 and 8, respectively. If test scores are normally distributed, what is the minimum score required for an interview?
36. A financial advisor informs a client that the expected return on a portfolio is 8% with a standard deviation of 12%. There is a 15% chance that the return would be above 16%. If the advisor is right about her assessment, is it reasonable to assume that the underlying return distribution is normal?
37. A packaging system fills boxes to an average weight of 18 ounces with a standard deviation of 0.2 ounce. It is reasonable to assume that the weights are normally distributed. Calculate the 1st, 2nd, and 3rd quartiles of the box weight.
38. According to the Bureau of Labor Statistics, it takes an average of 22 weeks for someone over 55 to find a new job, compared with 16 weeks for younger workers (*The Wall Street Journal*, September 2, 2008). Assume that the probability distributions are normal and that the standard deviation is 2 weeks for both distributions.
    a. What is the probability that it takes a worker over the age of 55 more than 19 weeks to find a job?
    b. What is the probability that it takes a younger worker more than 19 weeks to find a job?
    c. What is the probability that it takes a worker over the age of 55 between 23 and 25 weeks to find a job?
    d. What is the probability that it takes a younger worker between 23 and 25 weeks to find a job?
39. Loans that are 60 days or more past due are considered seriously delinquent. The Mortgage Bankers Association reported that the rate of seriously delinquent loans has an average of 9.1% (*The Wall Street Journal*, August 26, 2010). Let the rate of seriously delinquent loans follow a normal distribution with a standard deviation of 0.80%.
    a. What is the probability that the rate of seriously delinquent loans is above 8%?
    b. What is the probability that the rate of seriously delinquent loans is between 9.5% and 10.5%?
40. The time required to assemble an electronic component is normally distributed with a mean and a standard deviation of 16 minutes and 4 minutes, respectively.
    a. Find the probability that a randomly picked assembly takes between 10 and 20 minutes.
    b. It is unusual for the assembly time to be above 24 minutes or below 6 minutes. What proportion of assembly times fall in these unusual categories?

41. Research suggests that Americans make an average of 10 phone calls per day (*CNN*, August 26, 2010). Let the number of calls be normally distributed with a standard deviation of 3 calls.
    a. What is the probability that an average American makes between 4 and 12 calls per day?
    b. What is the probability that an average American makes more than 6 calls per day?
    c. What is the probability that an average American makes more than 16 calls per day?
42. The manager of a night club in Boston stated that 95% of the customers are between the ages of 22 and 28 years. If the age of customers is normally distributed with a mean of 25 years, calculate its standard deviation.
43. An estimated 1.8 million students take on student loans to pay ever-rising tuition and room and board (*The New York Times*, April 17, 2009). It is also known that the average cumulative debt of recent college graduates is about $22,500. Let the cumulative debt among recent college graduates be normally distributed with a standard deviation of $7,000. Approximately how many recent college graduates have accumulated student loans of more than $30,000?
44. Scores on a marketing exam are known to be normally distributed with a mean and a standard deviation of 60 and 20, respectively.
    a. Find the probability that a randomly selected student scores between 50 and 80.
    b. Find the probability that a randomly selected student scores between 20 and 40.
    c. The syllabus suggests that the top 15% of the students will get an A in the course. What is the minimum score required to get an A?
    d. What is the passing score if 10% of the students will fail the course?
45. On average, an American professional football game lasts about three hours, even though the ball is actually in play only 11 minutes (www.sportsgrid.com, January 14, 2014). Assume that game times are normally distributed with a standard deviation of 0.4 hour.
    a. Find the probability that a game lasts less than 2.5 hours.
    b. Find the probability that a game lasts either less than 2.5 hours or more than 3.5 hours.
    c. Find the maximum value for the game time that will place it in the bottom 1% of the distribution.
46. A young investment manager tells his client that the probability of making a positive return with his suggested portfolio is 90%. If it is known that returns are normally distributed with a mean of 5.6%, what is the risk, measured by standard deviation, that this investment manager assumes in his calculation?
47. A construction company in Naples, Florida, is struggling to sell condominiums. In order to attract buyers, the company has made numerous price reductions and better financing offers. Although condominiums were once listed for $300,000, the company believes that it will be able to get an average sale price of $210,000. Let the price of these condominiums in the next quarter be normally distributed with a standard deviation of $15,000.
    a. What is the probability that the condominium will sell at a price (i) below $200,000? (ii) Above $240,000?
    b. The company is also trying to sell an artist's condo. Potential buyers will find the unusual features of this condo either pleasing or objectionable. The manager expects the average sale price of this condo to be the same as others at $210,000, but with a higher standard deviation of $20,000. What is the probability that this condo will sell at a price (i) below $200,000? (ii) Above $240,000?
48. You are considering the risk-return profile of two mutual funds for investment. The relatively risky fund promises an expected return of 8% with a standard deviation of 14%. The relatively less risky fund promises an expected return and standard deviation of 4% and 5%, respectively. Assume that the returns are approximately normally distributed.
    a. Which mutual fund will you pick if your objective is to minimize the probability of earning a negative return?
    b. Which mutual fund will you pick if your objective is to maximize the probability of earning a return above 8%?
49. First introduced in Los Angeles, the concept of Korean-style tacos sold from a catering truck has been gaining popularity nationally (*The New York Times*, July 27, 2010). This taco is an interesting mix of corn tortillas with Korean-style beef, garnished with onion, cilantro, and a hash of chili-soy-dressed lettuce. Suppose one such taco truck operates in the Detroit area. The owners have estimated that the daily consumption of beef is normally distributed with a mean of 24 pounds and a standard deviation of 6 pounds. While purchasing too much beef results in wastage, purchasing too little can disappoint customers.
    a. Determine the amount of beef the owners should buy so that it meets demand on 80% of the days.
    b. How much should the owners buy if they want to meet demand on 95% of the days?
50. While Massachusetts is no California when it comes to sun, the solar energy industry is flourishing in this state (*The Boston Globe*, May 27, 2012). The state's capital, Boston, averages 211.7 sunny days per year. Assume that the number of sunny days follows a normal distribution with a standard deviation of 20 days.
    a. What is the probability that Boston has less than 200 sunny days in a given year?
    b. Los Angeles averages 266.5 sunny days per year. What is the probability that Boston has at least as many sunny days as Los Angeles?

c. Suppose a dismal year in Boston is one where the number of sunny days is in the bottom 10% for that year. At most, how many sunny days must occur annually for it to be a dismal year in Boston?

d. In 2012, Boston experienced unusually warm, dry, and sunny weather. Suppose this occurs only 1% of the time. What is the minimum number of sunny days that would satisfy the criteria for being an unusually warm, dry, and sunny year in Boston?

51. A new car battery is sold with a two-year warranty whereby the owner gets the battery replaced free of cost if it breaks down during the warranty period. Suppose an auto store makes a net profit of $20 on batteries that stay trouble-free during the warranty period; it makes a net loss of $10 on batteries that break down. The life of batteries is known to be normally distributed with a mean and a standard deviation of 40 and 16 months, respectively.

a. What is the probability that a battery will break down during the warranty period?

b. What is the expected profit of the auto store on a battery?

c. What is the expected monthly profit on batteries if the auto store sells an average of 500 batteries a month?

52. A certain brand of refrigerators has a length of life that is normally distributed with a mean and a standard deviation of 15 years and 2 years, respectively.

a. What is the probability a refrigerator will last less than 6.5 years?

b. What is the probability that a refrigerator will last more than 23 years?

c. What length of life should the retailer advertise for these refrigerators so that only 3% of the refrigerators fail before the advertised length of life?

LO 6.5

Calculate and interpret probabilities for a random variable that follows the exponential distribution.

6.3 THE EXPONENTIAL DISTRIBUTION

As discussed earlier, the normal distribution is the most extensively used probability distribution in statistical work. One reason that this occurs is because the normal distribution accurately describes numerous random variables of interest. However, there are applications where other continuous distributions are more appropriate.

A useful nonsymmetric continuous probability distribution is the **exponential distribution**. The exponential distribution is related to the Poisson distribution, even though the Poisson distribution deals with discrete random variables. Recall from Chapter 5 that the Poisson random variable counts the number of occurrences of an event over a given interval of time or space. For instance, the Poisson distribution is used to calculate the likelihood of a specified number of cars arriving at a McDonald's drive-thru over a particular time period or the likelihood of a specified number of defects in a 50-yard roll of fabric. Sometimes we are less interested in the *number* of occurrences over a given interval of time or space, but rather in the time that has elapsed or space encountered *between* such occurrences. For instance, we might be interested in the length of time that elapses between car arrivals at the McDonald's drive-thru or the distance between defects in a 50-yard roll of fabric. We use the exponential distribution for describing these times or distances. The exponential random variable is nonnegative; that is, the underlying variable X is defined for $x \geq 0$.

In order to better understand the connection between the Poisson and the exponential distributions, consider the introductory case of Chapter 5 where Anne was concerned about staffing needs at the Starbucks that she managed. Recall that Anne believed that the typical Starbucks customer averaged 18 visits to the store over a 30-day period. The Poisson random variable appropriately captures the number of visits, with the expected value (mean), over a 30-day period, as

$$\mu_{\text{Poisson}} = 18.$$

Since the number of visits follows the Poisson distribution, the time between visits has an exponential distribution. In addition, given the expected number of 18 visits over a 30-day month, the expected time between visits is derived as

$$\mu_{\text{Exponential}} = \frac{30}{18} = 1.67.$$

It is common to define the exponential probability distribution in terms of its *rate parameter* λ (the Greek letter lambda), which is the inverse of its mean. In the above example,

$$\lambda = \frac{1}{\mu} = \frac{1}{1.67} = 0.60.$$

We can think of the mean of the exponential distribution as the average time between arrivals, whereas the rate parameter measures the average number of arrivals per unit of time. Note that the rate parameter is the same as the mean of the Poisson distribution, when defined per unit of time. For a Poisson process, the mean of 18 visits over a 30-day period is equivalent to a mean of 18/30 = 0.60 per day, which is the same as the rate parameter λ.

The probability density function for the exponential distribution is defined as follows.

THE EXPONENTIAL DISTRIBUTION

A random variable X follows the exponential distribution if its probability density function is

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0,$$

where λ is a rate parameter and $e \approx 2.718$ is the base of the natural logarithm.

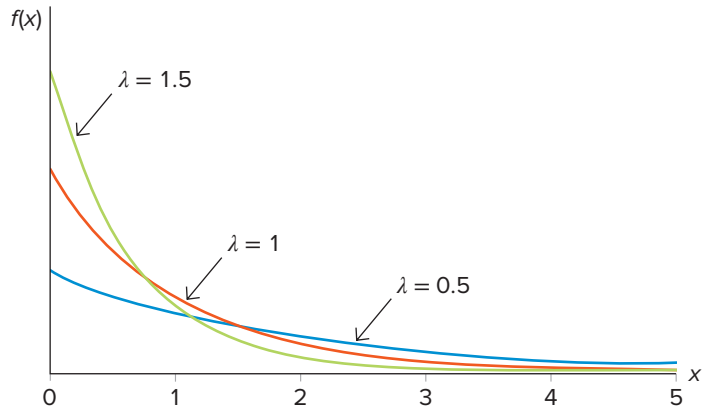
The mean and the standard deviation of X are equal: $E(X) = SD(X) = \frac{1}{\lambda}$. For $x \geq 0$, the cumulative distribution function of X is

$$P(X \leq x) = 1 - e^{-\lambda x}.$$

Therefore, $P(X > x) = 1 - P(X \leq x) = e^{-\lambda x}$.

The graphs in Figure 6.22 show the shapes of the exponential probability density function based on various values of the rate parameter λ.

FIGURE 6.22 Exponential probability density function for various values of λ



EXAMPLE 6.9

Let the time between e-mail messages during work hours be exponentially distributed with a mean of 25 minutes.

a. Calculate the rate parameter λ.

b. What is the probability that you do not get an e-mail for more than one hour?

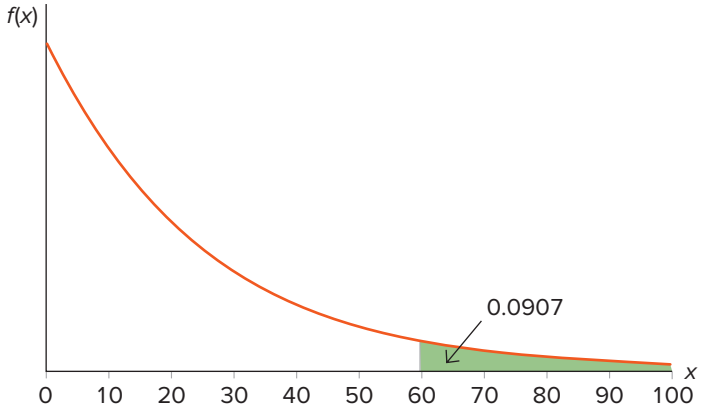
c. What is the probability that you get an e-mail within 10 minutes?

SOLUTION:

a. Since the mean $E(X)$ equals $\frac{1}{\lambda}$, we compute $\lambda = \frac{1}{E(X)} = \frac{1}{25} = 0.04$.

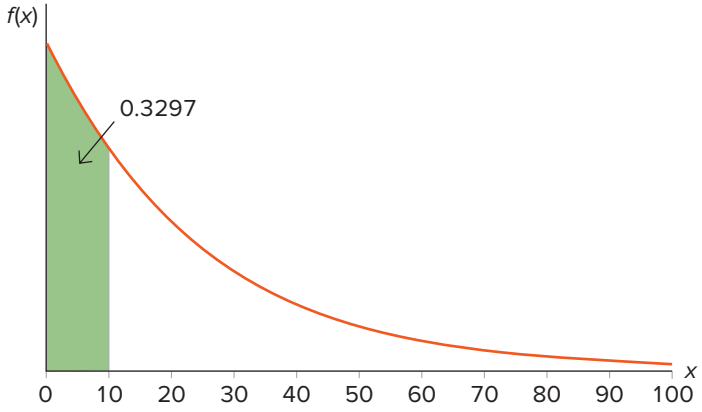
b. The probability that you do not get an e-mail for more than an hour is $P(X > 60)$. We use $P(X > x) = e^{-\lambda x}$ to compute $P(X > 60) = e^{-0.04(60)} = 0.0907$. The probability of not getting an e-mail for more than one hour is 0.0907. Figure 6.23 highlights this probability.

FIGURE 6.23 Finding $P(X > 60)$



c. The probability that you get an e-mail within 10 minutes is $P(X \leq 10) = 1 - e^{-0.04(10)} = 1 - 0.6703 = 0.3297$. Figure 6.24 highlights this probability.

FIGURE 6.24 Finding $P(X \leq 10)$



The exponential distribution is also used in modeling lifetimes or failure times. For example, an electric bulb with a rated life of 1,000 hours is expected to fail after about 1,000 hours of use. However, the bulb may burn out either before or after 1,000 hours. Thus, the lifetime of an electric bulb is a random variable with an expected value of 1,000. A noted feature of the exponential distribution is that it is "memoryless," thus implying a constant failure rate. In the electric bulb example, it implies that the probability that the bulb will burn out on a given day is independent of whether the bulb has already been used for 10, 100, or 1,000 hours.

Using Excel for the Exponential Distribution

Table 6.4 from the last section shows the Excel function that we can use to solve problems associated with the exponential distribution. Example 6.10 illustrates the use of this function.

EXAMPLE 6.10

A barbershop has an average of 12 customers between 8:00 am and 9:00 am every Saturday. Customers arrive according to the Poisson distribution. What is the probability that the time between consecutive arrivals (customers) will fall between 3 and 6 minutes?

SOLUTION: Since the number of arrivals follows the Poisson distribution, the time between arrivals has an exponential distribution. Also, if $\mu_{\text{Poisson}} = 12$ (12 arrivals over a 60-minute interval), then $\mu_{\text{Exponential}} = \frac{60}{12} = 5$ and $\lambda = \frac{12}{60} = \frac{1}{5} = 0.20$ (per minute). Let X represent the time between arrivals.

We use Excel's **EXPON.DIST** function to solve for probabilities associated with the exponential distribution. In order to find $P(X \leq x)$, we enter "=EXPON.DIST(x, λ, 1)", where x is the value for which we want to evaluate the cumulative probability and λ is the rate parameter. (If we enter "0" for the last argument in the function, then Excel returns the height of the exponential distribution at the point x.) With respect to the exponential distribution, Excel does not provide a function if we want to find a particular x value for a given cumulative probability.

In order to find the probability that the time between consecutive arrivals (customers) will fall between 3 and 6 minutes, $P(3 \leq X \leq 6)$, we enter "=EXPON.DIST(6, 0.20, 1) − EXPON.DIST(3, 0.20, 1)". Excel returns 0.2476.

EXERCISES 6.3

Mechanics

53. Assume a Poisson random variable has a mean of 6 successes over a 120-minute period.
    a. Find the mean of the random variable, defined by the time between successes.
    b. What is the rate parameter of the appropriate exponential distribution?
    c. Find the probability that the time to success will be more than 60 minutes.
54. Assume a Poisson random variable has a mean of four arrivals over a 10-minute interval.
    a. What is the mean of the random variable, defined by the time between arrivals?
    b. Find the probability that the next arrival would be within the mean time.
    c. Find the probability that the next arrival would be between one and two minutes.
55. A random variable X is exponentially distributed with a mean of 0.1.
    a. What is the rate parameter λ? What is the standard deviation of X?
    b. Compute $P(X > 0.20)$.
    c. Compute $P(0.10 \leq X \leq 0.20)$.

56. A random variable X is exponentially distributed with a probability density function of $f(x) = 5e^{-5x}$. Calculate the mean and the standard deviation of X.

57. A random variable X is exponentially distributed with an expected value of 25.
    a. What is the rate parameter λ? What is the standard deviation of X?
    b. Compute $P(20 \leq X \leq 30)$.
    c. Compute $P(15 \leq X \leq 35)$.

58. Let X be exponentially distributed with $\mu = 1.25$. Compute the following values.
    a. $P(X < 2.3)$
    b. $P(1.5 \leq X \leq 5.5)$
    c. $P(X > 7)$

59. Let X be exponentially distributed with $\lambda = 0.5$. Use Excel's function options to find the following values.
    a. $P(X \leq 1)$
    b. $P(2 < X < 4)$
    c. $P(X > 10)$

Applications

60. Studies have shown that bats can consume an average of 10 mosquitoes per minute (berkshiremuseum.org). Assume that the number of mosquitoes consumed per minute follows a Poison distribution.
    a. What is the mean time between eating mosquitoes?
    b. Find the probability that the time between eating mosquitoes is more than 15 seconds.
    c. Find the probability that the time between eating mosquitoes is between 15 and 20 seconds.

61. According to the *Daily Mail* (February 28, 2012), there was an average of one complaint every 12 seconds against Britain's biggest banks in 2011. It is reasonable to assume that the time between complaints is exponentially distributed.
    a. What is the mean time between complaints?
    b. What is the probability that the next complaint will take less than the mean time?
    c. What is the probability that the next complaint will take between 5 and 10 seconds?

62. A tollbooth operator has observed that cars arrive randomly at an average rate of 360 cars per hour.
    a. What is the mean time between car arrivals at this tollbooth?
    b. What is the probability that the next car will arrive within ten seconds?

63. Customers make purchases at a convenience store, on average, every six minutes. It is fair to assume that the time between customer purchases is exponentially distributed. Jack operates the cash register at this store.
    a. What is the rate parameter λ? What is the standard deviation of this distribution?
    b. Jack wants to take a five-minute break. He believes that if he goes right after he has serviced a customer, he will lower the probability of someone showing up during his five-minute break. Is he right in this belief?
    c. What is the probability that a customer will show up in less than five minutes?
    d. What is the probability that nobody shows up for over half an hour?

64. A hospital administrator worries about the possible loss of electric power as a result of a power blackout. The hospital, of course, has a standby generator, but it, too, is subject to failure, having a mean time between failures of 500 hours. It is reasonable to assume that the time between failures is exponentially distributed.
    a. What is the probability that the standby generator fails during the next 24-hour blackout?
    b. Suppose the hospital owns two standby generators that work independently of one another. What is the probability that both generators fail during the next 24-hour blackout?

65. When crossing the Golden Gate Bridge traveling into San Francisco, all drivers must pay a toll. Suppose the amount of time (in minutes) drivers wait in line to pay the toll follows an exponential distribution with a probability density function of $f(x) = 0.2e^{-0.2x}$.
    a. What is the mean waiting time that drivers face when entering San Francisco via the Golden Gate Bridge?
    b. What is the probability that a driver spends more than the average time to pay the toll?
    c. What is the probability that a driver spends more than 10 minutes to pay the toll?
    d. What is the probability that a driver spends between 4 and 6 minutes to pay the toll?

66. On average, the state police catch eight speeders per hour at a certain location on Interstate 90. Assume that the number of speeders per hour follows the Poisson distribution.
    a. What is the probability that the state police wait less than 10 minutes for the next speeder?
    b. What is the probability that the state police wait between 15 and 20 minutes for the next speeder?
    c. What is the probability that the state police wait more than 25 minutes for the next speeder?

67. Motorists arrive at a Gulf station at the rate of two per minute during morning hours. Assume that the arrival of motorists at the station follows a Poisson distribution.
    a. What is the probability that the next car's arrival is in less than one minute?
    b. What is the probability that the next car's arrival is in more than five minutes?

WRITING WITH STATISTICS

Professor Lang is a professor of economics at Salem State University. She has been teaching a course in Principles of Economics for over 25 years. Professor Lang has never graded on a curve since she believes that relative grading may unduly penalize (benefit) a good (poor) student in an unusually strong (weak) class. She always uses an absolute scale for making grades, as shown in the two left columns of Table 6.5.



©Image Source, all rights reserved.

TABLE 6.5 Grading Scales with Absolute Grading versus Relative Grading

Absolute Grading		Relative Grading	
Grade	Score	Grade	Probability
A	92 and above	A	0.10
B	78 up to 92	B	0.35
C	64 up to 78	C	0.40
D	58 up to 64	D	0.10
F	Below 58	F	0.05

A colleague of Professor Lang's has convinced her to move to relative grading, since it corrects for unanticipated problems. Professor Lang decides to experiment with grading based on the relative scale as shown in the two right columns of Table 6.5. Using this relative grading scheme, the top 10% of students will get A's, the next 35% B's, and so on. Based on her years of teaching experience, Professor Lang believes that the scores in her course follow a normal distribution with a mean of 78.6 and a standard deviation of 12.4.

Professor Lang wants to use the above information to

1. Calculate probabilities based on the absolute scale. Compare these probabilities to the relative scale.
2. Calculate the range of scores for various grades based on the relative scale. Compare these ranges to the absolute scale.
3. Determine which grading scale makes it harder to get higher grades.

Sample Report—Absolute Grading versus Relative Grading

Many teachers would confess that grading is one of the most difficult tasks of their profession. Two common grading systems used in higher education are relative and absolute. Relative grading systems are norm-referenced or curve-based, in which a grade is based on the student's relative position in class. Absolute grading systems, on the other hand, are criterion-referenced, in which a grade is related to the student's absolute performance in class. In short, with absolute grading, the student's score is compared to a predetermined scale, whereas with relative grading, the score is compared to the scores of other students in the class.

Let X represent a grade in Professor Lang's class, which is normally distributed with a mean of 78.6 and a standard deviation of 12.4. This information is used to derive the grade probabilities based on the absolute scale. For instance, the probability of receiving an A is derived as $P(X \geq 92) = P(Z \geq 1.08) = 0.14$. Other probabilities, derived similarly, are presented in Table 6.A.

TABLE 6.A Probabilities Based on Absolute Scale and Relative Scale

Grade	Probability Based on Absolute Scale	Probability Based on Relative Scale
A	0.14	0.10
B	0.38	0.35
C	0.36	0.40
D	0.07	0.10
F	0.05	0.05

The second column of Table 6.A shows that 14% of students are expected to receive A's, 38% B's, and so on. Although these numbers are generally consistent with the relative scale restated in the third column of Table 6.A, it appears that the relative scale makes it harder for students to get higher grades. For instance, 14% get A's with the absolute scale compared to only 10% with the relative scale.

Alternatively, we can compare the two grading methods on the basis of the range of scores for various grades. The second column of Table 6.B restates the range of scores based on absolute grading. In order to obtain the range of scores based on relative grading, it is once again necessary to apply concepts from the normal distribution. For instance, the minimum score required to earn an A with relative grading is derived by solving for x in $P(X \geq x) = 0.10$. Since $P(X \geq x) = 0.10$ is equivalent to $P(Z \geq z) = 0.10$, it follows that $z = 1.28$. Inserting the proper values of the mean, the standard deviation, and z into $x = \mu + z\sigma$ yields a value of x equal to 94.47. Ranges for other grades, derived similarly, are presented in the third column of Table 6.B.

TABLE 6.B Range of Scores with Absolute Grading versus Relative Grading

Grade	Range of Scores Based on Absolute Grading	Range of Scores Based on Relative Grading
A	92 and above	94.47 and above
B	78 up to 92	80.21 up to 94.47
C	64 up to 78	65.70 up to 80.21
D	58 up to 64	58.20 up to 65.70
F	Below 58	Below 58.20

Once again comparing the results in Table 6.B, the use of the relative scale makes it harder for students to get higher grades in Professor Lang's courses. For instance, in order to receive an A with relative grading, a student must have a score of at least 94.47 versus a score of at least 92 with absolute grading. Both absolute and relative grading methods have their merits and teachers often make the decision on the basis of their teaching philosophy. However, if Professor Lang wants to keep the grades consistent with her earlier absolute scale, she should base her relative scale on the probabilities computed in the second column of Table 6.A.

CONCEPTUAL REVIEW

LO 6.1 Describe a continuous random variable.

A **continuous random variable** is characterized by uncountable values because it can take on any value within an interval. The probability that a continuous random variable X assumes a particular value x is zero; that is, $P(X = x) = 0$. Thus, for a continuous random variable, we calculate the probability within a specified interval. Moreover, the following equalities hold: $P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$.

The **probability density function** $f(x)$ of a continuous random variable X is nonnegative and the entire area under this function equals one. The probability $P(a \leq X \leq b)$ is the area under $f(x)$ between points a and b.

For any value x of the random variable X, the **cumulative distribution function** $F(x)$ is defined as $F(x) = P(X \leq x)$.

LO 6.2 Calculate and interpret probabilities for a random variable that follows the continuous uniform distribution.

The **continuous uniform distribution** describes a random variable that has an equally likely chance of assuming a value within a specified range. The probability is essentially

the area of a rectangle, which is the base times the height; that is, the length of a specified interval times the probability density function $f(x) = \frac{1}{b-a}$, where a and b are the lower and upper bounds of the interval, respectively.

LO 6.3 Explain the characteristics of the normal distribution.

The **normal distribution** is the most extensively used continuous probability distribution and is the cornerstone of statistical inference. It is the familiar bell-shaped distribution, which is symmetric around the mean. The normal distribution is completely described by two parameters: the population mean μ and the population variance σ^2.

The **standard normal distribution**, also referred to as the z **distribution**, is a special case of the normal distribution, with mean equal to zero and standard deviation (or variance) equal to one. The **standard normal table**, also called the z **table**, provides **cumulative probabilities** $P(Z \leq z)$; this table appears on two pages in Table 1 of Appendix A. The left-hand page provides cumulative probabilities for z values less than or equal to zero. The right-hand page shows cumulative probabilities for z values greater than or equal to zero. We also use the table to compute z values for given cumulative probabilities.

LO 6.4 Calculate and interpret probabilities for a random variable that follows the normal distribution.

Any normally distributed random variable X with mean μ and standard deviation σ can be transformed into the standard normal random variable Z as $Z = \frac{X-\mu}{\sigma}$. This standard transformation implies that any value x has a corresponding value z given by $z = \frac{x-\mu}{\sigma}$.

The standard normal variable Z can be transformed to the normally distributed random variable X with mean μ and standard deviation σ as $X = \mu + Z\sigma$. This inverse transformation implies that any value z has a corresponding value x given by $x = \mu + z\sigma$.

LO 6.5 Calculate and interpret probabilities for a random variable that follows the exponential distribution.

A useful nonsymmetric continuous probability distribution is the **exponential distribution**. A random variable X follows the exponential distribution if its probability density function is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, where λ is a rate parameter and $e \approx 2.718$ is the base of the natural logarithm. The mean and the standard deviation of the distribution are both equal to $1/\lambda$. For $x \geq 0$, the **cumulative probability** is computed as $P(X \leq x) = 1 - e^{-\lambda x}$.

ADDITIONAL EXERCISES AND CASE STUDIES

Exercises

68. A florist makes deliveries between 1:00 pm and 5:00 pm daily. Assume delivery times follow the continuous uniform distribution.
    a. Calculate the mean and the variance of this distribution.
    b. Determine the percentage of deliveries that are made after 4:00 pm.
    c. Determine the percentage of deliveries that are made prior to 2:30 pm.

69. A worker at a landscape design center uses a machine to fill bags with potting soil. Assume that the quantity put in each bag follows the continuous uniform distribution with low and high filling weights of 10 pounds and 12 pounds, respectively.
    a. Calculate the expected value and the standard deviation of this distribution.
    b. Find the probability that the weight of a randomly selected bag is no more than 11 pounds.
    c. Find the probability that the weight of a randomly selected bag is at least 10.5 pounds.

70. The revised guidelines from the National High Blood Pressure Education Program define normal blood pressure as readings below 120/80 millimeters of mercury (*The New York Times*, May 14, 2003). Prehypertension is suspected when the top number (systolic) is between 120 and 139 or when the bottom number (diastolic) is between 80 and 90. A recent survey reported that the mean systolic reading of Canadians is 125 with a standard deviation of 17 and the mean diastolic reading is 79 with a standard deviation of 10. Assume that diastolic as well as systolic readings are normally distributed.
    a. What proportion of Canadians are suffering from prehypertension caused by high diastolic readings?
    b. What proportion of Canadians are suffering from prehypertension caused by high systolic readings?

71. U.S. consumers are increasingly viewing debit cards as a convenient substitute for cash and checks. The average amount spent annually on a debit card is $7,790 (*Kiplinger's*, August 2007). Assume that the average amount spent on a debit card is normally distributed with a standard deviation of $500.
    a. A consumer advocate comments that the majority of consumers spend over $8,000 on a debit card. Find a flaw in this statement.
    b. Compute the 25th percentile of the amount spent on a debit card.
    c. Compute the 75th percentile of the amount spent on a debit card.
    d. What is the interquartile range of this distribution?

72. On St. Patrick's Day, men spend an average of $43.87 while women spend an average of $29.54 (*USA TODAY*, March 17, 2009). Assume the standard deviations of spending for men and women are $2.50 and $9.00, respectively, and that both distributions are normally distributed.
    a. What is the probability that men spend over $50 on St. Patrick's Day?
    b. What is the probability that women spend over $50 on St. Patrick's Day?
    c. Are men or women more likely to spend over $50 on St. Patrick's Day?

73. Lisa Mendes and Brad Lee work in the sales department of an AT&T Wireless store. Lisa has been signing up an average of 148 new cell phone customers every month with a standard deviation of 22, while Brad signs up an average of 156 new customers with a standard deviation of 17. The store manager offers both Lisa and Brad a $100 incentive bonus if they can sign up more than 200 new customers in a month. Assume a normal distribution to answer the following questions.
    a. What is the probability that Lisa will earn the $100 incentive bonus?
    b. What is the probability that Brad will earn the $100 incentive bonus?
    c. Are you surprised by the results? Explain.

74. The car speeds on a certain stretch of the interstate highway I-95 are known to be normally distributed with a mean of 72 and a standard deviation of 15. You have just heard a policeman comment that about 3% of the drivers drive at extremely dangerous speeds. What is the minimum speed that the policeman considers extremely dangerous?

75. The average household income in a community is known to be $80,000. Also, 20% of the households have an income below $60,000 and another 20% have an income above $90,000. Is it reasonable to use the normal distribution to model the household income in this community?

76. The length of components produced by a company is normally distributed with a mean of 6 cm and a standard deviation of 0.02 cm. Calculate the first, second, and third quartiles of the component length.

77. Entrance to a prestigious MBA program in India is determined by a national test where only the top 10% of the examinees are admitted to the program. Suppose it is known that the scores on this test are normally distributed with a mean of 420 and a standard deviation of 80. Parul Monga is trying desperately to get into this program. What is the minimum score that she must earn to get admitted?

78. A new water filtration system is sold with a 10-year warranty that includes all parts and repairs. Suppose the life of this water filtration system is normally distributed with mean and standard deviation of 16 and 5 years, respectively.
    a. What is the probability that the water filtration system will require a repair during the warranty period?
    b. Suppose the water filtration firm makes a $300 profit for every new system it installs. This profit, however, is reduced to $50 if the system requires repair during the warranty period. Find the expected profit of the firm if it installs 1,000 new water filtration systems.

79. Suppose that the average IQ score is normally distributed with a mean of 100 and a standard deviation of 16.
    a. What is the probability that a randomly selected person will have an IQ score of less than 80?
    b. What is the probability that a randomly selected person will have an IQ score greater than 125?
    c. What minimum IQ score does a person have to achieve to be in the top 2.5% of IQ scores?
80. Suppose that the annual household income in a small Midwestern community is normally distributed with a mean of $55,000 and a standard deviation of $4,500.
    a. What is the probability that a randomly selected household will have an income between $50,000 and $65,000?
    b. What is the probability that a randomly selected household will have an income of more than $70,000?
    c. What minimum income does a household need to earn to be in the top 5% of incomes?
    d. What maximum income does a household need to earn to be in the bottom 40% of incomes?
81. On a particularly busy section of the Garden State Parkway in New Jersey, police use radar guns to detect speeders. Assume the time that elapses between successive speeders is exponentially distributed with a mean of 15 minutes.
    a. Calculate the rate parameter λ.
    b. What is the probability of a waiting time less than 10 minutes between successive speeders?
    c. What is the probability of a waiting time in excess of 25 minutes between successive speeders?
82. In a local law office, jobs to a printer are sent at a rate of 8 jobs per hour. Suppose that the number of jobs sent to a printer follows the Poisson distribution.
    a. What is the expected time between successive jobs?
    b. What is the probability that the next job will be sent within five minutes?
83. According to the Federal Bureau of Investigation, there is a violent crime in the United States every 22 seconds (*ABC News*, September 25, 2007). Assume that the time between successive violent crimes is exponentially distributed.
    a. What is the probability that there is a violent crime in the United States in the next one minute?
    b. If there has not been a violent crime in the previous minute, what is the probability that there will be a violent crime in the subsequent minute?
84. Disturbing news regarding Scottish police concerns the number of crashes involving vehicles on operational duties (*BBC News*, March 10, 2008). Statistics showed that Scottish forces' vehicles had been involved in traffic accidents at the rate of 1,000 per year. Suppose the number of crashes involving vehicles on operational duties follows a Poisson distribution.
    a. What is the average number of days between successive crashes?
    b. What is the rate parameter of the appropriate exponential distribution?
    c. What is the probability that the next vehicle will crash within a day?
85. The mileage (in 1,000s of miles) that car owners get with a certain kind of radial tire is a random variable having an exponential distribution with a mean of 50.
    a. What is the probability that a tire will last at most 40,000 miles?
    b. What is the probability that a tire will last at least 65,000 miles?
    c. What is the probability that a tire will last between 70,000 and 80,000 miles?
86. A large technology firm receives an average of 12 new job applications every 10 days for positions that are not even advertised. Suppose the number of job applications received follows a Poisson distribution.
    a. What is the average number of days between successive job applications?
    b. What is the probability that the next job application is received within a day?
    c. What is the probability that the next job application is received between the next 1 and 2 days?
87. On average, a certain kind of kitchen appliance requires repairs once every four years. Assume that the times between repairs are exponentially distributed.
    a. What is the probability that the appliance will work no more than three years without requiring repairs?
    b. What is the probability that the appliance will work at least six years without requiring repairs?

CASE STUDIES

CASE STUDY 6.1 Body mass index (BMI) is a reliable indicator of body fat for most children and teens. BMI is calculated from a child's weight and height and is used as an easy-to-perform method of screening for weight categories that may lead to health problems. For children and teens, BMI is age- and sex-specific and is often referred to as BMI-for-age.

The Centers for Disease Control and Prevention (CDC) reports BMI-for-age growth charts for girls as well as boys to obtain a percentile ranking. Percentiles are the most commonly used indicator to assess the size and growth patterns of individual children in the United States.

The following table provides weight status categories and the corresponding percentiles and BMI ranges for 10-year-old boys in the United States.

Weight Status Category	Percentile Range	BMI Range
Underweight	Less than 5th	Less than 14.2
Healthy Weight	Between 5th and 85th	Between 14.2 and 19.4
Overweight	Between 85th and 95th	Between 19.4 and 22.2
Obese	More than 95th	More than 22.2

Health officials of a Midwestern town are concerned about the weight of children in their town. They believe that the BMI of their 10-year-old boys is normally distributed with mean 19.2 and standard deviation 2.6.

In a report, use the sample information to

1. Compute the proportion of 10-year-old boys in this town that are in the various weight status categories given the BMI ranges.
2. Discuss whether the concern of health officials is justified.

CASE STUDY 6.2 Vanguard's Precious Metals and Mining fund (Metals) and Fidelity's Strategic Income fund (Income) were two top-performing mutual funds for the years 2000 through 2009. An analysis of annual return data for these two funds provided important information for any type of investor. Over the past 10 years, the Metals fund posted a mean return of 24.65% with a standard deviation of 37.13%. On the other hand, the mean and the standard deviation of return for the Income fund were 8.51% and 11.07%, respectively. It is reasonable to assume that the returns of the Metals and the Income funds are both normally distributed, where the means and the standard deviations are derived from the 10-year sample period.

In a report, use the sample information to compare and contrast the Metals and Income funds from the perspective of an investor whose objective is to

1. Minimize the probability of earning a negative return.
2. Maximize the probability of earning a return between 0% and 10%.
3. Maximize the probability of earning a return greater than 10%.

CASE STUDY 6.3 A variety of packaging solutions exist for products that must be kept within a specific temperature range. A cold chain distribution is a temperature-controlled supply chain. An unbroken cold chain is an uninterrupted series of storage and distribution activities that maintain a given temperature range. Cold chains are particularly useful in the food and pharmaceutical industries. A common suggested temperature range for a cold chain distribution in pharmaceutical industries is between 2 and 8 degrees Celsius.

Gopal Vasudeva works in the packaging branch of Merck & Co. He is in charge of analyzing a new package that the company has developed. With repeated trials, Gopal has determined that the mean temperature that this package is able to maintain during its use is 5.6°C with a standard deviation of 1.2°C. He is not sure if the distribution of temperature is symmetric or skewed to the right.

In a report, use the sample information to

1. Calculate and interpret the probability that temperature goes (a) below 2°C and (b) above 8°C using a normal distribution approximation.
2. Calculate and interpret the 5th and the 95th percentiles of the temperatures that the package maintains.

APPENDIX 6.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, JMP, and R.

Minitab

The Uniform Distribution

A. (Replicating Example 6.1b) From the menu, choose **Calc > Probability Distributions > Uniform**.

B. Select **Cumulative probability**. Enter 2,500 as the **Lower endpoint** and 5,000 as the **Upper endpoint**. Select **Input constant** and enter 4,000. Minitab returns $P(X \leq 4{,}000)$. Subtract this probability from 1.0 to get the answer.

The Normal Distribution

The Standard Transformation

A. (Replicating Example 6.8a) From the menu, choose **Calc > Probability Distributions > Normal**.

B. Select **Cumulative probability**. Enter 7.49 for the **Mean** and 6.41 for the **Standard deviation**. Select **Input constant** and enter 10. Minitab returns $P(X \leq 10)$. Perform similar steps to find $P(X \leq 5)$, and then find the difference between the probabilities.

The Inverse Transformation

A. (Replicating Example 6.8b) From the menu, choose **Calc > Probability Distributions > Normal**.

B. Select **Inverse cumulative probability**. Enter 7.49 for the **Mean** and 6.41 for the **Standard deviation**. Select **Input constant** and enter 0.90.

The Exponential Distribution

A. (Replicating Example 6.10) Choose **Calc > Probability Distributions > Exponential**.

B. Select **Cumulative probability**. Enter 5 for **Scale** (since Scale = $E(X) = 5$) and 0.0 for **Threshold**. Select **Input constant** and enter 6. Minitab returns $P(X \leq 6)$. Perform similar steps to find $P(X \leq 3)$, and then find the difference between the probabilities.

SPSS

Note: In order for the calculated probability to be seen on the spreadsheet, SPSS must first "view" data on the spreadsheet. For this purpose, enter a value of zero in the first cell of the first column.

The Uniform Distribution

A. (Replicating Example 6.1b) From the menu, choose **Transform > Compute Variable**.

B. Under **Target Variable**, type cdfuniform. Under **Function group**, select **CDF & Noncentral CDF** and under **Functions and Special Variables**, double-click on **Cdf.Uniform**. In the Numeric Expression box, enter 4,000 for **quant**, 2,500 for **min**, and 5,000 for **max**. SPSS returns $P(X \leq 4{,}000)$. Subtract this probability from 1.0 to get the answer.

The Normal Distribution

The Standard Transformation

A. (Replicating Example 6.8a) From the menu, choose **Transform > Compute Variable**.

B. Under **Target Variable**, type cdfnorm. Under **Function group**, select **CDF & Noncentral CDF**, and under **Functions and Special Variables**, double-click on **Cdf.Normal**. In the **Numeric Expression** box, enter 10 for **quant**, 7.49 for **mean**, and 6.41 for **stddev**. SPSS returns $P(X \leq 10)$. Perform similar steps to find $P(X \leq 5)$, and then find the difference between the probabilities.

The Inverse Transformation

A. (Replicating Example 6.8b) From the menu, choose **Transform > Compute Variable**.

B. Under **Target Variable**, type invnorm. Under **Function group**, select **Inverse DF**, and under **Functions and Special Variables**, double-click on **Idf.Normal**. In the **Numeric Expression** box, enter 0.9 for **prob**, 7.49 for **mean**, and 6.41 for **stddev**.

The Exponential Distribution

A. (Replicating Example 6.10) From the menu, choose **Transform > Compute Variable**.

B. Under **Target Variable**, type cdfexp. Under **Function group**, select **CDF & Noncentral CDF**, and under **Functions and Special Variables**, double-click on **Cdf.Exp**. In the **Numeric Expression** box, enter 6 for **quant** and 0.2 for **scale**. SPSS returns $P(X \leq 6)$. Perform similar steps to find $P(X \leq 3)$, and then find the difference between the probabilities.

JMP

Note: In order for the calculated probability to be seen on the spreadsheet, JMP must first "view" data on the spreadsheet. For this purpose, enter a value of zero in the first cell of the first column.

The Normal Distribution

The Standard Transformation

A. (Replicating Example 6.8a) Right-click on the header at the top of the column in the spreadsheet view and select **Formula**. Under **Functions** (**grouped**), choose **Probability > Normal Distribution**.

B. Put the insertion marker on the box for **x** and click the insert button (shown as a caret ^ with the mathematical operations) until you see **mean** and **std dev** next to **x**. Enter 10 for **x**, 7.49 for **mean**, and 6.41 for **std dev**. JMP returns $P(X \leq 10)$. Perform similar steps to find $P(X \leq 5)$, and then find the difference between the probabilities.

The Inverse Transformation

A. (Replicating Example 6.8b) Right-click on the header at the top of the column in the spreadsheet view and select **Formula**. Under **Functions** (**grouped**), choose **Probability > Normal Quantile**.

B. Put the insertion marker on the box for **p** and click the insert button (shown as a caret ^ with the mathematical operations) until you see **mean** and **std dev** next to **p**. Enter 0.90 for **p**, 7.49 for **mean**, and 6.41 for **std dev**.

The Exponential Distribution

A. (Replicating Example 6.10) Right-click on the header at the top of the column in the spreadsheet view and select **Formula**. Under **Functions** (**grouped**), choose **Probability > Weibull Distribution**. (The exponential distribution is a special case of the Weibull distribution, when the shape parameter, see next step, equals 1.)

B. Put the insertion marker on the box for **x** and click the insert button (shown as a caret ^ with the mathematical operations) until you see **shape** and **scale** next to **x**. Enter 6 for **x**, 1 for **shape**, and 5 for **scale**. JMP returns $P(X \leq 6)$. Perform similar steps to find $P(X \leq 3)$, and then find the difference between the probabilities.

R

The Uniform Distribution

(Replicating Example 6.1b) Use the **punif** function to calculate uniform probabilities. In order to find $P(X \leq x)$, enter "punif(x, a, b, lower.tail=TRUE)", where x is the value for which the cumulative probability is evaluated, a represents the lower limit, and b represents the upper limit. Thus, to find $P(X > 4{,}000)$ with $a = 2{,}500$ and $b = 5{,}000$ enter:

```
> punif(4000, 2500, 5000, lower.tail=FALSE)
```

The Normal Distribution

The Standard Transformation

(Replicating Example 6.8a) Use the **pnorm** function to calculate normal probabilities. In order to find $P(X \leq x)$, enter "pnorm(x, μ, σ, lower.tail = TRUE)", where x is the value for which the cumulative probability is evaluated, μ is the mean of the distribution, and σ is the standard deviation of the distribution. Thus, to find $P(5 \leq X \leq 10)$ with $\mu = 7.49$ and $\sigma = 6.41$, enter:

```
> pnorm(10, 7.49, 6.41, lower.tail=TRUE) - pnorm(5, 7.49, 6.41, lower.
  tail=TRUE)
```

The Inverse Transformation

(Replicating Example 6.8b) Use the **qnorm** function to find a particular x value for a given cumulative probability. In order to find x for a given cumulative probability (*cumulprob*), enter "qnorm(*cumulprob*, μ, σ)". Thus, to find x to satisfy $P(X > x) = 0.10$ with $\mu = 7.49$ and $\sigma = 6.41$, enter:

```
> qnorm(0.90, 7.49, 6.41)
```

The Exponential Distribution

(Replicating Example 6.10) Use the **pexp** function to calculate exponential probabilities. In order to find $P(X \leq x)$, enter "pexp(x, λ, lower.tail=TRUE)", where x is the value for which the cumulative probability is evaluated and λ is the rate parameter. Thus, to find $P(3 \leq X \leq 6)$ with $\lambda = 0.2$, enter:

```
> pexp(6, 0.2, lower.tail=TRUE) - pexp(3, 0.2, lower.tail=TRUE)
```

7 Sampling and Sampling Distributions

Learning Objectives

After reading this chapter you should be able to:

LO **7.1** Explain common sample biases.

LO **7.2** Describe various sampling methods.

LO **7.3** Describe the sampling distribution of the sample mean.

LO **7.4** Explain the importance of the central limit theorem.

LO **7.5** Describe the sampling distribution of the sample proportion.

LO **7.6** Use a finite population correction factor.

LO **7.7** Construct and interpret control charts for quantitative and qualitative data.

In the last few chapters, we had information on the population parameters, such as the population proportion and the population mean, for the analysis of discrete and continuous random variables. In many instances, we do not know the population parameters, so we make statistical inferences on the basis of sample statistics. The credibility of any statistical inference depends on the quality of the sample on which it is based. In this chapter, we discuss various ways to draw a good sample and also highlight cases in which the sample misrepresents the population. It is important to note that any given statistical problem involves only one population, but many possible samples, from which a statistic can be derived. Therefore, while the population parameter is a constant, the sample statistic is a random variable whose value depends on the choice of the random sample. We will discuss how to evaluate the properties of sample statistics. In particular, we will study the probability distributions of the sample mean and the sample proportion based on simple random sampling. Finally, we will use these distributions to construct control charts, which are popular statistical tools for monitoring and improving quality.



©KPG_Payless/Shutterstock

Introductory Case

Marketing Iced Coffee

Although hot coffee is still Americans' drink of choice, the market share of iced coffee is growing steadily. Thirty percent of coffee drinkers had at least one iced, frozen, or blended coffee drink in 2009, up from 28% in 2008 (*The Boston Globe,* April 6, 2010). In response to this growing change in taste, the coffee chains have ramped up their offerings: Starbucks recently introduced an upgraded Frappuccino; Dunkin' Donuts launched a new iced dark roast; and McDonald's unveiled new blended coffee iced drinks and smoothies.

In order to capitalize on this trend, Starbucks advertised a Happy Hour from May 7 through May 16 when customers enjoyed a half-price Frappuccino beverage between 3 pm and 5 pm (www.starbucks.com). Anne Jones, a manager at a local Starbucks, wonders how this marketing campaign has affected her business. She knows that women and teenage girls comprise the majority of the iced-coffee market, since they are willing to spend more on indulgences. In fact, Anne reviews her records prior to the promotion and finds that 43% of iced-coffee customers were women and 21% were teenage girls. She also finds that customers spent an average of $4.18 on iced coffee with a standard deviation of $0.84.

One month after the marketing period ends, Anne surveys 50 of her iced-coffee customers and finds that they had spent an average of $4.26. In addition, 23 (46%) of the customers were women and 17 (34%) were teenage girls. Anne wants to determine if the marketing campaign has had a lingering effect on the amount of money customers spend on iced coffee and on the proportion of customers who are women and teenage girls. Anne wonders if Starbucks would have gotten such business if it had chosen not to pursue the marketing campaign.

Anne wants to use the above survey information to

1. Calculate the probability that customers spend an average of $4.26 or more on iced coffee.
2. Calculate the probability that 46% or more of iced-coffee customers are women.
3. Calculate the probability that 34% or more of iced-coffee customers are teenage girls.

A synopsis of this case is provided at the end of Section 7.3.

LO **7.1**

Explain common sample biases.

7.1 SAMPLING

A major portion of statistics is concerned with statistical inference, where we examine the problem of estimating population parameters or testing hypotheses about such parameters. Recall that a population consists of all items of interest in the statistical problem. If we had access to data that encompass the entire population, then the values of the parameters would be known and no statistical inference would be needed. Since it is generally not feasible to gather data on an entire population, we use a subset of the population, or a sample, and use this information to make statistical inference. We can think of a census and survey data as representative of population and sample data, respectively. While a census captures almost everyone in the country, a survey captures a small number of people who fit a particular category. We regularly use survey data to analyze government and business activities.

POPULATION VERSUS SAMPLE

A population consists of all items of interest in a statistical problem, whereas a sample is a subset of the population. We use a sample statistic, or simply statistic, to make inferences about the unknown population parameter.

In later chapters, we explore estimation and hypothesis testing, which are based on sample information. It is important to note that no matter how sophisticated the statistical methods are, the credibility of statistical inference depends on the quality of the sample on which it is based. A primary requisite for a "good" sample is that it be representative of the population we are trying to describe. When the information from a sample is not typical of information in the population in a systematic way, we say that **bias** has occurred.

Bias refers to the tendency of a sample statistic to systematically overestimate or underestimate a population parameter. It is often caused by samples that are not representative of the population.

Classic Case of a "Bad" Sample: The *Literary Digest* Debacle of 1936

In theory, drawing conclusions about a population based on a good sample sounds logical; however, in practice, what constitutes a "good" sample? Unfortunately, there are many ways to collect a "bad" sample. One way is to inadvertently pick a sample that represents only a portion of the population. The *Literary Digest*'s attempt to predict the 1936 presidential election is a classic example of an embarrassingly inaccurate poll.

In 1932 and amid the Great Depression, Herbert Hoover was voted out of the White House and Franklin Delano Roosevelt (FDR) was elected the 32nd president of the United States. Although FDR's attempts to end the Great Depression within four years were largely unsuccessful, he retained the general public's faith. In 1936, FDR ran for reelection against Alf Landon, the governor of Kansas and the Republican nominee. The *Literary Digest*, an influential, general-interest weekly magazine, wanted to predict the next U.S. president, as it had done successfully five times before.

After conducting the largest poll in history, the *Literary Digest* predicted a landslide victory for Alf Landon: 57% of the vote to FDR's 43%. Moreover, the *Literary Digest* claimed that its prediction would be within a fraction of 1% of the actual vote. Instead, FDR won in a landslide: 62% to 38%. So what went wrong?

The *Literary Digest* sent postcards to 10 million people (one-quarter of the voting population at the time) and received responses from 2.4 million people. The response rate of 24% (2.4 million/10 million) might seem low to some, but in reality it is a reasonable response rate given this type of polling. What was atypical of the poll is the manner in which the *Literary Digest* obtained the respondents' names. The *Literary Digest* randomly sampled its own subscriber list, club membership rosters, telephone directories, and automobile registration rolls. This sample reflected predominantly middle- and upper-class people; that is, the vast majority of those polled were wealthier people, who were more inclined to vote for the Republican candidate. Back in the 1930s, owning a phone, for instance, was far from universal. Only 11 million residential phones were in service in 1936, and these homes were disproportionately well-to-do and in favor of Landon. The sampling methodology employed by the *Literary Digest* suffered from **selection bias**. Selection bias occurs when portions of the population are underrepresented in the sample. FDR's support came from lower-income classes whose opinion was not reflected in the poll. The sample, unfortunately, misrepresented the general electorate.

Selection bias refers to a systematic underrepresentation of certain groups from consideration for the sample.

What should the *Literary Digest* have done differently? At a minimum, most would agree that names should have been obtained from voter registration lists rather than telephone directory lists and car registrations.

In addition to selection bias, the *Literary Digest* survey also had a great deal of **nonresponse bias**. This occurs when those responding to a survey or poll differ systematically from the nonrespondents. In the survey, a larger percentage of educated people mailed back the questionnaires. During that time period, the more educated tended to come from affluent families that again favored the Republican candidate.

Nonresponse bias refers to a systematic difference in preferences between respondents and nonrespondents to a survey or a poll.

The most effective way to deal with nonresponse bias is to reduce nonresponse rates. Paying attention to survey design, wording, and ordering of the questions can increase the response rate. Sometimes, rather than sending out a very large number of surveys, it may be preferable to use a smaller representative sample for which the response rate is likely to be high.

It turns out that someone did accurately predict the 1936 presidential election. From a sample of 50,000 with a response rate of 10% (5,000 respondents), a young pollster named George Gallup predicted that FDR would win 56% of the vote to Landon's 44%. Despite using a far smaller sample with a lower response rate, it was far more *representative* of the true voting population. Gallup later founded the Gallup Organization, one of the leading polling companies of all time.

Trump's Stunning Victory in 2016

The results of the U.S. presidential election in 2016 came as a surprise to nearly everyone who had been following the national and state election polling, which consistently projected Hillary Clinton as defeating Donald Trump (www.pewresearch.org, November 9, 2016). It appears that problems with selection bias and nonresponse bias persist today. Many pollsters and strategists believe that rural white voters, who were

a key demographic for Trump on Election Day, eluded polling altogether. It is also believed that the frustration and anti-institutional feelings that drove the campaign may also have aligned these same voters with an unwillingness to respond to surveys (www.politico.com, March 27, 2017).

Another theory that has gained some traction in explaining the polling missteps in the 2016 election was the presence of **social-desirability bias**. This bias occurs when voters provide incorrect answers to a survey or poll because they think that others will look unfavorably on their ultimate choices.

Social-desirability bias refers to a systematic difference between a group's "socially acceptable" responses to a survey or poll and this group's ultimate choice.

Due to Trump's inflammatory comments, many voters did not want to be associated with him by their peers. This was perfectly exemplified by the fact that Trump consistently performed better in online polling. For example, in one aggregation of telephone polls, Clinton led Trump by nine percentage points; however, in a similar aggregation of online polls, Clinton's lead was only four percentage points (*The New York Times*, May 11, 2016). This seems to suggest that one way to battle social-desirability bias is to use online surveys. Despite their flaws, online surveys resemble an anonymous voting booth and remove the human factor of the pollsters.

LO **7.2**

Describe various sampling methods.

Sampling Methods

As mentioned earlier, a primary requisite for a "good" sample is that it be representative of the population you are trying to describe. The basic type of sample that can be used to draw statistically sound conclusions about a population is a **simple random sample**.

SIMPLE RANDOM SAMPLE

A simple random sample is a sample of n observations that has the same probability of being selected from the population as any other sample of n observations. Most statistical methods presume simple random samples.

While a simple random sample is the most commonly used sampling method, in some situations, other sampling methods have an advantage over simple random samples. Two alternative methods for forming a sample are stratified random sampling and cluster sampling.

Political pollsters often employ **stratified random sampling** in an attempt to ensure that each area of the country, each ethnic group, each religious group, and so forth, is appropriately represented in the sample. With stratified random sampling, the population is divided into groups (strata) based on one or more classification criteria. Simple random samples are then drawn from each stratum in sizes proportional to the relative size of each stratum in the population. These samples are then pooled.

STRATIFIED RANDOM SAMPLING

In stratified random sampling, the population is first divided up into mutually exclusive and collectively exhaustive groups, called *strata*. A stratified sample includes randomly selected observations from each stratum. The number of observations per stratum is proportional to the stratum's size in the population. The data for each stratum are eventually pooled.

Stratified random sampling has two advantages. First, it guarantees that the population subdivisions of interest are represented in the sample. Second, the estimates of parameters produced from stratified random sampling have greater precision than estimates obtained from simple random sampling.

Even stratified random sampling, however, can fall short with its predictive ability. One of the nagging mysteries of the 2008 Democratic presidential primaries was: why were the polls so wrong in New Hampshire? All nine major polling groups predicted that Barack Obama would beat Hillary Clinton in the New Hampshire primary, by an average of 8.3 percentage points. When the votes were counted, Clinton won by 2.6%. Several factors contributed to the wrong prediction by the polling industry. First, pollsters overestimated the turnout of young voters, who overwhelmingly favored Obama in exit polls but did not surge to vote as they had in the Iowa caucus. Second, Clinton's campaign made a decision to target female Democrats, especially single women. This focus did not pay off in Iowa, but it did in New Hampshire. Finally, on the eve of the primary, a woman in Portsmouth asked Clinton: "How do you do it?" Clinton's teary response was powerful and warm. Voters, who rarely saw Clinton in such an emotional moment, found her response humanizing and appealing. Most polls had stopped phoning voters over the weekend, too soon to catch the likely voter shift.

Cluster sampling is another method for forming a representative sample. A cluster sample is formed by dividing the population into groups (clusters), such as geographic areas, and then selecting a sample of the groups for the analysis. The technique works best when most of the variation in the population is within the groups and not between the groups. In such instances, a cluster is a miniversion of the population.

CLUSTER SAMPLING

In cluster sampling, the population is first divided up into mutually exclusive and collectively exhaustive groups, called *clusters*. A cluster sample includes observations from randomly selected clusters.

In general, cluster sampling is cheaper as compared to other sampling methods. However, for a given sample size, it provides less precision than either simple random sampling or stratified sampling. Cluster sampling is useful in applications where the population is concentrated in natural clusters such as city blocks, schools, and other geographic areas. It is especially attractive when constructing a complete list of the population members is difficult and/or costly. For example, since it may not be possible to create a full list of customers who go to Walmart, we can form a sample that includes customers only from selected stores.

STRATIFIED VERSUS CLUSTER SAMPLING

In stratified sampling, the sample consists of observations from each group, whereas in cluster sampling, the sample consists of observations from the selected groups. Stratified sampling is preferred when the objective is to increase precision, and cluster sampling is preferred when the objective is to reduce costs.

In practice, it is extremely difficult to obtain a truly random sample that is representative of the underlying population. As researchers, we need to be aware of the population from which the sample was selected and then limit our conclusions to that population. For the remainder of the text, we assume that the sample data are void of "human error." That is, we have sampled from the correct population (no selection bias); we have no nonresponse or social-desirability biases; and we have collected, analyzed, and reported the data properly.

Using Excel to Generate a Simple Random Sample

Excel provides functions that we can use to draw simple random samples. Example 7.1 illustrates the use of one of these functions.

EXAMPLE 7.1

There has been an increase in students working their way through college to offset rising tuition costs (*US News*, January 11, 2017). Work helps them to lower student loans and learn crucial time management skills. A dean at the Orfalea College of Business (OCOB) wants to analyze performance of her students who work while they are enrolled. For the analysis, use Excel to generate a random sample of 100 students drawn from 2,750 OCOB students.

SOLUTION: Since each student has a unique student identification number, we start by creating an ordered list using 1 and 2,750 as the smallest and largest student identification numbers, respectively. We then generate 100 random integers (numbers) between these values and use them to identify students based on their order on the list.

We use Excel's **RANDBETWEEN** function to generate random integers within some interval. In general, we enter "=RANDBETWEEN(lower, upper)" where lower and upper refer to the smallest and largest integers in the interval, respectively. In this example, we enter "=RANDBETWEEN(1, 2750)." Suppose Excel returns 983. The student whose order on the list is 983 is then selected for the sample. To generate the remaining 99 numbers, we select the cell with the value 983, drag it down 99 additional cells, and from the menu, choose **Home > Fill > Down**. (Note: Since Excel may generate the same number more than once, we may have to generate more than 100 numbers in order to obtain 100 unique numbers.)

EXERCISES 7.1

1. In 2010, Apple introduced the iPad, a tablet-style computer that its former CEO Steve Jobs called a "a truly magical and revolutionary product" (*CNN,* January 28, 2010). Suppose you are put in charge of determining the age profile of people who purchased the iPad in the United States. Explain in detail the following sampling strategies that you could use to select a representative sample.
   a. Simple random sampling
   b. Stratified random sampling
   c. Cluster sampling
2. A marketing firm opens a small booth at a local mall over the weekend, where shoppers are asked how much money they spent at the food court. The objective is to determine the average monthly expenditure of shoppers at the food court. Has the marketing firm committed any sampling bias? Discuss.
3. Natalie Min is an undergraduate in the Haas School of Business at Berkeley. She wishes to pursue an MBA from Berkeley and wants to know the profile of other students who are likely to apply to the Berkeley MBA program. In particular, she wants to know the GPA of students with whom she might be competing. She randomly surveys 40 students from her accounting class for the analysis. Discuss in detail whether or not Natalie's analysis is based on a representative sample.
4. Vons, a large supermarket in Grover Beach, California, is considering extending its store hours from 7:00 am to midnight, seven days a week, to 6:00 am to midnight. Discuss the sampling bias in the following sampling strategies:
   a. Mail a prepaid envelope to randomly selected residents in the Grover Beach area, asking for their preference for the store hours.
   b. Ask the customers who frequent the store in the morning if they would prefer an earlier opening time.
   c. Place an ad in the local newspaper, requesting people to submit their preference for store hours on the store's website.
5. In the previous question regarding Vons' store hours, explain how you can obtain a representative sample based on the following sampling strategies:
   a. Simple random sampling.
   b. Stratified random sampling.
   c. Cluster sampling.

7.2 THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

As mentioned earlier, we are generally interested in the characteristics of a population. For instance, a student is interested in the average starting salary (population mean) of business graduates. Similarly, a banker is interested in the default probability (population proportion) of mortgage holders. Recall that the population mean and the population proportion are parameters that describe quantitative and qualitative data, respectively. Since it is cumbersome, if not impossible, to analyze the entire population, we generally make inferences about the characteristics of the population on the basis of a random sample drawn from the population.

It is important to note that there is only one population, but many possible samples of a given size can be drawn from the population. Therefore, a population parameter is a constant, even though its value may be unknown. On the other hand, a statistic, such as the sample mean or the sample proportion, is a random variable whose value depends on the particular sample that is randomly drawn from the population.

A parameter is a constant, although its value may be unknown. A statistic is a random variable whose value depends on the chosen random sample.

Consider the starting salary of business graduates as the variable of interest. If you decide to make inferences about the population mean salary on the basis of a random draw of 38 recent business graduates, then the sample mean $\overline{X}$ is the relevant statistic. Note that the value of $\overline{X}$ will change if you choose a different random sample of 38 business graduates. In other words, $\overline{X}$ is a random variable whose value depends on the chosen random sample. The sample mean is commonly referred to as the **estimator**, or the **point estimator**, of the population mean.

In the starting salary example, the sample mean $\overline{X}$ is the estimator of the mean starting salary of business graduates. If the average derived from a specific sample is \$54,000, then $\bar{x} = 54,000$ is the **estimate** of the population mean. Similarly, if the variable of interest is the default probability of mortgage holders, then the sample proportion of defaults, denoted by $\overline{P}$, from a random sample of 80 mortgage holders is the estimator of the population proportion. If 10 out of 80 mortgage holders in a given sample default, then $\bar{p} = 10/80 = 0.125$ is the estimate of the population proportion.

ESTIMATOR AND ESTIMATE

When a statistic is used to estimate a parameter, it is referred to as an estimator. A particular value of the estimator is called an estimate.

In this section, we will focus on the probability distribution of the sample mean $\overline{X}$, which is also referred to as the **sampling distribution** of $\overline{X}$. Since $\overline{X}$ is a random variable, its sampling distribution is simply the probability distribution derived from all possible samples of a given size from the population. Consider, for example, a mean derived from a sample of n observations. Another mean can similarly be derived from a different sample of n observations. If we repeat this process a very large number of times, then the frequency distribution of the sample means can be thought of as its sampling distribution. In particular, we will discuss the expected value and the standard deviation of the sample mean. We will also study the conditions under which the sampling distribution of the sample mean is normally distributed.

LO **7.3**

Describe the sampling distribution of the sample mean.

The Expected Value and the Standard Error of the Sample Mean

Let the random variable X represent a certain characteristic of a population under study, with an expected value, $E(X) = \mu$, and a variance, $Var(X) = \sigma^2$. For example, X could represent the salary of business graduates or the return on an investment. We can think of μ and σ^2 as the mean and the variance of an individual observation drawn randomly from the population of interest, or simply as the population mean and the population variance. Let the sample mean $\overline{X}$ be based on a random sample of n observations from this population. It is easy to derive the expected value and the variance of $\overline{X}$; see Appendix 7.1 for the derivations.

The expected value of $\overline{X}$ is the same as the expected value of the individual observation—that is, $E(\overline{X}) = E(X) = \mu$. In other words, if we were to sample repeatedly from a given population, the average value of the sample means will equal the average value of all individual observations in the population, or simply, the population mean. This is an important property of an estimator, called unbiasedness, that holds irrespective of whether the sample mean is based on a small or a large sample. An estimator is **unbiased** if its expected value equals the population parameter. Other desirable properties of an estimator are described in Appendix 7.2.

THE EXPECTED VALUE OF THE SAMPLE MEAN

The expected value of the sample mean $\overline{X}$ equals the population mean, or $E(\overline{X}) = \mu$. In other words, the sample mean is an unbiased estimator of the population mean.

It is important to note that we estimate the population mean on the basis of just one sample. The above result shows that we are not systematically underestimating or overestimating the population parameter.

The variance of $\overline{X}$ is equal to $Var(\overline{X}) = \frac{\sigma^2}{n}$. In other words, if we were to sample repeatedly from a given population, the variance of the sample mean will equal the variance of all individual observations in the population, divided by the sample size, n. Note that $Var(\overline{X})$ is smaller than the variance of X, which is equal to $Var(X) = \sigma^2$. This is an intuitive result, suggesting that the variability between sample means is less than the variability between observations. Since each sample is likely to contain both high and low observations, the highs and lows cancel one another, making the variance of $\overline{X}$ smaller than the variance of X. As usual, the standard deviation of $\overline{X}$ is calculated as the positive square root of the variance. However, in order to distinguish the variability between samples from the variability between individual observations, we refer to the standard deviation of $\overline{X}$ as the **standard error** of the sample mean, computed as $se(\overline{X}) = \frac{\sigma}{\sqrt{n}}$.

THE STANDARD ERROR OF THE SAMPLE MEAN

The standard deviation of the sample mean $\overline{X}$ is referred to as the standard error of the sample mean. It equals the population standard deviation divided by the square root of the sample size; that is, $se(\overline{X}) = \frac{\sigma}{\sqrt{n}}$.

In Chapter 8, we will discuss that the exact standard error of an estimator is often not known and, therefore, must be estimated from the given sample data. For convenience, we use "*se*" to denote both the exact and the estimated standard errors of an estimator.

EXAMPLE 7.2

The chefs at a local pizza chain in Cambria, California, strive to maintain the suggested size of their 16-inch pizzas. Despite their best efforts, they are unable to make every pizza exactly 16 inches in diameter. The manager has determined that the size of the pizzas is normally distributed with a mean of 16 inches and a standard deviation of 0.8 inch.

a. What are the expected value and the standard error of the sample mean derived from a random sample of 2 pizzas?

b. What are the expected value and the standard error of the sample mean derived from a random sample of 4 pizzas?

c. Compare the expected value and the standard error of the sample mean with those of an individual pizza.

SOLUTION: We know that the population mean $\mu = 16$ and the population standard deviation $\sigma = 0.8$. We use $E(\overline{X}) = \mu$ and $se(\overline{X}) = \frac{\sigma}{\sqrt{n}}$ to calculate the following results.

a. With the sample size $n = 2$, $E(\overline{X}) = 16$ and $se(\overline{X}) = \frac{0.8}{\sqrt{2}} = 0.57$.

b. With the sample size $n = 4$, $E(\overline{X}) = 16$ and $se(\overline{X}) = \frac{0.8}{\sqrt{4}} = 0.40$.

c. The expected value of the sample mean for both sample sizes is identical to the expected value of the individual pizza. However, the standard error of the sample mean with $n = 4$ is lower than the one with $n = 2$. For both sample sizes, the standard error of the sample mean is lower than the standard deviation of the individual pizza. This result confirms that averaging reduces variability.

Sampling from a Normal Population

An important feature of the sampling distribution of the sample mean $\overline{X}$ is that, irrespective of the sample size n, $\overline{X}$ is normally distributed if the population X from which the sample is drawn is normal. In other words, if X is normal with expected value μ and standard deviation σ, then $\overline{X}$ is also normal with expected value μ and standard error $\sigma/\sqrt{n}$.

SAMPLING FROM A NORMAL POPULATION

For any sample size n, the sampling distribution of $\overline{X}$ is normal if the population X from which the sample is drawn is normally distributed.

If $\overline{X}$ is normally distributed, then it can be transformed into a standard normal random variable as

$$Z = \frac{\overline{X} - E(\overline{X})}{se(\overline{X})} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

Therefore, any value $\bar{x}$ has a corresponding value z given by $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$.

EXAMPLE 7.3

Use the information in Example 7.2 to answer the following questions:

a. What is the probability that a randomly selected pizza is less than 15.5 inches?

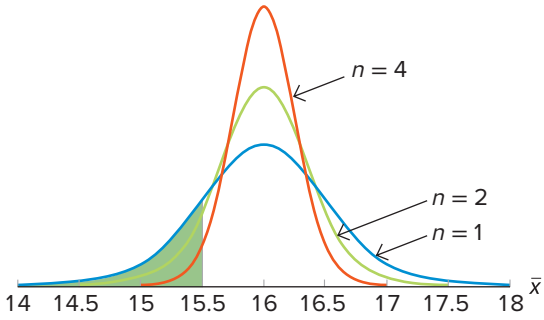
b. What is the probability that 2 randomly selected pizzas average less than 15.5 inches?

c. What is the probability that 4 randomly selected pizzas average less than 15.5 inches?

d. Comment on the computed probabilities.

SOLUTION: Since the population is normally distributed, the sampling distribution of the sample mean is also normal. Figure 7.1 depicts the shapes of the three distributions based on the population mean $\mu = 16$ and the population standard deviation $\sigma = 0.8$.

FIGURE 7.1
Normal distribution of the sample mean



Note that when the sample size $n = 1$, the sample mean $\bar{x}$ is the same as the individual observation x.

a. We use the standard transformation to derive $P(X < 15.5) = P\left(Z < \frac{15.5 - 16}{0.8}\right) = P(Z < -0.63) = 0.2643$. There is a 26.43% chance that an individual pizza is less than 15.5 inches.

b. Here we use the standard transformation to derive $P(\overline{X} < 15.5) = P\left(Z < \frac{15.5 - 16}{0.8/\sqrt{2}}\right) = P(Z < -0.88) = 0.1894$. In a random sample of 2 pizzas, there is an 18.94% chance that the average size is less than 15.5 inches.

c. Again we find $P(\overline{X} < 15.5)$, but now $n = 4$. Therefore, $P(\overline{X} < 15.5) = P\left(Z < \frac{15.5 - 16}{0.8/\sqrt{4}}\right) = P(Z < -1.25) = 0.1056$. In a random sample of 4 pizzas, there is a 10.56% chance that the average size is less than 15.5 inches.

d. The probability that the average size is under 15.5 inches, for 4 randomly selected pizzas, is less than half of that for an individual pizza. This is due to the fact that while X and $\overline{X}$ have the same expected value of 16, the variance of $\overline{X}$ is less than that of X.

LO **7.4**

Explain the importance of the central limit theorem.

The Central Limit Theorem

For making statistical inferences, it is essential that the sampling distribution of $\overline{X}$ is normally distributed. So far we have only considered the case where $\overline{X}$ is normally distributed because the population X from which the sample is drawn is normal. What if the underlying population is not normal? Here we present the **central limit theorem** (**CLT**), which perhaps is the most remarkable result of probability theory. The CLT states that the sum or the average of a large number of independent observations from the same underlying distribution has an approximate normal distribution. The approximation steadily improves as the number of observations increases. In other words, irrespective of whether or not the population X is normal, the sample mean $\overline{X}$ computed from a random sample of size n will be approximately normally distributed as long as n is sufficiently large.

THE CENTRAL LIMIT THEOREM FOR THE SAMPLE MEAN

For any population X with expected value μ and standard deviation σ, the sampling distribution of $\overline{X}$ will be approximately normal if the sample size n is sufficiently large. As a general guideline, the normal distribution approximation is justified when $n \geq 30$.

As before, if $\overline{X}$ is approximately normally distributed, then any value $\bar{x}$ can be transformed to its corresponding value z given by $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$.

Figure 7.1, discussed in Example 7.3, is not representative of the CLT principle because, for a normal population, the sampling distribution of $\overline{X}$ is normal irrespective of the sample size. Figures 7.2 and 7.3, however, illustrate the CLT by using random samples of various sizes drawn from nonnormal populations. The relative frequency polygon of $\overline{X}$, which essentially represents its distribution, is generated from repeated draws (computer simulations) from the continuous uniform distribution (Figure 7.2) and the exponential distribution (Figure 7.3). Both of these nonnormal distributions were discussed in Chapter 6.

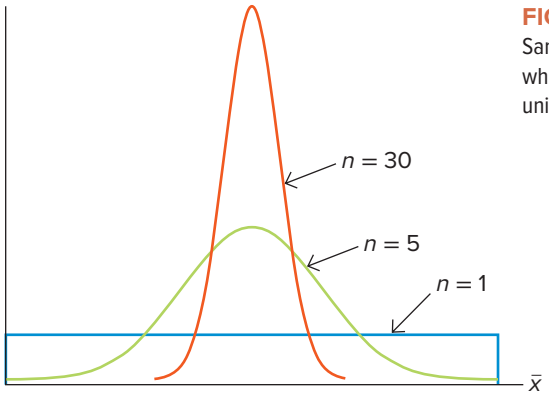


FIGURE 7.2
Sampling distribution of $\overline{X}$ when the population has a uniform distribution

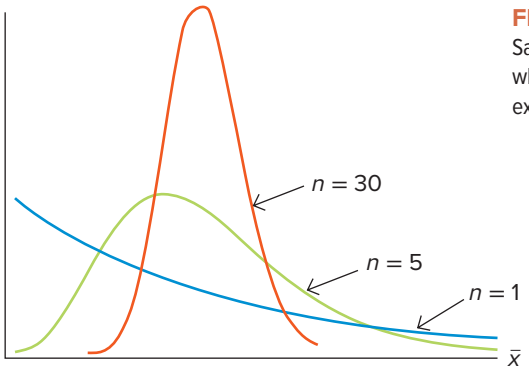


FIGURE 7.3
Sampling distribution of $\overline{X}$ when the population has an exponential distribution

Note that when the sample size $n = 1$, the sampling distribution of $\overline{X}$ is the same as the sampling distribution of X with the familiar uniform and exponential shapes. With $n = 5$, the sampling distribution of $\overline{X}$ is already approximately normal when the population has the uniform distribution. With $n = 30$, the shape of the sampling distribution of $\overline{X}$ is approximately normal when the population has the exponential distribution. The CLT can similarly be illustrated with other distributions of the population. How large a

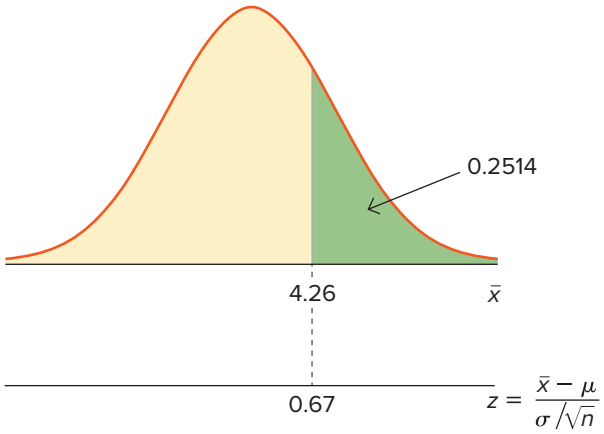
sample is necessary for normal convergence depends on the magnitude of the departure of the population from normality. As mentioned earlier, practitioners often use the normal distribution approximation when $n \geq 30$.

EXAMPLE 7.4

Consider the information presented in the introductory case of this chapter. Recall that Anne wants to determine if the marketing campaign has had a lingering effect on the amount of money customers spend on iced coffee. Before the campaign, customers spent an average of \$4.18 on iced coffee with a standard deviation of \$0.84. Anne reports that the average amount, based on 50 customers sampled after the campaign, is \$4.26. If Starbucks chose not to pursue the marketing campaign, how likely is it that customers will spend an average of \$4.26 or more on iced coffee?

SOLUTION: If Starbucks did not pursue the marketing campaign, spending on iced coffee would still have mean $\mu = 4.18$ and standard deviation $\sigma = 0.84$. Anne needs to calculate the probability that the sample mean is at least 4.26, or, $P(\overline{X} \geq 4.26)$. The population from which the sample is drawn is not known to be normally distributed. However, since $n \geq 30$, from the central limit theorem we know that $\overline{X}$ is approximately normally distributed. Therefore, as shown in Figure 7.4, $P(\overline{X} \geq 4.26) = P\left(Z \geq \frac{4.26 - 4.18}{0.84/\sqrt{50}}\right) = P(Z \geq 0.67) = 1 - 0.7486 = 0.2514$. It is quite plausible (probability = 0.2514) that in a sample of 50 customers, the sample mean is \$4.26 or more even if Starbucks did not pursue the marketing campaign.

FIGURE 7.4 Finding $P(\overline{X} \geq 4.26)$



EXERCISES 7.2

Mechanics

6. A random sample is drawn from a normally distributed population with mean $\mu = 12$ and standard deviation $\sigma = 1.5$.
   a. Comment on the sampling distribution of the sample mean with $n = 20$ and $n = 40$.
   b. Can you use the standard normal distribution to calculate the probability that the sample mean is less than 12.5 for both sample sizes?
   c. Report the probability if you answered yes to the previous question for either sample size.

7. A random sample is drawn from a population with mean $\mu = 66$ and standard deviation $\sigma = 5.5$.
   a. Comment on the sampling distribution of the sample mean with $n = 16$ and $n = 36$.
   b. Can you use the standard normal distribution to calculate the probability that the sample mean falls between 66 and 68 for both sample sizes?
   c. Report the probability if you answered yes to the previous question for either sample size.
8. A random sample of size $n = 100$ is taken from a population with mean $\mu = 80$ and standard deviation $\sigma = 14$.
   a. Calculate the expected value and the standard error for the sampling distribution of the sample mean.
   b. What is the probability that the sample mean falls between 77 and 85?
   c. What is the probability that the sample mean is greater than 84?
9. A random sample of size $n = 50$ is taken from a population with mean $\mu = -9.5$ and standard deviation $\sigma = 2$.
   a. Calculate the expected value and the standard error for the sampling distribution of the sample mean.
   b. What is the probability that the sample mean is less than −10?
   c. What is the probability that the sample mean falls between −10 and −9?

Applications

10. According to a survey, high school girls average 100 text messages daily (*The Boston Globe*, April 21, 2010). Assume the population standard deviation is 20 text messages. Suppose a random sample of 50 high school girls is taken.
    a. What is the probability that the sample mean is more than 105?
    b. What is the probability that the sample mean is less than 95?
    c. What is the probability that the sample mean is between 95 and 105?
11. Beer bottles are filled so that they contain an average of 330 ml of beer in each bottle. Suppose that the amount of beer in a bottle is normally distributed with a standard deviation of 4 ml.
    a. What is the probability that a randomly selected bottle will have less than 325 ml of beer?
    b. What is the probability that a randomly selected 6-pack of beer will have a mean amount less than 325 ml?
    c. What is the probability that a randomly selected 12-pack of beer will have a mean amount less than 325 ml?
    d. Comment on the sample size and the corresponding probabilities.
12. Despite its nutritional value, seafood is only a tiny part of the American diet, with the average American eating just 16 pounds of seafood per year. Janice and Nina both work in the seafood industry and they decide to create their own random samples and document the average seafood diet in their sample. Let the standard deviation of the American seafood diet be 7 pounds.
    a. Janice samples 42 Americans and finds an average seafood consumption of 18 pounds. How likely is it to get an average of 18 pounds or more if she had a representative sample?
    b. Nina samples 90 Americans and finds an average seafood consumption of 17.5 pounds. How likely is it to get an average of 17.5 pounds or more if she had a representative sample?
    c. Which of the two women is likely to have used a more representative sample? Explain.
13. The weight of people in a small town in Missouri is known to be normally distributed with a mean of 180 pounds and a standard deviation of 28 pounds. On a raft that takes people across the river, a sign states, "Maximum capacity 3,200 pounds or 16 persons." What is the probability that a random sample of 16 persons will exceed the weight limit of 3,200 pounds?
14. The weight of turkeys is known to be normally distributed with a mean of 22 pounds and a standard deviation of 5 pounds.
    a. Discuss the sampling distribution of the sample mean based on a random draw of 16 turkeys.
    b. Find the probability that the mean weight of 16 randomly selected turkeys is more than 25 pounds.
    c. Find the probability that the mean weight of 16 randomly selected turkeys is between 18 and 24 pounds.
15. A small hair salon in Denver, Colorado, averages about 30 customers on weekdays with a standard deviation of 6. It is safe to assume that the underlying distribution is normal. In an attempt to increase the number of weekday customers, the manager offers a $2 discount on 5 consecutive weekdays. She reports that her strategy has worked since the sample mean of customers during this 5 weekday period jumps to 35.
    a. How unusual would it be to get a sample average of 35 or more customers if the manager had not offered the discount?
    b. Do you feel confident that the manager's discount strategy has worked? Explain.
16. The typical college student graduates with $27,200 in debt (*The Boston Globe*, May 27, 2012). Let debt among recent college graduates be normally distributed with a standard deviation of $7,000.
    a. What is the probability that the average debt of four recent college graduates is more than $25,000?
    b. What is the probability that the average debt of four recent college graduates is more than $30,000?
17. Forty families gathered for a fund-raising event. Suppose the individual contribution for each family is normally distributed with a mean and a standard deviation of $115 and $35,

respectively. The organizers would call this event a success if the total contributions exceed $5,000. What is the probability that this fund-raising event is a success?

18. A doctor is getting sued for malpractice by four of her former patients. It is believed that the amount that each patient will sue her for is normally distributed with a mean of $800,000 and a standard deviation of $250,000.
    a. What is the probability that a given patient sues the doctor for more than $1,000,000?
    b. If the four patients sue the doctor independently, what is the probability that the total amount they sue for is over $4,000,000?
19. Suppose that the miles-per-gallon (mpg) rating of passenger cars is normally distributed with a mean and a standard deviation of 33.8 and 3.5 mpg, respectively.
    a. What is the probability that a randomly selected passenger car gets more than 35 mpg?
    b. What is the probability that the average mpg of four randomly selected passenger cars is more than 35 mpg?
    c. If four passenger cars are randomly selected, what is the probability that all of the passenger cars get more than 35 mpg?
20. Suppose that IQ scores are normally distributed with a mean of 100 and a standard deviation of 16.
    a. What is the probability that a randomly selected person will have an IQ score of less than 90?
    b. What is the probability that the average IQ score of four randomly selected people is less than 90?
    c. If four people are randomly selected, what is the probability that all of them have an IQ score of less than 90?

LO **7.5**

Describe the sampling distribution of the sample proportion.

7.3 THE SAMPLING DISTRIBUTION OF THE SAMPLE PROPORTION

Our discussion thus far has focused on the population mean, but many business, socioeconomic, and political matters are concerned with the population proportion. For instance, a banker is interested in the default probability of mortgage holders; a superintendent may note the proportion of students suffering from the flu when determining whether to keep school open; an incumbent seeking reelection cares about the proportion of constituents that will ultimately cast a vote for him/her. In all of these examples, the parameter of interest is the population proportion p. As in the case of the population mean, we almost always make inferences about the population proportion on the basis of sample data. Here, the relevant statistic (estimator) is the sample proportion, $\overline{P}$; a particular value (estimate) is denoted by $\bar{p}$. Since $\overline{P}$ is a random variable, we need to discuss its sampling distribution.

The Expected Value and the Standard Error of the Sample Proportion

We first introduced the population proportion p in Chapter 5, when we discussed the binomial distribution. It turns out that the sampling distribution of $\overline{P}$ is closely related to the binomial distribution. Recall that the binomial distribution describes the number of successes X in n trials of a Bernoulli process where p is the probability of success; thus, $\overline{P} = \frac{X}{n}$ is the number of successes X divided by the sample size n. We can derive the expected value and the variance of the sampling distribution of $\overline{P}$ as $E(\overline{P}) = p$ and $Var(\overline{P}) = \frac{p(1-p)}{n}$, respectively. (See Appendix 7.1 for the derivations.) Note that since $E(\overline{P}) = p$, it implies that $\overline{P}$ is an unbiased estimator of p.

THE EXPECTED VALUE OF THE SAMPLE PROPORTION

The expected value of the sample proportion $\overline{P}$ is equal to the population proportion, or, $E(\overline{P}) = p$. In other words, the sample proportion is an unbiased estimator of the population proportion.

Analogous to our discussion in the last section, we refer to the standard deviation of the sample proportion as the standard error of the sample proportion; that is, $se(\overline{P}) = \sqrt{\frac{p(1-p)}{n}}$.

THE STANDARD ERROR OF THE SAMPLE PROPORTION

The standard deviation of the sample proportion $\overline{P}$ is referred to as the standard error of the sample proportion. It equals $se(\overline{P}) = \sqrt{\frac{p(1-p)}{n}}$.

EXAMPLE 7.5

Many people apply for jobs to serve as paramedics or firefighters, yet they cannot complete basic physical fitness standards. A study found that 77% of all candidates for paramedic and firefighter positions were overweight or obese (*Obesity*, March 19, 2009).

a. What are the expected value and the standard error of the sample proportion derived from a random sample of 100 candidates for paramedic or firefighter positions?

b. What are the expected value and the standard error of the sample proportion derived from a random sample of 200 candidates for paramedic or firefighter positions?

c. Comment on the value of the standard error as the sample size gets larger.

SOLUTION: Given that $p = 0.77$, we can derive the expected value and the standard error of $\overline{P}$ as follows.

a. With $n = 100$, $E(\overline{P}) = 0.77$ and $se(\overline{P}) = \sqrt{\frac{0.77(1-0.77)}{100}} = 0.042$.

b. With $n = 200$, $E(\overline{P}) = 0.77$ and $se(\overline{P}) = \sqrt{\frac{0.77(1-0.77)}{200}} = 0.030$.

c. As in the case of the sample mean, while the expected value of the sample proportion is unaffected by the sample size, the standard error of the sample proportion is reduced as the sample size increases.

In this text, we make statistical inferences about the population proportion only when the sampling distribution of $\overline{P}$ is approximately normal. From the discussion of the central limit theorem (CLT) in Section 7.2, we can conclude that $\overline{P}$ is approximately normally distributed when the sample size is sufficiently large.

THE CENTRAL LIMIT THEOREM FOR THE SAMPLE PROPORTION

For any population proportion p, the sampling distribution of $\overline{P}$ is approximately normal if the sample size n is sufficiently large. As a general guideline, the normal distribution approximation is justified when $np \geq 5$ and $n(1-p) \geq 5$.

If $\overline{P}$ is normally distributed, we can transform it into the standard normal random variable as

$$Z = \frac{\overline{P} - E(\overline{P})}{se(\overline{P})} = \frac{\overline{P} - p}{\sqrt{\frac{p(1-p)}{n}}}.$$

Therefore, any value $\overline{p}$ has a corresponding value z given by

$$z = \frac{\overline{p} - p}{\sqrt{\frac{p(1-p)}{n}}}.$$

According to the CLT, the sampling distribution of $\overline{P}$ approaches the normal distribution as the sample size increases. However, as the population proportion deviates from $p = 0.50$, we need a larger sample size for the approximation. We illustrate these results by generating the sampling distribution of $\overline{P}$ from repeated draws from a population with various values for the population proportion and sample sizes. As in the case of $\overline{X}$, we use the relative frequency polygon to represent the distribution of $\overline{P}$. The simulated sampling distribution of $\overline{P}$ is based on the population proportion $p = 0.10$ (Figure 7.5) and $p = 0.30$ (Figure 7.6).

FIGURE 7.5 Sampling distribution of $\overline{P}$ when the population proportion is $p = 0.10$

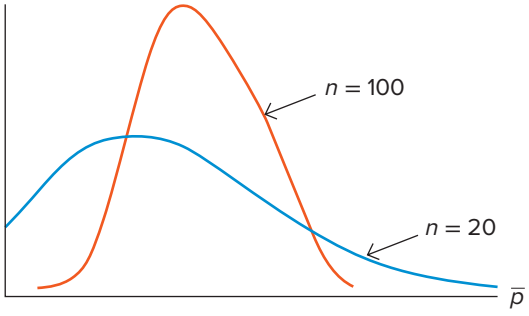
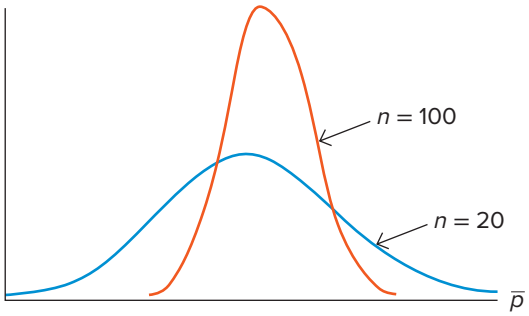


FIGURE 7.6 Sampling distribution of $\overline{P}$ when the population proportion is $p = 0.30$



When $p = 0.10$, the sampling distribution of $\overline{P}$ does not resemble the bell shape of the normal distribution with $n = 20$ since the approximation condition $np \geq 5$ and $n(1 - p) \geq 5$ is not satisfied. However, the graph becomes somewhat bell-shaped with $n = 100$. When $p = 0.30$, the shape of the sampling distribution of $\overline{P}$ is approximately normal since the approximation condition is satisfied with both sample sizes. In empirical work, it is common to work with large survey data, and, as a result, the normal distribution approximation is justified.

EXAMPLE 7.6

Consider the information presented in the introductory case of this chapter. Recall that Anne Jones wants to determine if the marketing campaign has had a lingering effect on the proportion of customers who are women and teenage girls. Prior to the campaign, 43% of the customers were women and 21% were teenage girls. Based on a random sample of 50 customers after the campaign, these proportions increase to 46% for women and 34% for teenage girls. Anne has the following questions.

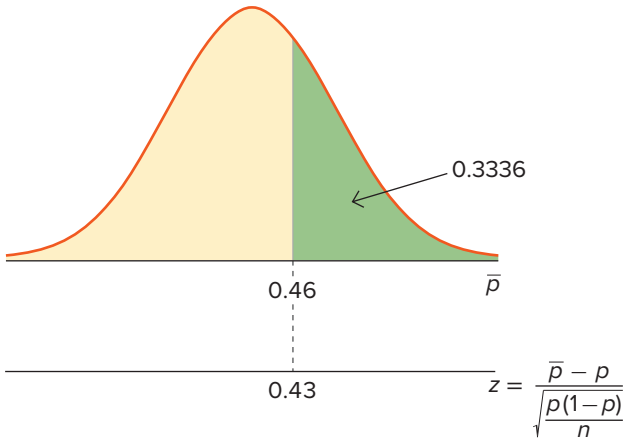
a. If Starbucks chose not to pursue the marketing campaign, how likely is it that 46% or more of iced-coffee customers are women?

b. If Starbucks chose not to pursue the marketing campaign, how likely is it that 34% or more of iced-coffee customers are teenage girls?

SOLUTION: If Starbucks had not pursued the marketing campaign, the proportion of customers would still be $p = 0.43$ for women and $p = 0.21$ for teenage girls. With $n = 50$, the normal approximation for the sample proportion is justified for both population proportions.

a. As shown in Figure 7.7, we find that $P(\overline{P} \geq 0.46) = P\left(Z \geq \dfrac{0.46 - 0.43}{\sqrt{\frac{0.43(1-0.43)}{50}}}\right) =$ $P(Z \geq 0.43) = 1 - 0.6664 = 0.3336$.

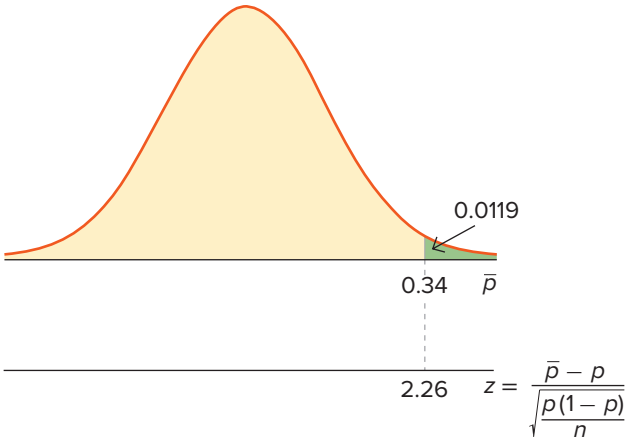
FIGURE 7.7 Finding $P(\overline{P} \geq 0.46)$



With a chance of 33.36%, it is quite plausible that the proportion of iced coffee purchased by women is at least 0.46 even if Starbucks did not pursue the marketing campaign.

b. As shown in Figure 7.8, we find $P(\overline{P} \geq 0.34) = P\left(Z \geq \dfrac{0.34 - 0.21}{\sqrt{\frac{0.21(1-0.21)}{50}}}\right) =$ $P(Z \geq 2.26) = 1 - 0.9881 = 0.0119$.

FIGURE 7.8 Finding $P(\overline{P} \geq 0.34)$



With only a 1.19% chance, it is unlikely that the proportion of iced coffee purchased by teenage girls is at least 0.34 if Starbucks did not pursue the marketing campaign.

Therefore, Anne can use this sample information to infer that the increase in the proportion of iced-coffee sales to women may not necessarily be due to the marketing campaign. However, the marketing campaign may have been successful in increasing the proportion of iced-coffee sales to teenage girls.

SYNOPSIS OF INTRODUCTORY CASE



©ARIRIYAWAT/Shutterstock

Iced coffee, traditionally a warm-weather and warm-region drink, has broadened its appeal over the years. According to a May 13, 2010, report in *Bloomberg Businessweek,* the number of servings of iced coffee surged from 300 million in 2001 to 1.2 billion in 2009. Large corporations have taken notice and have engaged in various strategies to capitalize on the growing trend. Starbucks, for instance, recently promoted a happy hour where customers paid half-price for a Frappuccino beverage between 3:00 pm and 5:00 pm for a 10-day period in May. One month after the marketing period ended, Anne Jones, the manager at a local Starbucks, surveys 50 of her customers. She reports an increase in spending in the sample, as well as an increase in the proportion of customers who are women and teenage girls. Anne wants to determine if the increase is due to chance or due to the marketing campaign. Based on an analysis with probabilities, Anne finds that higher spending in a sample of 50 customers is plausible even if Starbucks had not pursued the marketing campaign. Using a similar analysis with proportions, she infers that while the marketing campaign may not have necessarily increased the proportion of women customers, it seems to have attracted more teenage girls. The findings are consistent with current market research, which has shown that teenage girls have substantial income of their own to spend and often purchase items that are perceived as indulgences.

EXERCISES 7.3

Mechanics

21. Consider a population proportion $p = 0.68$.
    a. Calculate the expected value and the standard error of $\overline{P}$ with $n = 20$. Is it appropriate to use the normal distribution approximation for $\overline{P}$? Explain.
    b. Calculate the expected value and the standard error of $\overline{P}$ with $n = 50$. Is it appropriate to use the normal distribution approximation for $\overline{P}$? Explain.
22. Consider a population proportion $p = 0.12$.
    a. Discuss the sampling distribution of the sample proportion with $n = 20$ and $n = 50$.
    b. Can you use the normal approximation to calculate the probability that the sample proportion is between 0.10 and 0.12 for both sample sizes?
    c. Report the probabilities if you answered yes to the previous question.
23. A random sample of size $n = 200$ is taken from a population with a population proportion $p = 0.75$.
    a. Calculate the expected value and the standard error for the sampling distribution of the sample proportion.
    b. What is the probability that the sample proportion is between 0.70 and 0.80?
    c. What is the probability that the sample proportion is less than 0.70?

Applications

24. Europeans are increasingly upset at their leaders for making deep budget cuts to many social programs that are becoming too expensive to sustain. For example, the popularity of then-President Nicolas Sarkozy of France plummeted in 2010, giving him an approval rating of just 26% (*The Wall Street Journal,* July 2, 2010).
    a. What is the probability that fewer than 60 of 200 French people gave President Sarkozy a favorable rating?
    b. What is the probability that more than 150 of 200 French people gave President Sarkozy an *unfavorable* rating?
25. A study by Allstate Insurance Co. finds that 82% of teenagers have used cell phones while driving (*The Wall Street Journal,* May 5, 2010). Suppose a random sample of 100 teen drivers is taken.

a. Discuss the sampling distribution of the sample proportion.
b. What is the probability that the sample proportion is less than 0.80?
c. What is the probability that the sample proportion is within ±0.02 of the population proportion?

26. According to a FCC survey, one in six cell phone users has experienced "bill shock" from unexpectedly high cell phone bills (*Tech Daily Dose,* May 26, 2010).
a. Discuss the sampling distribution of the sample proportion based on a sample of 200 cell phone users. Is it appropriate to use the normal distribution approximation for the sample proportion?
b. What is the probability that more than 20% of cell phone users in the sample have experienced "bill shock"?

27. A car manufacturer is concerned about poor customer satisfaction at one of its dealerships. The management decides to evaluate the satisfaction surveys of its next 40 customers. The dealer will be fined if the number of customers who report favorably is between 22 and 26. The dealership will be dissolved if fewer than 22 customers report favorably. It is known that 70% of the dealer's customers report favorably on satisfaction surveys.
a. What is the probability that the dealer will be fined?
b. What is the probability that the dealership will be dissolved?

28. At a new exhibit in the Museum of Science, people are asked to choose between 50 or 100 random draws from a machine. The machine is known to have 60 green balls and 40 red balls. After each draw, the color of the ball is noted and the ball is put back for the next draw. You win a prize if more than 70% of the draws result in a green ball. Would you choose 50 or 100 draws for the game? Explain.

29. After years of rapid growth, illegal immigration into the United States has declined, perhaps owing to the recession and increased border enforcement by the United States (*Los Angeles Times,* September 1, 2010). While its share has declined, California still accounts for 23% of the nation's estimated 11.1 million undocumented immigrants.
a. In a sample of 50 illegal immigrants, what is the probability that more than 20% live in California?
b. In a sample of 200 illegal immigrants, what is the probability that more than 20% live in California?
c. Comment on the reason for the difference between the computed probabilities in parts a and b.

7.4 THE FINITE POPULATION CORRECTION FACTOR

LO 7.6
Use a finite population correction factor.

One of the implicit assumptions we have made thus far is that the sample size n is much smaller than the population size N. In many applications, the size of the population is not even known. For instance, we do not have information on the total number of pizzas made at a local pizza chain in Cambria (Examples 7.2 and 7.3) or the total number of customers at the local Starbucks store (Examples 7.4 and 7.6). If the sample size is large relative to the population size, then the standard errors of the estimators must be multiplied by a correction factor. This correction factor, called the **finite population correction factor**, accounts for the added precision gained by sampling a larger percentage of the population. As a general guideline, we use the finite population correction factor $\sqrt{\frac{N-n}{N-1}}$ when the sample constitutes at least 5% of the population—that is, $n \geq 0.05N$.

THE FINITE POPULATION CORRECTION FACTOR FOR THE SAMPLE MEAN

When the sample size is large relative to the population size ($n \geq 0.05N$), the finite population correction factor is used to reduce the sampling variation of the sample mean $\overline{X}$. The resulting standard error of $\overline{X}$ is $se(\overline{X}) = \frac{\sigma}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)$. The transformation for any value $\bar{x}$ to its corresponding z value is made accordingly.

Note that the correction factor is always less than one; when N is large relative to n, the correction factor is close to one and the difference between the formulas with and without the correction is negligible.

EXAMPLE 7.7

A large introductory marketing class has 340 students. The class is divided up into groups for the final course project. Connie is in a group of 34 students. These students had averaged 72 on the midterm, when the class as a whole had an average score of 73 with a standard deviation of 10.

a. Calculate the expected value and the standard error of the sample mean based on a random sample of 34 students.

b. How likely is it that a random sample of 34 students will average 72 or lower?

SOLUTION: The population mean is $\mu = 73$ and the population standard deviation is $\sigma = 10$.

a. The expected value of the sample mean is $E(\overline{X}) = \mu = 73$. We use the finite population correction factor because the sample size $n = 34$ is more than 5% of the population size $N = 340$. Therefore, the standard error of the sample mean is $se(\overline{X}) = \frac{\sigma}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right) = \frac{10}{\sqrt{34}}\left(\sqrt{\frac{340-34}{340-1}}\right) = 1.6294$. Note that without the correction factor, the standard error would be higher at $se(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{34}} = 1.7150$.

b. We find $P(\overline{X} \le 72) = P\left(Z \le \frac{72-73}{1.6294}\right) = P(Z \le -0.61) = 0.2709$. That is, the likelihood of 34 students averaging 72 or lower is 27.09%.

We use a similar finite population correction factor for the sample proportion when the sample size is at least 5% of the population size.

THE FINITE POPULATION CORRECTION FACTOR FOR THE SAMPLE PROPORTION

When the sample size is large relative to the population size ($n \ge 0.05N$), the finite population correction factor is used to reduce the sampling variation of the sample proportion $\overline{P}$. The resulting standard error of $\overline{P}$ is $se(\overline{P}) = \sqrt{\frac{P(1-p)}{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)$. The transformation for any value $\overline{p}$ to its corresponding z value is made accordingly.

EXAMPLE 7.8

The home ownership rate during 2009 declined to approximately 67%, becoming comparable to the rate in early 2000 (*U.S. Census Bureau News*, February 2, 2010). A random sample of 80 households is taken from a small island community with 1,000 households. The home ownership rate on the island is equivalent to the national home ownership rate of 67%.

a. Calculate the expected value and the standard error for the sampling distribution of the sample proportion. Is it necessary to apply the finite population correction factor? Explain.

b. What is the probability that the sample proportion is within 0.02 of the population proportion?

SOLUTION:

a. We must apply the finite population correction factor because the sample size $n = 80$ is at least 5% of the population size $N = 1{,}000$. Therefore, $E(\overline{P}) = p = 0.67$ and

$$se(\overline{P}) = \sqrt{\frac{p(1-p)}{n}}\left(\sqrt{\frac{N-n}{N-1}}\right) = \sqrt{\frac{0.67(1-0.67)}{80}}\left(\sqrt{\frac{1{,}000-80}{1{,}000-1}}\right) = 0.0505.$$

b. The probability that the sample proportion is within 0.02 of the population proportion is $P(0.65 \le \overline{P} \le 0.69)$. We find that $P(0.65 \le \overline{P} \le 0.69) = P\left(\frac{0.65-0.67}{0.0505} \le Z \le \frac{0.69-0.67}{0.0505}\right) = P(-0.40 \le Z \le 0.40) = 0.6554 - 0.3446 = 0.3108$. The likelihood that the home ownership rate is within 0.02 of the population proportion is 31.08%.

EXERCISES 7.4

Mechanics

30. A random sample of size $n = 100$ is taken from a population of size $N = 2{,}500$ with mean $\mu = -45$ and variance $\sigma^2 = 81$.
    a. Is it necessary to apply the finite population correction factor? Explain. Calculate the expected value and the standard error of the sample mean.
    b. What is the probability that the sample mean is between -47 and -43?
    c. What is the probability that the sample mean is greater than -44?

31. A random sample of size $n = 70$ is taken from a finite population of size $N = 500$ with mean $\mu = 220$ and variance $\sigma^2 = 324$.
    a. Is it necessary to apply the finite population correction factor? Explain. Calculate the expected value and the standard error of the sample mean.
    b. What is the probability that the sample mean is less than 210?
    c. What is the probability that the sample mean lies between 215 and 230?

32. A random sample of size $n = 100$ is taken from a population of size $N = 3{,}000$ with a population proportion of $p = 0.34$.
    a. Is it necessary to apply the finite population correction factor? Explain. Calculate the expected value and the standard error of the sample proportion.
    b. What is the probability that the sample proportion is greater than 0.37?

33. A random sample of size $n = 80$ is taken from a population of size $N = 600$ with a population proportion $p = 0.46$.
    a. Is it necessary to apply the finite population correction factor? Explain. Calculate the expected value and the standard error of the sample proportion.
    b. What is the probability that the sample mean is less than 0.40?

Applications

34. A study finds that companies are setting aside a large chunk of their IT spending for green technology projects (*BusinessWeek,* March 5, 2009). Two out of three of the large companies surveyed by Deloitte said they have at least 5% of their IT budget earmarked for green IT projects. Suppose that the survey was based on 1,000 large companies. What is the probability that more than 75 of 120 large companies will have at least 5% of their IT expenditure earmarked for green IT projects?

35. The issues surrounding the levels and structure of executive compensation have gained added prominence in the wake of the financial crisis that erupted in the fall of 2008. Based on the 2006 compensation data obtained from the Securities and Exchange Commission (SEC) website, it was determined that the mean and the standard deviation of compensation for the 500 highest paid CEOs in publicly traded U.S. companies are $10.32 million and $9.78 million, respectively. An analyst randomly chooses 32 CEO compensations for 2006.
    a. Is it necessary to apply the finite population correction factor? Explain.
    b. Is the sampling distribution of the sample mean approximately normally distributed? Explain.
    c. Calculate the expected value and the standard error of the sample mean.
    d. What is the probability that the sample mean is more than $12 million?

36. Suppose in the previous question that the analyst had randomly chosen 12 CEO compensations for 2006.
    a. Is it necessary to apply the finite population correction factor? Explain.
    b. Is the sampling distribution of the sample mean approximately normally distributed? Explain.

c. Calculate the expected value and the standard error of the sample mean.
d. Can you use the normal approximation to calculate the probability that the sample mean is more than $12 million? Explain.

37. It is expected that only 60% in a graduating class of 250 will find employment in the first round of a job search. You have 20 friends who have recently graduated.
a. Discuss the sampling distribution of the sample proportion of your friends who will find employment in the first round of a job search.
b. What is the probability that less than 50% of your friends will find employment in the first round of a job search?

LO **7.7**

Construct and interpret control charts for quantitative and qualitative data.

7.5 STATISTICAL QUALITY CONTROL

Now more than ever, a successful firm must focus on the quality of the products and services it offers. Global competition, technological advances, and consumer expectations are all factors contributing to the quest for quality. In order to ensure the production of high-quality goods and services, a successful firm implements some form of quality control. In this section, we give a brief overview of the field of **statistical quality control**.

Statistical quality control involves statistical techniques used to develop and maintain a firm's ability to produce high-quality goods and services.

In general, two approaches are used for statistical quality control. A firm uses **acceptance sampling** if it produces a product (or offers a service) and at the completion of the production process, the firm then inspects a portion of the products. If a particular product does not conform to certain specifications, then it is either discarded or repaired. There are several problems with this approach to quality control. First, it is costly to discard or repair a product. Second, the detection of all defective products is not guaranteed. Defective products may be delivered to customers, thus damaging the firm's reputation.

A preferred approach to quality control is the **detection approach**. A firm using the detection approach inspects the production process and determines at which point the production process does not conform to specifications. The goal is to determine whether the production process should be continued or adjusted before a large number of defects are produced. In this section, we focus on the detection approach to quality control.

In general, no two products or services are identical. In any production process, variation in the quality of the end product is inevitable. Two types of variation occur. **Chance variation** is caused by a number of randomly occurring events that are part of the production process. This type of variation is not generally considered to be under the control of the individual worker or machine. For example, suppose a machine fills one-gallon jugs of milk. It is unlikely that the filling weight of each jug is exactly 128 ounces. Very slight differences in the production process lead to minor differences in the weights of one jug to the next. Chance variation is expected and is not a source of alarm in the production process so long as its magnitude is tolerable and the end product meets acceptable specifications.

The other source of variation is referred to as **assignable variation**. This type of variation in the production process is caused by specific events or factors that can usually

be identified and eliminated. Suppose in the milk example that the machine is "drifting" out of alignment. This causes the machine to overfill each jug—a costly expense for the firm. Similarly, it is bad for the firm in terms of its reputation if the machine begins to underfill each jug. The firm wants to identify and correct these types of variations in the production process.

Control Charts

Walter A. Shewhart, a researcher at Bell Telephone Laboratories during the 1920s, is often credited as being the first to apply statistics to improve the quality of output. He developed the **control chart**—a tool used to monitor the behavior of a production process.

THE CONTROL CHART

The most commonly used statistical tool in quality control is the control chart, a plot of calculated statistics of the production process over time. If the calculated statistics fall in an expected range, then the production process is in control. If the calculated statistics reveal an undesirable trend, then adjustment of the production process is likely necessary.

We can construct a number of different control charts where each differs by either the variable of interest and/or the type of data that are available. For quantitative data, examples of control charts include

- The $\bar{x}$ **chart**, which monitors the *central tendency* of a production process, and
- The ***R* chart** and the ***s* chart**, which monitor the *variability* of a production process.

For qualitative data, examples of control charts include

- The $\bar{p}$ **chart**, which monitors the *proportion* of defectives (or some other characteristic) in a production process, and
- The ***c* chart**, which monitors the *count* of defects per item, such as the number of blemishes on a sampled piece of furniture.

In general, all of these control charts (and others that we have not mentioned) have the following characteristics:

1. A control chart plots the sample estimates, such as $\bar{x}$ or $\bar{p}$. So as more and more samples are taken, the resulting control chart provides one type of safeguard when assessing if the production process is operating within predetermined guidelines.
2. All sample estimates are plotted with reference to a **centerline**. The centerline represents the variable's expected value when the production process is in control.
3. In addition to the centerline, all control charts include an **upper control limit (UCL)** and a **lower control limit (LCL)**. These limits indicate excessive deviation above or below the expected value of the variable of interest. A control chart is valid only if the sampling distribution of the relevant estimator is (approximately) normal. Under this assumption, the control limits are generally set at three standard deviations from the centerline. The area under the normal curve that corresponds to ± 3 standard deviations from the expected value is 0.9973. Thus, there is

only a $1 - 0.9973 = 0.0027$ chance that the sample estimates will fall outside the limit boundaries. In general, we define the upper and lower control limits as follows:

$$\text{UCL: Expected Value} + (3 \times \text{Standard Error})$$
$$\text{LCL: Expected Value} - (3 \times \text{Standard Error})$$

If the sample estimates fall randomly within the upper and lower control limits, then the production process is deemed in control. Any sample estimate that falls above the upper control limit or below the lower control limit is considered evidence that the production process is out of control and should be adjusted. In addition, any type of patterns within the control limits may suggest possible problems with the process. One indication of a process that is potentially heading out of control is unusually long runs above or below the centerline. Another possible problem is any evidence of a trend within the control limits.

In the next example, we focus on quantitative data and illustrate the $\bar{x}$ chart. We then turn to qualitative data and construct the $\bar{p}$ chart.

EXAMPLE 7.9

A firm that produces one-gallon jugs of milk wants to ensure that the machine is operating properly. Every two hours, the company samples 25 jugs and calculates the following sample mean filling weights (in ounces):

$\bar{x}_1 = 128.7$	$\bar{x}_2 = 128.4$	$\bar{x}_3 = 128.0$	$\bar{x}_4 = 127.8$	$\bar{x}_5 = 127.5$	$\bar{x}_6 = 126.9$

Assume that when the machine is operating properly, $\mu = 128$ and $\sigma = 2$, and that filling weights follow the normal distribution. Can the firm conclude that the machine is operating properly? Should the firm have any concerns with respect to this machine?

SOLUTION: Here the firm is interested in monitoring the population mean. To answer these questions, we construct an $\bar{x}$ chart. As mentioned earlier, this chart relies on the normal distribution for the sampling distribution of the estimator $\overline{X}$. Recall that if we are sampling from a normal population, then $\overline{X}$ is normally distributed even for small sample sizes. In this example, we are told that filling weights follow the normal distribution, a common assumption in the literature on quality control.

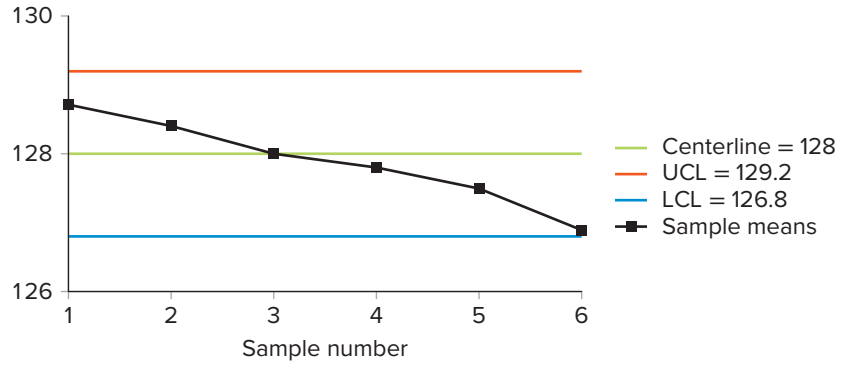
For the $\bar{x}$ chart, the centerline is the mean when the process is in control. Here, we are given that $\mu = 128$. We then calculate the UCL as three standard deviations above the mean and the LCL as three standard deviations below the mean:

$$\text{UCL: } \mu + 3\frac{\sigma}{\sqrt{n}} = 128 + 3\frac{2}{\sqrt{25}} = 129.2$$
$$\text{LCL: } \mu - 3\frac{\sigma}{\sqrt{n}} = 128 - 3\frac{2}{\sqrt{25}} = 126.8$$

Figure 7.9 shows the centerline and the control limits as well as the sample means.

All of the sample means fall within the upper control and the lower control limits, which indicates, at least initially, that the production process is in control. However, the sample means should be randomly spread between these limits; there should be no pattern. In this example, there is clearly a downward trend in the sample means. It appears as though the machine is beginning to underfill the one-gallon jugs. So even though none of the sample means lies beyond the control limits, the production process is likely veering out of control and the firm would be wise to inspect the machine sooner rather than later.

FIGURE 7.9 Mean chart for milk production process



A firm may be interested in the stability of the proportion of its goods or services possessing a certain attribute or characteristic. For example, most firms strive to produce high-quality goods (or services) and thus hope to keep the proportion of defects at a minimum. When a production process is to be assessed based on sample proportions—here, the proportion of defects—then a $\bar{p}$ chart proves quite useful. Since the primary purpose of the $\bar{p}$ chart is to track the proportion of defects in a production process, it is also referred to as a fraction defective chart or a percent defective chart. Consider the next example.

EXAMPLE 7.10

A production process has a 5% defective rate. A quality inspector takes 6 samples of $n = 500$. The following sample proportions are obtained:

$\bar{p}_1 = 0.065$	$\bar{p}_2 = 0.075$	$\bar{p}_3 = 0.082$	$\bar{p}_4 = 0.086$	$\bar{p}_5 = 0.090$	$\bar{p}_6 = 0.092$

a. Construct a $\bar{p}$ chart. Plot the sample proportions on the $\bar{p}$ chart.

b. Is the production process in control? Explain.

SOLUTION:

a. The $\bar{p}$ chart relies on the central limit theorem for the normal approximation for the sampling distribution of the estimator $\overline{P}$. Recall that so long as np and $n(1-p)$ are greater than or equal to five, then the sampling distribution of $\overline{P}$ is approximately normal. This condition is satisfied in this example. Since the expected proportion of defects is equal to 0.05, we set the centerline at $p = 0.05$. We then calculate the UCL and the LCL as follows.

$$\text{UCL: } p + 3\sqrt{\frac{p(1-p)}{n}} = 0.05 + 3\sqrt{\frac{0.05(1-0.05)}{500}} = 0.079$$

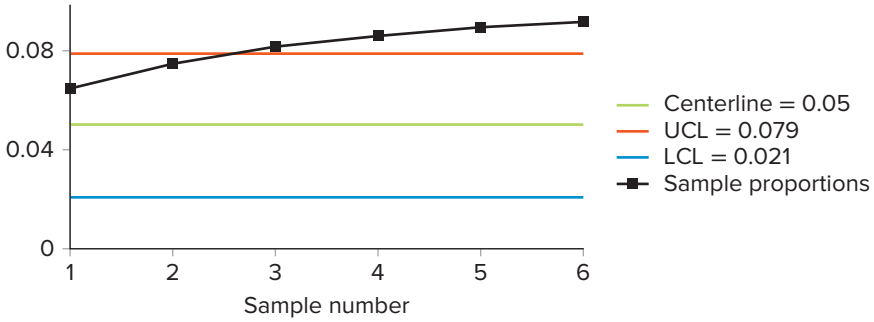
$$\text{LCL: } p - 3\sqrt{\frac{p(1-p)}{n}} = 0.05 - 3\sqrt{\frac{0.05(1-0.05)}{500}} = 0.021$$

We note that if the UCL is a value greater than one, then we reset the UCL to one in the control chart. Similarly, if the LCL is a negative value, we reset the LCL to zero in the control chart.

Plotting the values for the centerline, the UCL, the LCL, as well as the sample proportions, yields Figure 7.10.

b. Four of the most recent sample proportions fall above the upper control limit. This provides evidence that the process is out of control and needs adjustment.

FIGURE 7.10 Proportion of defects



Using Excel to Create a Control Chart

Even though Excel does not have a built-in function to create a control chart, it is still relatively easy to construct one. If we are not given values for the centerline, the UCL, the LCL, and the sample means, then we first must provide these values in an Excel spreadsheet. Other software packages do these calculations for us. We will illustrate the construction of an $\bar{x}$ chart using Example 7.11.

EXAMPLE 7.11

JK Paints manufactures various kinds of paints in 4-liter cans. The cans are filled on an assembly line with an automatic valve regulating the amount of paint. To ensure that the correct amount of paint goes into each can, the quality control manager draws a random sample of four cans each hour and measures their amounts of paint. Since past experience has produced a standard deviation of $\sigma = 0.25$, the quality control manager has been able to calculate lower and upper control limits of $3.625\ (=4-3\times 0.25/\sqrt{4})$ and $4.375\ (=4+3\times 0.25/\sqrt{4})$, respectively. Table 7.1 shows a portion of the results from the last 25 hours. The table also includes the sample mean of the four randomly selected cans, the LCL, the centerline, and the UCL. Create an $\bar{x}$ chart to determine whether the cans are being filled properly.

TABLE 7.1 Data for Example 7.11

FILE *Paint*

Sample	Obs. 1	Obs. 2	Obs. 3	Obs. 4	$\bar{x}$	LCL	Centerline	UCL
1	4.175	3.574	3.795	4.211	3.939	3.625	4	4.375
2	4.254	4.012	4.119	3.866	4.063	3.625	4	4.375
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
25	4.104	4.107	4.236	3.505	3.988	3.625	4	4.375

SOLUTION: As mentioned earlier, if only the first five columns of Table 7.1 were provided, we would have had to first populate the rest of the table by finding values for $\bar{x}$, LCL, Centerline, and UCL to make the $\bar{x}$ chart in Excel.

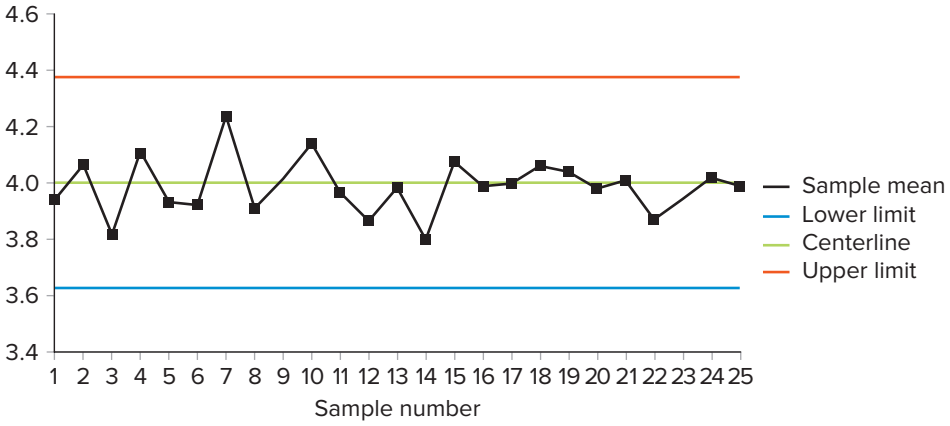
a. Open the data file ***Paint***.

b. Simultaneously select the headings and values in the $\bar{x}$, LCL, Centerline, and UCL columns and choose **Insert > Line Chart > 2-D Line**. Then, choose the option on the top left.

c. Formatting (regarding axis titles, colors, etc.) can be done by selecting **Format > Add Chart Element** from the menu.

Figure 7.11 shows the control chart that Excel produces. All sample means fall within the lower and upper limits, and they also fall randomly above and below the centerline. This indicates that the cans are being filled properly.

FIGURE 7.11 Using Excel to create a control chart



EXERCISES 7.5

Mechanics

38. Consider a normally distributed population with mean $\mu = 80$ and standard deviation $\sigma = 14$.
    a. Construct the centerline and the upper and lower control limits for the $\bar{x}$ chart if samples of size 5 are used.
    b. Repeat the analysis with samples of size 10.
    c. Discuss the effect of the sample size on the control limits.

39. Random samples of size $n = 250$ are taken from a population with $p = 0.04$.
    a. Construct the centerline and the upper and lower control limits for the $\bar{p}$ chart.
    b. Repeat the analysis with $n = 150$.
    c. Discuss the effect of the sample size on the control limits.

40. Random samples of size $n = 25$ are taken from a normally distributed population with mean $\mu = 20$ and standard deviation $\sigma = 10$.
    a. Construct the centerline and the upper and lower control limits for the $\bar{x}$ chart.
    b. Suppose six samples of size 25 produced the following sample means: 18, 16, 19, 24, 28, and 30. Plot these values on the $\bar{x}$ chart.
    c. Are any points outside the control limits? Does it appear that the process is under control? Explain.

41. Random samples of size $n = 36$ are taken from a population with mean $\mu = 150$ and standard deviation $\sigma = 42$.
    a. Construct the centerline and the upper and lower control limits for the $\bar{x}$ chart.
    b. Suppose five samples of size 36 produced the following sample means: 133, 142, 150, 165, and 169. Plot these values on the $\bar{x}$ chart.
    c. Are any points outside the control limits? Does it appear that the process is under control? Explain.

42. Random samples of size $n = 500$ are taken from a population with $p = 0.34$.
    a. Construct the centerline and the upper and lower control limits for the $\bar{p}$ chart.
    b. Suppose six samples of size 500 produced the following sample proportions: 0.28, 0.30, 0.33, 0.34, 0.37, and 0.39. Plot these values on the $\bar{p}$ chart.
    c. Are any points outside the control limits? Does it appear that the process is under control? Explain.

43. Random samples of size $n = 400$ are taken from a population with $p = 0.10$.
    a. Construct the centerline and the upper and lower control limits for the $\bar{p}$ chart.

b. Suppose six samples of size 400 produced the following sample proportions: 0.06, 0.11, 0.09, 0.08, 0.14, and 0.16. Plot these values on the $\bar{p}$ chart.

c. Is the production process under control? Explain.

Applications

44. Major League Baseball Rule 1.09 states that "the baseball shall weigh not less than 5 or more than 5¼ ounces" (www.mlb.com). Use these values as the lower and the upper control limits, respectively. Assume the centerline equals 5.125 ounces. Periodic samples of 50 baseballs produce the following sample means:

$\bar{x}_1 = 5.05$	$\bar{x}_2 = 5.10$	$\bar{x}_3 = 5.15$	$\bar{x}_4 = 5.20$	$\bar{x}_5 = 5.22$	$\bar{x}_6 = 5.24$

a. Construct an $\bar{x}$ chart. Plot the sample means on the $\bar{x}$ chart.

b. Are any points outside the control limits? Does it appear that the process is under control? Explain.

45. A production process is designed to fill boxes with an average of 14 ounces of cereal. The population of filling weights is normally distributed with a standard deviation of 2 ounces. Inspectors take periodic samples of 10 boxes. The following sample means are obtained.

$\bar{x}_1 = 13.7$	$\bar{x}_2 = 14.2$	$\bar{x}_3 = 13.9$	$\bar{x}_4 = 14.1$	$\bar{x}_5 = 14.5$	$\bar{x}_6 = 13.9$

a. Construct an $\bar{x}$ chart. Plot the sample means on the $\bar{x}$ chart.

b. Can the firm conclude that the production process is operating properly? Explain.

46. FILE ***Cricket.*** Fast bowling, also known as pace bowling, is an important component of the bowling attack in the sport of cricket. The objective is to bowl at a high speed and make the ball turn in the air and off the ground so that it becomes difficult for the batsman to hit it cleanly. Kalwant Singh is a budding Indian cricketer in a special bowling camp. While his coach is happy with Kalwant's average bowling speed, he feels that Kalwant lacks consistency. He records his bowling speed on the next four overs, where each over consists of six balls.

Over 1	Over 2	Over 3	Over 4
96.8	99.2	88.4	98.4
99.5	100.2	97.8	91.4
88.8	90.1	82.8	85.5
81.9	98.7	91.2	87.6
100.1	96.4	94.2	90.3
96.8	98.8	89.8	85.9

It is fair to assume that Kalwant's bowling speed is normally distributed with a mean and a standard deviation of 94 miles per hour and 2.8 miles per hour, respectively.

a. Construct the centerline and the upper and lower control limits for the $\bar{x}$ chart. Plot the average speed of Kalwant's four overs on the $\bar{x}$ chart.

b. Is there any pattern in Kalwant's bowling that justifies his coach's concerns that he is not consistent in bowling? Explain.

47. A manufacturing process produces steel rods in batches of 1,000. The firm believes that the percent of defective items generated by this process is 5%.

a. Construct the centerline and the upper and lower control limits for the $\bar{p}$ chart.

b. An engineer inspects the next batch of 1,000 steel rods and finds that 6.2% are defective. Is the manufacturing process under control? Explain.

48. A firm produces computer chips for personal computers. From past experience, the firm knows that 4% of the chips are defective. The firm collects a sample of the first 500 chips manufactured at 1:00 pm for the past two weeks. The following sample proportions are obtained:

$\bar{p}_1 = 0.044$	$\bar{p}_2 = 0.052$	$\bar{p}_3 = 0.060$	$\bar{p}_4 = 0.036$	$\bar{p}_5 = 0.028$
$\bar{p}_6 = 0.042$	$\bar{p}_7 = 0.034$	$\bar{p}_8 = 0.054$	$\bar{p}_9 = 0.048$	$\bar{p}_{10} = 0.025$

a. Construct a $\bar{p}$ chart. Plot the sample proportions on the $\bar{p}$ chart.

b. Can the firm conclude that the process is operating properly?

49. The college admissions office at a local university usually admits 750 students and knows from previous experience that 25% of these students choose not to enroll at the university.

a. Construct the centerline and the upper and lower control limits for the $\bar{p}$ chart.

b. Assume that this year the university admits 750 students and 240 choose not to enroll at the university. Should the university be concerned? Explain.

50. Following customer complaints about the quality of service, Dell stopped routing corporate customers to a technical support call center in Bangalore, India (*USA TODAY,* November 24, 2003). Suppose Dell's decision to direct customers to call centers outside of India was based on consumer complaints in the last six months. Let the number of complaints per month for 80 randomly selected customers be given below.

Month	Number of Complaints
1	20
2	12
3	24
4	14
5	25
6	22

a. Construct the centerline and the upper and lower control limits for the $\bar{p}$ chart if management allows a 15% complaint rate.

b. Can you justify Dell's decision to direct customers to call centers outside of India?

WRITING WITH STATISTICS

Barbara Dwyer, the manager at Lux Hotel, makes every effort to ensure that customers attempting to make phone reservations wait an average of only 60 seconds to speak with a reservations specialist. She knows that this is likely to be the customer's first impression of the hotel and she wants the initial interaction to be a positive one. Since the hotel accepts phone reservations 24 hours a day, Barbara wonders if the quality of service is consistently maintained throughout the day. She takes six samples of $n = 4$ calls during each of four shifts over one 24-hour period and records the wait time of each call. A portion of the data, in seconds, is presented in Table 7.2.

Barbara assumes that wait times are normally distributed with a mean and standard deviation of 60 seconds and 30 seconds, respectively. She wants to use the sample information to

1. Prepare a control chart for wait times.
2. Determine, using the control chart, whether the quality of service is consistently maintained throughout the day.



©Ryan McVay/Getty Images

TABLE 7.2 Wait times for phone reservations

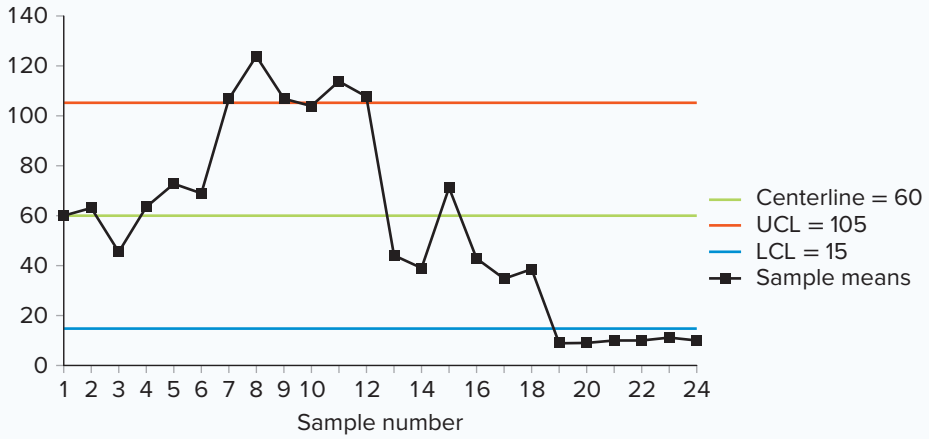
Shift	Sample	Wait Time (in seconds)				Sample Mean, $\bar{x}$
Shift 1: 12:00 am–6:00 am	1	67	48	52	71	60
	2	57	68	60	66	63
	3	37	41	60	41	45
	4	83	59	49	66	64
	5	82	63	64	83	73
	6	87	53	66	69	69
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Shift 4: 6:00 pm–12:00 am	19	6	11	8	9	9
	20	10	8	10	9	9
	21	11	7	14	7	10
	22	8	9	9	12	10
	23	9	12	9	14	11
	24	5	8	15	11	10

FILE
Lux_Hotel

Sample Report—Customer Wait Time

When a potential customer phones Lux Hotel, it is imperative for the reservations specialist to set a tone that relays the high standard of service that the customer will receive if he/she chooses to stay at the Lux. For this reason, management at the Lux strives to minimize the time that elapses before a potential customer speaks with a reservations specialist; however, management also recognizes the need to use its resources wisely. If too many reservations specialists are on duty, then resources are wasted due to idle time; yet if too few reservations specialists are on duty, the result might mean angry first-time customers or, worse, lost customers. In order to ensure customer satisfaction as well as an efficient use of resources, a study is conducted to determine whether a typical customer waits an average of 60 seconds to speak with a reservations specialist. Before data are collected, a control chart is constructed. The upper control limit (UCL) and the lower control limit (LCL) are set three standard deviations from the desired average of 60 seconds. In Figure 7.A, the desired average of 60 seconds is denoted as the centerline and the upper and lower control limits amount to 105 seconds and 15 seconds ($\mu \pm 3\frac{\sigma}{\sqrt{n}} = 60 \pm 3\frac{30}{\sqrt{4}} = 60 \pm 45$), respectively. The reservation process is deemed under control if the sample means fall randomly within the upper and lower control limits; otherwise the process is out of control and adjustments should be made.

FIGURE 7.A Sample mean wait times



During each of four shifts, six samples of $n = 4$ calls are randomly selected over one 24-hour period and the average wait time of each sample is recorded. All six sample means from the first shift (1st shift: 12:00 am–6:00 am, sample numbers 1 through 6) fall within the control limits, indicating that the reservation process is in control. However, five sample means from the second shift (2nd shift: 6:00 am–12:00 pm, sample numbers 7 through 12) lie above the upper control limit. Customers calling during the second shift are waiting too long before they speak with a specialist. In terms of quality standards, this is unacceptable from the hotel's perspective. All six sample means from the third shift fall within the control limits (3rd shift: 12:00 pm–6:00 pm, sample numbers 13 through 18), yet all sample means for the fourth shift fall below the lower control limit (4th shift: 6:00 pm–12:00 am, sample numbers 19 through 24). Customers are waiting for very short periods of time to speak with a reservations specialist, but reservations specialists may have too much idle time. Perhaps one solution is to shift some reservations specialists from shift four to shift two.

CONCEPTUAL REVIEW

LO 7.1 Explain common sample biases.

A sampling **bias** occurs when the information from a sample is not typical of that in the population in a systematic way. It is often caused by samples that are not representative of the population. **Selection bias** refers to a systematic underrepresentation of certain groups from consideration for the sample. **Nonresponse bias** refers to a systematic difference in preferences between respondents and nonrespondents to a survey or a poll. **Social-desirability bias** refers to a systematic difference between a group's "socially acceptable" responses to a survey or poll and this group's ultimate choice.

LO 7.2 Describe various sampling methods.

A **simple random sample** is a sample of n observations that has the same probability of being selected from the population as any other sample of n observations. Most statistical methods presume a simple random sample.

A **stratified random sample** is formed when the population is divided into groups (strata) based on one or more classification criteria. A stratified random sample includes randomly selected observations from each stratum. The number of observations per

stratum is proportional to the stratum's size in the population. The data for each stratum are eventually pooled. A **cluster sample** is formed when the population is divided into groups (clusters) based on geographic areas. Whereas a stratified random sample consists of elements from each group, a cluster sample includes observations from randomly selected clusters. Stratified random sampling is preferred when the objective is to increase precision and cluster sampling is preferred when the objective is to reduce costs.

LO 7.3 Describe the sampling distribution of the sample mean.

A particular characteristic of a population, such as the mean or the proportion, is called a parameter, which is a constant even though its value may be unknown. A statistic, such as the sample mean or the sample proportion, is a random variable whose value depends on the chosen random sample. When a statistic is used to estimate a parameter, it is referred to as an **estimator**. A particular value of the estimator is called an **estimate**.

Since the statistic $\overline{X}$ is a random variable, its sampling distribution is the probability distribution of sample means derived from all possible samples of a given size from the population. The expected value of the sample mean $\overline{X}$ equals $E(\overline{X}) = \mu$, and the standard deviation, commonly referred to as the standard error of the sample mean, equals $se(\overline{X}) = \frac{\sigma}{\sqrt{n}}$. For any sample size, the sampling distribution of $\overline{X}$ is normal if the population is normally distributed. If $\overline{X}$ is normally distributed, then any value $\bar{x}$ can be transformed to its corresponding z value as $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$.

LO 7.4 Explain the importance of the central limit theorem.

The **central limit theorem** (**CLT**) is used when the random sample is drawn from an unknown or a nonnormal population. It states that for any population X with expected value μ and standard deviation σ, the sampling distribution of $\overline{X}$ is approximately normal if the sample size n is sufficiently large. As a general guideline, the normal distribution approximation is justified when $n \geq 30$.

LO 7.5 Describe the sampling distribution of the sample proportion.

The expected value of the sample proportion $\overline{P}$ equals $E(\overline{P}) = p$ and its standard error equals $se(\overline{P}) = \sqrt{\frac{p(1-p)}{n}}$. From the CLT, we can conclude that for any population proportion p, the sampling distribution of $\overline{P}$ is approximately normal if the sample size n is sufficiently large. As a general guideline, the normal distribution approximation is justified when $np \geq 5$ and $n(1-p) \geq 5$. If $\overline{P}$ is normally distributed, then any value $\bar{p}$ can be transformed to its corresponding z value as $z = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$.

LO 7.6 Use a finite population correction factor.

If the sample size is large relative to the population size, then the standard errors of the estimators must be multiplied by a correction factor. This correction factor, called the **finite population correction factor**, is used when the sample constitutes at least 5% of the population—that is, $n \geq 0.05N$. With the correction factor, $se(\overline{X}) = \frac{\sigma}{\sqrt{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)$ and $se(\overline{P}) = \sqrt{\frac{p(1-p)}{n}}\left(\sqrt{\frac{N-n}{N-1}}\right)$. The transformation to the corresponding z value is made accordingly.

LO 7.7 Construct and interpret control charts for quantitative and qualitative data.

Statistical quality control involves statistical techniques used to develop and maintain a firm's ability to produce high-quality goods and services. The most commonly used

statistical tool in quality control is the **control chart**. A control chart specifies a centerline as well as an upper control limit (UCL) and a lower control limit (LCL). In general, the UCL and the LCL are set within three standard deviations of the centerline.

The UCL and the LCL for the $\bar{x}$ **chart** are defined as $\mu + 3\frac{\sigma}{\sqrt{n}}$ and $\mu - 3\frac{\sigma}{\sqrt{n}}$, respectively. For the $\bar{p}$ **chart**, these limits are defined as $p + 3\sqrt{\frac{p(1-p)}{n}}$ and $p - 3\sqrt{\frac{p(1-p)}{n}}$, respectively. In general, if the sample means or the sample proportions fall within the control limits, then the process is under control; otherwise it is out of control and adjustment is necessary. However, even if these sample estimates fall within the control limits, they must be randomly spread between the limits. If there is a trend or unusually long runs above or below the centerline, then the process may be veering out of control.

ADDITIONAL EXERCISES AND CASE STUDIES

Exercises

51. A seminal study conducted by scientists at the University of Illinois found evidence of improved memory and reasoning for those who took three vigorous 40-minute walks a week over six months (*Newsweek*, June 28–July 5, 2010). As an assistant manager working for a public health institute based in Florida, you would like to estimate the proportion of adults in Miami, Florida, who follow such a walking regimen. Discuss the sampling bias in the following strategies where people are asked if they walk regularly:
    a. Randomly selected adult beachgoers in Miami.
    b. Randomly selected Miami residents who are requested to disclose the information in prepaid envelopes.
    c. Randomly selected Miami residents who are requested to disclose the information on the firm's website.
    d. Randomly selected adult patients at all hospitals in Miami.
52. In the previous question regarding walking regimens of the residents of Miami, explain how you can obtain a representative sample based on the following sampling strategies:
    a. Simple random sampling.
    b. Stratified random sampling.
    c. Cluster sampling.
53. According to the Bureau of Labor Statistics, it takes an average of 22 weeks for someone over 55 to find a new job, compared with 16 weeks for younger workers (*The Wall Street Journal*, September 2, 2008). Assume that the probability distributions are normal and that the standard deviation is 2 weeks for both distributions.
    a. What is the probability that 8 workers over the age of 55 take an average of more than 20 weeks to find a job?
    b. What is the probability that 20 younger workers average less than 15 weeks to find a job?
54. Presidential job approval is the most-watched statistic in American politics. According to the June 2010 NBC/*Wall Street Journal* public opinion poll, President Barack Obama had reached his lowest approval rating since taking office in January of 2009. The poll showed that 48% of people disapproved of the job Obama was doing as president of the United States, while only 45% approved. Experts attributed the drop in approval ratings to a poor economy and the government's reaction to the massive oil spill in the Gulf of Mexico. Use the June 2010 approval and disapproval ratings to answer the following questions.
    a. What is the probability that President Obama gets a majority support in a random sample of 50 Americans?
    b. What is the probability that President Obama gets a majority disapproval in a random sample of 50 Americans?
55. While starting salaries have fallen for college graduates in many of the top hiring fields, there is some good news for business undergraduates

with concentrations in accounting and finance (*Bloomberg Businessweek*, July 1, 2010). According to the National Association of Colleges and Employers' Summer 2010 Salary Survey, accounting graduates commanded the second highest salary at $50,402, followed by finance graduates at $49,703. Let the standard deviation for accounting and finance graduates be $6,000 and $10,000, respectively.

a. What is the probability that 100 randomly selected accounting graduates will average more than $52,000 in salary?
b. What is the probability that 100 randomly selected finance graduates will average more than $52,000 in salary?
c. Comment on the above probabilities.

56. An automatic machine in a manufacturing process is operating properly if the length of an important subcomponent is normally distributed with a mean $\mu = 80$ cm and a standard deviation $\sigma = 2$ cm.
    a. Find the probability that the length of one randomly selected unit is less than 79 cm.
    b. Find the probability that the average length of 10 randomly selected units is less than 79 cm.
    c. Find the probability that the average length of 30 randomly selected units is less than 79 cm.

57. Trader Joe's is a privately held chain of specialty grocery stores in the United States. Starting out as a small chain of convenience stores, it has expanded to over 340 stores as of June 2010 (www.traderjoes.com). It has developed a reputation as a unique grocery store selling products such as gourmet foods, beer and wine, bread, nuts, cereal, and coffee. One of their best-selling nuts is Raw California Almonds, which are priced at $4.49 for 16 ounces. Since it is impossible to pack exactly 16 ounces in each packet, a researcher has determined that the weight of almonds in each packet is normally distributed with a mean and a standard deviation equal to 16.01 ounces and 0.08 ounces, respectively.
    a. Discuss the sampling distribution of the sample mean based on any given sample size.
    b. Find the probability that a random sample of 20 bags of almonds will average less than 16 ounces.
    c. Suppose your cereal recipe calls for no less than 48 ounces of almonds. What is the probability that three packets of almonds will meet your requirement?

58. Georgia residents spent an average of $470.73 on the lottery in 2010, or 1% of their personal income (www.msn.com, May 23, 2012). Suppose the amount spent on the lottery follows a normal distribution with a standard deviation of $50.
    a. What is the probability that a randomly selected Georgian spent more than $500 on the lottery?
    b. If four Georgians are randomly selected, what is the probability that the average amount spent on the lottery was more than $500?
    c. If four Georgians are randomly selected, what is the probability that all of them spent more than $500 on the lottery?

59. Data from the Bureau of Labor Statistics' Consumer Expenditure Survey show that the average annual expenditure for cellular phone services per consumer unit increased from $210 in 2001 to $608 in 2007. Let the standard deviation of annual cellular expenditure be $48 in 2001 and $132 in 2007.
    a. What is the probability that the average annual expenditure of 100 cellular customers in 2001 exceeded $200?
    b. What is the probability that the average annual expenditure of 100 cellular customers in 2007 exceeded $600?

60. According to a report, scientists in New England say they have identified a set of genetic variants that predicts extreme longevity with 77% accuracy (*The New York Times*, July 1, 2010). Assume 150 patients decide to get their genomes sequenced.
    a. If the claim by scientists is accurate, what is the probability that more than 120 patients will get a correct diagnosis for extreme longevity?
    b. If the claim by scientists is accurate, what is the probability that fewer than 70% of the patients will get a correct diagnosis for extreme longevity?

61. American workers are increasingly planning to delay retirement (*U.S. News & World Report*, June 30, 2010). According to a Pew Research Center comprehensive survey, 35% of employed adults of age 62 and older say they have pushed back their retirement date.
    a. What is the probability that in a sample of 100 employed adults of age 62 and older, more than 40% have pushed back their retirement date?
    b. What is the probability that in a sample of 200 employed adults of age 62 and older, more than 40% have pushed back their retirement date?
    c. Comment on the difference between the two estimated probabilities.

62. FILE ***Packaging.*** A variety of packaging solutions exist for products that must be kept within a specific temperature range. Cold chain distribution is particularly useful in the food and

pharmaceutical industries. A packaging company strives to maintain a constant temperature for its packages. It is believed that the temperature of its packages follows a normal distribution with a mean of 5 degrees Celsius and a standard deviation of 0.3 degree Celsius. Inspectors take weekly samples for 5 weeks of eight randomly selected boxes and report the temperatures in degrees Celsius. A portion of the data is given below.

Week 1	Week 2	Week 3	Week 4	Week 5
3.98	5.52	5.79	3.98	5.14
4.99	5.52	6.42	5.79	6.25
⋮	⋮	⋮	⋮	⋮
4.95	4.95	5.44	5.95	4.28

a. Construct an $\bar{x}$ chart for quality control. Plot the five weekly sample means on the $\bar{x}$ chart.
b. Are any points outside the control limits? Does it appear that the process is in control? Explain.

63. The producer of a particular brand of soup claims that its sodium content is 50% less than that of its competitor. The food label states that the sodium content measures 410 milligrams per serving. Assume the population of sodium content is normally distributed with a standard deviation of 25 milligrams. Inspectors take periodic samples of 25 cans and measure the sodium content. The following sample means are obtained.

$\bar{x}_1 = 405$	$\bar{x}_2 = 412$	$\bar{x}_3 = 399$
$\bar{x}_4 = 420$	$\bar{x}_5 = 430$	$\bar{x}_6 = 428$

a. Construct an $\bar{x}$ chart. Plot the sample means on the $\bar{x}$ chart.
b. Can the inspectors conclude that the producer is advertising the sodium content accurately? Explain.

64. Acceptance sampling is an important quality control technique, where a batch of data is tested to determine if the proportion of units having a particular attribute exceeds a given percentage. Suppose that 10% of produced items are known to be nonconforming. Every week a batch of items is evaluated and the production machines are adjusted if the proportion of nonconforming items exceeds 15%.
a. What is the probability that the production machines will be adjusted if the batch consists of 50 items?
b. What is the probability that the production machines will be adjusted if the batch consists of 100 items?

65. In the previous question, suppose that the management decides to use a $\bar{p}$ chart for the analysis. As noted earlier, 10% of produced items are known to be nonconforming. The firm analyzes a batch of production items for 6 weeks and computes the following percentages of nonconforming items.

Week	Nonconforming Percentage
1	5.5
2	13.1
3	16.8
4	13.6
5	19.8
6	2.0

a. Suppose weekly batches consisted of 50 items. Construct a $\bar{p}$ chart and determine if the machine needs adjustment in any of the weeks.
b. Suppose weekly batches consisted of 100 items. Construct a $\bar{p}$ chart and determine if the machine needs adjustment in any of the weeks.

CASE STUDIES

CASE STUDY 7.1 The significant decline of savings in the United States from the 1970s and 1980s to the 1990s and 2000s has been widely discussed by economists (www.money.cnn.com, June 30, 2010). According to the Bureau of Economic Analysis, the savings rate of American households, defined as a percentage of the disposable personal income, was 4.20% in 2009. The reported savings rate is not uniform across the country. A public policy institute conducts two of its own surveys to compute the savings rate in the Midwest. In the first survey, a sample of 160 households is taken and the average savings rate is found to be 4.48%. Another sample of 40 households finds an average savings rate of 4.60%. Assume that the population standard deviation is 1.4%.

In a report, use the above information to

1. Compute the probability of obtaining a sample mean that is at least as high as the one computed in each of the two surveys.

2. Use these probabilities to decide which of the two samples is likely to be more representative of the United States as a whole.

CASE STUDY 7.2 According to a report, college graduates in 2010 were likely to face better job prospects than 2009 graduates (*The New York Times*, May 24, 2010). Many employers who might have been pessimistic at the start of the 2009–2010 academic year were making more offers than expected. Despite the improvement in job prospects, the Bureau of Labor Statistics reported that the current jobless rate for college graduates under age 25 was still 8%. For high school graduates under age 25 who did not enroll in college, the current jobless rate was 24.5%. Cindy Chan works in the sales department of a trendy apparel company and has recently been relocated to a small town in Iowa. She finds that there are a total of 220 college graduates and 140 high school graduates under age 25 who live in this town. Cindy wants to gauge the demand for her products by the number of youths in this town who are employed.

In a report, use the above information to

1. Compute the expected number of college and high school graduates who are employed.
2. Report the probabilities that at least 200 college graduates and at least 100 high school graduates under age 25 are employed.

CASE STUDY 7.3 Hockey pucks used by the National Hockey League (NHL) and other professional leagues weigh an average of 163 grams (5.75 ounces). A quality inspector monitors the manufacturing process for hockey pucks. She takes eight samples of $n = 10$. It is believed that puck weights are normally distributed, and when the production process is in control, $\mu = 163$ and $\sigma = 7.5$. A portion of the data, measured in grams, is shown in the accompanying table.

Data for Case Study 7.3 Hockey Puck Weights (in grams)

FILE
Hockey_Puck

#1	#2	#3	#4	#5	#6	#7	#8
162.2	165.8	156.4	165.3	168.6	167.0	186.8	178.3
159.8	166.2	156.4	173.3	175.8	171.4	160.4	163.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
160.3	160.6	152.2	166.4	168.2	168.4	176.8	171.3

In a report, use the above information to

1. Prepare a control chart that specifies a centerline as well as an upper control limit (UCL) and a lower control limit (LCL).
2. Determine, using the control chart, whether the process is in control.

APPENDIX 7.1 Derivation of the Mean and the Variance for $\overline{X}$ and $\overline{P}$

Sample Mean, $\overline{X}$

Let the expected value and the variance of the population X be denoted by $E(X) = \mu$ and $Var(X) = \sigma^2$, respectively. The sample mean $\overline{X}$ based on a random draw of n observations, $X_1, X_2, \ldots, X_n$, from the population is computed as $\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$.

We use the properties of the sum of random variables to derive

$$E(\overline{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n}$$

$$= \frac{\mu + \mu + \cdots + \mu}{n} = \frac{n\mu}{n} = \mu.$$

Since the sample mean is based on n independent draws from the population, the covariance terms drop out and the variance of the sample mean is thus derived as

$$Var(\overline{X}) = Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2} Var(X_1 + X_2 + \cdots + X_n)$$

$$= \frac{1}{n^2}(Var(X_1) + Var(X_2) + \cdots + Var(X_n))$$

$$= \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Sample Proportion, $\overline{P}$

Let X be a binomial random variable representing the number of successes in n trials. Recall from Chapter 5 that $E(X) = np$ and $Var(X) = np(1 - p)$ where p is the probability of success. For the sample proportion $\overline{P} = \frac{X}{n}$,

$$E(\overline{P}) = E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{np}{n} = p, \text{ and}$$

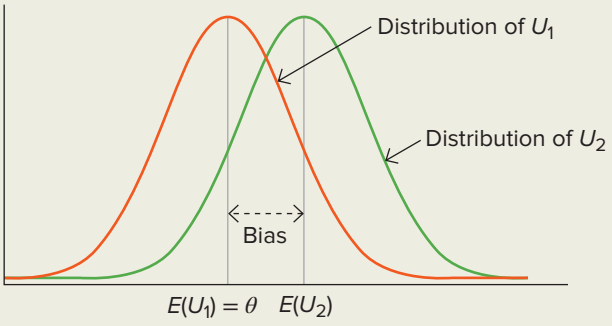
$$Var(\overline{P}) = Var\left(\frac{X}{n}\right) = \frac{Var(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

APPENDIX 7.2 Properties of Point Estimators

We generally discuss the performance of an estimator in terms of its statistical properties. Some of the desirable properties of a point estimator include unbiasedness, consistency, and efficiency. An estimator is **unbiased** if, based on repeated sampling from the population, the average value of the estimator equals the population parameter. In other words, for an unbiased estimator, the expected value of the point estimator equals the population parameter.

Figure A7.1 shows the sampling distributions for two estimators U_1 and U_2, which are assumed to be normally distributed. Let θ (the Greek letter read as theta) be the true parameter value of the population. Estimator U_1 is unbiased because its expected value $E(U_1)$ equals θ. Estimator U_2 is biased because $E(U_2) \neq \theta$; the degree of bias is given by the difference between $E(U_2)$ and θ.

FIGURE A7.1 The distributions of unbiased (U_1) and biased (U_2) estimators



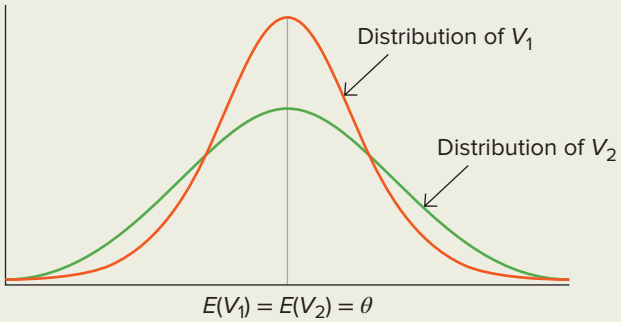
Since $E(\overline{X}) = \mu$ and $E(\overline{P}) = p$, $\overline{X}$ and $\overline{P}$ are the unbiased estimators of μ and p, respectively. This property is independent of the sample size. For instance, the expected value of the sample mean is equal to the population mean irrespective of the sample size.

We often compare the performance of unbiased estimators in terms of their relative **efficiency**. An estimator is deemed efficient if its variability between samples is smaller than that of other unbiased estimators. Recall that the variability is often measured by the standard error of the estimator. For an unbiased estimator to be efficient, its standard

error must be lower than that of other unbiased estimators. It is well documented that the estimators $\overline{X}$ and $\overline{P}$ are not only unbiased, but also efficient.

Figure A7.2 shows the sampling distributions for two unbiased estimators, V_1 and V_2, for the true population parameter θ. Again, for illustration, V_1 and V_2 follow the normal distribution. While both V_1 and V_2 are unbiased ($E(V_1) = E(V_2) = \theta$), V_1 is more efficient because it has less variability.

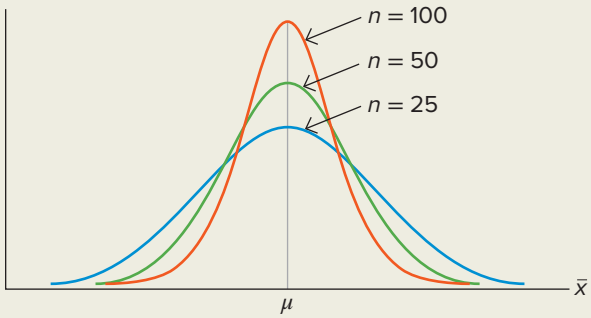
FIGURE A7.2 The distributions of efficient (V_1) and less efficient (V_2) estimators



Another desirable property, which is often considered a minimum requirement for an estimator, is **consistency**. An estimator is consistent if it approaches the population parameter of interest as the sample size increases. Consistency implies that we will get the inference right if we take a large enough sample. The estimators $\overline{X}$ and $\overline{P}$ are not only unbiased, but also consistent. For instance, the sample mean collapses to the population mean ($\overline{X} \to \mu$) as the sample size approaches infinity ($n \to \infty$). An unbiased estimator is consistent if its standard error collapses to zero as the sample size increases.

The consistency of $\overline{X}$ is illustrated in Figure A7.3.

FIGURE A7.3 The distribution of a consistent estimator $\overline{X}$ for various sample sizes



As the sample size n increases, the variability of $\overline{X}$ decreases. In particular as $n \to \infty$, $SD(\overline{X}) = \sigma/\sqrt{n} \to 0$, thus implying that $\overline{X}$ is a consistent estimator of μ.

APPENDIX 7.3 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, JMP, and R. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands. When importing data into R, use the menu-driven option: File > Import Dataset > From Excel.

Minitab

Generating a Random Sample

A. (Replicating Example 7.1) From the menu, choose **Calc > Random Data > Integer**.

B. Enter 100 as the **Number of rows of data to generate**; enter C1 for **Store in column**; enter 1 for **Minimum value** and 2,750 as **Maximum value**.

Constructing an $\overline{X}$ Chart

FILE *Paint*

A. (Replicating Figure 7.11). Stack all the observations (columns 2–5) into one column, and label it "observations."

B. From the menu, choose **Stat > Control Charts > Variables Charts for Subgroups > Xbar**.

C. Choose **All observations for a chart are in one column**, and in the box directly under this one, select observations. For **Subgroup sizes**, enter the number 4. Choose **Xbar Options** and enter 4 for **Mean** and 0.25 for **Standard deviation**.

SPSS

Constructing an $\overline{X}$ Chart

FILE *Paint*

A. (Replicating Figure 7.11). Stack the observations (columns 2–5) into one column, and label it "observations." In an adjacent column, indicate how the data are grouped and label this column "group." For instance, the first four observations are given the value 1; the next four observations are given the value 2, and so on.

B. From the menu, select **Analyze > Quality Control > Control Charts > X-bar, R, s**.

C. Under **Process Measurement**, select observations, and under **Subgroups Defined by** select group. Under **Charts**, select **X-bar using standard deviation**. Choose **Options**. After **Number of Sigmas**, enter 3, and after **Minimum subgroup size**, enter 4. Choose **Statistics**. Under **Specification Limits**, enter 4.375 for **Upper**, 3.625 for **Lower**, and 4 for **Target**.

JMP

Generating a Random Sample

A. (Replicating Example 7.1) Right-click on the header at the top of the column in the spreadsheet view and select **Formula**. Under **Functions (grouped)**, choose **Random > Random Integer**.

B. Put the insertion marker on the box for **n1** and click the insert button (shown as a caret ^ with the mathematical operations) until you see **n2** next to **n1**. Enter 1 for **n1** and 2,750 for **n2**.

Constructing the $\overline{X}$ Chart

FILE *Paint*

A. (Replicating Figure 7.11). Stack all the observations (columns 2–5) into one column, and label it "observations."

B. From the menu, choose **Analyze > Quality and Process > Control Chart > X-bar**.

C. Under **Select Columns**, select observations, and under **Cast Columns into Roles**, select Process. Under **Parameters**, select **KSigma** and enter 3. Under **Sample Size**, select **Sample Size Constant** and enter 4. Select **Specify Stats**. Enter 0.25 for **Sigma** and 4 for **Mean(measure)**.

R

Generating a Random Sample

(Replicating Example 7.1) Use the **sample** function to generate random integers within some interval. In general, enter "sample(lower:upper, *n*)", where lower and upper refer to the smallest and largest integers in the interval, respectively, and *n* denotes the sample size. Thus, to generate a sample of 100 integers, labeled Sample_draw, with values between 1 and 2,750, enter:

```
> Sample_draw <- sample(1:2750, 100)
> list(Sample_draw)
```

Constructing the $\overline{X}$ Chart

A. (Replicating Example 7.11) Install and load the *qcc* package (where *qcc* stands for Quality Control Charts). Enter:

FILE
Paint

```
> install.packages("qcc")
> library(qcc)
```

B. Use the **qcc** function from the *qcc* package. Within the function, extract the data in columns 2 through 5 with the use of square brackets. Then, for options, use *type* to designate the type of control chart, *center* to denote the centerline, *std.dev* to denote the standard deviation, *nsigmas* to denote the number of standard deviations from the centerline, and *title* to specify a main title for the chart. Enter:

```
> qcc(Paint[ ,2:5], type="xbar", center=4, st.dev.=0.25,
  nsigmas=3, title="Control Chart for Paint").
```

8 Interval Estimation

Learning Objectives

After reading this chapter you should be able to:

LO **8.1** Explain a confidence interval.

LO **8.2** Calculate a confidence interval for the population mean when the population standard deviation is known.

LO **8.3** Describe the factors that influence the width of a confidence interval.

LO **8.4** Discuss features of the *t* distribution.

LO **8.5** Calculate a confidence interval for the population mean when the population standard deviation is not known.

LO **8.6** Calculate a confidence interval for the population proportion.

LO **8.7** Select a sample size to estimate the population mean and the population proportion.

In earlier chapters, we made a distinction between the population parameters, such as the population mean and the population proportion, and the corresponding sample statistics. The sample statistics are used to make statistical inferences regarding the unknown values of the population parameters. In general, two basic methodologies emerge from the inferential branch of statistics: estimation and hypothesis testing. As discussed in Chapter 7, a point estimator uses sample data to produce a single value as an estimate for the unknown population parameter of interest. A confidence interval, on the other hand, produces a range of values that estimate the unknown population parameter. In this chapter, we develop and interpret confidence intervals for the population mean and the population proportion. Since obtaining a sample is one of the first steps in making statistical inferences, we also learn how an appropriate sample size is determined in order to achieve a certain level of precision in the estimates.



©1000 Words/Shutterstock

INTRODUCTORY CASE

Fuel Usage of "Ultra-Green" Cars

A car manufacturer advertises that its new "ultra-green" car obtains an average of 100 miles per gallon (mpg) and, based on its fuel emissions, is one of the few cars that earns an A+ rating from the Environmental Protection Agency. Jared Beane, an analyst at Pinnacle Research, records the mpg for a sample of 25 "ultra-green" cars after the cars were driven equal distances under identical conditions. Table 8.1 shows each car's mpg.

TABLE 8.1 MPG for a Sample of 25 "Ultra-Green" Cars

97	117	93	79	97
87	78	83	94	96
102	98	82	96	113
113	111	90	101	99
112	89	92	96	98

FILE
MPG

Jared has already used tabular and graphical methods to summarize the data in his report. He would like to make statistical inferences regarding key population parameters. In particular, he wants to use the above sample information to

1. Estimate the mean mpg of all ultra-green cars with 90% confidence.
2. Estimate the proportion of all ultra-green cars that obtain over 100 mpg with 90% confidence.
3. Determine the sample size that will enable him to achieve a specified level of precision in his mean and proportion estimates.

A synopsis of this case is provided at the end of Section 8.4.

LO 8.1

Explain a confidence interval.

8.1 CONFIDENCE INTERVAL FOR THE POPULATION MEAN WHEN σ IS KNOWN

Recall that a population consists of all items of interest in a statistical problem, whereas a sample is a subset of the population. Given sample data, we use sample statistics to make inferences about unknown population parameters, such as the population mean and the population proportion. Two basic methodologies emerge from the inferential branch of statistics: estimation and hypothesis testing. Although sample statistics are based on a portion of the population, they contain useful information to estimate the population parameters and to conduct tests regarding the population parameters. In this chapter, we focus on estimation.

As discussed in Chapter 7, when a statistic is used to estimate a parameter, it is referred to as a point estimator, or simply an estimator. A particular value of an estimator is called a point estimate or an estimate. Recall that the sample mean $\overline{X}$ is the estimator of the population mean μ, and the sample proportion $\overline{P}$ is the estimator of the population proportion p. Let us consider the introductory case where Jared Beane records the mpg for a sample of 25 ultra-green cars. We use the sample information in Table 8.1 to compute the mean mpg of the cars as $\bar{x} = 96.52$ mpg. Similarly, since Jared is also interested in the proportion of these cars that get an mpg greater than 100, and seven of the cars in the sample satisfied this criterion, we compute the relevant sample proportion as $\bar{p} = 7/25 = 0.28$. Therefore, our estimate for the mean mpg of all ultra-green cars is 96.52 mpg, and our estimate for the proportion of all ultra-green cars with mpg greater than 100 is 0.28.

It is important to note that the above estimates are based on a sample of 25 cars and, therefore, are likely to vary between samples. For instance, the values will change if another sample of 25 cars is used. What Jared really wishes to estimate are the mean and the proportion (parameters) of all ultra-green cars (population), not just those comprising the sample. We now examine how we can extract useful information from a single sample to make inferences about these population parameters.

So far we have only discussed point estimators. Often it is more informative to provide a range of values—an interval—rather than a single point estimate for the unknown population parameter. This range of values is called a **confidence interval**, also referred to as an **interval estimate**, for the population parameter.

CONFIDENCE INTERVAL

A confidence interval, or interval estimate, provides a range of values that, with a certain level of confidence, contains the population parameter of interest.

In order to construct a confidence interval for the population mean μ or the population proportion p, it is essential that the sampling distributions of $\overline{X}$ and $\overline{P}$ follow, or approximately follow, a normal distribution. Other methods that do not require the normality condition are not discussed in this text. Recall from Chapter 7 that $\overline{X}$ follows a normal distribution when the underlying population is normally distributed; this result holds irrespective of the sample size n. If the underlying population is not normally distributed, then by the central limit theorem, the sampling distribution of $\overline{X}$ will be approximately normal if the sample size is sufficiently large—that is, when $n \geq 30$. Similarly, the sampling distribution of $\overline{P}$ is approximately normal if the sample size is sufficiently large—that is, when $np \geq 5$ and $n(1 - p) \geq 5$.

The main ingredient for developing a confidence interval is the sampling distribution of the underlying statistic. The sampling distribution of $\overline{X}$, for example, describes how the sample mean varies between samples. Recall that the variability between samples is measured by the standard error of $\overline{X}$. If the standard error is small, it implies that the sample means are not only close to one another, they are also close to the unknown population mean μ.

A confidence interval is generally associated with a **margin of error** that accounts for the standard error of the estimator and the desired confidence level of the interval. As we have just stressed, the sampling distributions of the estimators for the population mean and the population proportion must be approximately normal. The symmetry implied by the normal distribution allows us to construct a confidence interval by adding and subtracting the same margin of error to the point estimate.

GENERAL FORMAT OF THE CONFIDENCE INTERVAL FOR μ AND ρ

The confidence interval for the population mean and the population proportion is constructed as

point estimate $\pm$ margin of error.

An analogy to a simple weather example is instructive. If you feel that the outside temperature is about 50 degrees, then perhaps you can, with a certain level of confidence, suggest that the actual temperature is between 40 and 60 degrees. In this example, 50 degrees is analogous to a point estimate of the actual temperature, and 10 degrees is the margin of error that is added to and subtracted from this point estimate.

We know from the introductory case study that the point estimate for the population mean mpg of all ultra-green cars is 96.52 mpg; that is, $\bar{x} = 96.52$. We can construct a confidence interval by using the point estimate as a base to which we add and subtract the margin of error.

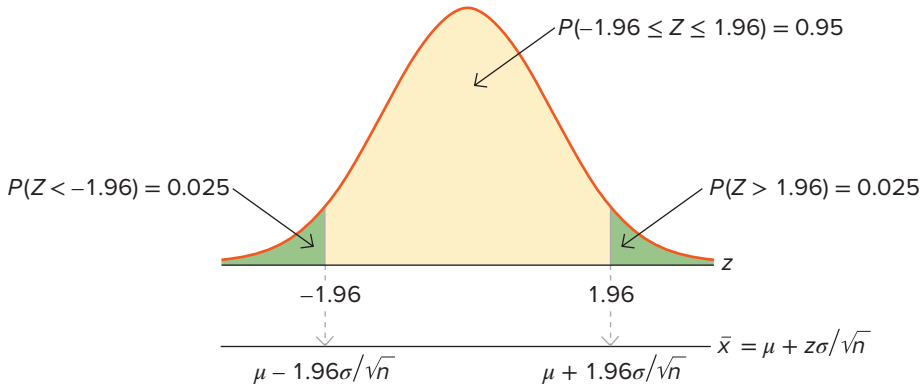
Constructing a Confidence Interval for μ When σ Is Known

LO **8.2**

Calculate a confidence interval for the population mean when the population standard deviation is known.

Let us construct the 95% confidence interval for μ when the sampling distribution of $\overline{X}$ is normal. Consider the standard normal random variable Z. See Figure 8.1; using the symmetry of Z, we can compute $P(Z > 1.96) = P(Z < -1.96) = 0.025$. Remember that $z = 1.96$ is easily determined from the z table given the probability of 0.025 in the upper tail of the distribution. Therefore, we formulate the probability statement $P(-1.96 \le Z \le 1.96) = 0.95$.

FIGURE 8.1 Graphical depiction of $P(Z < -1.96) = 0.025$ and $P(Z > 1.96) = 0.025$



Since $Z = \frac{X-\mu}{\sigma/\sqrt{n}}$, for a normally distributed $\overline{X}$ with mean μ and standard error $\sigma/\sqrt{n}$, we get

$$P\left(-1.96 \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right) = 0.95.$$

We multiply by $\sigma/\sqrt{n}$ and add μ to obtain

$$P(\mu - 1.96\sigma/\sqrt{n} \le \overline{X} \le \mu + 1.96\sigma/\sqrt{n}) = 0.95.$$

This equation (see also the lower portion of Figure 8.1) implies that there is a 0.95 probability that the sample mean $\overline{X}$ will fall between $\mu - 1.96\sigma/\sqrt{n}$ and $\mu + 1.96\sigma/\sqrt{n}$, that is, within the interval $\mu \pm 1.96\sigma/\sqrt{n}$. If samples of size n are drawn repeatedly from

a given population, 95% of the computed sample means, $\bar{x}$'s, will fall within the interval and the remaining 5% will fall outside the interval.

We do not know the population mean μ and, therefore, cannot determine if a particular $\bar{x}$ falls within the interval or not. However, we do know that $\bar{x}$ will fall within the interval $\mu \pm 1.96\sigma/\sqrt{n}$ if, and only if, μ falls within the interval $\bar{x} \pm 1.96\sigma/\sqrt{n}$. This will happen 95% of the time given how the interval is constructed. Therefore, we call the interval $\bar{x} \pm 1.96\sigma/\sqrt{n}$ the 95% confidence interval for the population mean, where $1.96\sigma/\sqrt{n}$ is its margin of error.

Confidence intervals are often misinterpreted; we need to exercise care in characterizing them. For instance, the above 95% confidence interval does *not* imply that the probability that μ falls in the confidence interval is 0.95. Remember that μ is a constant, although its value is not known. It either falls in the interval (probability equals one) or does not fall in the interval (probability equals zero). The randomness comes from $\overline{X}$, not μ, since many possible sample means can be derived from a population. Therefore, it is incorrect to say that the probability that μ falls in the $\bar{x} \pm 1.96\sigma/\sqrt{n}$ interval is 0.95. The 95% confidence interval simply implies that if numerous samples of size n are drawn from a given population, then 95% of the intervals formed by the preceding procedure (formula) will contain μ. Keep in mind that we only use a single sample to derive the estimates. Since there are many possible samples, we will be right 95% of the time, thus giving us 95% confidence.

INTERPRETING THE 95% CONFIDENCE INTERVAL

Technically, the 95% confidence interval for the population mean μ implies that for 95% of the samples, the procedure (formula) produces an interval that contains μ. Informally, we can report with 95% confidence that μ lies in the given interval. It is not correct to say that there is a 95% chance that μ lies in the given interval.

EXAMPLE 8.1

A sample of 25 cereal boxes of Granola Crunch, a generic brand of cereal, yields a mean weight of 1.02 pounds of cereal per box. Construct the 95% confidence interval for the mean weight of all cereal boxes. Assume that the weight is normally distributed with a population standard deviation of 0.03 pound.

SOLUTION: Note that the normality condition of $\overline{X}$ is satisfied since the underlying population is normally distributed. The 95% confidence interval for the population mean is computed as

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}} = 1.02 \pm 1.96\frac{0.03}{\sqrt{25}} = 1.02 \pm 0.012.$$

With 95% confidence, we can report that the mean weight of all cereal boxes falls between 1.008 and 1.032 pounds.

While it is common to report the 95% confidence interval, in theory we can construct an interval of any level of confidence ranging from 0 to 100%. Let's now extend the analysis to include intervals of any confidence level. Let the Greek letter α (alpha) denote the allowed probability of error; in Chapter 9 this is referred to as the significance level. This is the probability that the estimation procedure will generate an interval that does not contain μ. The **confidence coefficient** $(1 - \alpha)$ is interpreted as the probability that

the estimation procedure will generate an interval that contains μ. Thus, the probability of error α is related to the confidence coefficient and the confidence level as follows:

- Confidence coefficient $= 1 - \alpha$, and
- Confidence level $= 100(1 - \alpha)\%$.

For example, the confidence coefficient of 0.95 implies that the probability of error α equals $1 - 0.95 = 0.05$ and the confidence level equals $100(1 - 0.05)\% = 95\%$. Similarly, for the 90% confidence interval, the confidence coefficient equals 0.90 and $\alpha = 1 - 0.90 = 0.10$. The following statement generalizes the construction of a confidence interval for μ when σ is known.

CONFIDENCE INTERVAL FOR μ WHEN σ IS KNOWN

A $100(1 - \alpha)\%$ confidence interval for the population mean μ when the population standard deviation σ is known is computed as

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

This formula is valid only if $\overline{X}$ (approximately) follows a normal distribution.

The notation $z_{\alpha/2}$ is the z value associated with the probability of $\alpha/2$ in the upper tail of the standard normal probability distribution. In other words, if Z is a standard normal random variable and α is any probability, then $z_{\alpha/2}$ represents the z value such that the area under the z curve to the right of $z_{\alpha/2}$ is $\alpha/2$, that is, $P(Z \geq z_{\alpha/2}) = \alpha/2$. Figure 8.2 depicts the notation $z_{\alpha/2}$.

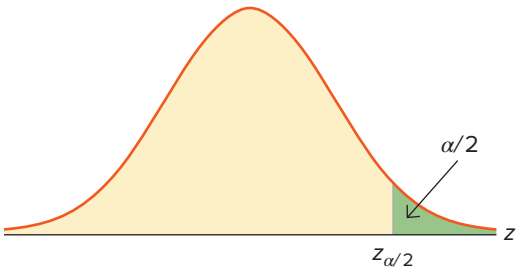


FIGURE 8.2 Graphical depiction of the notation $z_{\alpha/2}$

As discussed earlier, for the 95% confidence interval, $\alpha = 0.05$ and $\alpha/2 = 0.025$. Therefore, $z_{\alpha/2} = z_{0.025} = 1.96$. Similarly, using the z table, we can derive the following:

- For the 90% confidence interval, $\alpha = 0.10$, $\alpha/2 = 0.05$, and $z_{\alpha/2} = z_{0.05} = 1.645$.
- For the 99% confidence interval, $\alpha = 0.01$, $\alpha/2 = 0.005$, and $z_{\alpha/2} = z_{0.005} = 2.576$.

These values can also be obtained using Excel's **NORM.S.INV** function, which was first discussed in Chapter 6.

The Width of a Confidence Interval

LO 8.3

Describe the factors that influence the width of a confidence interval.

The margin of error used in the computation of the confidence interval for the population mean, when the population standard deviation is known, is $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Since we are basically adding and subtracting this quantity from $\bar{x}$, the width of the confidence interval is two times the margin of error. In Example 8.1, the margin of error for the 95% confidence interval is 0.012 and the width of the interval is $1.032 - 1.008 = 0.024 = 2(0.012)$. Now let's examine how the width of a confidence interval is influenced by various factors.

I. For a given confidence level $100(1-\alpha)\%$ and sample size n, the larger the population standard deviation σ, the wider the confidence interval.

EXAMPLE 8.1b

Let the standard deviation of the population in Example 8.1 be 0.05 instead of 0.03. Compute the 95% confidence interval based on the same sample information.

SOLUTION: We use the same formula as before, but we substitute 0.05 for the standard deviation instead of 0.03:

$$1.02 \pm 1.96 \frac{0.05}{\sqrt{25}} = 1.02 \pm 0.020.$$

The width has increased from 0.024 to $2(0.020) = 0.040$.

II. For a given confidence level $100(1-\alpha)\%$ and population standard deviation σ, the smaller the sample size n, the wider the confidence interval.

EXAMPLE 8.1c

Instead of 25 observations, let the sample in Example 8.1 be based on 16 observations. Compute the 95% confidence interval using the same sample mean of 1.02 pounds and the same population standard deviation of 0.03.

SOLUTION: Again, we use the same formula as before, but this time we substitute 16 for n instead of 25:

$$1.02 \pm 1.96 \frac{0.03}{\sqrt{16}} = 1.02 \pm 0.015.$$

The width has increased from 0.024 to $2(0.015) = 0.030$.

III. For a given sample size n and population standard deviation σ, the greater the confidence level $100(1-\alpha)\%$, the wider the confidence interval.

EXAMPLE 8.1d

Compute the 99%, instead of the 95%, confidence interval based on the information in Example 8.1.

SOLUTION: Now we use the same formula and substitute the value 2.576 for $z_{\alpha/2}$ instead of 1.96:

$$1.02 \pm 2.576 \frac{0.03}{\sqrt{25}} = 1.02 \pm 0.015.$$

The width has increased from 0.024 to $2(0.015) = 0.030$.

The precision is directly linked with the width of the confidence interval—the wider the interval, the lower its precision. Continuing with the weather analogy, a temperature estimate of 40 to 80 degrees is imprecise because the interval is too wide to be of value. We lose precision when the sample does not reveal a great deal about the population, resulting in a wide confidence interval. Examples 8.1b and 8.1c suggest that the estimate will be less precise if the variability of the underlying population is high (σ is high) or a small segment of the population is sampled (n is small). Example 8.1d relates the width with the confidence level. For given sample information, the only way we can gain confidence is by making the interval wider. If you are 95% confident that the outside temperature is between 40 and 60, then you can increase your confidence level to 99% only by using a wider range, say between 35 and 65. This result also helps us understand the difference between precision (width of the interval) and the confidence level. There is a trade-off between the amount of confidence we have in an interval and its width.

EXAMPLE 8.2

IQ tests are designed to yield scores that are approximately normally distributed. A reporter is interested in estimating the average IQ of employees in a large high-tech firm in California. She gathers the IQ scores from 22 employees of this firm and records the sample mean IQ as 106. She assumes that the population standard deviation is 15.

a. Compute 90% and 99% confidence intervals for the average IQ in this firm.

b. Use these results to infer if the mean IQ in this firm is different from the national average of 100.

SOLUTION:

a. For the 90% confidence interval, $z_{\alpha/2} = z_{0.05} = 1.645$. Similarly, for the 99% confidence interval, $z_{\alpha/2} = z_{0.005} = 2.576$.

The 90% confidence interval is $106 \pm 1.645 \frac{15}{\sqrt{22}} = 106 \pm 5.26$.

The 99% confidence interval is $106 \pm 2.576 \frac{15}{\sqrt{22}} = 106 \pm 8.24$.

Note that the 99% interval is wider than the 90% interval.

b. With 90% confidence, the reporter can infer that the average IQ of this firm's employees differs from the national average, since the value 100 falls outside the 90% confidence interval, [100.74, 111.26]. However, she cannot infer the same result with 99% confidence, since the wider range of the interval, [97.76, 114.24], includes the value 100. We will study the link between estimation and testing in more detail in the next chapter.

Using Excel to Construct a Confidence Interval for μ When σ Is Known

We can use functions in Excel to construct confidence intervals. These functions are particularly useful with large data sets. Consider the following example.

EXAMPLE 8.3

Table 8.2 lists a portion of the weights (in grams) for a sample of 80 hockey pucks. Construct the 90% confidence interval for the population mean weight assuming that the population standard deviation is 7.5 grams.

TABLE 8.2 Hockey Puck Weights, $n = 80$

FILE *Hockey_Pucks*

Weight
162.2
159.8
⋮
171.3

SOLUTION: We use Excel to find the lower and upper limits of the confidence interval: $[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$. We are given $\sigma = 7.5$ and $n = 80$.

a. Open the ***Hockey_Pucks*** data file. Note that the values for weights are in cells A2 through A81.

b. Recall from Chapter 6 that Excel's **NORM.S.INV** function finds a particular z value for a given cumulative probability. For the 90% confidence interval, $\alpha = 0.10$ and $z_{\alpha/2} = z_{0.05}$. To find the z value such that the area under the z curve to the right of $z_{0.05}$ is 0.05 (and area to the left of $z_{0.05}$ is 0.95), we use "NORM.S.INV(0.95)". For the lower limit of the confidence level, we enter "=AVERAGE(A2:A81) − NORM.S.INV(0.95) * 7.5/SQRT(80)", and Excel returns 165.33. For the upper limit of the confidence level, we enter "=AVERAGE(A2:A81) + NORM.S.INV(0.95) * 7.5/SQRT(80)", and Excel returns 168.09. With 90% confidence, we conclude that the mean weight of all hockey pucks falls between 165.33 and 168.09 grams.

EXERCISES 8.1

Mechanics

1. Find $z_{\alpha/2}$ for each of the following confidence levels used in estimating the population mean.
   a. 90%
   b. 98%
   c. 88%
2. Find $z_{\alpha/2}$ for each of the following confidence levels used in estimating the population mean.
   a. 89%
   b. 92%
   c. 96%
3. A simple random sample of 25 observations is derived from a normally distributed population with a known standard deviation of 8.2.
   a. Is the condition that $\bar{X}$ is normally distributed satisfied? Explain.
   b. Compute the margin of error with 80% confidence.
   c. Compute the margin of error with 90% confidence.
   d. Which of the two margins of error will lead to a wider interval?
4. Consider a population with a known standard deviation of 26.8. In order to compute an interval estimate for the population mean, a sample of 64 observations is drawn.
   a. Is the condition that $\bar{X}$ is normally distributed satisfied? Explain.
   b. Compute the margin of error at the 95% confidence level.
   c. Compute the margin of error at the 95% confidence level based on a larger sample of 225 observations.
   d. Which of the two margins of error will lead to a wider confidence interval?
5. Discuss the factors that influence the margin of error for the confidence interval for the population mean. What can a practitioner do to reduce the margin of error?
6. FILE ***Excel_1.*** Given the accompanying sample data, use Excel's formula options to find the 95% confidence interval for the population mean. Assume that the population is normally distributed and that the population standard deviation equals 12.
7. FILE ***Excel_2.*** Given the accompanying sample data, use Excel's formula options to find the 90% confidence interval for the population mean. Assume that the population is normally distributed and that the population standard deviation equals 5.

Applications

8. The average life expectancy for Bostonians is 78.1 years (*The Boston Globe,* August 16, 2010). Assume that this average was based on a sample of 50 Bostonians and that the population standard deviation is 4.5 years.
   a. What is the point estimate of the population mean?
   b. At 90% confidence, what is the margin of error?
   c. Construct the 90% confidence interval for the population average life expectancy of Bostonians.

9. In order to estimate the mean 30-year fixed mortgage rate for a home loan in the United States, a random sample of 28 recent loans is taken. The average calculated from this sample is 5.25%. It can be assumed that 30-year fixed mortgage rates are normally distributed with a population standard deviation of 0.50%. Compute 90% and 99% confidence intervals for the population mean 30-year fixed mortgage rate.

10. An article in the *National Geographic News* ("U.S. Racking Up Huge Sleep Debt," February 24, 2005) argues that Americans are increasingly skimping on their sleep. A researcher in a small Midwestern town wants to estimate the mean weekday sleep time of its adult residents. He takes a random sample of 80 adult residents and records their weekday mean sleep time as 6.4 hours. Assume that the population standard deviation is fairly stable at 1.8 hours.
    a. Calculate the 95% confidence interval for the population mean weekday sleep time of all adult residents of this Midwestern town.
    b. Can we conclude with 95% confidence that the mean sleep time of all adult residents in this Midwestern town is not 7 hours?

11. A family is relocating from St. Louis, Missouri, to California. Due to an increasing inventory of houses in St. Louis, it is taking longer than before to sell a house. The wife is concerned and wants to know when it is optimal to put their house on the market. Her realtor friend informs them that the last 26 houses that sold in their neighborhood took an average time of 218 days to sell. The realtor also tells them that based on her prior experience, the population standard deviation is 72 days.
    a. What assumption regarding the population is necessary for making an interval estimate for the population mean?
    b. Construct the 90% confidence interval for the mean sale time for all homes in the neighborhood.

12. U.S. consumers are increasingly viewing debit cards as a convenient substitute for cash and checks. The average amount spent annually on a debit card is $7,790 (*Kiplinger's,* August 2007). Assume that this average was based on a sample of 100 consumers and that the population standard deviation is $500.
    a. At 99% confidence, what is the margin of error?
    b. Construct the 99% confidence interval for the population mean amount spent annually on a debit card.

13. Suppose the 95% confidence interval for the mean salary of college graduates in a town in Mississippi is given by [$36,080, $43,920]. The population standard deviation used for the analysis is known to be $12,000.
    a. What is the point estimate of the mean salary for all college graduates in this town?
    b. Determine the sample size used for the analysis.

14. A manager is interested in estimating the mean time (in minutes) required to complete a job. His assistant uses a sample of 100 observations to report the confidence interval as [14.355, 17.645]. The population standard deviation is known to be equal to 10 minutes.
    a. Find the sample mean time used to compute the confidence interval.
    b. Determine the confidence level used for the analysis.

15. FILE ***CT_Undergrad_Debt.*** A study reports that recent college graduates from New Hampshire face the highest average debt of $31,048 (*The Boston Globe,* May 27, 2012). A researcher from Connecticut wants to determine how recent undergraduates from that state fare. He collects data on debt from 40 recent undergraduates. A portion of the data is shown in the accompanying table. Assume that the population standard deviation is $5,000.

Debt
24040
19153
⋮
29329

    a. Construct the 95% confidence interval for the mean debt of all undergraduates from Connecticut.
    b. Use the 95% confidence interval to determine if the debt of Connecticut undergraduates differs from that of New Hampshire undergraduates.

16. FILE ***Hourly_Wage.*** An economist wants to estimate the mean hourly wage (in $) of all workers. She collects data on 50 hourly wage earners. A portion of the data is shown in the accompanying table. Assume that the population standard deviation is $6. Construct and interpret 90% and 99% confidence intervals for the mean hourly wage of all workers.

Hourly Wage
37.85
21.72
⋮
24.18

17. FILE ***Highway_Speeds.*** A safety officer is concerned about speeds on a certain section of the New Jersey Turnpike. He records the speeds of 40 cars on a Saturday afternoon. The accompanying table shows a portion of the results. Assume that the population standard deviation is 5 mph. Construct the 95% confidence interval for the mean speed of all cars on that section of the turnpike. Are the safety officer's concerns valid if the speed limit is 55 mph? Explain.

Highway Speeds
70
60
⋮
65

8.2 CONFIDENCE INTERVAL FOR THE POPULATION MEAN WHEN σ IS UNKNOWN

So far we have considered confidence intervals for the population mean when the population standard deviation σ is known. In reality, σ is rarely known. Recall from Chapter 3 that the population variance and the population standard deviation are calculated as $\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$ and $\sigma = \sqrt{\sigma^2}$, respectively. It is highly unlikely that σ is known when μ is not. However, there are instances when the population standard deviation is considered fairly stable and, therefore, can be determined from prior experience. In these cases, the population standard deviation is treated as known.

Recall that the margin of error in a confidence interval depends on the standard error of the estimator and the desired confidence level. With σ unknown, the standard error of $\overline{X}$, given by $\sigma/\sqrt{n}$, can be conveniently estimated by $s/\sqrt{n}$, where s denotes the sample standard deviation. For convenience, we denote this estimate of the standard error of $\overline{X}$ also by $se(\overline{X}) = s/\sqrt{n}$.

LO **8.4**

Discuss features of the *t* distribution.

The *t* Distribution

As discussed earlier, in order to derive a confidence interval for μ, it is essential that $\overline{X}$ be normally distributed. A normally distributed $\overline{X}$ is standardized as $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ where Z follows the z distribution. Another standardized statistic, which uses the estimator S in place of σ, is computed as $T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$. The random variable T follows the **Student's *t* distribution**, more commonly known as the ***t* distribution**.[1]

THE *t* DISTRIBUTION

If a random sample of size n is taken from a normal population with a finite variance, then the statistic $T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$ follows the t distribution with $(n - 1)$ degrees of freedom, *df*.

The t distribution is actually a family of distributions, which are similar to the z distribution in that they are all bell-shaped and symmetric around zero. However, all t distributions have slightly broader tails than the z distribution. Each t distribution is identified by the **degrees of freedom**, or simply, *df*. The degrees of freedom determine the extent of the broadness of the tails of the distribution; the fewer the degrees of freedom, the broader the tails. Since the t distribution is defined by the degrees of freedom, it is common to refer to it as the t_{df} distribution.

Specifically, the degrees of freedom refer to the number of independent pieces of information that go into the calculation of a given statistic and, in this sense, can be "freely chosen." Consider the number of independent observations that enter into the calculation of the sample mean. If it is known that $\bar{x} = 20$, $n = 4$, and three of the observations have values of $x_1 = 16$, $x_2 = 24$, and $x_3 = 18$, then there is no choice but for the fourth observation to have a value of 22. In other words, three degrees of freedom are involved in computing $\bar{x} = 20$ if $n = 4$; in effect, one degree of freedom is lost.

Summary of the t_{df} Distribution

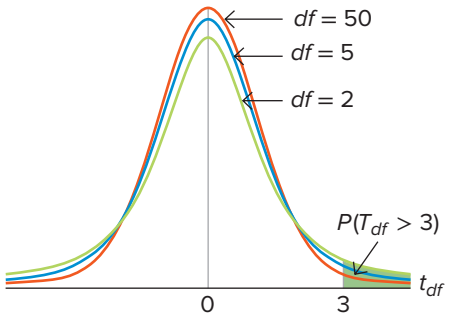
- Like the z distribution, the t_{df} distribution is bell-shaped and symmetric around 0 with asymptotic tails (the tails get closer and closer to the horizontal axis but never touch it).

[1]William S. Gossett (1876–1937) published his research concerning the t distribution under the pen name "Student" because his employer, the Guinness Brewery, did not allow employees to publish their research results.

- The t_{df} distribution has slightly broader tails than the z distribution.
- The t_{df} distribution consists of a family of distributions where the actual shape of each one depends on the degrees of freedom *df*. As *df* increases, the t_{df} distribution becomes similar to the z distribution; it is identical to the z distribution when *df* approaches infinity.

From Figure 8.3 we note that the tails of the t_2 and t_5 distributions are broader than the tails of the t_{50} distribution. For instance, for t_2 and t_5, the area exceeding a value of 3, or $P(T_{df} > 3)$, is greater than that for t_{50}. In addition, the t_{50} resembles the z distribution.

FIGURE 8.3
The t_{df} distribution with various degrees of freedom



Locating t_{df} Values and Probabilities

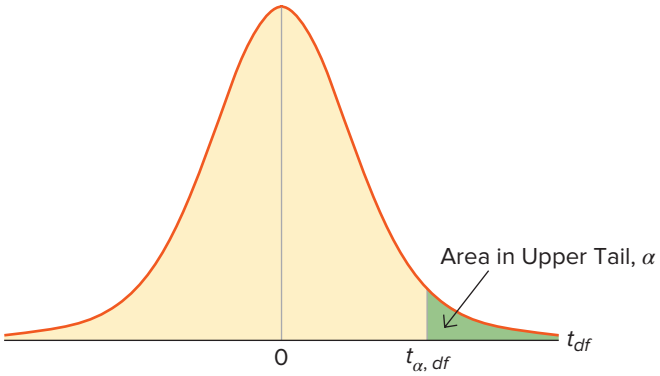
Table 8.3 lists t_{df} values for selected upper-tail probabilities and degrees of freedom *df*. Table 2 of Appendix A provides a more complete table. Since the t_{df} distribution is a family of distributions identified by the *df* parameter, the t table is not as comprehensive as the z table. It only lists probabilities corresponding to a limited number of values. Also, unlike the cumulative probabilities in the z table, the t table provides the probabilities in the upper tail of the distribution.

TABLE 8.3 Portion of the t Table

	Area in Upper Tail, α					
df	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	1.376	3.078	6.314	12.706	31.821	63.657
⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	0.879	1.372	**1.812**	2.228	2.764	3.169
⋮	⋮	⋮	⋮	⋮	⋮	⋮
∞	0.842	1.282	1.645	1.960	2.326	2.576

We use the notation $t_{\alpha,df}$ to denote a value such that the area in the upper tail equals α for a given *df*. In other words, for a random variable T_{df}, the notation $t_{\alpha,df}$ represents a value such that $P(T_{df} \geq t_{\alpha,df}) = \alpha$. Figure 8.4 illustrates the notation. Similarly, $t_{\alpha/2,df}$ represents a value such that $P(T_{df} \geq t_{\alpha/2,df}) = \alpha/2$.

FIGURE 8.4
Graphical depiction of $P(T_{df} \geq t_{\alpha,df}) = \alpha$



When determining the value $t_{\alpha,df}$, we need two pieces of information: (a) the sample size n or, analogously, $df = n - 1$, and (b) α. For instance, suppose we want to find the value $t_{\alpha,df}$ with $\alpha = 0.05$ and $df = 10$; that is, $t_{0.05,10}$. Using Table 8.3, we look at the first column labeled df and find the row 10. We then continue along this row until we reach the column $\alpha = 0.05$. The value 1.812 suggests that $P(T_{10} \geq 1.812) = 0.05$. Due to the symmetry of the t distribution, we also get $P(T_{10} \leq -1.812) = 0.05$. Figure 8.5 shows these results graphically. Also, since the area under the entire t_{df} distribution sums to one, we deduce that $P(T_{10} < 1.812) = 1 - 0.05 = 0.95$, which also equals $P(T_{10} > -1.812)$.

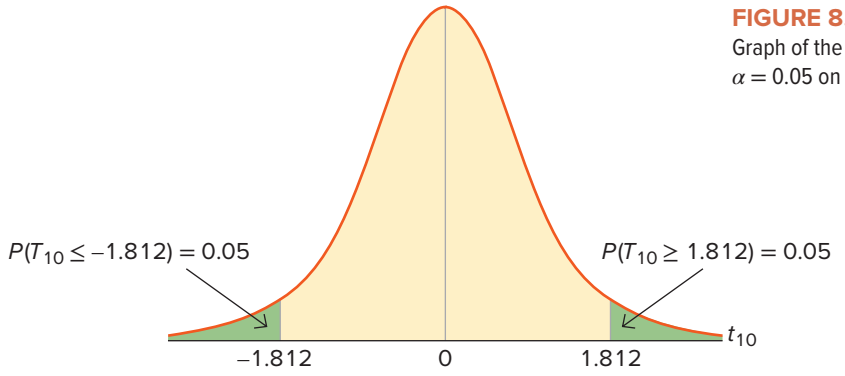


FIGURE 8.5
Graph of the probability $\alpha = 0.05$ on both sides of T_{10}

Sometimes the exact probability cannot be determined from the t table. For example, given $df = 10$, the exact probability $P(T_{10} \geq 1.562)$ is not included in the table. However, this probability is between 0.05 and 0.10 because the value 1.562 falls between 1.372 and 1.812. Similarly, $P(T_{10} < 1.562)$ is between 0.90 and 0.95. We can use Excel to find exact probabilities.

EXAMPLE 8.4

Compute $t_{\alpha,df}$ for $\alpha = 0.025$ using 2, 5, and 50 degrees of freedom.

SOLUTION:

- For $df = 2$, $t_{0.025,2} = 4.303$.
- For $df = 5$, $t_{0.025,5} = 2.571$.
- For $df = 50$, $t_{0.025,50} = 2.009$.

Note that the t_{df} values change with the degrees of freedom. Moreover, as df increases, the t_{df} distribution begins to resemble the z distribution. In fact, with $df = \infty$, $t_{0.025,\infty} = 1.96$, which is identical to the corresponding z value; recall that $P(Z \geq 1.96) = 0.025$.

LO 8.5

Calculate a confidence interval for the population mean when the population standard deviation is not known.

Constructing a Confidence Interval for μ When σ Is Unknown

We can never stress enough the importance of the requirement that $\overline{X}$ follows a normal distribution in estimating the population mean. Recall that $\overline{X}$ follows the normal distribution when the underlying population is normally distributed or when the sample size is sufficiently large ($n \geq 30$). We still construct the confidence interval for μ as point estimate $\pm$ margin of error. However, when the population standard deviation is unknown, we now use the t_{df} distribution to calculate the margin of error.

CONFIDENCE INTERVAL FOR μ WHEN σ IS NOT KNOWN

A $100(1-\alpha)\%$ confidence interval for the population mean μ when the population standard deviation σ is not known is computed as

$$\bar{x} \pm t_{\alpha/2,df}\frac{s}{\sqrt{n}} \quad \text{or} \quad \left[\bar{x} - t_{\alpha/2,df}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,df}\frac{s}{\sqrt{n}}\right],$$

where s is the sample standard deviation. This formula is valid only if $\overline{X}$ (approximately) follows a normal distribution.

As before, $100(1-\alpha)\%$ is the confidence level and $t_{\alpha/2,df}$ is the t_{df} value associated with the probability $\alpha/2$ in the upper tail of the distribution with $df = n - 1$. In other words, $P(T_{df} > t_{\alpha/2,df}) = \alpha/2$. It is important to note that uncertainty is increased when we estimate the population standard deviation with the sample standard deviation, making the confidence interval wider, especially for smaller samples. This is appropriately captured by the wider tail of the t_{df} distribution.

EXAMPLE 8.5

FILE
MPG

In the introductory case of this chapter, Jared Beane wants to estimate the mean mpg for all ultra-green cars. Table 8.1 lists the mpg of a sample of 25 cars. Use this information to construct the 90% confidence interval for the population mean. Assume that mpg follows a normal distribution.

SOLUTION: The condition that $\overline{X}$ follows a normal distribution is satisfied since we assumed that mpg is normally distributed. Thus, we construct the confidence interval as $\bar{x} \pm t_{\alpha/2,df}\frac{s}{\sqrt{n}}$. This is a classic example where a statistician has access only to sample data. Since the population standard deviation is not known, the sample standard deviation has to be computed from the sample. From the sample data in Table 8.1, we find that $\bar{x} = \frac{\Sigma x_i}{n} = \frac{2{,}413}{25} = 96.52$ mpg and $s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{2{,}746.24}{25-1}} = 10.70$. For the 90% confidence interval, $\alpha = 0.10$, $\alpha/2 = 0.05$, and, given $n = 25$, $df = 25 - 1 = 24$. Thus, $t_{0.05,24} = 1.711$.

The 90% confidence interval for μ is computed as

$$\bar{x} \pm t_{\alpha/2,df}\frac{s}{\sqrt{n}} = 96.52 \pm 1.711\frac{10.70}{\sqrt{25}} = 96.52 \pm 3.66.$$

Thus, Jared concludes with 90% confidence that the average mpg of all ultra-green cars is between 92.86 mpg and 100.18 mpg. Note that the manufacturer's claim that the ultra-green car will average 100 mpg cannot be rejected by the sample data since the value 100 falls within the 90% confidence interval.

Using Excel to Construct a Confidence Interval for μ When σ Is Unknown

Again we find that functions in Excel are quite useful when constructing confidence intervals. Consider the following example.

EXAMPLE 8.6

A recent article found that Massachusetts residents spent an average of $860.70 on the lottery in 2010 (www.businessweek.com, March 14, 2012). In order to verify the results, a researcher at a Boston think tank surveys 100 Massachusetts residents and asks them about their annual lottery expenditures (in $). Table 8.4 shows a portion of the results. Construct the 95% confidence interval for the average annual expenditures on the lottery for all Massachusetts residents. Do the results dispute the article's claim? Explain.

TABLE 8.4 Massachusetts Residents' Annual Lottery Expenditures, $n = 100$

FILE *Lottery*

Expenditures
790
594
⋮
759

SOLUTION: We use Excel to find the lower and upper limits of the confidence interval: $[\bar{x} - t_{\alpha/2,df}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,df}\frac{s}{\sqrt{n}}]$. We are given $n = 100$.

a. Open the ***Lottery*** data file. Note that the values are in cells A2 through A101.

b. We apply a procedure analogous to the one used in Section 8.1, except we use Excel's **T.INV** function instead of the **NORM.S.INV** function. The **T.INV** function finds a particular t_{df} value for a given cumulative probability (*cumulprob*). If we want to find t_{df} to satisfy $P(T_{df} < t_{df}) = cumulprob$, we use "T.INV(*cumulprob*, *df*)". For the 95% confidence interval, $\alpha/2 = 0.025$ and $df = n - 1 = 99$. Since we want to find $t_{\alpha/2,df}$ such that the area under the t_{df} curve to the right of this value is $\alpha/2$ (and the area to the left of this value is $1 - \alpha/2$), we use "T.INV(0.975,99)". In order to find the lower limit of the confidence interval, we enter "=AVERAGE(A2:A101) − T.INV(0.975,99) * STDEV.S(A2:A101)/SQRT(100)", and Excel returns 798.85. For the upper limit of the confidence interval, we enter "=AVERAGE(A2:A101) + T.INV(0.975,99) * STDEV.S(A2:A101)/SQRT(100)", and Excel returns 885.03. With 95% confidence, we conclude that the average annual expenditures on the lottery for all Massachusetts residents fall between $798.85 and $885.03. The results do not dispute the article's claim since the interval includes the reported mean value of $860.70.

EXERCISES 8.2

Mechanics

18. Find $t_{\alpha,df}$ from the following information.
    a. $\alpha = 0.025$ and $df = 12$
    b. $\alpha = 0.10$ and $df = 12$
    c. $\alpha = 0.025$ and $df = 25$
    d. $\alpha = 0.10$ and $df = 25$

19. We use the *t* distribution to construct a confidence interval for the population mean when the underlying population standard deviation is not known. Under the assumption that the population is normally distributed, find $t_{\alpha/2,df}$ for the following scenarios.
    a. A 90% confidence level and a sample of 28 observations.
    b. A 95% confidence level and a sample of 28 observations.
    c. A 90% confidence level and a sample of 15 observations.
    d. A 95% confidence level and a sample of 15 observations.

20. A random sample of 24 observations is used to estimate the population mean. The sample mean and the sample standard deviation are calculated as 104.6 and 28.8, respectively. Assume that the population is normally distributed.
    a. Construct the 90% confidence interval for the population mean.

b. Construct the 99% confidence interval for the population mean.
c. Use your answers to discuss the impact of the confidence level on the width of the interval.

21. Consider a normal population with an unknown population standard deviation. A random sample results in $\bar{x} = 48.68$ and $s^2 = 33.64$.
    a. Compute the 95% confidence interval for μ if $\bar{x}$ and s^2 were obtained from a sample of 16 observations.
    b. Compute the 95% confidence interval for μ if $\bar{x}$ and s^2 were obtained from a sample of 25 observations.
    c. Use your answers to discuss the impact of the sample size on the width of the interval.

22. Let the following sample of 8 observations be drawn from a normal population with unknown mean and standard deviation: 22, 18, 14, 25, 17, 28, 15, 21.
    a. Calculate the sample mean and the sample standard deviation.
    b. Construct the 80% confidence interval for the population mean.
    c. Construct the 90% confidence interval for the population mean.
    d. What happens to the margin of error as the confidence level increases from 80% to 90%?

23. FILE ***Excel_1.*** Given the accompanying sample data, use Excel's formula options to find the 90% confidence interval for the population mean. Assume that the population is normally distributed.

24. FILE ***Excel_2.*** Given the accompanying sample data, use Excel's formula options to find the 99% confidence interval for the population mean. Assume that the population is normally distributed.

Applications

25. A random sample of eight drugstores shows the following prices (in $) of a popular pain reliever:

3.50	4.00	2.00	3.00	2.50	3.50	2.50	3.00

    Assume the normal distribution for the underlying population to construct the 90% confidence interval for the population mean.

26. A popular weight loss program claims that with its recommended healthy diet regimen, clients lose significant weight within a month. In order to estimate the mean weight loss of all clients, a nutritionist takes a sample of 18 clients and records their weight loss one month after joining the program. He computes the sample mean and the sample standard deviation of weight loss as 12.5 pounds and 9.2 pounds, respectively. He believes that weight loss is likely to be normally distributed.
    a. Calculate the margin of error with 95% confidence.
    b. Calculate the 95% confidence interval for the population mean.

27. The manager of The Cheesecake Factory in Boston reports that on six randomly selected weekdays, the number of customers served was 120, 130, 100, 205, 185, and 220. She believes that the number of customers served on weekdays follows a normal distribution. Construct the 90% confidence interval for the average number of customers served on weekdays.

28. According to a recent survey, high school girls average 100 text messages daily (*The Boston Globe,* April 21, 2010). Assume that the survey was based on a random sample of 36 high school girls. The sample standard deviation is computed as 10 text messages daily.
    a. Calculate the margin of error with 99% confidence.
    b. What is the 99% confidence interval for the population mean texts that all high school girls send daily?

29. The Chartered Financial Analyst (CFA) designation is fast becoming a requirement for serious investment professionals. Although it requires a successful completion of three levels of grueling exams, it also entails promising careers with lucrative salaries. A student of finance is curious about the average salary of a CFA charterholder. He takes a random sample of 36 recent charterholders and computes a mean salary of $158,000 with a standard deviation of $36,000. Use this sample information to determine the 95% confidence interval for the average salary of a CFA charterholder.

30. The sudoku puzzle has become very popular all over the world. It is based on a 9×9 grid and the challenge is to fill in the grid so that every row, every column, and every 3×3 box contains the digits 1 through 9. A researcher is interested in estimating the average time taken by a college student to solve the puzzle. He takes a random sample of 8 college students and records their solving times (in minutes) as 14, 7, 17, 20, 18, 15, 19, 28.
    a. Construct the 99% confidence interval for the average time taken by a college student to solve a sudoku puzzle.
    b. What assumption is necessary to make this inference?

31. Executive compensation has risen dramatically compared to the rising levels of an average worker's wage over the years. Sarah is an MBA student who decides to use her statistical skills to estimate the mean CEO compensation in 2010 for all large companies in the United States. She takes a random sample of six CEO compensations (in $ millions) as shown in the accompanying table.

Firm	Compensation
Intel	8.20
Coca-Cola	2.76
Wells Fargo	6.57
Caterpillar	3.88
McDonald's	6.56
U.S. Bancorp	4.10

Source: www.finance.yahoo.com.

a. Help Sarah use the information to construct the 90% confidence interval for the mean CEO compensation for all large companies in the United States.
b. What assumption is necessary for deriving the interval estimate?
c. How can the margin of error reported in part a be reduced?

32. As reported by tradingeconomics.com on September 2, 2012, the unemployment rates (in %) in major economies around the world were as follows:

Country	Unemployment Rate
Australia	5.2
China	4.1
France	10.0
Germany	6.8
India	3.8
United Kingdom	8.0
United States	8.3

a. Calculate the margin of error used in the 95% confidence level for the population mean unemployment rate. Explain the assumption made for the analysis.
b. How can we reduce the margin of error for the 95% confidence interval?

33. A price-earnings ratio, or P/E ratio, is calculated as a firm's share price compared to the income or profit earned by the firm per share. Generally, a high P/E ratio suggests that investors are expecting higher earnings growth in the future compared to companies with a lower P/E ratio. The following table shows the P/E ratios for a sample of firms in the footwear industry:

Firm	P/E Ratio
Brown Shoe Co., Inc.	26
CROCS, Inc.	13
DSW, Inc.	21
Foot Locker, Inc.	16
Nike, Inc.	21

Source: www.finance.yahoo.com, data retrieved September 2, 2012.

Let these ratios represent a random sample drawn from a normally distributed population. Construct the 90% confidence interval for the mean P/E ratio for the entire footwear industry.

34. Suppose the 90% confidence interval for the mean SAT scores of applicants at a business college is given by [1690, 1810]. This confidence interval uses the sample mean and the sample standard deviation based on 25 observations. What are the sample mean and the sample standard deviation used when computing the interval?

35. The monthly closing stock prices (rounded to the nearest dollar) for Panera Bread Co. for the first six months of 2010 are reported in the following table.

Month	Closing Stock Price
January	71
February	73
March	76
April	78
May	81
June	75

Source: www.finance.yahoo.com.

a. Calculate the sample mean and the sample standard deviation.
b. Calculate the 90% confidence interval for the mean stock price of Panera Bread Co., assuming that the stock price is normally distributed.
c. What happens to the margin of error if a higher confidence level is used for the interval estimate?

36. The accompanying table shows the annual returns (in percent) for Fidelity's Electronics and Utilities funds.
a. Derive the 99% confidence intervals for the mean returns for Fidelity's Electronics and Utilities funds.
b. What did you have to assume to make the inferences in part a?

Year	Electronics	Utilities
2010	17	11
2011	–8	13
2012	4	7
2013	39	21
2014	38	22

Source: www.finance.yahoo.com, data retrieved April 3, 2015.

37. A teacher wants to estimate the mean time (in minutes) that students take to go from one classroom to the next. His research assistant uses the sample time of 36 students to report the confidence interval as [8.20, 9.80].
a. Find the sample mean time used to compute the confidence interval.
b. Determine the confidence level if the sample standard deviation used for the interval is 2.365.

38. In order to attract more Millennial customers (those born between 1980 and 2000), a new clothing store offers free gourmet coffee and pastry to its customers. The average daily revenue over the past five-week period has been $1,080 with a standard deviation of $260. Use this sample information to construct the 95% confidence interval for the average daily revenue. The store manager believes that the coffee and pastry strategy would lead to an average daily revenue of $1,200. Use the 95% interval to determine if the manager is wrong.

39. FILE ***Debt_Payments.*** A study found that consumers are making average monthly debt payments of $983 (Experian.com, November 11, 2010). The accompanying table shows a portion of average monthly debt payments (Debt in $) for 26 metropolitan areas. Construct 90% and 95% confidence intervals for the population mean. Comment on the width of the intervals.

City	Debt
Washington, D.C.	1285
Seattle	1135
⋮	⋮
Pittsburgh	763

Source: www.Experian.com, November 11, 2010.

40. FILE ***Economics.*** An associate dean of a university wishes to compare the means on the standardized final exams in microeconomics and macroeconomics. He has access to a random sample of 40 scores from each of these two courses. A portion of the data is shown in the accompanying table.

Micro	Macro
85	48
78	79
⋮	⋮
75	74

a. Construct 95% confidence intervals for the mean score in microeconomics and the mean score in macroeconomics.
b. Explain why the widths of the two intervals are different.

41. FILE ***Math_Scores.*** For decades, people have believed that boys are innately more capable than girls in math. In other words, due to the intrinsic differences in brains, boys are better suited for doing math than girls. Recent research challenges this stereotype, arguing that gender differences in math performance have more to do with culture than innate aptitude. Others argue, however, that while the average may be the same, there is more variability in math ability for boys than girls, resulting in some boys with soaring math skills. A portion of the data on math scores of boys and girls is shown in the accompanying table.

Boys	Girls
74	83
89	76
⋮	⋮
66	74

a. Construct 95% confidence intervals for the mean scores of boys and the mean scores of girls. Explain your assumptions.
b. Explain why the widths of the two intervals are different.

42. FILE ***Startups.*** Many of today's leading companies, including Google, Microsoft, and Facebook, are based on technologies developed within universities. Lisa Fisher is a business school professor who believes that a university's research expenditure (Research in $ millions) and the age of its technology transfer office (Duration in years) are major factors that enhance innovation. She wants to know what the average values are for the Research and the Duration variables. She collects data from 143 universities on these variables for the academic year 2008. A portion of the data is shown in the accompanying table.

Research	Duration
145.52	23
237.52	23
⋮	⋮
154.38	9

Source: Association of University Managers and National Science Foundation.

a. Construct and interpret the 95% confidence interval for the mean research expenditure of all universities.
b. Construct and interpret the 95% confidence interval for the mean duration of all universities.

8.3 CONFIDENCE INTERVAL FOR THE POPULATION PROPORTION

LO 8.6
Calculate a confidence interval for the population proportion.

Sometimes the parameter of interest describes a population that is qualitative rather than quantitative. Recall that while the population mean μ and the population variance σ^2 describe quantitative data, the population proportion p is the essential descriptive measure when the data type is qualitative. The parameter p represents the proportion of successes in the population, where success is defined by a particular outcome. Examples of population proportions include the proportion of women students at a university, the proportion of defective items in a manufacturing process, and the default probability on a mortgage loan.

As in the case of the population mean, we estimate the population proportion on the basis of its sample counterpart. In particular, we use the sample proportion $\overline{P}$ as the point

estimator of the population proportion p. Also, although the sampling distribution of $\overline{P}$ is based on a binomial distribution, we can approximate it by a normal distribution for large samples, according to the central limit theorem. This approximation is valid when the sample size n is such that $np \geq 5$ and $n(1-p) \geq 5$.

Using the normal approximation for $\overline{P}$ with $E(\overline{P}) = p$ and $se(\overline{P}) = \sqrt{p(1-p)/n}$, and analogous to the derivation of the confidence interval for the population mean, a $100\,(1-\alpha)\%$ confidence interval for the population proportion is

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \quad \text{or} \quad \left[\bar{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \bar{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right].$$

This confidence interval is theoretically sound; however, it cannot be implemented because it uses p in the derivation, which is unknown. Since we always use large samples for the normal distribution approximation, we can also conveniently replace p with its estimate $\bar{p}$ in the construction of the interval. Therefore, for $\sqrt{\frac{p(1-p)}{n}}$, we substitute $\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$. This substitution yields a feasible confidence interval for the population proportion.

CONFIDENCE INTERVAL FOR p

A $100(1-\alpha)\%$ confidence interval for the population proportion p is computed as

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad \text{or} \quad \left[\bar{p} - z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \bar{p} + z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}\right].$$

This formula is valid only if $\overline{P}$ (approximately) follows a normal distribution.

The normality condition is evaluated at the sample proportion $\bar{p}$. In other words, for constructing a confidence interval for the population proportion p, we require that $n\bar{p} \geq 5$ and $n(1-\bar{p}) \geq 5$.

EXAMPLE 8.7

FILE
MPG

In the introductory case of this chapter, Jared Beane wants to estimate the proportion of all ultra-green cars that obtain over 100 mpg. Use the information in Table 8.1 to construct 90% and 99% confidence intervals for the population proportion.

SOLUTION: As shown in Table 8.1, 7 of the 25 cars obtain over 100 mpg; thus, the point estimate of the population proportion is $\bar{p} = 7/25 = 0.28$. Note that the normality condition is satisfied, since $np \geq 5$ and $n(1-p) \geq 5$, where p is evaluated at $\bar{p} = 0.28$. With the 90% confidence level, $\alpha/2 = 0.10/2 = 0.05$; thus, we find $z_{\alpha/2} = z_{0.05} = 1.645$. Substituting the appropriate values into $\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$ yields

$$0.28 \pm 1.645\sqrt{\frac{0.28(1-0.28)}{25}} = 0.28 \pm 0.148.$$

With 90% confidence, Jared reports that the percentage of cars that obtain over 100 mpg is between 13.2% and 42.8%.

If Jared had wanted the 99% confidence level, we would use $\alpha/2 = 0.01/2 = 0.005$ and $z_{\alpha/2} = z_{0.005} = 2.576$ to obtain

$$0.28 \pm 2.576\sqrt{\frac{0.28(1-0.28)}{25}} = 0.28 \pm 0.231.$$

At a higher confidence level of 99%, the interval for the percentage of cars that obtain over 100 mpg becomes 4.9% to 51.1%. Given the current sample size of 25 cars, Jared can gain confidence (from 90% to 99%) at the expense of precision, as the corresponding margin of error increases from 0.148 to 0.231.

EXERCISES 8.3

Mechanics

43. A random sample of 80 observations results in 50 successes.
    a. Construct the 95% confidence interval for the population proportion of successes.
    b. Construct the 95% confidence interval for the population proportion of failures.
44. Assume $\bar{p} = 0.6$ in a sample of size $n = 50$.
    a. Construct the 95% confidence interval for the population proportion.
    b. What happens to the margin of error if the above sample proportion is based on $n = 200$ instead of $n = 50$?
45. A sample of 80 results in 30 successes.
    a. Calculate the point estimate for the population proportion of successes.
    b. Construct 90% and 99% confidence intervals for the population proportion.
    c. Can we conclude at 90% confidence that the population proportion differs from 0.5?
    d. Can we conclude at 99% confidence that the population proportion differs from 0.5?
46. A random sample of 100 observations results in 40 successes.
    a. What is the point estimate for the population proportion of successes?
    b. Construct 90% and 99% confidence intervals for the population proportion.
    c. Can we conclude at 90% confidence that the population proportion differs from 0.5?
    d. Can we conclude at 99% confidence that the population proportion differs from 0.5?
47. In a sample of 30 observations, the number of successes equals 18.
    a. Construct the 88% confidence interval for the population proportion of successes.
    b. Construct the 98% confidence interval for the population proportion of successes.
    c. What happens to the margin of error as you move from the 88% confidence interval to the 98% confidence interval?

Applications

48. A poll of 1,079 adults found that 51% of Americans support Arizona's stringent new immigration enforcement law, even though it may lead to racial profiling (*The New York Times*/CBS News, April 28–May 2, 2010). Use the sample information to compute the 95% confidence interval for the population parameter of interest.
49. A survey of 1,026 people asked: "What would you do with an unexpected tax refund?" Forty-seven percent responded that they would pay off debts (*Vanity Fair,* June 2010).
    a. At 95% confidence, what is the margin of error?
    b. Construct the 95% confidence interval for the population proportion of people who would pay off debts with an unexpected tax refund.
50. In a CNNMoney.com poll conducted on July 13, 2010, a sample of 5,324 Americans were asked about what matters most to them in a place to live. Thirty-seven percent of the respondents felt job opportunities matter most.
    a. Construct the 90% confidence interval for the proportion of Americans who feel that good job opportunities matter most in a place to live.
    b. Construct the 99% confidence interval for the proportion of Americans who feel that good job opportunities matter most in a place to live.
    c. Which of the above two intervals has a higher margin of error? Explain why.
51. An economist reports that 560 out of a sample of 1,200 middle-income American households actively participate in the stock market.
    a. Construct the 90% confidence interval for the proportion of middle-income Americans who actively participate in the stock market.
    b. Can we conclude that the percentage of middle-income Americans who actively participate in the stock market is not 50%?
52. In an NBC News/*Wall Street Journal* poll of 1,000 American adults conducted August 5–9, 2010, 44% of respondents approved of the job that Barack Obama was doing in handling the economy.
    a. Compute the 90% confidence interval for the proportion of Americans who approved of Barack Obama's handling of the economy.
    b. What is the resulting margin of error?
    c. Compute the margin of error associated with the 99% confidence level.

53. In a recent poll of 760 homeowners in the United States, one in five homeowners reports having a home equity loan that he or she is currently paying off. Using a confidence coefficient of 0.90, derive the interval estimate for the proportion of all homeowners in the United States that hold a home equity loan.

54. Obesity is generally defined as 30 or more pounds over a healthy weight. A recent study of obesity reports 27.5% of a random sample of 400 adults in the United States to be obese.
    a. Use this sample information to compute the 90% confidence interval for the adult obesity rate in the United States.
    b. Is it reasonable to conclude with 90% confidence that the adult obesity rate in the United States differs from 30%?

55. An accounting professor is notorious for being stingy in giving out good letter grades. In a large section of 140 students in the fall semester, she gave out only 5% A's, 23% B's, 42% C's, and 30% D's and F's. Assuming that this was a representative class, compute the 95% confidence interval of the probability of getting at least a B from this professor.

56. A survey conducted by CBS News asked 1,026 respondents: "What would you do with an unexpected tax refund?" The responses are summarized in the following table.

Response	Frequency
Pay off debts	482
Put it in the bank	308
Spend it	112
I never get a refund	103
Other	21

Source: *Vanity Fair*, June 2010.

    a. Construct the 90% confidence interval for the population proportion of those who would put the tax refund in the bank.
    b. Construct the 90% confidence interval for the population proportion of those who never get a refund.

57. A recent survey asked 5,324 individuals: What's most important to you when choosing where to live? The responses are shown by the following frequency distribution.

Response	Frequency
Good jobs	1,969
Affordable homes	799
Top schools	586
Low crime	1,225
Things to do	745

Source: CNNMoney.com, July 13, 2010.

    a. Calculate the margin of error used in the 95% confidence level for the population proportion of those who believe that low crime is most important.
    b. Calculate the margin of error used in the 95% confidence level for the population proportion of those who believe that good jobs or affordable homes are most important.
    c. Explain why the margins of error in parts a and b are different.

58. One in five 18-year-old Americans has not graduated from high school (*The Wall Street Journal,* April 19, 2007). A mayor of a Northeastern city comments that its residents do not have the same graduation rate as the rest of the country. An analyst from the Department of Education decides to test the mayor's claim. In particular, she draws a random sample of 80 18-year-olds in the city and finds that 20 of them have not graduated from high school.
    a. Compute the point estimate for the proportion of 18-year-olds who have not graduated from high school in this city.
    b. Use this point estimate to derive the 95% confidence interval for the population proportion.
    c. Can the mayor's comment be justified at 95% confidence?

LO **8.7**

Select a sample size to estimate the population mean and the population proportion.

8.4 SELECTING THE REQUIRED SAMPLE SIZE

So far we have discussed how a confidence interval provides useful information on an unknown population parameter. We compute the confidence interval by adding and subtracting the margin of error to/from the point estimate. If the margin of error is very large, the confidence interval becomes too wide to be of much value. For instance, little useful information can be gained from a confidence interval that suggests that the average annual starting salary of a business graduate is between $16,000 and $64,000. Similarly, an interval estimate that 10% to 60% of business students pursue an MBA is not very informative.

Statisticians like precision in their interval estimates, which is implied by a low margin of error. If we are able to increase the size of the sample, the larger n reduces the margin of error for the interval estimates. Although a larger sample size improves precision, it also entails the added cost in terms of time and money. Before getting into data collection, it is important that we first decide on the sample size that is adequate for what we wish to accomplish. In this section, we examine the required sample size, for a desired margin of error, in the confidence intervals for the population mean μ and the population proportion p. In order to be conservative, we always round up noninteger values for the required sample size.

Selecting n to Estimate μ

Consider a confidence interval for μ with a known population standard deviation σ. In addition, let E denote the desired margin of error. In other words, you do not want the sample mean to deviate from the population mean by more than E for a given level of confidence. Since $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, we rearrange this equation to derive the formula for the required sample size as $n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$. The sample size can be computed if we specify the population standard deviation σ, the value of $z_{\alpha/2}$ based on the confidence level $100(1-\alpha)\%$, and the desired margin of error E.

This formula is based on a knowledge of σ. However, in most cases σ is not known and, therefore, has to be estimated. Note that the sample standard deviation s cannot be used as an estimate for σ because s can be computed only after a sample of size n has been selected. In such cases, we replace σ with its reasonable estimate $\hat{\sigma}$.

THE REQUIRED SAMPLE SIZE WHEN ESTIMATING THE POPULATION MEAN

For a desired margin of error E, the minimum sample size n required to estimate a $100(1-\alpha)\%$ confidence interval for the population mean μ is

$$n = \left(\frac{z_{\alpha/2}\hat{\sigma}}{E}\right)^2,$$

where $\hat{\sigma}$ is a reasonable estimate of σ in the planning stage.

If σ is known, we replace $\hat{\sigma}$ with σ. Sometimes we use the sample standard deviation from a preselected sample as $\hat{\sigma}$ in the planning stage. Another choice for $\hat{\sigma}$ is to use an estimate of the population standard deviation from prior studies. Finally, if the minimum and maximum values of the population are available, a rough approximation for the population standard deviation is given by $\hat{\sigma} = \text{range}/4$.

EXAMPLE 8.8

Let us revisit Example 8.5, where Jared Beane wants to construct the 90% confidence interval for the mean mpg of all ultra-green cars. Suppose Jared would like to constrain the margin of error to within 2 mpg. Further, Jared knows that the minimum mpg in the population is 76 mpg, whereas the maximum is 118 mpg. How large a sample does Jared need to compute the 90% confidence interval for the population mean?

SOLUTION: For the 90% confidence level, Jared computes $z_{\alpha/2} = z_{0.05} = 1.645$. He estimates the population standard deviation as $\hat{\sigma} = \text{range}/4 = (118 - 76)/4 = 10.50$. Given $E = 2$, the required sample size is

$$n = \left(\frac{z_{\alpha/2}\hat{\sigma}}{E}\right)^2 = \left(\frac{1.645 \times 10.50}{2}\right)^2 = 74.54,$$

which is rounded up to 75. Therefore, Jared needs a random sample of at least 75 ultra-green cars to provide a more precise interval estimate of the mean mpg.

Selecting *n* to Estimate *p*

The margin of error E for the confidence interval for the population proportion p is $E = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$, where $\bar{p}$ represents the sample proportion. By rearranging, we derive the formula for the required sample size as $n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \bar{p}(1-\bar{p})$. Analogous to the case of the population mean, this formula is not feasible because it uses $\bar{p}$, which cannot be computed unless a sample of size n has already been selected. We replace $\bar{p}$ with a reasonable estimate $\hat{p}$ of the population proportion p.

THE REQUIRED SAMPLE SIZE WHEN ESTIMATING THE POPULATION PROPORTION

For a desired margin of error E, the minimum sample size n required to estimate a $100(1-\alpha)\%$ confidence interval for the population proportion p is

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1-\hat{p}),$$

where $\hat{p}$ is a reasonable estimate of p in the planning stage.

Sometimes we use the sample proportion from a preselected sample as $\hat{p}$ in the planning stage. Another choice for $\hat{p}$ is to use an estimate of the population proportion from prior studies. If no other reasonable estimate of the population proportion is available, we can use $\hat{p} = 0.5$ as a conservative estimate to derive the optimal sample size; note that the required sample is the largest when $\hat{p} = 0.5$.

EXAMPLE 8.9

Let us revisit Example 8.7, where Jared Beane wants to construct the 90% confidence interval for the proportion of all ultra-green cars that obtain over 100 mpg. Jared does not want the margin of error to be more than 0.10. How large a sample does Jared need for his analysis of the population proportion?

SOLUTION: For the 90% confidence level, Jared computes $z_{\alpha/2} = z_{0.05} = 1.645$. Since no estimate for the population proportion is readily available, Jared uses a conservative estimate of $\hat{p} = 0.50$. Given $E = 0.10$, the required sample size is

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1-\hat{p}) = \left(\frac{1.645}{0.10}\right)^2 0.50(1-0.50) = 67.65,$$

which is rounded up to 68. Therefore, Jared needs to find another random sample of at least 68 ultra-green cars to provide a more precise interval estimate for the proportion of all ultra-green cars that obtain over 100 mpg.

SYNOPSIS OF INTRODUCTORY CASE

Jared Beane, an analyst at a research firm, prepares to write a report on the new ultra-green car that boasts an average of 100 mpg. Based on a sample of 25 cars, Jared reports with 90% confidence that the average mpg of all ultra-green cars is between 92.86 mpg and 100.18 mpg. Jared also constructs the 90% confidence interval for the proportion of cars that obtain more than 100 mpg and reports the interval between 0.132 and 0.428. Jared wishes to increase the precision of his confidence intervals by reducing the margin of error. If his desired margin of error is 2 mpg for the population mean, he must use a sample of at least 75 cars for the analysis. Jared also wants to reduce the margin of error to 0.10 for the proportion of cars that obtain more than 100 mpg. Using a conservative estimate, he calculates that a sample of at least 68 cars is needed to achieve this goal. Thus, in order to gain precision in the interval estimate for both the mean and the proportion with 90% confidence, Jared's sample must contain at least 75 cars.



©McGraw-Hill Education/Mark Dierker, photographer

EXERCISES 8.4

Mechanics

59. The minimum and maximum observations in a population are 20 and 80, respectively. What is the minimum sample size n required to estimate μ with 80% confidence if the desired margin of error is $E = 2.6$? What happens to n if you decide to estimate μ with 95% confidence?

60. Find the required sample size for estimating the population mean in order to be 95% confident that the sample mean is within 10 units of the population mean. Assume that the population standard deviation is 40.

61. You need to compute the 99% confidence interval for the population mean. How large a sample should you draw to ensure that the sample mean does not deviate from the population mean by more than 1.2? (Use 6.0 as an estimate of the population standard deviation from prior studies.)

62. What is the minimum sample size n required to estimate μ with 90% confidence if the desired margin of error is $E = 1.2$? The population standard deviation is estimated as $\hat{\sigma} = 3.5$. What happens to n if the desired margin of error decreases to $E = 0.7$?

63. In the planning stage, a sample proportion is estimated as $\hat{p} = 40/50 = 0.80$. Use this information to compute the minimum sample size n required to estimate p with 99% confidence if the desired margin of error $E = 0.12$. What happens to n if you decide to estimate p with 90% confidence?

64. What is the minimum sample size n required to estimate p with 95% confidence if the desired margin of error $E = 0.08$? The population proportion is estimated as $\hat{p} = 0.36$ from prior studies. What happens to n if the desired margin of error increases to $E = 0.12$?

65. You wish to compute the 95% confidence interval for the population proportion. How large a sample should you draw to ensure that the sample proportion does not deviate from the population proportion by more than 0.06? No prior estimate for the population proportion is available.

Applications

66. Mortgage lenders often use FICO scores to check the credit worthiness of consumers applying for real estate loans. In general, FICO scores range from 300 to 850 with higher scores representing a better credit profile. A lender in a Midwestern town would like to estimate the mean credit score of its residents. What is the required number of sample FICO scores needed if the lender does not want the margin of error to exceed 20, with 95% confidence?

67. An analyst from an energy research institute in California wishes to estimate the 99% confidence interval for the average price of unleaded gasoline in the state. In particular, she does not want the sample mean to deviate from the population mean by more than \$0.06. What is the minimum number of gas stations that she should include in her sample if she uses the standard deviation estimate of \$0.32, as reported in the popular press?

68. An analyst would like to construct 95% confidence intervals for the mean stock returns in two industries. Industry A is a high-risk industry with a known population standard deviation of 20.6%, whereas Industry B is a low-risk industry with a known population standard deviation of 12.8%.
    a. What is the minimum sample size required by the analyst if she wants to restrict the margin of error to 4% for Industry A?
    b. What is the minimum sample size required by the analyst if she wants to restrict the margin of error to 4% for Industry B?
    c. Why do the results differ if they use the same margin of error?

69. The manager of a pizza chain in Albuquerque, New Mexico, wants to determine the average size of their advertised 16-inch pizzas. She takes a random sample of 25 pizzas and records their mean and standard deviation as 16.10 inches and 1.8 inches, respectively. She subsequently computes the 95% confidence interval of the mean size of all pizzas as [15.36, 16.84]. However, she finds this interval to be too broad to implement quality control and decides to reestimate the mean based on a bigger sample. Using the standard deviation estimate of 1.8 from her earlier analysis, how large a sample must she take if she wants the margin of error to be under 0.5 inch?

70. The manager of a newly opened Target store wants to estimate the average expenditure of his customers. From a preselected sample, the standard deviation was determined to be $18. The manager would like to construct the 95% confidence interval for the mean customer expenditure.
    a. Find the appropriate sample size necessary to achieve a margin of error of $5.
    b. Find the appropriate sample size necessary to achieve a margin of error of $3.

71. A budget airline wants to estimate what proportion of customers would consider paying $12 for in-flight wireless access. Given that the airline has no prior knowledge of the proportion, how many customers would it have to sample to ensure a margin of error of no more than 0.05 for the 90% confidence interval?

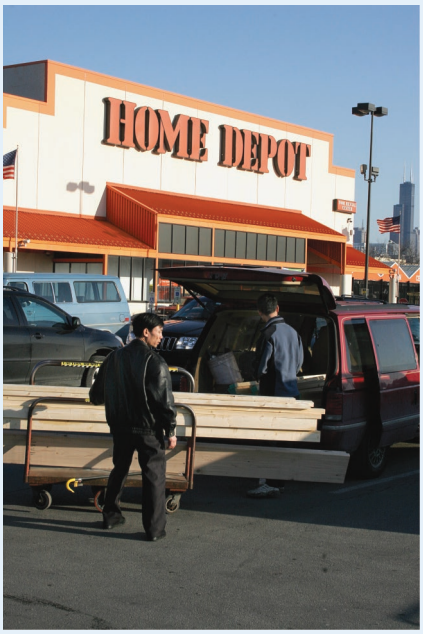
72. Newscasters wish to estimate the proportion of registered voters who support the incumbent candidate in the mayoral election. In an earlier poll of 240 registered voters, 110 had supported the incumbent candidate. Find the sample size required to construct the 90% confidence interval if newscasters do not want the margin of error to exceed 0.02.

73. A survey by AARP (*Money,* June 2007) reported that approximately 70% of people in the 50 to 64 age bracket have tried some type of alternative therapy (for instance, acupuncture or the use of nutrition supplements). Assume this survey was based on a sample of 400 people.
    a. Identify the relevant parameter of interest for these qualitative data and compute its point estimate as well as the margin of error with 90% confidence.
    b. You decide to redo the analysis with the margin of error reduced to 2%. How large a sample do you need to draw? State your assumptions in computing the required sample size.

74. Subprime lending was big business in the United States in the mid-2000s, when lenders provided mortgages to people with poor credit. However, subsequent increases in interest rates coupled with a drop in home values necessitated many borrowers to default. Suppose a recent report finds that two in five subprime mortgages are likely to default nationally. A research economist is interested in estimating default rates in Illinois with 95% confidence. How large a sample is needed to restrict the margin of error to within 0.06, using the reported national default rate?

75. A business student is interested in estimating the 99% confidence interval for the proportion of students who bring laptops to campus. He wishes a precise estimate and is willing to draw a large sample that will keep the sample proportion within five percentage points of the population proportion. What is the minimum sample size required by this student, given that no prior estimate of the population proportion is available?

WRITING WITH STATISTICS



©McGraw-Hill Education/Andrew Resek



©Todd A. Merport/Shutterstock

Callie Fitzpatrick, a research analyst with an investment firm, has been asked to write a report summarizing the weekly stock performance of Home Depot and Lowe's. Her manager is trying to decide whether or not to include one of these stocks in a client's portfolio and the average stock performance is one of the factors influencing this decision. Callie decides to use descriptive measures to summarize stock returns in her report, as well as provide confidence intervals for the average return for Home Depot and Lowe's. She collects weekly returns for each firm for the first eight months of 2010. A portion of the return data is shown in Table 8.5.

TABLE 8.5 Weekly Returns (in percent) for Home Depot and Lowe's

FILE
Weekly_Returns

Date	Home Depot	Lowe's
1/11/2010	−1.44	−1.59
1/19/2010	−2.98	−3.53
⋮	⋮	⋮
8/30/2010	−2.61	−3.89

Source: www.finance.yahoo.com.

Callie would like to use the sample information to

1. Summarize weekly returns for Home Depot and Lowe's.
2. Provide confidence intervals for the average weekly returns.
3. Make recommendations for further analysis.

Sample Report—Weekly Stock Performance: Home Depot vs. Lowe's

Grim news continues to distress the housing sector. On August 24, 2010, Reuters reported that the sales of previously owned U.S. homes took a record plunge in July to the slowest pace in 15 years. Combine this fact with the continued fallout from the subprime mortgage debacle, a sluggish economy, and high unemployment, and the housing sector appears quite unstable. Have these unfavorable events managed to trickle down and harm the financial performance of Home Depot and Lowe's, the two largest home improvement retailers in the United States?

One way to analyze their financial stability is to observe their stock performance during this period. In order to make valid statements concerning the reward of holding these stocks, weekly return data for each firm were gathered from January through August of 2010. Table 8.A summarizes the important descriptive statistics.

TABLE 8.A Descriptive Statistics for Weekly Returns of Home Depot and Lowe's ($n = 34$)

	Home Depot (in %)	Lowe's (in %)
Mean	0.00	−0.33
Median	0.76	−0.49
Minimum	−8.08	−7.17
Maximum	5.30	7.71
Standard deviation	3.59	3.83
Margin of error with 95% confidence	1.25	1.34

Over the past 34 weeks, Home Depot posted both a higher average return and median return of 0.00% and 0.76%, respectively. Lowe's return over the same period was negative, whether the central tendency was measured by its mean (−0.33%) or its median (−0.49%). In terms of dispersion, Lowe's return data had the higher standard deviation (3.83% > 3.59%). In terms of descriptive measures, an investment in Home Depot's stock not only provided higher returns, but also was less risky than an investment in Lowe's stock.

Table 8.A also shows the margins of error for 95% confidence intervals for the mean returns. With 95% confidence, the mean return for Home Depot fell in the range [−1.25%, 1.25%], while that for Lowe's fell in the range [−1.67%, 1.01%]. Given that these two intervals overlap, one cannot conclude that Home Depot delivered the higher reward over this period—a conclusion one may have arrived at had only the point estimates been evaluated. It is not possible to recommend one stock over the other for inclusion in a client's portfolio based solely on the mean return performance. Other factors, such as the correlation between the stock and the existing portfolio, must be analyzed before this decision can be made.

CONCEPTUAL REVIEW

LO 8.1 Explain a confidence interval.

The sample mean $\overline{X}$ is the point estimator for the population mean μ, and the sample proportion $\overline{P}$ is the point estimator for the population proportion p. Sample values of the point estimators represent the point estimates for the population parameter of interest; $\bar{x}$ and $\bar{p}$ are the point estimates for μ and p, respectively. While a point estimator provides a single value that approximates the unknown parameter, a **confidence interval**, or an **interval estimate**, provides a range of values that, with a certain level of confidence, will contain the population parameter of interest.

Often, we construct a confidence interval as point estimate ± margin of error. The **margin of error** accounts for the variability of the estimator and the desired confidence level of the interval.

LO 8.2 Calculate a confidence interval for the population mean when the population standard deviation is known.

A $100(1 - \alpha)\%$ confidence interval for the population mean μ when the population standard deviation σ is known is computed as $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, where $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is the margin of error. This formula is valid only if $\overline{X}$ (approximately) follows a normal distribution.

LO 8.3 Describe the factors that influence the width of a confidence interval.

The precision of a confidence interval is directly linked with the width of the interval: the wider the interval, the lower its precision. A confidence interval is wider (a) the greater the population standard deviation σ, (b) the smaller the sample size n, and (c) the greater the confidence level.

LO 8.4 Discuss features of the *t* distribution.

The ***t* distribution** is a family of distributions that are similar to the z distribution, in that they are all symmetric and bell-shaped around zero with asymptotic tails. However, the t distribution has broader tails than does the z distribution. Each t distribution is identified by a parameter known as the **degrees of freedom *df***. The *df* determine the extent of broadness—the smaller the *df*, the broader the tails. Since the t distribution is defined by the degrees of freedom, it is common to refer to it as the t_{df} distribution.

LO 8.5 Calculate a confidence interval for the population mean when the population standard deviation is not known.

A $100(1 - \alpha)\%$ confidence interval for the population mean μ when the population standard deviation σ is not known is computed as $\bar{x} \pm t_{\alpha/2,df}\frac{s}{\sqrt{n}}$, where s is the sample standard deviation. This formula is valid only if $\overline{X}$ (approximately) follows a normal distribution.

LO 8.6 Calculate a confidence interval for the population proportion.

A $100(1 - \alpha)\%$ confidence interval for the population proportion p is computed as $\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$, where $\bar{p}$ is the sample proportion. This formula is valid only if $\overline{P}$ (approximately) follows a normal distribution.

LO 8.7 Select a sample size to estimate the population mean and the population proportion.

For a desired margin of error E, the minimum n required to estimate μ with $100(1-\alpha)\%$ confidence is $n = \left(\frac{z_{\alpha/2}\hat{\sigma}}{E}\right)^2$, where $\hat{\sigma}$ is a reasonable estimate of σ in the planning stage. If σ is known, we replace $\hat{\sigma}$ with σ. Other choices for $\hat{\sigma}$ include an estimate from a preselected sample, prior studies, or $\hat{\sigma} =$ range/4.

For a desired margin of error E, the minimum n required to estimate p with $100(1-\alpha)\%$ confidence is $n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1-\hat{p})$, where $\hat{p}$ is a reasonable estimate of p in the planning stage. Choices for $\hat{p}$ include an estimate from a preselected sample or prior studies; a conservative estimate of $\hat{p} = 0.5$ is used when no other reasonable estimate is available.

ADDITIONAL EXERCISES AND CASE STUDIES

Exercises

76. Over a 10-year sample period, the mean and the standard deviation of annual returns on a portfolio you are analyzing were 10% and 15%, respectively. You assume that returns are normally distributed. Construct the 95% confidence interval for the population mean.

77. A hair salon in Cambridge, Massachusetts, reports that on seven randomly selected weekdays, the number of customers who visited the salon were 40, 30, 28, 22, 36, 16, and 50. It can be assumed that weekday customer visits follow a normal distribution.
    a. Construct the 90% confidence interval for the average number of customers who visit the salon on weekdays.
    b. Construct the 99% confidence interval for the average number of customers who visit the salon on weekdays.
    c. What happens to the width of the interval as the confidence level increases?

78. According to data from the Organization for Economic Cooperation and Development, the average U.S. worker takes 16 days of vacation each year (*The Wall Street Journal*, June 20, 2007). Assume that these data were based on a sample of 225 workers and that the sample standard deviation is 12 days.
    a. Construct the 95% confidence interval for the population mean.
    b. At the 95% confidence level, can we conclude that the average U.S. worker does not take 14 days of vacation each year?

79. Recently, six single-family homes in San Luis Obispo County in California sold at the following prices (in $1,000s): 549, 449, 705, 529, 639, and 609.
    a. Construct the 95% confidence interval for the mean sale price of homes in San Luis Obispo County.
    b. What assumption have you made when constructing this confidence interval?

80. Students who graduated from college in 2010 owed an average of $25,250 in student loans (*The New York Times*, November 2, 2011). An economist wants to determine if average debt has changed. She takes a sample of 40 recent graduates and finds that their average debt was $27,500 with a standard deviation of $9,120. Use the 90% confidence interval to determine if average debt has changed.

81. A machine that is programmed to package 1.20 pounds of cereal is being tested for its accuracy. In a sample of 36 cereal boxes, the sample mean filling weight is calculated as 1.22 pounds. The population standard deviation is known to be 0.06 pound.
    a. Identify the relevant parameter of interest for these quantitative data and compute its point estimate as well as the margin of error with 95% confidence.
    b. Can we conclude that the packaging machine is operating improperly?
    c. How large a sample must we take if we want the margin of error to be at most 0.01 pound with 95% confidence?

82. The SAT is the most widely used test in the undergraduate admissions process. Scores on

the math portion of the SAT are believed to be normally distributed and range from 200 to 800. A researcher from the admissions department at the University of New Hampshire is interested in estimating the mean math SAT scores of the incoming class with 90% confidence. How large a sample should she take to ensure that the margin of error is below 15?

83. A study by Allstate Insurance Co. finds that 82% of teenagers have used cell phones while driving (*The Wall Street Journal*, May 5, 2010). Suppose this study was based on a random sample of 50 teen drivers.
    a. Construct the 99% confidence interval for the proportion of all teenagers that have used cell phones while driving.
    b. What is the margin of error with 99% confidence?

84. The following table shows the annual returns (in percent) for the Vanguard Energy Fund.

Year	Return
2010	13
2011	−2
2012	3
2013	18
2014	−14

Source: www.finance.yahoo.com, data retrieved April 4, 2015.

    a. Calculate the point estimate for μ.
    b. Construct the 95% confidence interval for μ.
    c. What assumption did you make when constructing the interval?

85. FILE ***MV_Houses.*** A realtor wants to estimate the mean price of houses in Mission Viejo, California. She collects a sample of 36 recent house sales (in $1,000s), a portion of which is shown in the accompanying table. Assume that the population standard deviation is 100 (in $1,000s). Construct and interpret 95% and 98% confidence intervals for the mean price of all houses in Mission Viejo, CA.

Prices
430
520
⋮
430

86. FILE ***MI_Life_Expectancy.*** Residents of Hawaii have the longest life expectancies, averaging 81.48 years (www.worldlifeexpectancy.com, data retrieved June 4, 2012). A sociologist collects data on the age at death for 50 recently deceased Michigan residents. A portion of the data is shown in the accompanying table. Assume that the population standard deviation is 5 years.

Age at Death
76.4
76.0
⋮
73.6

    a. Construct the 95% confidence interval for the mean life expectancy of all residents of Michigan.
    b. Use the 95% confidence interval to determine if the mean life expectancy of Michigan residents differs from that for Hawaii residents.

87. FILE ***Fastballs.*** The manager of a minor league baseball team wants to estimate the average fastball speed of two pitchers. He clocks 50 fastballs, in miles per hour, for each pitcher. A portion of the data is shown in the accompanying table.

Pitcher 1	Pitcher 2
87	82
86	92
⋮	⋮
86	93

    a. Construct 95% confidence intervals for the mean speed for each pitcher.
    b. Explain why the widths of the two intervals are different.

88. FILE ***Theater.*** The new manager of a theater would like to offer discounts to increase the number of tickets sold for shows on Monday and Tuesday evenings. She uses a sample of 30 weeks to record the number of tickets sold on these two days. A portion of the data is shown in the accompanying table.

Monday	Tuesday
221	208
187	199
⋮	⋮
194	180

    a. Compare the margin of error for the 95% confidence intervals for the mean number of tickets sold for shows on Monday and Tuesday evenings.
    b. Construct the 95% confidence intervals for the mean number of tickets sold for shows on Monday and Tuesday evenings.
    c. Determine if the population mean differs from 200 for shows on Monday and Tuesday evenings.

89. FILE ***AnnArbor_Rental.*** Real estate investment in college towns continues to promise good returns (*The Wall Street Journal*, September 24, 2010). Marcela Treisman works for an investment firm in Michigan. Her assignment is to analyze the rental market in Ann Arbor, which is home to the University of Michigan. She gathers data on monthly rents for 2011 along with the square footage of 40 homes. A portion of the data is shown in the accompanying table.

Monthly Rent	Square Footage
645	500
675	648
⋮	⋮
2400	2700

Source: www.zillow.com.

a. Construct 90% and 95% confidence intervals for the mean rent for all rental homes in Ann Arbor, Michigan.
b. Construct 90% and 95% confidence intervals for the mean square footage for all rental homes in Ann Arbor, Michigan.

90. According to a survey of 1,235 businesses by IDC, a market-research concern in Framingham, Massachusetts, 12.1% of sole proprietors are engaging in e-commerce (*The Wall Street Journal*, July 26, 2007).
a. With 95% confidence, what is the margin of error when estimating the proportion of sole proprietors that engage in e-commerce?
b. Construct the 95% confidence interval for the population proportion.

91. A Monster.com poll of 3,057 individuals asked: "What's the longest vacation you plan to take this summer?" The following relative frequency distribution summarizes the results.

Response	Relative Frequency
A few days	0.21
A few long weekends	0.18
One week	0.36
Two weeks	0.22

Source: *The Boston Globe*, June 12, 2007.

a. Construct the 95% confidence interval for the proportion of people who plan to take a one-week vacation this summer.
b. Construct the 99% confidence interval for the proportion of people who plan to take a one-week vacation this summer.
c. Which of the two confidence intervals is wider?

92. Linda Barnes has learned from prior studies that one out of five applicants gets admitted to top MBA programs in the country. She wishes to construct her own 90% confidence interval for the acceptance rate in top MBA programs. How large a sample should she take if she does not want the acceptance rate of the sample to deviate from that of the population by more than five percentage points? State your assumptions in computing the required sample size.

93. FILE ***Field_Choice.*** There is a declining interest among teenagers to pursue a career in science and health care (*U.S. News & World Report*, May 23, 2011). Thirty college-bound students in Portland, Oregon, are asked about the field they would like to pursue in college. The choices offered in the questionnaire are science, business, and other. The gender information also is included in the questionnaire. A portion of the data is shown.

Field Choice	Gender
Business	Male
Other	Female
⋮	⋮
Science	Female

a. Compare the 95% confidence interval for the proportion of students who would like to pursue science with the proportion who would like to pursue business.
b. Construct and interpret the 90% confidence interval for the proportion of female students who are college bound.

94. FILE ***Pedestrians.*** A study examined "sidewalk rage" in an attempt to find insight into anger's origins and offer suggestions for anger-management treatments (*The Wall Street Journal*, February 15, 2011). "Sidewalk ragers" tend to believe that pedestrians should behave in a certain way. One possible strategy for sidewalk ragers is to avoid walkers who are distracted by other activities such as smoking and tourism. Sample data were obtained from 50 pedestrians in Lower Manhattan. It was noted if the pedestrian was smoking (equaled 1 if smoking, 0 otherwise) or was a tourist (equaled 1 if tourist, 0 otherwise). The accompanying table shows a portion of the data.

Smoking	Tourist
0	1
0	1
⋮	⋮
0	0

a. Construct and interpret the 95% confidence interval for the proportion of pedestrians in Lower Manhattan who smoke while walking.
b. Construct and interpret the 95% confidence interval for the proportion of pedestrians in Lower Manhattan who are tourists.

95. An economist would like to estimate the 95% confidence interval for the average real estate taxes collected by a small town in California. In a prior analysis, the standard deviation of real estate taxes was reported as $1,580. What is the minimum sample size required by the economist if he wants to restrict the margin of error to $500?

96. An employee of the Bureau of Transportation Statistics has been given the task of estimating the proportion of on-time arrivals of a budget airline. A prior study had estimated this on-time arrival rate as 78.5%. What is the minimum number of arrivals this employee must include in the sample to ensure that the margin of error for the 95% confidence interval is no more than 0.05?

97. According to a report by the PEW Research Center, 85% of adults under 30 feel optimistic about the economy, but the optimism is shared by only 45% of those who are over 50 (*Newsweek*, September 13, 2010). A research analyst would like to construct 95% confidence intervals for the proportion patterns in various regions of the country. She uses the reported rates by the PEW Research Center to determine the sample size that would restrict the margin of error to within 0.05.
a. How large a sample is required to estimate the proportion of adults under 30 who feel optimistic about the economy?
b. How large a sample is required to estimate the proportion of adults over 50 who feel optimistic about the economy?

CASE STUDIES

CASE STUDY 8.1 Texas is home to more than one million undocumented immigrants, and most of them are stuck in low-paying jobs. Meanwhile, the state also suffers from a lack of skilled workers. The Texas Workforce Commission estimates that 133,000 jobs are currently unfilled, many because employers cannot find qualified applicants (*The Boston Globe*, September 29, 2011). Texas was the first state to pass a law that allows children of undocumented immigrants to pay in-state college tuition rates if they have lived in Texas for three years and plan to become permanent residents. The law passed easily back in 2001 because most legislators believed that producing college graduates and keeping them in Texas benefits the business community. In addition, since college graduates earn more money, they also provide the state with more revenue. Carol Capaldo wishes to estimate the mean hourly wage of workers with various levels of education. She collects a sample of the hourly wages (in $) of 30 workers with a bachelor's degree or higher, 30 workers with only a high school diploma, and 30 workers who did not finish high school. A portion of the data is shown in the accompanying table.

FILE
Texas_Wage

Data for Case Study 8.1 Hourly Wages of Texas Workers by Education Level (in $)

Bachelor's Degree or Higher	High School Diploma	No High School Diploma
22.50	12.68	11.21
19.57	11.23	8.54
⋮	⋮	⋮
21.44	7.47	10.27

In a report, use the information to

1. Calculate descriptive statistics to compare the hourly wages for the three education levels.
2. Construct and interpret 95% confidence intervals for the mean hourly wage at each education level.

CASE STUDY 8.2 The following table presents a portion of the annual returns for two mutual funds offered by the investment giant Fidelity. The *Fidelity Select Automotive Fund* invests primarily in companies engaged in the manufacturing, marketing, or sales of automobiles, trucks, specialty vehicles, parts, tires, and related services. The *Fidelity Gold Fund* invests primarily in companies engaged in exploration, mining, processing, or dealing in gold and, to a lesser degree, in other precious metals and minerals.

Data for Case Study 8.2 Annual Total Return (%) History

FILE
Fidelity_Returns

Year	Annual Total Return History (in %)	
	Automotive	Gold
2001	22.82	24.99
2002	−6.48	64.28
⋮	⋮	⋮
2016	−5.83	47.28

Source: www.finance.yahoo.com, data retrieved April 2, 2017.

In a report, use the above information to

1. Calculate descriptive statistics to compare the returns of the mutual funds.
2. Assess reward by constructing and interpreting 95% confidence intervals for the population mean return. What assumption did you make for the interval estimates?

CASE STUDY 8.3 The information gathered from opinion polls and political surveys is becoming so increasingly important for candidates on the campaign trail that it is hard to imagine an election that lacks extensive polling. An NBC News/*Wall Street Journal* survey (August 5–9, 2010) of 1,000 adults asked people's preferences on candidates and issues prior to the midterm 2010 elections. Some of the responses to the survey are shown below, as well as responses from prior surveys. (Copyright © 2010 Dow Jones & Co., Inc.)

Question: In general, do you approve or disapprove of the way Barack Obama is handling the aftermath of the Gulf Coast oil spill in August 2010 (and George W. Bush's handling of Katrina in March 2006)?

	August 2010 (%)	March 2006 (%)
Approve	50	36
Disapprove	38	53
Not sure	12	11

Question: Which are more important to you in your vote for Congress this November: domestic issues such as the economy, health care, and immigration; or international issues such as Afghanistan, Iran, and terrorism?

	August 2010 (%)	September 2006 (%)
Domestic issues	73	43
International issues	12	28
Both equally important	15	28

In a report, construct 95% confidence intervals for the relevant population proportions to

1. Compare the approval rates of President Obama's handling of the Gulf Coast oil spill and President George W. Bush's handling of the Hurricane Katrina crisis.
2. Compare the importance of domestic issues in August 2010 and in September 2006.

APPENDIX 8.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, JMP, and R. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands. When importing data into R, use the menu-driven option: File > Import Dataset > From Excel.

Minitab

Estimating μ, σ Known

FILE *Hockey_Pucks*

A. (Replicating Example 8.3) From the menu, choose **Stat > Basic Statistics > 1-Sample Z**.

B. Select **One or more samples, each in a column**, and then select Weight. After **Known standard deviation**, enter 7.5. Choose **Options**. Enter 90.0 for **Confidence Level**.

Estimating μ, σ Unknown

FILE *Lottery*

A. (Replicating Example 8.6) From the menu, choose **Stat > Basic Statistics > 1-Sample t**.

B. Select **One or more samples**, **each in a column**, and then select Expenditures. Choose **Options**. Enter 95.0 for **Confidence Level**.

Estimating p

A. (Replicating Example 8.7) From the menu, choose **Stat > Basic Statistics > 1 Proportion**.

B. Select **Summarized data**. Enter 7 for **Number of events** and 25 for **Number of trials**. Choose **Options**. Enter 90.0 for **Confidence Level** and check **Use test and interval based on normal distribution**.

SPSS

Estimating μ, σ Unknown

FILE *Lottery*

(Replicating Example 8.6) From the menu, choose **Analyze > Compare Means > One-Sample T Test**. Under **Test Variable(s)**, select Expenditures. Choose **Options**. After **Confidence Interval Percentage** enter 95.

JMP

Estimating μ, σ Known

FILE *Hockey_Pucks*

A. (Replicating Example 8.3) From the menu, choose **Analyze > Distribution**.

B. Under **Select Columns**, select Weight, and then under **Cast Selected Columns into Roles**, select **Y, Columns**.

C. Click on the red triangle in the output window beside Weight. Choose **Confidence Interval > Other**, and after **Enter (1-alpha for Confidence level)**, enter 0.90. Select **Use known sigma** and enter 7.5.

Estimating μ, σ Unknown

A. (Replicating Example 8.6) From the menu, choose **Analyze > Distribution**.

FILE *Lottery*

B. Under **Select Columns**, select Expenditures, and then under **Cast Selected Columns into Roles**, select **Y, Columns**.

C. Click on the red triangle in the output window beside Expenditures. Choose **Confidence Interval > 0.95**.

R

Estimating μ, σ Known

A. (Replicating Example 8.3) First find the margin of error, labeled Error. In order to find $z_{\alpha/2} = z_{0.05}$, use the **qnorm** function with $\mu = 0$ and $\sigma = 1$. The margin of error for a 90% confidence interval with $\sigma = 7.5$ and $n = 80$ is found by entering:

FILE *Hockey_Pucks*

```
> Error <- qnorm(0.95, 0, 1)*7.5/sqrt(80)
```

B. Find the lower and upper limits of the confidence interval by adding and subtracting the Error from the mean. Enter:

```
> Lower <- mean(Hockey_Pucks$`Weight') - Error
> list(Lower)
> Upper <- mean(Hockey_Pucks$`Weight') + Error
> list(Upper)
```

Estimating μ, σ Unknown

A. (Replicating Example 8.6) First find the margin of error, labeled Error. In order to find $t_{\alpha/2df}$, use the **qt** function. The margin of error for a 95% confidence interval with $n = 100$ is found by entering:

FILE *Lottery*

```
> Error <- qt(0.975, 99,lower.tail = TRUE)*sd(Lottery$'Expenditures')/
  sqrt(100)
```

B. Find the lower and upper limits of the confidence interval by adding and subtracting the Error from the mean. Enter:

```
> Lower <- mean(Lottery$`Expenditures') - Error
> list(Lower)
> Upper <- mean(Lottery$`Expenditures') + Error
> list(Upper).
```

9 Hypothesis Testing

Learning Objectives

After reading this chapter you should be able to:

LO **9.1** Define the null hypothesis and the alternative hypothesis.

LO **9.2** Distinguish between Type I and Type II errors.

LO **9.3** Conduct a hypothesis test for the population mean when σ is known.

LO **9.4** Conduct a hypothesis test for the population mean when σ is unknown.

LO **9.5** Conduct a hypothesis test for the population proportion.

In Chapter 8, we used confidence intervals to estimate an unknown population parameter of interest. In this chapter, we will focus on the second major area of statistical inference: hypothesis testing. We use a hypothesis test to challenge the status quo, or some belief about an underlying population parameter, based on sample data. In particular, we develop hypothesis tests for the population mean and the population proportion. For instance, we may wish to test whether the average age of MBA students in the United States is less than 30 years or whether the percentage of defective items in a production process differs from 5%. In either case, since we do not have access to the entire population, we have to perform statistical inference on the basis of limited sample information. If the sample information is not consistent with the status quo, we use the hypothesis testing framework to determine if the inconsistency is real (that is, we contradict the status quo) or due to chance (that is, we do not contradict the status quo).



Introductory Case

Undergraduate Study Habits

Are today's college students studying hard or hardly studying? A study asserts that, over the past five decades, the number of hours that the average college student studies each week has been steadily dropping (*The Boston Globe,* July 4, 2010). In 1961, students invested 24 hours per week in their academic pursuits, whereas today's students study an average of 14 hours per week.

Susan Knight is a dean at a large university in California. She wonders if the study trend is reflective of the students at her university. She randomly selects 35 students and asks their average study time per week (in hours). The responses are shown in Table 9.1.

TABLE 9.1 Average Hours Studied per Week for a Sample of 35 College Students

25	17	8	14	17	7	11
19	16	9	15	12	17	19
26	14	22	17	14	35	24
11	21	6	20	27	17	6
29	10	10	4	25	13	16

FILE
Study_Hours

Summary measures: $\bar{x} = 16.3714$ hours and $s = 7.2155$ hours.

Susan wants to use the sample information to

1. Determine if the mean study time of students at her university is below the 1961 national average of 24 hours per week.
2. Determine if the mean study time of students at her university differs from today's national average of 14 hours per week.

A synopsis of this case is provided at the end of Section 9.3.

LO **9.1**

Define the null hypothesis and the alternative hypothesis.

9.1 INTRODUCTION TO HYPOTHESIS TESTING

Every day people make decisions based on their beliefs about the true state of the world. They hold certain things to be true and others to be false, and then act accordingly. For example, an engineer believes that a certain steel cable has a breaking strength of 5,000 pounds or more, and then permits its use at a construction site; a manufacturer believes that a certain process yields capsules that contain precisely 100 milligrams of a drug, and then ships the capsules to a pharmacy; a manager believes that an incoming shipment contains 2%, or fewer, of defects, and then accepts the shipment. In these cases, and many more, the formation of these beliefs may have started as a mere conjecture, an informed guess, or a proposition tentatively advanced as true. When people formulate a belief in this way, we refer to it as a hypothesis. Sooner or later, however, every hypothesis eventually confronts evidence that either substantiates or refutes it. Determining the validity of a belief is called hypothesis testing.

We use hypothesis testing to resolve conflicts between two competing hypotheses on a particular population parameter of interest. We refer to one hypothesis as the **null hypothesis**, denoted H_0, and the other as the **alternative hypothesis**, denoted H_A. We think of the null hypothesis as corresponding to a presumed default state of nature or status quo. The alternative hypothesis, on the other hand, contradicts the default state or status quo.

NULL HYPOTHESIS VERSUS ALTERNATIVE HYPOTHESIS

When constructing a hypothesis test, we define a null hypothesis, denoted H_0, and an alternative hypothesis, denoted H_A. We conduct a hypothesis test to determine whether or not sample evidence contradicts H_0.

In statistics, we use sample information to make inferences regarding the unknown population parameters of interest. In this chapter, our goal is to determine if the null hypothesis can be rejected in favor of the alternative hypothesis. An analogy can be drawn with applications in the medical and legal fields, where we can define the null hypothesis as "an individual is free of a particular disease" or "an accused is innocent." In both cases, the verdict is based on limited evidence, which in statistics translates into making a decision based on limited sample information.

The Decision to "Reject" or "Not Reject" the Null Hypothesis

The hypothesis testing procedure enables us to make one of two decisions. If sample evidence is inconsistent with the null hypothesis, we reject the null hypothesis. Conversely, if sample evidence is not inconsistent with the null hypothesis, then we do not reject the null hypothesis. It is not correct to conclude that "we accept the null hypothesis" because while the sample information may not be inconsistent with the null hypothesis, it does not necessarily prove that the null hypothesis is true.

On the basis of sample information, we either "reject the null hypothesis" or "do not reject the null hypothesis."

Consider the example just referenced where the null is defined as "an individual is free of a particular disease." Suppose a medical procedure does not detect this disease. On the basis of this limited information, we can only conclude that we are unable to detect the

disease (do not reject the null hypothesis). It does not necessarily prove that the person does not have the disease (accept the null hypothesis). Similarly, in the court example where the null hypothesis is defined as "an accused is innocent," we can conclude that the person is guilty (reject the null hypothesis) or that there is not enough evidence to convict (do not reject the null hypothesis).

Defining the Null and the Alternative Hypotheses

As mentioned earlier, we use a hypothesis test to contest the status quo, or some belief about an underlying population parameter, based on sample data. A very crucial step concerns the formulation of the two competing hypotheses, since the conclusion of the test depends on how the hypotheses are stated. As a general guideline, whatever we wish to establish is placed in the alternative hypothesis, whereas the null hypothesis includes the status quo. If we are unable to reject the null hypothesis, then we maintain the status quo or "business as usual." However, if we reject the null hypothesis, this establishes that the evidence supports the alternative hypothesis, which may require that we take some kind of action. For instance, if we reject the null hypothesis that an individual is free of a particular disease, then we conclude that the person is sick, for which treatment is prescribed. Similarly, if we reject the null hypothesis that an accused is innocent, we conclude that the person is guilty and is suitably punished.

In most applications, we require some form of the equality sign in the null hypothesis. (The justification for the equality sign will be provided later.) In general, any statement including one of the three signs "=", "≤", or "≥" is valid for the null hypothesis. Given that the alternative hypothesis states the opposite of the null hypothesis, the alternative hypothesis is then specified with a "≠", ">", or "<" sign.

As a general guideline, we use the alternative hypothesis as a vehicle to establish something new—that is, contest the status quo. In most applications, the null hypothesis regarding a particular population parameter of interest is specified with one of the following signs: =, ≤, or ≥; the alternative hypothesis is then specified with the corresponding opposite sign: ≠, >, or <.

A hypothesis test can be **one-tailed** or **two-tailed**. A two-tailed test is defined when the alternative hypothesis includes the sign "≠ ". For example, H_0: $\mu = \mu_0$ versus H_A: $\mu \neq \mu_0$ and H_0: $p = p_0$ versus H_A: $p \neq p_0$ are examples of two-tailed tests, where μ_0 and p_0 represent hypothesized values of the population mean and the population proportion, respectively. If the null hypothesis is rejected, it suggests that the true parameter does not equal the hypothesized value.

A one-tailed test, on the other hand, involves a null hypothesis that can only be rejected on one side of the hypothesized value. For example, consider H_0: $\mu \leq \mu_0$ versus H_A: $\mu > \mu_0$. Here we can reject the null hypothesis only when there is substantial evidence that the population mean is greater than μ_0. It is also referred to as a right-tailed test since rejection of the null hypothesis occurs on the right side of the hypothesized mean. Another example is a left-tailed test, H_0: $\mu \geq \mu_0$ versus H_A: $\mu < \mu_0$, where the null hypothesis can only be rejected on the left side of the hypothesized mean. One-tailed tests for the population proportion are defined similarly.

ONE-TAILED VERSUS TWO-TAILED HYPOTHESIS TESTS

Hypothesis tests can be one-tailed or two-tailed. In a one-tailed test, we can reject the null hypothesis only on one side of the hypothesized value of the population parameter. In a two-tailed test, we can reject the null hypothesis on either side of the hypothesized value of the population parameter.

In general, we follow three steps when formulating the competing hypotheses:

1. Identify the relevant population parameter of interest.
2. Determine whether it is a one- or two-tailed test.
3. Include some form of the equality sign in the null hypothesis and use the alternative hypothesis to establish a claim.

The following examples highlight one- and two-tailed tests for the population mean and the population proportion. In each example, we want to state the appropriate competing hypotheses.

EXAMPLE 9.1

A trade group predicts that back-to-school spending will average $606.40 per family this year. A different economic model is needed if the prediction is wrong. Specify the null and the alternative hypotheses to determine if a different economic model is needed.

SOLUTION: Given that we are examining average back-to-school spending, the parameter of interest is the population mean μ. Since we want to be able to determine if the population mean differs from $606.40 ($\mu \neq 606.40$), we need a two-tailed test and formulate the null and alternative hypotheses as

$$H_0: \mu = 606.40$$

$$H_A: \mu \neq 606.40$$

The trade group is advised to use a different economic model if the null hypothesis is rejected.

EXAMPLE 9.2

An advertisement for a popular weight-loss clinic suggests that participants in its new diet program experience an average weight loss of more than 10 pounds. A consumer activist wants to determine if the advertisement's claim is valid. Specify the null and the alternative hypotheses to validate the advertisement's claim.

SOLUTION: The advertisement's claim concerns average weight loss; thus, the parameter of interest is again the population mean μ. This is an example of a one-tailed test because we want to determine if the mean weight loss is more than 10 pounds ($\mu > 10$). We specify the competing hypotheses as

$$H_0: \mu \leq 10 \text{ pounds}$$

$$H_A: \mu > 10 \text{ pounds}$$

The underlying claim that the mean weight loss is more than 10 pounds is true if our decision is to reject the null hypothesis. Conversely, if we do not reject the null hypothesis, we cannot support the claim.

EXAMPLE 9.3

A television research analyst wishes to test a claim that more than 50% of the households will tune in for a TV episode. Specify the null and the alternative hypotheses to test the claim.

SOLUTION: This is an example of a one-tailed test regarding the population proportion p. Given that the analyst wants to determine whether $p > 0.50$, this claim is placed in the alternative hypothesis, whereas the null hypothesis is just its opposite.

$$H_0: p \leq 0.50$$

$$H_A: p > 0.50$$

The claim that more than 50% of the households will tune in for a TV episode is valid only if the null hypothesis is rejected.

EXAMPLE 9.4

It is generally believed that at least 60% of the residents in a small town in Texas are happy with their lives. A sociologist wonders whether recent economic woes have adversely affected the happiness level in this town. Specify the null and the alternative hypotheses to determine if the sociologist's concern is valid.

SOLUTION: This is also a one-tailed test regarding the population proportion p. While the population proportion has been at least 0.60 ($p \geq 0.60$), the sociologist wants to establish that the current population proportion is below 0.60 ($p < 0.60$). Therefore, the hypotheses are formulated as

$$H_0: p \geq 0.60$$

$$H_A: p < 0.60$$

In this case, the sociologist's concern is valid if the null hypothesis is rejected. Nothing new is established if the null hypothesis is not rejected.

Type I and Type II Errors

LO 9.2

Distinguish between Type I and Type II errors.

Since the decision of a hypothesis test is based on limited sample information, we are bound to make errors. Ideally, we would like to be able to reject the null hypothesis when the null hypothesis is false and not reject the null hypothesis when the null hypothesis is true. However, we may end up rejecting or not rejecting the null hypothesis erroneously. In other words, sometimes we reject the null hypothesis when we should not, or not reject the null hypothesis when we should.

We consider two types of errors in the context of hypothesis testing: a **Type I error** and a **Type II error**. A Type I error is committed when we reject the null hypothesis when the null hypothesis is actually true. On the other hand, a Type II error is made when we do not reject the null hypothesis when the null hypothesis is actually false.

Table 9.2 summarizes the circumstances surrounding Type I and Type II errors. Two correct decisions are possible: not rejecting the null hypothesis when the null hypothesis is true and rejecting the null hypothesis when the null hypothesis is false. Conversely, two incorrect decisions (errors) are also possible: rejecting the null hypothesis when the null hypothesis is true (Type I error) and not rejecting the null hypothesis when the null hypothesis is false (Type II error).

TABLE 9.2 Type I and Type II Errors

Decision	Null hypothesis is true	Null hypothesis is false
Reject the null hypothesis	Type I error	Correct decision
Do not reject the null hypothesis	Correct decision	Type II error

EXAMPLE 9.5

Consider the following hypotheses that relate to the medical example mentioned earlier.

H_0: A person is free of a particular disease

H_A: A person has a particular disease

Suppose the person takes a medical test that attempts to detect this disease. Discuss the consequences of a Type I error and a Type II error.

SOLUTION: A Type I error occurs when the medical test indicates that the person has the disease (reject H_0), but, in reality, the person is free of the disease. We often refer to this type of result as a false positive. If the medical test shows that the person is free of the disease (do not reject H_0), when the person actually has the disease, then a Type II error occurs. We often call this type of result a false negative. Arguably, the consequences of a Type II error in this example are more serious than those of a Type I error.

EXAMPLE 9.6

Consider the following competing hypotheses that relate to the court of law.

H_0: An accused person is innocent

H_A: An accused person is guilty

Suppose the accused person is judged by a jury of her peers. Discuss the consequences of a Type I error and a Type II error.

SOLUTION: A Type I error is a verdict that finds that the accused is guilty (reject H_0) when she is actually innocent. A Type II error is a verdict that finds that the accused is innocent (do not reject H_0) when she is actually guilty. In this example, it is not clear which of the two errors is more costly to society.

As noted in Example 9.6, it is not always easy to determine which of the two errors has more serious consequences. For given evidence, there is a trade-off between these errors; by reducing the likelihood of a Type I error, we implicitly increase the likelihood of a Type II error, and vice versa. The only way we can reduce both errors is by collecting more evidence. Let us denote the probability of a Type I error by α, the probability of a Type II error by β, and the strength of the evidence by the sample size n. Therefore, we can conclude that the only way we can lower both α and β is by increasing n. For a given n, however, we can reduce α only at the expense of a higher β and reduce β only at the expense of a higher α. The optimal choice of α and β depends on the relative cost of these two types of errors, and determining these costs is not always easy. Typically, the decision regarding the optimal level of Type I and Type II errors is made by the management of a firm where the job of a statistician is to conduct the hypothesis test for a chosen value of α.

EXERCISES 9.1

1. Explain why the following hypotheses are not constructed correctly.
   a. $H_0: \mu \leq 10; H_A: \mu \geq 10$
   b. $H_0: \mu \neq 500; H_A: \mu = 500$
   c. $H_0: p \leq 0.40; H_A: p > 0.42$
   d. $H_0: \overline{X} \leq 128; H_A: \overline{X} > 128$
2. Which of the following statements are valid null and alternative hypotheses? If they are invalid hypotheses, explain why.
   a. $H_0: \overline{X} \leq 210; H_A: \overline{X} > 210$
   b. $H_0: \mu = 120; H_A: \mu \neq 120$
   c. $H_0: p \leq 0.24; H_A: p > 0.24$
   d. $H_0: \mu < 252; H_A: \mu > 252$
3. Explain why the following statements are not correct.
   a. "With my methodological approach, I can reduce the Type I error with the given sample information without changing the Type II error."
   b. "I have already decided how much of the Type I error I am going to allow. A bigger sample will not change either the Type I error or the Type II error."
   c. "I can reduce the Type II error by making it difficult to reject the null hypothesis."
   d. "By making it easy to reject the null hypothesis, I am reducing the Type I error."
4. Which of the following statements are correct? Explain if incorrect.
   a. "I accept the null hypothesis since sample evidence is not inconsistent with the null hypothesis."
   b. "Since sample evidence cannot be supported by the null hypothesis, I reject the null hypothesis."
   c. "I can establish a given claim if sample evidence is consistent with the null hypothesis."
   d. "I cannot establish a given claim if the null hypothesis is not rejected."
5. Construct the null and the alternative hypotheses for the following tests:
   a. Test if the mean weight of cereal in a cereal box differs from 18 ounces.
   b. Test if the stock price increases on more than 60% of the trading days.
   c. Test if Americans get an average of less than seven hours of sleep.
6. Define the consequences of Type I and Type II errors for each of the tests considered in the preceding question.
7. Construct the null and the alternative hypotheses for the following claims:
   a. "I am going to get the majority of the votes to win this election."
   b. "I suspect that your 10-inch pizzas are, on average, less than 10 inches in size."
   c. "I will have to fine the company since its tablets do not contain an average of 250 mg of ibuprofen as advertised."
8. Discuss the consequences of Type I and Type II errors for each of the claims considered in the preceding question.
9. A polygraph (lie detector) is an instrument used to determine if an individual is telling the truth. These tests are considered to be 95% reliable. In other words, if an individual lies, there is a 0.95 probability that the test will detect a lie. Let there also be a 0.005 probability that the test erroneously detects a lie even when the individual is actually telling the truth. Consider the null hypothesis, "the individual is telling the truth," to answer the following questions.
   a. What is the probability of a Type I error?
   b. What is the probability of a Type II error?
   c. What are the consequences of Type I and Type II errors?
   d. What is wrong with the statement, "I can prove that the individual is telling the truth on the basis of the polygraph result"?
10. The manager of a large manufacturing firm is considering switching to new and expensive software that promises to reduce its assembly costs. Before purchasing the software, the manager wants to conduct a hypothesis test to determine if the new software does reduce its assembly costs.
    a. Would the manager of the manufacturing firm be more concerned about a Type I error or a Type II error? Explain.
    b. Would the software company be more concerned about a Type I error or a Type II error? Explain.
11. The screening process for detecting a rare disease is not perfect. Researchers have developed a blood test that is considered fairly reliable. It gives a positive reaction in 98% of the people who have that disease. However, it erroneously gives a positive reaction in 3% of the people who do not have the disease. Consider the null hypothesis "the individual does not have the disease" to answer the following questions.
    a. What is the probability of a Type I error?
    b. What is the probability of a Type II error?
    c. What are the consequences of Type I and Type II errors?
    d. What is wrong with the nurse's analysis, "The blood test result has proved that the individual is free of disease"?
12. A consumer group has accused a restaurant of using higher fat content than what is reported on its menu. The group has been asked to conduct a hypothesis test to substantiate its claims.
    a. Is the manager of the restaurant more concerned about a Type I error or a Type II error? Explain.
    b. Is the consumer group more concerned about a Type I error or a Type II error? Explain.

9.2 HYPOTHESIS TEST FOR THE POPULATION MEAN WHEN σ IS KNOWN

In order to introduce the basic methodology for hypothesis testing, we first conduct a hypothesis test regarding the population mean μ under the assumption that the population standard deviation σ is known. While it is true that σ is rarely known, there are instances when σ is considered fairly stable and, therefore, can be determined from prior experience. In such cases, σ is treated as known. Fortunately, this assumption has no bearing on the overall procedure of conducting a hypothesis test, a procedure we use throughout the remainder of the text.

A hypothesis test regarding the population mean μ is based on the sampling distribution of the sample mean $\overline{X}$. In particular, it uses the fact that $E(\overline{X}) = \mu$ and $se(\overline{X}) = \sigma/\sqrt{n}$. Also, in order to implement the test, it is essential that $\overline{X}$ is normally distributed. Recall that $\overline{X}$ is normally distributed when the underlying population is normally distributed. If the underlying population is not normally distributed, then, by the central limit theorem, $\overline{X}$ is approximately normally distributed if the sample size is sufficiently large—that is, $n \geq 30$.

The basic principle of hypothesis testing is to first assume that the null hypothesis is true and then determine if sample evidence contradicts this assumption. This principle is analogous to the scenario in the court of law where the null hypothesis is defined as "the individual is innocent" and the decision rule is best described by "innocent until proven guilty."

There are two approaches to implementing a hypothesis test—the p-value approach and the critical value approach. The critical value approach is attractive when a computer is unavailable and all calculations must be done by hand. We discuss this approach in the appendix to this chapter. Most researchers and practitioners, however, favor the p-value approach since virtually every statistical software package reports p-values. In this text, we too will focus on the p-value approach. We implement a four-step procedure that is valid for one- and two-tailed tests regarding the population mean, the population proportion, or any other population parameter of interest.

LO **9.3**

Conduct a hypothesis test for the population mean when σ is known.

The *p*-Value Approach

Suppose a sociologist wants to establish that the mean retirement age is greater than 67 ($\mu > 67$). It is assumed that retirement age is normally distributed with a known population standard deviation of 9 years ($\sigma = 9$). We can investigate the sociologist's belief by specifying the competing hypotheses as

$$H_0: \mu \leq 67$$
$$H_A: \mu > 67$$

Let a random sample of 25 retirees produce an average retirement age of 71—that is, $\bar{x} = 71$. This sample evidence casts doubt on the validity of the null hypothesis, since the sample mean is greater than the hypothesized value, $\mu_0 = 67$. However, the discrepancy between $\bar{x}$ and μ_0 does not necessarily imply that the null hypothesis is false. Perhaps the discrepancy can be explained by pure chance. It is common to evaluate this discrepancy in terms of the appropriate **test statistic**.

TEST STATISTIC FOR μ WHEN σ IS KNOWN

The value of the test statistic for the hypothesis test of the population mean μ when the population standard deviation σ is known is computed as

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}},$$

where μ_0 is the hypothesized value of the population mean. This formula is valid only if $\overline{X}$ (approximately) follows a normal distribution.

Note that the value of the test statistic z is evaluated at $\mu = \mu_0$, which explains why we need some form of the equality sign in the null hypothesis. Given that the population is normally distributed with a known standard deviation, $\sigma = 9$, we compute the value of the test statistic as $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{71 - 67}{9/\sqrt{25}} = 2.22$. Therefore, comparing $\bar{x} = 71$ with 67 is identical to comparing $z = 2.22$ with 0, where 67 and 0 are the means of $\overline{X}$ and Z, respectively.

We now find the ***p*-value**, which is the likelihood of obtaining a sample mean that is at least as extreme as the one derived from the given sample, under the assumption that the null hypothesis is true as an equality—that is, $\mu_0 = 67$. Since in this example $\bar{x} = 71$, we define the extreme value as a sample mean of 71 or higher and use the z table to find the p-value as $P(\overline{X} \geq 71) = P(Z \geq 2.22) = 1 - 0.9868 = 0.0132$. Figure 9.1 shows the computed p-value.

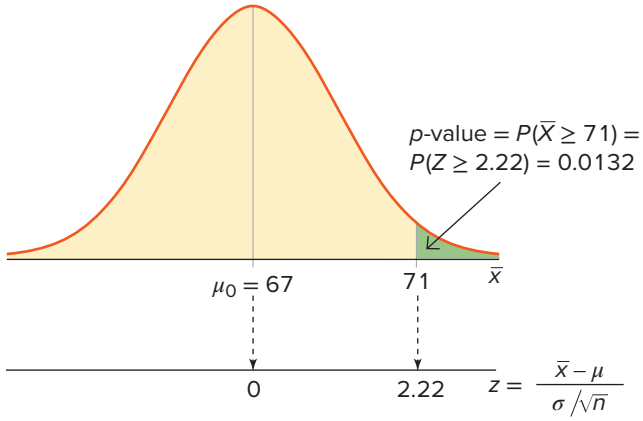


FIGURE 9.1 The p-value for a right-tailed test with $z = 2.22$

Note that when the null hypothesis is true, there is only a 1.32% chance that the sample mean will be 71 or more. This seems like a very small chance, but is it small enough to allow us to reject the null hypothesis in favor of the alternative hypothesis? Let's see how we define "small enough."

Remember that a Type I error occurs when we reject the null hypothesis when it is actually true. We define the *allowed* probability of making a Type I error as α; we refer to $100\alpha\%$ as the **significance level**. The p-value, on the other hand, is referred to as the *observed* probability of making a Type I error. When using the p-value approach, **the decision rule is to reject the null hypothesis if the *p*-value $< \alpha$ and not reject the null hypothesis if the *p*-value $\geq \alpha$.**

We generally choose a value for α *before* implementing a hypothesis test; that is, we set the rules of the game before playing. Care must be exercised in choosing α because important decisions are often based on the results of a hypothesis test, which in turn depend on α. Most hypothesis tests are conducted using a significance level of 1%, 5%, or 10%, using $\alpha = 0.01$, 0.05, or 0.10, respectively. For example, $\alpha = 0.05$ means that we allow a 5% chance of rejecting a true null hypothesis. It is customary to interpret these conventional significance levels as follows:

- If we reject a null hypothesis at the 10% significance level ($\alpha = 0.10$), then we have *some evidence* that the null hypothesis is false;
- If we reject a null hypothesis at the 5% significance level ($\alpha = 0.05$), then we have *strong evidence* that the null hypothesis is false; and
- If we reject a null hypothesis at the 1% significance level ($\alpha = 0.01$), then we have *very strong evidence* that the null hypothesis is false.

In our example, given the p-value of 0.0132, if we decide to reject the null hypothesis, then there is a 1.32% chance that our decision will be erroneous.

Suppose we had chosen $\alpha = 0.05$ to conduct the above test. At this significance level, we reject the null hypothesis because $0.0132 < 0.05$. This means that the sample data support the sociologist's claim that the average retirement age is greater than 67 years

old. Individuals may be working past the normal retirement age of 67 because of poor savings and/or because this generation is expected to outlive any previous generation and needs jobs to pay the bills. We should note that if α had been set at 0.01, then the findings would have been different. At this smaller significance level, the evidence does not allow us to reject the null hypothesis ($0.0132 > 0.01$). At the 1% significance level, we cannot conclude that the mean retirement age is greater than 67.

In the retirement age example, we calculate the p-value as $P(Z \geq z)$ since it is a right-tailed test. Analogously, for a left-tailed test, the p-value is given by $P(Z \leq z)$. For a two-tailed test, the extreme values exist on both sides of the distribution of the test statistic. Given the symmetry of the z distribution, the p-value for a two-tailed test is twice that of the p-value for a one-tailed test. It is calculated as $2P(Z \geq z)$ if $z > 0$ or as $2P(Z \leq z)$ if $z < 0$.

THE p-VALUE APPROACH

Under the assumption that $\mu = \mu_0$, the p-value is the likelihood of observing a sample mean that is at least as extreme as the one derived from the given sample. Its calculation depends on the specification of the alternative hypothesis.

Alternative Hypothesis	p-value
H_A: $\mu > \mu_0$	Right-tail probability: $P(Z \geq z)$
H_A: $\mu < \mu_0$	Left-tail probability: $P(Z \leq z)$
H_A: $\mu \neq \mu_0$	Two-tail probability: $2P(Z \geq z)$ if $z > 0$ or $2P(Z \leq z)$ if $z < 0$

The decision rule is:

- Reject H_0 if the p-value $< \alpha$, or
- Do not reject H_0 if p-value $\geq \alpha$.

Figure 9.2 shows the three different scenarios of determining the p-value depending on the specification of the competing hypotheses.

FIGURE 9.2 The p-values for one- and two-tailed tests

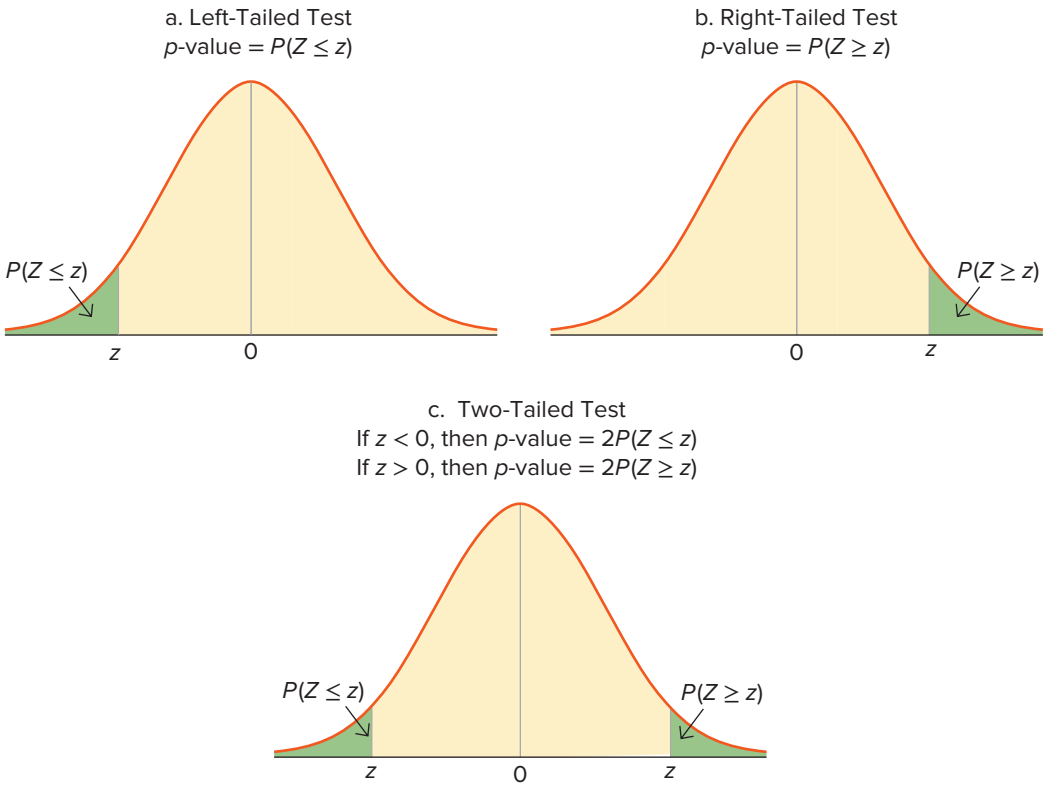


Figure 9.2a shows the p-value for a left-tailed test. Since the appropriate test statistic follows the standard normal distribution, we compute the p-value as $P(Z \leq z)$. When calculating the p-value for a right-tailed test (see Figure 9.2b), we find the area to the right of z or, equivalently, $P(Z \geq z)$. Figure 9.2c shows the p-value for a two-tailed test, calculated as $2P(Z \leq z)$ when $z < 0$ or as $2P(Z \geq z)$ when $z > 0$.

It is important to note that we *cannot* reject H_0 for a right-tailed test if $\bar{x} \leq \mu_0$ or, equivalently, $z \leq 0$. Consider, for example, a right-tailed test with the hypotheses specified as H_0: $\mu \leq 67$ versus H_A: $\mu > 67$. Here, if $\bar{x} = 65$, there is no need for formal testing since we have no discrepancy between the sample mean and the hypothesized value of the population mean. Similarly, we *cannot* reject H_0 for a left-tailed test if $\bar{x} \geq \mu_0$ or, equivalently, $z \geq 0$. We will now summarize the four-step procedure using the p-value approach.

THE FOUR-STEP PROCEDURE USING THE p-VALUE APPROACH

Step 1. Specify the null and the alternative hypotheses. It is important to (1) identify the relevant population parameter, (2) determine whether a one- or a two-tailed test is appropriate, and (3) include some form of the equality sign in the null hypothesis and use the alternative hypothesis to establish a claim.

Step 2. Specify the significance level. Before implementing a hypothesis test, the significance level α is specified, which specifies the *allowed* probability of making a Type I error.

Step 3. Calculate the value of the test statistic and the p-value. When testing the population mean μ and the population standard deviation σ is known, the value of the test statistic is $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$, where μ_0 is the hypothesized value of the population mean. For a right-tailed test, the p-value is $P(Z \geq z)$, and for a left-tailed test, the p-value is $P(Z \leq z)$. For a two-tailed test, the p-value is $2P(Z \geq z)$ if $z > 0$, or $2P(Z \leq z)$ if $z < 0$.

Step 4. State the conclusion and interpret results. The decision rule is to reject the null hypothesis when the p-value $< \alpha$ and not reject the null hypothesis when the p-value $\geq \alpha$. Clearly interpret the results in the context of the problem.

EXAMPLE 9.7

A research analyst disputes a trade group's prediction that back-to-school spending will average $606.40 per family this year. She believes that average back-to-school spending will differ from this amount. She decides to conduct a test on the basis of a random sample of 30 households with school-age children. She calculates the sample mean as $622.85. She also believes that back-to-school spending is normally distributed with a population standard deviation of $65. She wants to conduct the test at the 5% significance level.

a. Specify the competing hypotheses in order to test the research analyst's claim.

b. What is the allowed probability of a Type I error?

c. Calculate the value of the test statistic and the p-value.

d. At the 5% significance level, does average back-to-school spending differ from $606.40?

SOLUTION:

a. Since we want to determine if the average is different from the predicted value of \$606.40, we specify the hypotheses as

$$H_0: \mu = 606.40$$
$$H_A: \mu \neq 606.40$$

b. The allowed probability of a Type I error is equivalent to the significance level of the test, which in this example is given as $\alpha = 0.05$.

c. Note that $\overline{X}$ is normally distributed since it is computed from a random sample drawn from a normal population. Since σ is known, the test statistic follows the standard normal distribution, and its value is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{622.85 - 606.40}{65/\sqrt{30}} = 1.39.$$

For a two-tailed test with a positive value for the test statistic, we find the p-value as $2P(Z \geq 1.39)$. From the z table, we first find $P(Z \geq 1.39) = 1 - 0.9177 = 0.0823$; so the p-value $= 2 \times 0.0823 = 0.1646$.

d. The decision rule is to reject the null hypothesis if the p-value is less than α. Since $0.1646 > 0.05$, we do not reject H_0. Therefore, at the 5% significance level, we cannot conclude that average back-to-school spending differs from \$606.40 per family this year. The sample data do not support the research analyst's claim.

Confidence Intervals and Two-Tailed Hypothesis Tests

A confidence interval for the population parameter is sometimes used as an alternative method for conducting a two-tailed hypothesis test. Informally, we had used this procedure when discussing confidence intervals in Chapter 8. Given that we conduct the hypothesis test at the α significance level, we can use the sample data to determine a corresponding $100(1 - \alpha)\%$ confidence interval for the population mean μ. If the confidence interval does not contain the hypothesized value of the population mean μ_0, then we reject the null hypothesis. If the confidence interval contains μ_0, then we do not reject the null hypothesis.

IMPLEMENTING A TWO-TAILED TEST USING A CONFIDENCE INTERVAL

The general specification for a $100(1 - \alpha)\%$ confidence interval for the population mean μ when the population standard deviation σ is known is computed as

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Given a hypothesized value of the population mean μ_0, the decision rule is:

- Reject H_0 if μ_0 does not fall within the confidence interval, or
- Do not reject H_0 if μ_0 falls within the confidence interval.

EXAMPLE 9.8

Use the confidence interval approach to conduct the hypothesis test described in Example 9.7.

SOLUTION: We are testing H_0: $\mu = 606.40$ versus H_A: $\mu \neq 606.40$ at the 5% significance level. We use $n = 30$, $\bar{x} = 622.85$, and $\sigma = 65$, along with $\alpha = 0.05$, to determine the 95% confidence interval for μ. We find $z_{\alpha/2} = z_{0.025} = 1.96$ and compute

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 622.85 \pm 1.96\,\frac{65}{\sqrt{30}} = 622.85 \pm 23.26,$$

resulting in the interval [599.59, 646.11]. Since the hypothesized value of the population mean $\mu_0 = 606.40$ falls within the 95% confidence interval, we do not reject H_0. Thus, we arrive at the same conclusion as with the p-value approach; that is, the sample data do not support the research analyst's claim that average back-to-school spending differs from \$606.40 per family this year.

As shown above, we use the confidence interval as an alternative method for conducting a two-tailed test. It is possible to adjust the confidence interval to accommodate a one-tailed test, but we do not discuss this adjustment in this text.

Using Excel to Test μ When σ Is Known

Excel provides various functions that simplify the steps of conducting a hypothesis test. We use the following example to demonstrate some of these functions.

EXAMPLE 9.9

A report in *The New York Times* (August 7, 2010) suggests that consumers are spending less due to a realization that excessive spending does not make them happier. A researcher wants to use debit card data to contradict the generally held view that the average amount spent annually on a debit card is at least \$8,000. She surveys 20 consumers and asks them how much they spend annually on their debit cards. The results are given below.

FILE
Debit_Spending

7960	7700	7727	7704	8543	7661	7767	8761	7530	8128
7938	7771	7272	8113	7727	7697	7690	8000	8079	7547

It is assumed that the population standard deviation is \$500 and that spending on debit cards is normally distributed. Test the claim at the 1% level of significance.

SOLUTION: The researcher would like to establish that average spending on debit cards is less than \$8,000 or, equivalently, $\mu < 8{,}000$. Thus, we formulate the competing hypotheses as

$$H_0: \mu \geq 8{,}000$$
$$H_A: \mu < 8{,}000$$

The normality condition of $\overline{X}$ is satisfied since spending on debit cards is assumed to be normally distributed. Also, since the population standard deviation is known, here $\sigma = 500$, the test statistic is assumed to follow the z distribution.

a. Open the ***Debit_Spending*** data file. Note that the values for spending are in cells A2 through A21.

b. We use Excel's **AVERAGE** function to help in the calculation of the value of the test statistic $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$. We enter "=(AVERAGE(A2:A21) − 8000)/(500/SQRT(20))", and Excel returns −1.2008, so $z = -1.2008$.

c. We use Excel's **NORM.S.DIST** function, first discussed in Chapter 6, to find the p-value. Recall that in order to find $P(Z \leq z)$, we enter "=NORM.S.DIST(z, 1)" where z is the value for which we want to evaluate the cumulative probability. If we enter "=1 − NORM.S.DIST(z, 1)", then Excel returns $P(Z \geq z)$. In order to find $P(Z \leq -1.2008)$, we enter "=NORM.S.DIST(−1.2008, 1)". Excel returns 0.1149.

d. The hypothesis test is conducted at the 1% significance level. Thus, since the p-value of 0.1149 is not less than $\alpha = 0.01$, we do not reject the null hypothesis. In other words, at the 1% significance level, the researcher cannot conclude that annual spending on debit cards is less than $8,000. Perhaps these findings can be reconciled with a report that claims that individuals are shunning their credit cards and using debit cards to avoid incurring more debt (www.businessweek.com, September 8, 2010).

One Last Remark

An important component of any well-executed statistical analysis is to clearly communicate the results. Thus, it is not sufficient to end the analysis with a conclusion that you reject the null hypothesis or you do not reject the null hypothesis. You must interpret the results, clearly reporting whether or not the claim regarding the population parameter of interest can be justified on the basis of the sample information.

EXERCISES 9.2

Mechanics

13. Consider the following hypotheses:

$$H_0: \mu \leq 12.6$$
$$H_A: \mu > 12.6$$

A sample of 25 observations yields a sample mean of 13.4. Assume that the sample is drawn from a normal population with a population standard deviation of 3.2.

a. Calculate the p-value. What is the conclusion if $\alpha = 0.10$?
b. Calculate the p-value if the sample mean was based on a sample of 100 observations. What is the conclusion if $\alpha = 0.10$?

14. Consider the following hypotheses:

$$H_0: \mu = 100$$
$$H_A: \mu \neq 100$$

A sample of 16 observations yields a sample mean of 95. Assume that the sample is drawn from a normal population with a population standard deviation of 10.

a. Calculate the value of the test statistic.
b. Find the p-value.
c. At the 10% significance level, what is the conclusion?

15. Consider the following hypotheses:

$$H_0: \mu \geq 150$$
$$H_A: \mu < 150$$

A sample of 80 observations results in a sample mean of 144. The population standard deviation is known to be 28.

a. Calculate the value of the test statistic and the p-value.
b. Does the above sample evidence enable us to reject the null hypothesis at $\alpha = 0.01$?
c. Does the above sample evidence enable us to reject the null hypothesis at $\alpha = 0.05$?

16. A researcher wants to determine if the population mean is greater than 45. A random sample of 36 observations yields a sample mean of 47. Assume that the population standard deviation is 8.

a. Specify the competing hypotheses to test the researcher's claim.
b. Calculate the value of the test statistic.
c. Find the p-value.
d. At the 5% significance level, what is the conclusion?

17. Consider the following hypotheses:

$$H_0: \mu = 1{,}800$$
$$H_A: \mu \neq 1{,}800$$

The population is normally distributed with a population standard deviation of 440. Compute the value of the test statistic and the resulting p-value for each of the following sample results. For each sample, determine if you can reject the null hypothesis at the 10% significance level.

a. $\bar{x} = 1{,}850$; $n = 110$
b. $\bar{x} = 1{,}850$; $n = 280$
c. $\bar{x} = 1{,}650$; $n = 32$
d. $\bar{x} = 1{,}700$; $n = 32$

18. Consider the following hypothesis test:

$$H_0: \mu \leq -5$$
$$H_A: \mu > -5$$

A random sample of 50 observations yields a sample mean of −3. The population standard deviation is 10. Calculate the *p*-value. What is the conclusion to the test if $\alpha = 0.05$?

19. Consider the following hypothesis test:

$$H_0: \mu \leq 75$$
$$H_A: \mu > 75$$

A random sample of 100 observations yields a sample mean of 80. The population standard deviation is 30. Calculate the *p*-value. What is the conclusion to the test if $\alpha = 0.10$?

20. Consider the following hypothesis test:

$$H_0: \mu = -100$$
$$H_A: \mu \neq -100$$

A random sample of 36 observations yields a sample mean of −125. The population standard deviation is 42. Conduct the test at $\alpha = 0.01$.

21. Consider the following hypotheses:

$$H_0: \mu = 120$$
$$H_A: \mu \neq 120$$

The population is normally distributed with a population standard deviation of 46.

a. If $\bar{x} = 132$ and $n = 50$, what is the conclusion at the 5% significance level?

b. If $\bar{x} = 108$ and $n = 50$, what is the conclusion at the 10% significance level?

22. FILE ***Excel_1.*** Given the accompanying sample data, use Excel's formula options to determine if the population mean is less than 125 at the 5% significance level. Assume that the population is normally distributed and that the population standard deviation equals 12.

23. FILE ***Excel_2.*** Given the accompanying sample data, use Excel's formula options to determine if the population mean differs from 3 at the 5% significance level. Assume that the population is normally distributed and that the population standard deviation equals 5.

Applications

24. It is advertised that the average braking distance for a small car traveling at 65 miles per hour equals 120 feet. A transportation researcher wants to determine if the statement made in the advertisement is false. She randomly test drives 36 small cars at 65 miles per hour and records the braking distance. The sample average braking distance is computed as 114 feet. Assume that the population standard deviation is 22 feet.

a. State the null and the alternative hypotheses for the test.

b. Calculate the value of the test statistic and the *p*-value.

c. Use $\alpha = 0.01$ to determine if the average breaking distance differs from 120 feet.

25. Customers at Costco spend an average of $130 per trip (*The Wall Street Journal,* October 6, 2010). One of Costco's rivals would like to determine whether its customers spend more per trip. A survey of the receipts of 25 customers found that the sample mean was $135.25. Assume that the population standard deviation is $10.50 and that spending follows a normal distribution.

a. Specify the null and alternative hypotheses to test whether average spending at the rival's store is more than $130.

b. Calculate the value of the test statistic and the *p*-value.

c. At the 5% significance level, what is the conclusion to the test?

26. In May 2008, CNN reported that sports utility vehicles (SUVs) are plunging toward the "endangered" list. Due to the uncertainty of oil prices and environmental concerns, consumers are replacing gas-guzzling vehicles with fuel-efficient smaller cars. As a result, there has been a big drop in the demand for new as well as used SUVs. A sales manager of a used car dealership for SUVs believes that it takes more than 90 days, on average, to sell an SUV. In order to test his claim, he samples 40 recently sold SUVs and finds that it took an average of 95 days to sell an SUV. He believes that the population standard deviation is fairly stable at 20 days.

a. State the null and the alternative hypotheses for the test.

b. What is the *p*-value?

c. Is the sales manager's claim justified at $\alpha = 0.01$?

27. According to the *Centers for Disease Control and Prevention* (February 18, 2016), 1 in 3 American adults do not get enough sleep. A researcher wants to determine if Americans are sleeping less than the recommended 7 hours of sleep on weekdays. He takes a random sample of 150 Americans and computes the average sleep time of 6.7 hours on weekdays. Assume that the population is normally distributed with a known standard deviation of 2.1 hours. Test the researcher's claim at $\alpha = 0.01$.

28. A local bottler in Hawaii wishes to ensure that an average of 16 ounces of passion fruit juice is used to fill each bottle. In order to analyze the accuracy of the bottling process, he takes a random sample of 48 bottles. The mean weight of the passion fruit juice in the sample is 15.80 ounces. Assume that the population standard deviation is 0.8 ounce.

a. State the null and the alternative hypotheses to test if the bottling process is inaccurate.

b. What is the value of the test statistic and the *p*-value?

c. At $\alpha = 0.05$, what is the conclusion to the hypothesis test? Make a recommendation to the bottler.

29. FILE ***MV_Houses.*** A realtor in Mission Viejo, California, believes that the average price of a house is more than $500,000.

a. State the null and the alternative hypotheses for the test.

b. The data accompanying this exercise show house prices. (Data are in $1,000s.) Assume the population standard

deviation is $100 (in $1,000s). What is the value of the test statistic and the p-value?

c. At $\alpha = 0.05$, what is the conclusion to the test? Is the realtor's claim supported by the data?

30. FILE ***Home_Depot.*** The data accompanying this exercise show the weekly stock price for Home Depot. Assume that stock prices are normally distributed with a population standard deviation of $3.

a. State the null and the alternative hypotheses in order to test whether or not the average weekly stock price differs from $30.

b. Find the value of the test statistic and the p-value.

c. At $\alpha = 0.05$, can you conclude that the average weekly stock price does not equal $30?

31. FILE ***Hourly_Wage.*** An economist wants to test if the average hourly wage is less than $22. Assume that the population standard deviation is $6.

a. State the null and the alternative hypotheses for the test.

b. The data accompanying this exercise show hourly wages. Find the value of the test statistic and the p-value.

c. At $\alpha = 0.05$, what is the conclusion to the test? Is the average hourly wage less than $22?

32. FILE ***CT_Undergrad_Debt.*** On average, a college student graduates with $27,200 in debt (*The Boston Globe,* May 27, 2012). The data accompanying this exercise show the debt for 40 recent undergraduates from Connecticut. Assume that the population standard deviation is $5,000.

a. A researcher believes that recent undergraduates from Connecticut have less debt than the national average. Specify the competing hypotheses to test this belief.

b. Find the value of the test statistic and the p-value.

c. Do the data support the researcher's claim, at $\alpha = 0.10$?

LO **9.4**

Conduct a hypothesis test for the population mean when σ is unknown.

9.3 HYPOTHESIS TEST FOR THE POPULATION MEAN WHEN σ IS UNKNOWN

So far we have considered hypothesis tests for the population mean μ under the assumption that the population standard deviation σ is known. In most business applications, σ is not known and we have to replace σ with the sample standard deviation s to estimate the standard error of $\overline{X}$.

TEST STATISTIC FOR μ WHEN σ IS UNKNOWN

The value of the test statistic for the hypothesis test of the population mean μ when the population standard deviation σ is unknown is computed as

$$t_{df} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

where μ_0 is the hypothesized value of the population mean, s is the sample standard deviation, n is the sample size, and the degrees of freedom $df = n - 1$. This formula is valid only if $\overline{X}$ (approximately) follows a normal distribution.

The next two examples show how we use the four-step procedure for hypothesis testing when we are testing the population mean μ and the population standard deviation σ is unknown.

EXAMPLE 9.10

FILE
Study_Hours

In the introductory case to this chapter, the dean at a large university in California wonders if students at her university study less than the 1961 national average of 24 hours per week. She randomly selects 35 students and asks their average study time per week (in hours). From their responses, she calculates a sample mean of 16.3714 hours and a sample standard deviation of 7.2155 hours.

a. Specify the competing hypotheses to test the dean's concern.

b. Calculate the value of the test statistic.

c. Find the p-value.

d. At the 5% significance level, what is the conclusion to the hypothesis test?

SOLUTION:

a. This is an example of a one-tailed test where we would like to determine if the mean hours studied is less than 24; that is, $\mu < 24$. We formulate the competing hypotheses as

$$H_0: \mu \geq 24 \text{ hours}$$
$$H_A: \mu < 24 \text{ hours}$$

b. Recall that for any statistical inference regarding the population mean, it is essential that the sample mean $\overline{X}$ is normally distributed. This condition is satisfied because the sample size is greater than 30, specifically $n = 35$. The degrees of freedom, $df = n - 1 = 34$. Given $\bar{x} = 16.3714$ and $s = 7.2155$, we compute the value of the test statistic as

$$t_{34} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{16.3714 - 24}{7.2155/\sqrt{35}} = -6.255.$$

c. Since this is a left-tailed test, we compute the p-value as $P(T_{34} \leq t_{34})$. Table 9.3 shows a portion of the t table. Referencing Table 9.3 for $df = 34$, we find that the exact probability $P(T_{34} \leq -6.255)$, which is equivalent to $P(T_{34} \geq 6.255)$, cannot be determined. Since 6.255 is larger than any value in this row, it implies that the p-value is less than 0.005. In other words, we approximate the p-value as $P(T_{34} \leq -6.255) < 0.005$. In the next example, we will show how to use Excel to obtain exact p-values.

TABLE 9.3 Portion of the t Table

	Area in Upper Tail					
df	0.20	0.10	0.05	0.025	0.01	0.005
1	1.376	3.078	6.341	12.706	31.821	63.657
	⋮	⋮	⋮	⋮	⋮	⋮
34	0.852	1.307	1.691	2.032	2.441	2.728

d. We reject the null hypothesis since the p-value is less than $\alpha = 0.05$. At the 5% significance level, we conclude that the average study time at the university is less than the 1961 average of 24 hours per week.

Using Excel to Test μ When σ is Unknown

Again we find that functions in Excel are quite useful when calculating the value of the test statistic and the exact p-value. Consider the following example.

EXAMPLE 9.11

As the introductory case to this chapter mentions, research finds that today's undergraduates study an average of 14 hours per week. Using the sample data from Table 9.1, the dean would also like to test if the mean study time of students at her university differs from today's national average of 14 hours per week. At the 5% significance level, what is the conclusion to this test?

FILE
Study_Hours

SOLUTION: Since the dean would like to test if the mean study time of students at her university differs from 14 hours per week, we formulate the competing hypotheses for the test as

$$H_0: \mu = 14 \text{ hours}$$
$$H_A: \mu \neq 14 \text{ hours}$$

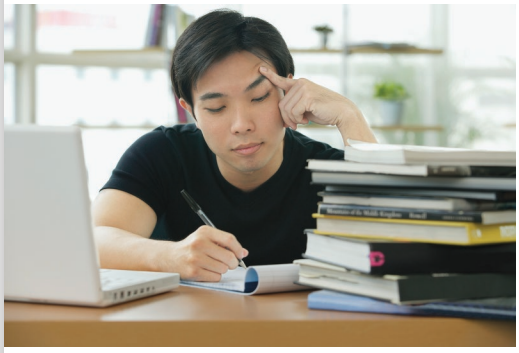
a. Open the ***Study_Hours*** data file. Note that the values for hours studied are in cells A2 through A36.

b. We use Excel's **AVERAGE** and **STDEV.S** functions to help in the calculation of the value of the test statistic $t_{df} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$. We enter "=(AVERAGE(A2:A36) − 14)/(STDEV.S(A2:A36)/SQRT(35))". Excel returns 1.9444, so $t_{34} = 1.9444$.

c. Even though Excel offers a number of functions that generate *p*-values, we use the **T.DIST.RT** function. If we enter "=T.DIST.RT(t_{df}, *df*)", where t_{df} is the value of the test statistic and *df* is the relevant degrees of freedom, then Excel returns $P(T_{df} \geq t_{df})$; this probability is the *p*-value for a right-tailed test. If we enter "=1 − T.DIST.RT (t_{df}, *df*)", then Excel returns $P(T_{df} \leq t_{df})$. Thus, in order to find the exact probability for this two-tailed hypothesis test where $t_{34} = 1.9444$, we enter "=2*T.DIST.RT(1.9444, 34)". Excel returns 0.0602.

d. Since the *p*-value of 0.0602 is not less than $\alpha = 0.05$, we do not reject the null hypothesis. At the 5% significance level, we cannot conclude that the mean study time of students at the university is different from today's national average of 14 hours per week.

SYNOPSIS OF INTRODUCTORY CASE



©Asia Images Group/Getty Images

A report claims that undergraduates are studying far less today as compared to five decades ago (*The Boston Globe,* July 4, 2010). The report finds that in 1961, students invested 24 hours per week in their academic pursuits, whereas today's students study an average of 14 hours per week. In an attempt to determine whether or not this national trend is present at a large university in California, 35 students are randomly selected and asked their average study time per week (in hours). The sample produces a mean of 16.37 hours with a standard deviation of 7.22 hours. Two hypothesis tests are conducted. The first test examines whether the mean study time of students at this university is below the 1961 national average of 24 hours per week. At the 5% significance level, the sample data suggest that the mean is less than 24 hours per week. The second test investigates whether the mean study time of students at this university differs from today's national average of 14 hours per week. At the 5% significance level, the results do not suggest that the mean study time is different from 14 hours per week. Thus, the sample results support the overall findings of the report: undergraduates study, on average, 14 hours per week, far below the 1961 average of 24 hours per week. The present analysis, however, does not explain why that might be the case. For instance, it cannot be determined whether students have just become lazier, or if with the advent of the computer, they can access information in less time.

EXERCISES 9.3

Mechanics

33. Consider the following hypotheses:

$$H_0: \mu \leq 210$$
$$H_A: \mu > 210$$

Find the *p*-value for this test based on the following sample information.

a. $\bar{x} = 216; s = 26; n = 40$
b. $\bar{x} = 216; s = 26; n = 80$
c. $\bar{x} = 216; s = 16; n = 40$
d. $\bar{x} = 214; s = 16; n = 40$

34. Which of the sample information in the preceding question enables us to reject the null hypothesis at $\alpha = 0.01$ and at $\alpha = 0.10$?

35. Consider the following hypotheses:

$$H_0: \mu = 12$$
$$H_A: \mu \neq 12$$

Find the *p*-value for this test based on the following sample information.

a. $\bar{x} = 11; s = 3.2; n = 36$
b. $\bar{x} = 13; s = 3.2; n = 36$
c. $\bar{x} = 11; s = 2.8; n = 36$
d. $\bar{x} = 11; s = 2.8; n = 49$

36. Which of the sample information in the preceding question enables us to reject the null hypothesis at $\alpha = 0.01$ and at $\alpha = 0.10$?

37. Consider the following hypotheses:

$$H_0: \mu = 50$$
$$H_A: \mu \neq 50$$

A sample of 16 observations yields a sample mean of 46. Assume that the sample is drawn from a normal population with a sample standard deviation of 10.

a. Calculate the value of the test statistic.
b. At the 5% significance level, does the population mean differ from 50? Explain.

38. In order to test if the population mean differs from 16, you draw a random sample of 32 observations and compute the sample mean and the sample standard deviation as 15.2 and 0.6, respectively. Conduct the test at the 1% level of significance.

39. In order to conduct a hypothesis test for the population mean, a random sample of 24 observations is drawn from a normally distributed population. The resulting sample mean and sample standard deviation are calculated as 4.8 and 0.8, respectively. Conduct the following tests at $\alpha = 0.05$.

a. $H_0: \mu \leq 4.5$ against $H_A: \mu > 4.5$
b. $H_0: \mu = 4.5$ against $H_A: \mu \neq 4.5$

40. Consider the following hypotheses:

$$H_0: \mu \geq -10$$
$$H_A: \mu < -10$$

A sample of 25 observations yields a sample mean of −12. Assume that the sample is drawn from a normal population with a sample standard deviation of 4.

a. Calculate the value of the test statistic.
b. At the 5% significance level, is the population mean less than −10? Explain.

41. Consider the following hypotheses:

$$H_0: \mu = 8$$
$$H_A: \mu \neq 8$$

The population is normally distributed. A sample produces the following observations:

6	9	8	7	7	11	10

Conduct the test at the 5% level of significance.

42. Consider the following hypotheses:

$$H_0: \mu \geq 100$$
$$H_A: \mu < 100$$

The population is normally distributed. A sample produces the following observations:

95	99	85	80	98	97

Conduct the test at the 1% level of significance.

43. FILE ***Excel_1.*** Given the accompanying sample data, use Excel's formula options to determine if the population mean is greater than 116 at the 5% significance level. Assume that the population is normally distributed.

44. FILE ***Excel_2.*** Given the accompanying sample data, use Excel's formula options to determine if the population mean differs from 1 at the 5% significance level. Assume that the population is normally distributed.

Applications

45. A machine that is programmed to package 1.20 pounds of cereal in each cereal box is being tested for its accuracy. In a sample of 36 cereal boxes, the mean and the standard deviation are calculated as 1.22 pounds and 0.06 pound, respectively.

a. Set up the null and the alternative hypotheses to determine if the machine is working improperly—that is, it is either underfilling or overfilling the cereal boxes.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 5% level of significance, can you conclude that the machine is working improperly? Explain.

46. The manager of a small convenience store does not want her customers standing in line for too long prior to a purchase. In

particular, she is willing to hire an employee for another cash register if the average wait time of the customers is more than five minutes. She randomly observes the wait time (in minutes) of customers during the day as:

3.5	5.8	7.2	1.9	6.8	8.1	5.4

a. Set up the null and the alternative hypotheses to determine if the manager needs to hire another employee.
b. Calculate the value of the test statistic and the *p*-value. What assumption regarding the population is necessary to implement this step?
c. Decide whether the manager needs to hire another employee at $\alpha = 0.10$.

47. Small, energy-efficient, Internet-centric, new computers are increasingly gaining popularity (*The New York Times,* July 20, 2008). Some of the biggest companies are wary of the new breed of computers because their low price could threaten PC makers' already thin profit margins. An analyst comments that the larger companies have a cause for concern since the mean price of these small computers has fallen below $350. She examines six popular brands of these small computers and records their retail prices as:

322	269	373	412	299	389

a. What assumption regarding the distribution of the price of small computers is necessary to test the analyst's claim?
b. Specify the null and alternative hypotheses to test the analyst's claim.
c. Calculate the value of the test statistic and the *p*-value.
d. At the 5% significance level, what is the conclusion to the test? Should the larger computer companies be concerned?

48. A local brewery wishes to ensure that an average of 12 ounces of beer is used to fill each bottle. In order to analyze the accuracy of the bottling process, the bottler takes a random sample of 48 bottles. The sample mean weight and the sample standard deviation of the bottles are 11.80 ounces and 0.8 ounce, respectively.
a. State the null and the alternative hypotheses to test if the accuracy of the bottling process is compromised.
b. Do you need to make any assumption regarding the population for testing?
c. Calculate the value of the test statistic and the *p*-value.
d. At $\alpha = 0.05$, what is the conclusion to the test? Make a recommendation to the bottler.

49. Based on the average predictions of 47 members of the National Association of Business Economists (NABE), the U.S. gross domestic product (GDP) will expand by 3.2% in 2011 (*The Wall Street Journal,* May 23, 2010). Suppose the sample standard deviation of their predictions was 1%. At the 5% significance level, test if the mean forecast GDP of all NABE members is greater than 3%.

50. In September 2007, U.S. home prices fell at a record pace, and price declines in Los Angeles and Orange counties in California outpaced other major metropolitan areas (*Los Angeles Times,* November 28, 2007). The report was based on the Standard & Poor's/Case-Shiller index that measures the value of single-family homes based on their sales histories. According to this index, the prices in San Diego dropped by an average of 9.6% from a year earlier. Assume that the survey was based on recent sales of 34 houses in San Diego that also resulted in a standard deviation of 5.2%. Can we conclude that the mean drop of all home prices in San Diego is greater than the 7% drop in Los Angeles? Use a 1% level of significance for the analysis.

51. A car manufacturer is trying to develop a new sports car. Engineers are hoping that the average amount of time that the car takes to go from 0 to 60 miles per hour is below 6 seconds. The manufacturer tested 12 of the cars and clocked their performance times. Three of the cars clocked in at 5.8 seconds, 5 cars at 5.9 seconds, 3 cars at 6.0 seconds, and 1 car at 6.1 seconds. At the 5% level of significance, test if the new sports car is meeting its goal to go from 0 to 60 miles per hour in less than 6 seconds. Assume a normal distribution for the analysis.

52. A mortgage specialist would like to analyze the average mortgage rates for Atlanta, Georgia. He collects data on the annual percentage rates (APR in %) for 30-year fixed loans as shown in the following table. If he is willing to assume that these rates are randomly drawn from a normally distributed population, can he conclude that the mean mortgage rate for the population exceeds 4.2%? Test the hypothesis at the 10% level of significance.

Financial Institution	APR
G Squared Financial	4.125
Best Possible Mortgage	4.250
Hersch Financial Group	4.250
Total Mortgages Services	4.375
Wells Fargo	4.375
Quicken Loans	4.500
Amerisave	4.750

Source: MSN Money.com; data retrieved October 1, 2010.

53. One of the consequences of the Great Recession was a free fall of the stock market's average price/earnings ratio, or P/E ratio (*The Wall Street Journal,* August 30, 2010). Generally, a high P/E ratio suggests that investors are expecting higher earnings growth in the future compared to companies with a

lower P/E ratio. An analyst wants to determine if the P/E ratio of firms in the footwear industry is different from the overall average of 14.9. The following table shows the P/E ratios for a sample of seven firms in the footwear industry.

Firm	P/E Ratio
Brown Shoe Co., Inc.	20.54
Collective Brands, Inc.	9.33
Crocs, Inc.	22.63
DSW, Inc.	14.42
Nike, Inc.	18.68
Skechers USA, Inc.	9.35
Timberland Co.	14.93

Source: biz.yahoo.com, data retrieved August 23, 2010.

a. State the null and the alternative hypotheses in order to test whether the P/E ratio of firms in the footwear industry differs from the overall average of 14.9.
b. What assumption regarding the population is necessary?
c. Calculate the value of the test statistic and the *p*-value.
d. At $\alpha = 0.10$, does the P/E ratio of firms in the footwear industry differ from the overall average of 14.9? Explain.

54. FILE ***MPG.*** The data accompanying this exercise show miles per gallon (MPG).
a. State the null and the alternative hypotheses in order to test whether the average MPG differs from 95.
b. Calculate the value of the test statistic and the *p*-value.
c. At $\alpha = 0.05$, can you conclude that the average MPG differs from 95?

55. FILE ***Debt_Payments.*** A study found that consumers are making average monthly debt payments of $983 (Experian.com, November 11, 2010). The data accompanying this exercise show the average debt payments (Debt, in $) for 26 metropolitan areas.
a. State the null and the alternative hypotheses in order to test whether average monthly debt payments are greater than $900.
b. What assumption regarding the population is necessary to implement this step?
c. Calculate the value of the test statistic and the *p*-value.
d. At $\alpha = 0.05$, are average monthly debt payments greater than $900? Explain.

56. FILE ***Highway_Speeds.*** A police officer is concerned about speeds on a certain section of Interstate 95. The data accompanying this exercise show the speeds of 40 cars on a Saturday afternoon.
a. The speed limit on this portion of Interstate 95 is 65 mph. Specify the competing hypotheses in order to determine if the average speed is greater than the speed limit.
b. Calculate the value of the test statistic and the *p*-value.
c. At $\alpha = 0.01$, are the officer's concerns warranted? Explain.

57. FILE ***Lottery.*** An article found that Massachusetts residents spent an average of $860.70 on the lottery in 2010, more than three times the U.S. average (www.businessweek.com, March 14, 2012). A researcher at a Boston think tank believes that Massachusetts residents spend less than this amount. The data accompanying this exercise show the annual lottery expenditures (Lottery, in $) for 100 Massachusetts residents.
a. Specify the competing hypotheses to test the researcher's claim.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 10% significance level, do the data support the researcher's claim? Explain.

9.4 HYPOTHESIS TEST FOR THE POPULATION PROPORTION

LO **9.5**

Conduct a hypothesis test for the population proportion.

As discussed earlier, sometimes the variable of interest is *qualitative* rather than *quantitative*. While the population mean μ and the population standard deviation σ describe quantitative data, the population proportion p is the essential descriptive measure when the data type is qualitative. The parameter p represents the proportion of observations with a particular attribute.

As in the case for the population mean, we estimate the population proportion on the basis of its sample counterpart. In particular, we use the sample proportion $\overline{P}$ to estimate the population proportion p. Recall that although $\overline{P}$ is based on a binomial distribution, it can be approximated by a normal distribution in large samples. This approximation is considered valid when $np \geq 5$ and $n(1-p) \geq 5$. Since p is not known, we typically test the

sample size requirement under the hypothesized value of the population proportion p_0. In most applications, the sample size is large and the normal distribution approximation is justified. However, when the sample size is not deemed large enough, the statistical methods suggested here for inference regarding the population proportion are no longer valid.

Recall from Chapter 7 that the mean and the standard error of the sample proportion $\overline{P}$ are given by $E(\overline{P}) = p$ and $se(\overline{P}) = \sqrt{p(1-p)/n}$, respectively. The test statistic for p is defined as follows.

TEST STATISTIC FOR p

The value of the test statistic for the hypothesis test of the population proportion p is computed as

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}},$$

where p_0 is the hypothesized value of the population proportion. This formula is valid only if $\overline{P}$ (approximately) follows a normal distribution.

The following examples elaborate on the four-step procedure for a hypothesis test for the population proportion.

EXAMPLE 9.12

A popular weekly magazine asserts that fewer than 40% of households in the United States have changed their lifestyles because of environmental concerns. A recent survey of 180 households finds that 67 households have made lifestyle changes due to environmental concerns.

a. Specify the competing hypotheses to test the magazine's claim.

b. Calculate the value of the test statistic and the p-value.

c. At the 5% level of significance, what is the conclusion to the test?

SOLUTION:

a. We wish to establish that the population proportion is less than 0.40—that is, $p < 0.40$. Thus, we construct the competing hypotheses as

$$H_0: p \geq 0.40$$
$$H_A: p < 0.40$$

b. We first ensure that the normality condition is satisfied. Since both np_0 and $n(1-p_0)$ exceed 5, the normal approximation is justified. We use the sample proportion, $\bar{p} = 67/180 = 0.3722$, to compute the value of the test statistic as

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.3722 - 0.40}{\sqrt{0.40(1-0.40)/180}} = -0.76.$$

Since this is a left-tailed test for the population proportion, we find the p-value as $P(Z \leq z) = P(Z \leq -0.76) = 0.2236$.

c. The p-value of 0.2236 is greater than the chosen $\alpha = 0.05$. Therefore, we do not reject the null hypothesis. This means that the magazine's claim that fewer than 40% of households in the United States have changed their lifestyles because of environmental concerns is not justified by the sample data at the 5% significance level. Such a conclusion may be welcomed by firms that have invested in alternative energy.

EXAMPLE 9.13

Driven by growing public support, the legalization of marijuana in America has been moving at a breakneck speed. Today, 57% of adults say the use of marijuana should be made legal (www.pewresearch.org, October 12, 2016). A health practitioner in Ohio collects data from 200 adults and finds that 102 of them favor marijuana legalization.

a. The health practitioner believes that the proportion of adults who favor marijuana legalization in Ohio is not representative of the national proportion. Specify the competing hypotheses to test her claim.

b. Calculate the value of the test statistic and the p-value.

c. At the 10% significance level, do the sample data support the health practitioner's belief?

SOLUTION:

a. The parameter of interest is again the population proportion p. The health practitioner wants to test if the population proportion of those who favor marijuana legalization in Ohio differs from the national proportion of 0.57. We construct the competing hypotheses as

$$H_0: p = 0.57$$
$$H_A: p \neq 0.57$$

b. When evaluated at $p_0 = 0.57$ with $n = 200$, the normality requirement that $np \geq 5$ and $n(1-p) \geq 5$ is easily satisfied. We use the sample proportion $\bar{p} = 102/200 = 0.51$ to compute the value of the test statistic as

$$z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.51 - 0.57}{\sqrt{0.57(1-0.57)/200}} = -1.71.$$

Given a two-tailed test and $z < 0$, we compute the p-value as $2P(Z \leq z) = 2P(Z \leq -1.71) = 0.0872$.

c. Since the p-value of 0.0872 is less than $\alpha = 0.10$, we reject the null hypothesis. Therefore, at the 10% significance level, the proportion of adults who favor marijuana legalization in Ohio differs from the national proportion of 0.57.

EXERCISES 9.4

Mechanics

58. Consider the following hypotheses:

$$H_0: p \geq 0.38$$
$$H_A: p < 0.38$$

Calculate the p-value based on the following sample information.

a. $x = 22; n = 74$
b. $x = 110; n = 300$
c. $\bar{p} = 0.34; n = 50$
d. $\bar{p} = 0.34; n = 400$

59. Which sample information in the preceding question enables us to reject the null hypothesis at $\alpha = 0.01$ and at $\alpha = 0.10$?

60. Consider the following hypotheses:

$$H_0: p = 0.32$$
$$H_A: p \neq 0.32$$

Calculate the p-value based on the following sample information

a. $x = 20; n = 66$
b. $x = 100; n = 264$
c. $\bar{p} = 0.40; n = 40$
d. $\bar{p} = 0.38; n = 180$

61. Which sample information in the preceding question enables us to reject the null hypothesis at $\alpha = 0.05$ and at $\alpha = 0.10$?

62. In order to test if the population proportion differs from 0.40, you draw a random sample of 100 observations and obtain a sample proportion of 0.48.
    a. Specify the competing hypotheses.
    b. Is the normality condition satisfied? Explain.
    c. Calculate the value of the test statistic and the *p*-value.
    d. At the 5% significance level, does the population proportion differ from 0.40? Explain.

63. In order to conduct a hypothesis test for the population proportion, you sample 320 observations that result in 128 successes. Conduct the following tests at $\alpha = 0.05$.
    a. $H_0: p \geq 0.45$; $H_A: p < 0.45$
    b. $H_0: p = 0.45$; $H_A: p \neq 0.45$

64. In order to test if the population proportion is greater than 0.65, you draw a random sample of 200 observations and obtain a sample proportion of 0.72.
    a. Specify the competing hypotheses.
    b. Is the normality condition satisfied? Explain.
    c. Calculate the value of the test statistic and the *p*-value.
    d. At the 5% significance level, is the population proportion greater than 0.65? Explain.

65. You would like to determine if the population probability of success differs from 0.70. You find 62 successes in 80 binomial trials. Implement the test at the 1% level of significance.

66. You would like to determine if more than 50% of the observations in a population are below 10. At $\alpha = 0.05$, conduct the test on the basis of the following 20 sample observations:

8	12	5	9	14	11	9	3	7	8
12	6	8	9	2	6	11	4	13	10

Applications

67. A study by Allstate Insurance Co. finds that 82% of teenagers have used cell phones while driving (*The Wall Street Journal,* May 5, 2010). In October 2010, Massachusetts enacted a law that forbids cell phone use by drivers under the age of 18. A policy analyst would like to determine whether the law has decreased the proportion of drivers under the age of 18 who use a cell phone.
    a. State the null and the alternative hypotheses to test the policy analyst's objective.
    b. Suppose a sample of 200 drivers under the age of 18 results in 150 who still use a cell phone while driving. What is the value of the test statistic? What is the *p*-value?
    c. At $\alpha = 0.05$, has the law been effective?

68. In order to endure financial hardships such as unemployment and medical emergencies, Americans have increasingly been raiding their already fragile retirement accounts (*MSN Money,* July 16, 2008). It is reported that between 1998 and 2004, about 12% of families with 401(k) plans borrowed from them. An economist is concerned that this percentage now exceeds 20%. He randomly surveys 190 households with 401(k) plans and finds that 50 are borrowing against them.
    a. Set up the null and the alternative hypotheses to test the economist's concern.
    b. Calculate the value of the test statistic and the *p*-value.
    c. Determine if the economist's concern is justifiable at $\alpha = 0.05$.

69. The margarita is one of the most common tequila-based cocktails, made with tequila mixed with triple sec and lime or lemon juice, often served with salt on the glass rim. A common ratio for a margarita is 2:1:1, which includes 50% tequila, 25% triple sec, and 25% fresh lime or lemon juice. A manager at a local bar is concerned that the bartender uses incorrect proportions in more than 50% of margaritas. He secretly observes the bartender and finds that he used the correct proportions in only 10 out of 30 margaritas. Test if the manager's suspicion is justified at $\alpha = 0.05$.

70. Research shows that many banks are unwittingly training their online customers to take risks with their passwords and other sensitive account information, leaving them more vulnerable to fraud (Yahoo.com, July 23, 2008). Even web-savvy surfers could find themselves the victims of identity theft because they have been conditioned to ignore potential signs about whether the banking site they are visiting is real or a bogus site served up by hackers. Researchers at the University of Michigan found design flaws in 78% of the 214 U.S. financial institution websites they studied. Is the sample evidence sufficient to conclude that more than three out of four financial institutions that offer online banking facilities are prone to fraud? Use a 5% significance level for the test.

71. Research commissioned by Vodafone suggests that older workers are the happiest employees (*BBC News,* July 21, 2008). The report documents that 70% of older workers in England feel fulfilled, compared with just 50% of younger workers. A demographer believes that an identical pattern does not exist in Asia. A survey of 120 older workers in Asia finds that 75 feel fulfilled. A similar survey finds that 58% of 210 younger workers feel fulfilled.
    a. At the 5% level of significance, test if older workers in Asia feel less fulfilled than their British counterparts.
    b. At the 5% level of significance, test if younger workers in Asia feel more fulfilled than their British counterparts.

72. A politician claims that he is supported by a clear majority of voters. In a recent survey, 24 out of 40 randomly selected voters indicated that they would vote for the politician. Is the politician's claim justified at the 5% level of significance?

73. A movie production company is releasing a movie with the hopes of many viewers returning to see the movie in the theater for a second time. Their target is to have 30 million viewers, and they want more than 30% of the viewers to want to see the movie again. They show the movie to a test

audience of 200 people, and after the movie they asked them if they would see the movie in theaters again. Of the test audience, 68 people said they would see the movie again.

a. At the 5% level of significance, test if more than 30% of the viewers will return to see the movie again.
b. Repeat the analysis at the 10% level of significance.
c. Interpret your results.

74. With increasing out-of-pocket healthcare costs, it is claimed that more than 60% of senior citizens are likely to make serious adjustments to their lifestyle. Test this claim at the 1% level of significance if in a survey of 140 senior citizens, 90 reported that they have made serious adjustments to their lifestyle.

75. FILE ***Silicon_Valley.*** According to a report on workforce diversity, about 60% of the employees in high-tech firms in Silicon Valley are white and about 20% are Asian (www.money.cnn.com, November 9, 2011). Women, along with blacks and Latinos, are highly underrepresented. Just about 30% of all employees are women, with blacks and Latinos accounting for only about 15% of the workforce. Tara Jones is a recent college graduate, working for a large high-tech firm in Silicon Valley. She wants to determine if her firm faces the same diversity as in the report. She collects gender and ethnicity information on 50 employees in her firm. A portion of the data is shown in the accompanying table.

Gender	Ethnicity
Female	White
Male	White
⋮	⋮
Male	Nonwhite

a. At the 5% level of significance, determine if the proportion of women in Tara's firm is different from 0.30.
b. At the 5% level of significance, determine if the proportion of whites in Tara's firm is more than 0.50.

WRITING WITH STATISTICS

The Associated Press reports that income inequality is at record levels in the United States (September 28, 2010). Over the years, the rich have become richer while working-class wages have stagnated. A local Latino politician has been vocal regarding his concern about the welfare of Latinos. In various speeches, he has stated that the mean salary of Latino households in his county has fallen below the 2008 mean of $49,000. He has also stated that the proportion of Latino households making less than $30,000 has risen above the 2008 level of 20%. Both of his statements are based on income data for 36 Latino households in the county, as shown in Table 9.4.



©Ariel Skelley/Blend Images LLC

FILE
Latino_Income

TABLE 9.4 Representative Sample of Latino Household Incomes in 2010

22	36	78	103	38	43
62	53	26	28	25	31
62	44	51	38	77	37
29	38	46	52	61	57
20	72	41	73	16	32
52	28	69	27	53	46

Incomes are measured in $1,000s and have been adjusted for inflation.

Trevor Jones is a newspaper reporter who is interested in verifying the concerns of the local politician.

Trevor wants to use the sample information to

1. Determine if the mean income of Latino households has fallen below the 2008 level of $49,000.
2. Determine if the percentage of Latino households making less than $30,000 has risen above 20%.

Sample Report—Income Inequality in the United States

One of the hotly debated topics in the United States is that of growing income inequality. Market forces such as increased trade and technological advances have made highly skilled and well-educated workers more productive, thus increasing their pay. Institutional forces, such as deregulation, the decline of unions, and the stagnation of the minimum wage, have contributed to income inequality. Arguably, this income inequality has been felt by minorities, especially African Americans and Latinos, since a very high proportion of both groups is working class. The condition has been further exacerbated by the Great Recession.

A sample of 36 Latino households resulted in a mean household income of $46,278 with a standard deviation of $19,524. The sample mean is below the 2008 level of $49,000. In addition, nine Latino households, or 25%, make less than $30,000; the corresponding percentage in 2008 was 20%. Based on these results, a politician concludes that current market conditions continue to negatively impact the welfare of Latinos. However, it is essential to provide statistically significant evidence to substantiate these claims. Toward this end, formal tests of hypotheses regarding the population mean and the population proportion are conducted. The results of the tests are summarized in Table 9.A.

TABLE 9.A Test Statistic Values and *p*-Values for Hypothesis Tests

Hypotheses	Test Statistic Value	*p*-value
$H_0: \mu \geq 49{,}000$ $H_A: \mu < 49{,}000$	$t_{35} = \dfrac{46{,}278 - 49{,}000}{19{,}524/\sqrt{36}} = -0.84$	0.2033
$H_0: p \leq 0.20$ $H_A: p > 0.20$	$z = \dfrac{0.25 - 0.20}{\sqrt{\dfrac{(0.20)\,(0.80)}{36}}} = 0.75$	0.2266

When testing whether the mean income of Latino households has fallen below the 2008 level of $49,000, a test statistic value of −0.84 is obtained. Given a *p*-value of 0.2033, the null hypothesis regarding the population mean, specified in Table 9.A, cannot be rejected at any reasonable level of significance. Similarly, given a *p*-value of 0.2266, the null hypothesis regarding the population proportion cannot be rejected. Therefore, sample evidence does not support the claims that the mean income of Latino households has fallen below $49,000 or that the proportion of Latino households making less than $30,000 has risen above 20%. Perhaps the politician's remarks were based on a cursory look at the sample statistics and not on a thorough statistical analysis.

CONCEPTUAL REVIEW

LO 9.1 Define the null hypothesis and the alternative hypothesis.

Every hypothesis test contains two competing hypotheses: the **null hypothesis**, denoted H_0, and the **alternative hypothesis**, denoted H_A. We can think of the null hypothesis as corresponding to a presumed default state of nature or status quo, whereas the alternative hypothesis contradicts the default state or status quo.

On the basis of sample information, we either reject H_0 or do not reject H_0. As a general guideline, whatever we wish to establish is placed in the alternative hypothesis. If we reject the null hypothesis, we are able to conclude that the alternative hypothesis is true.

Hypothesis tests can be **one-tailed** or **two-tailed.** A one-tailed test allows the rejection of the null hypothesis only on one side of the hypothesized value of the population parameter. In a two-tailed test, the null hypothesis can be rejected on both sides of the hypothesized value of the population parameter.

LO 9.2 Distinguish between Type I and Type II errors.

Since the statistical conclusion of a hypothesis test relies on sample data, there are two types of errors that may occur: a **Type I error** or a **Type II error**. A Type I error is committed when we reject the null hypothesis when it is actually true. On the other hand, a Type II error is made when we do not reject the null hypothesis when it is actually false. We denote the probability of a Type I error by α and the probability of a Type II error by β. For a given sample size n, a decrease (increase) in α will increase (decrease) β. However, both α and β will decrease if the sample size n increases.

LO 9.3 Conduct a hypothesis test for the population mean when σ is known.

Step 1. Specify the null and the alternative hypothesis. It is important to (1) identify the relevant population parameter (in this case, the population mean μ), (2) determine whether a one- or a two-tailed test is appropriate, and (3) include some form of the equality sign in the null hypothesis and use the alternative hypothesis to establish a claim.

Step 2. Specify the significance level. Before implementing a hypothesis test, the significance level α is specified. This is the *allowed* probability of making a Type I error.

Step 3. Calculate the value of the test statistic and the p-value. When testing the population mean μ when the population standard deviation σ is known, the value of the test statistic is calculated as $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$. The p-value is the probability that this test statistic is as extreme as its value computed from the given sample. The p-value is calculated as

- $P(Z \geq z)$ for a right-tailed test,
- $P(Z \leq z)$ for a left-tailed test, or
- $2P(Z \geq z)$ if $z > 0$ or $2P(Z \leq z)$ if $z < 0$ for a two-tailed test.

Step 4. State the conclusion and interpret the results. The decision rule is to reject the null hypothesis if the p-value $< \alpha$ and not reject the null hypothesis if the p-value $\geq \alpha$. Clearly interpret the results in the context of the problem.

LO 9.4 Conduct a hypothesis test for the population mean when σ is unknown.

Step 1 and Step 2. The first two steps are the same as those in the previous case.

Step 3. Calculate the value of the test statistic and the p-value. When testing the population mean μ when the population standard deviation σ is unknown, the value of the test statistic is calculated as $t_{df} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$, where the degrees of freedom $df = n - 1$. The p-value is calculated as

- $P(T_{df} \geq t_{df})$ for a right-tailed test,
- $P(T_{df} \leq t_{df})$ for a left-tailed test, or
- $2P(T_{df} \geq t_{df})$ if $t_{df} > 0$ or $2P(T_{df} \leq t_{df})$ if $t_{df} < 0$ for a two-tailed test.

Step 4. State the conclusion and interpret the results. The decision rule is to reject the null hypothesis if the p-value $< \alpha$ and not reject the null hypothesis if the p-value $\geq \alpha$. Clearly interpret the results in the context of the problem.

LO 9.5 Conduct a hypothesis test for the population proportion.

Step 1 and Step 2. The first two steps are the same as in the previous two cases except that the parameter of interest is now the population proportion p.

Step 3. Calculate the value of the test statistic and the p-value. When testing the population proportion p, the value of the test statistic is computed as $z = \frac{\bar{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$. The p-value is calculated as

- $P(Z \geq z)$ for a right-tailed test,
- $P(Z \leq z)$ for a left-tailed test, or
- $2P(Z \geq z)$ if $z > 0$ or $2P(Z \geq z)$ if $z < 0$ for a two-tailed test.

Step 4. State the conclusion and interpret the results. The decision rule is to reject the null hypothesis if the p-value $< \alpha$ and not reject the null hypothesis if the p-value $\geq \alpha$. Clearly interpret the results in the context of the problem.

ADDITIONAL EXERCISES AND CASE STUDIES

Exercises

76. A pharmaceutical company has developed a new drug for depression. There is a concern, however, that the drug also raises the blood pressure of its users. A researcher wants to conduct a test to validate this claim. Would the manager of the pharmaceutical company be more concerned about a Type I error or a Type II error? Explain.

77. A company has developed a new diet that it claims will lower one's weight by more than 10 pounds. Health officials decide to conduct a test to validate this claim.
    a. Would the manager of the company be more concerned about a Type I error or a Type II error? Explain.
    b. Would the consumers be more concerned about a Type I error or a Type II error? Explain.

78. An advertisement for a popular weight loss clinic suggests that participants in its new diet program lose, on average, more than 10 pounds. A consumer activist decides to test the authenticity of the claim. She follows the progress of 18 women who recently joined the weight reduction program. She calculates the mean weight loss of these participants as 10.8 pounds with a standard deviation of 2.4 pounds. Assume that weight loss is normally distributed.
    a. Set up the competing hypotheses to test the advertisement's claim.
    b. Calculate the value of the test statistic and the p-value.
    c. At the 5% significance level, what does the consumer activist conclude?

79. A phone manufacturer wants to compete in the touch screen phone market. He understands that the lead product has a battery life of just 5 hours. The manufacturer claims that while the new touch screen phone is more expensive, its battery life is more than twice as long as that of the leading product. In order to test the claim, a researcher samples 45 units of the new phone and finds that the sample battery life averages 10.5 hours with a sample standard deviation of 1.8 hours.
    a. Set up the competing hypotheses to test the manufacturer's claim.
    b. Calculate the value of the test statistic and the p-value.
    c. Test the phone manufacturer's claim at $\alpha = 0.05$.

80. A city council is deciding whether or not to spend additional money to reduce the amount of traffic. The council decides that it will increase the transportation budget if the amount of waiting time for drivers exceeds 20 minutes. A sample of 32 main roads results in a mean waiting time of 22.08 minutes with a standard deviation of 5.42 minutes. Conduct a hypothesis test at the 1% level of significance to determine whether or not the city should increase its transportation budget.

81. Rates on 30-year fixed mortgages continue to be at historic lows (*Chron Business News*, September 23, 2010). According to Freddie Mac, the average rate for 30-year fixed loans for the week was 4.37%. An economist wants to test if there is any change in the mortgage rates in the following week. She searches the Internet for 30-year fixed loans and reports the rates offered by seven banks as 4.25%, 4.125%, 4.375%, 4.50%, 4.75%, 4.375%, and 4.875%. Assume that rates are normally distributed.
    a. State the hypotheses to test if the average mortgage rate differs from 4.37%.
    b. Calculate the value of the test statistic and the p-value.
    c. At the 5% significance level, does the average mortgage rate differ from 4.37%? Explain.

82. The Great Recession cost America trillions of dollars in lost wealth and also levied a heavy toll on the national psyche (*The Wall Street Journal*, December 21, 2009). According to a poll, just 33% of those surveyed said America was headed in the right direction. Suppose this poll was based on a sample of 1,000 people. Does the sample evidence suggest that the proportion of Americans who feel that America is headed in the right direction is below 35%? Use a 5% level of significance for the analysis. What if the sample size was 2,000?

83. A retailer is looking to evaluate its customer service. Management has determined that if the retailer wants to stay competitive, then it will have to have at least a 90% satisfaction rate among its customers. Management will take corrective actions if the satisfaction rate falls below 90%. A survey of 1,200 customers showed that 1,068 were satisfied with their customer service.
    a. State the hypotheses to test if the retailer needs to improve its services.
    b. What is the value of the test statistic?
    c. Find the p-value.
    d. Interpret the results at $\alpha = 0.05$.

84. A national survey found that 33% of high school students said they texted or e-mailed while driving (*The Boston Globe*, June 8, 2012). These findings came a day after a Massachusetts teenager was convicted for causing a fatal crash while texting. A researcher wonders whether texting or e-mailing while driving is more prevalent among Massachusetts teens. He surveys 100 teens and 42% of them admitted that they texted or e-mailed while behind the wheel. Can he conclude at the 1% significance level that Massachusetts teens engage in this behavior at a rate greater than the national rate?

85. A television network is deciding whether or not to give its newest television show a spot during prime viewing time at night. For this to happen, it will have to move one of its most viewed shows to another slot. The network conducts a survey asking its viewers which show they would rather watch. The network will keep its current lineup of shows unless the majority of the customers want to watch the new show. The network receives 827 responses, of which 428 indicate that they would like to see the new show in the lineup.
    a. Set up the hypotheses to test if the television network should give its newest television show a spot during prime viewing time at night.
    b. Calculate the value of the test statistic and the *p*-value.
    c. At $\alpha = 0.01$, what should the television network do?

86. A survey finds that 17% of Americans cannot part with their landlines (*The Washington Post*, February 27, 2014). A researcher in the rural South collects data from 200 households and finds that 45 of them still have landlines.
    a. The researcher believes that the proportion of households with landlines in the rural South is not representative of the national proportion. Specify the competing hypotheses to test her claim.
    b. Calculate the value of the test statistic and the *p*-value.
    c. At the 5% significance level, do the sample data support the researcher's belief?

87. FILE ***Metals.*** Using data from the past 25 years, an investor wants to test whether the average return of Vanguard's Precious Metals and Mining Fund is greater than 12%. Assume returns are normally distributed with a population standard deviation of 30%.
    a. State the null and the alternative hypotheses for the test.
    b. Calculate the value of the test statistic and the *p*-value.
    c. At $\alpha = 0.05$, what is the conclusion? Is the return on Vanguard's Precious Metals and Mining Fund greater than 12%?

88. FILE ***Midwest_Drivers.*** On average, Americans drive 13,500 miles per year (*The Boston Globe*, June 7, 2012). An economist gathers data on the driving habits of 50 residents in the Midwest.
    a. The economist believes that the average number of miles driven annually by Midwesterners is different from the U.S. average. Specify the competing hypotheses to test the economist's claim.
    b. Calculate the value of the test statistic and the *p*-value.
    c. At the 10% significance level, do the data support the researcher's claim? Explain.

89. FILE ***Convenience_Stores.*** An entrepreneur examines monthly sales (in $1,000s) for 40 convenience stores in Rhode Island.
    a. State the null and the alternative hypotheses in order to test whether average sales differ from $130,000.
    b. Calculate the value of the test statistic and the *p*-value.
    c. At $\alpha = 0.05$, what is your conclusion to the test? Do average sales differ from $130,000?

90. FILE ***DJIA_Volume.*** The euro-zone crisis has wreaked havoc on U.S. stock markets (*The Wall Street Journal*, June 8, 2012). A portfolio analyst wonders if the average trading volume on the Dow Jones Industrial Average (DJIA) has decreased since the beginning of the year. She gathers data on daily trading volumes for 30 days.
    a. The average trading volume in the beginning of the year was about 4,000 shares (in millions). Specify the competing hypotheses to test her claim.
    b. Calculate the value of the test statistic and the *p*-value.
    c. At the 5% significance level, does it appear that the trading volume has decreased since the beginning of the year?

91. FILE ***Study_Hard.*** A report suggests that business majors spend the least amount of time on course work than do all other college students (*The New York Times*, November 17, 2011). A provost of a university conducts a survey of 50 business and 50 nonbusiness students. Students are asked if they study hard, defined as spending at least 20 hours per week on course work. The response shows "yes" if they study hard or "no" otherwise; a portion of the responses is shown in the following table.

Business Majors	Nonbusiness Majors
Yes	No
No	Yes
⋮	⋮
Yes	Yes

a. At the 5% level of significance, determine if the percentage of business majors who study hard is less than 20%.
b. At the 5% level of significance, determine if the percentage of nonbusiness majors who study hard is more than 20%.

92. FILE ***MI_Life.*** Residents of Hawaii have the longest life expectancies in the United States, averaging 81.48 years (www.worldlifeexpectancy.com; data retrieved June 4, 2012). A sociologist collects data on the age at death for 50 recently deceased Michigan residents.
a. The sociologist believes that the life expectancies of Michigan residents are less than those of Hawaii residents. Specify the competing hypotheses to test this belief.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 1% significance level, do the data support the sociologist's belief?

93. Thirty-three percent of children and teens in the United States are obese or overweight (*Health*, October 2010). A health practitioner in the Midwest collects data on 200 children and teens and finds that 84 of them are either obese or overweight.
a. The health practitioner believes that the proportion of obese and overweight children in the Midwest is not representative of the national proportion. Specify the competing hypotheses to test her claim.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 1% significance level, do the sample data support the health practitioner's belief?

CASE STUDIES

CASE STUDY 9.1 Harvard University revolutionized its financial aid policies, aimed at easing the financial strain on middle and upper-middle income families (*Newsweek*, August 18–25, 2008). The expected contribution of students who are admitted to Harvard has been greatly reduced. Many other elite private colleges are following suit to compete for top students. The motivation for these policy changes stems from competition from public universities as well as political pressure.

A spokesman from an elite college claims that elite colleges have been very responsive to financial hardships faced by families due to the rising costs of education. Now, he says, families with an income of $40,000 will have to spend less than $6,500 to send their children to prestigious colleges. Similarly, families with incomes of $80,000 and $120,000 will have to spend less than $20,000 and $35,000, respectively, for their children's education.

Although in general the cost of attendance has gone down at each family-income level, it still varies by thousands of dollars among prestigious schools. The accompanying table shows information on the cost of attendance by family income for 10 prestigious schools.

FILE
Family_Income

Data for Case Study 9.1 Cost of Attendance to Schools by Family Income

School	Family Income		
	40000	80000	120000
Amherst College	5302	19731	37558
Bowdoin College	5502	19931	37758
Columbia University	4500	12800	36845
Davidson College	5702	20131	37958
Harvard University	3700	8000	16000
Northwestern University	6311	26120	44146
Pomona College	5516	19655	37283
Princeton University	3887	11055	17792
Univ. of California system	10306	19828	25039
Yale University	4300	6048	13946

Source: *Newsweek*, August 18–25, 2008.

In a report, use the sample information to

1. Determine whether families with incomes of $40,000 will spend less than $6,500 to send their children to prestigious colleges. (Use $\alpha = 0.05$.)
2. Repeat the hypothesis test from part 1 by testing the spokesman's claims concerning college costs for families with incomes of $80,000 and $120,000, respectively. (Use $\alpha = 0.05$.)
3. Assess the validity of the spokesman's claims.

CASE STUDY 9.2 The effort to reward city students for passing Advanced Placement tests is part of a growing trend nationally and internationally. Financial incentives are offered in order to lift attendance and achievement rates. One such program in Dallas, Texas, offers $100 for every Advanced Placement test on which a student gets a score of 3 or higher (Reuters, September 20, 2010). A wealthy entrepreneur decides to experiment with the same idea of rewarding students to enhance performance, but in Chicago. He offers monetary incentives to students at an inner-city high school. Due to this incentive, 122 students take the Advancement Placement tests. Twelve students get a 5, the highest possible score. There are 49 students with scores of 3 and 4, and 61 students with failing scores of 1 and 2. Historically, about 100 of these tests are taken at this school each year, where 8% of students score 5, 38% score 3 and 4, and the remaining get failing scores of 1 and 2.

In a report, use the sample information to

1. Provide a descriptive analysis of student achievement on Advanced Placement before and after the monetary incentive is offered.
2. Conduct a hypothesis test that determines, at the 5% significance level, whether the monetary incentive has resulted in a higher proportion of scores of 5, the highest possible score.
3. Conduct a hypothesis test that determines, at the 5% significance level, whether the monetary incentive has decreased the proportion of failing scores of 1 and 2.
4. Assess the effectiveness of monetary incentives in improving student achievement.

CASE STUDY 9.3 The Gallup-Healthways Well-Being Index (www.well-beingindex.com) provides an assessment measure of health and well-being of U.S. residents. By collecting periodic data on life evaluation, physical health, emotional health, healthy behavior, work environment, and basic access, this assessment measure is of immense value to researchers in diverse fields such as business, medical sciences, and journalism. The overall composite score, as well as a score in each of the above six categories, is calculated on a scale from 0 to 100, where 100 represents fully realized well-being. In 2009, the overall well-being index score of American residents was reported as 65.9. Let the following table represent the overall well-being score of a random sample of 35 residents in Hawaii.

Data for Case Study 9.3 Overall Well-being of Hawaiians, $n = 35$

FILE
Hawaiians

20	40	40	100	60	20	40
90	90	60	60	90	90	90
80	100	90	80	80	80	100
70	90	80	100	20	70	90
80	30	80	90	90	80	30

In a report, use the sample information to

1. Determine whether the well-being score of Hawaiians is more than the national average of 65.9 at the 5% significance level.
2. Determine if fewer than 40% of Hawaiians report a score below 50 at the 5% significance level.
3. Comment on the well-being of Hawaiians given your results.

APPENDIX 9.1 The Critical Value Approach

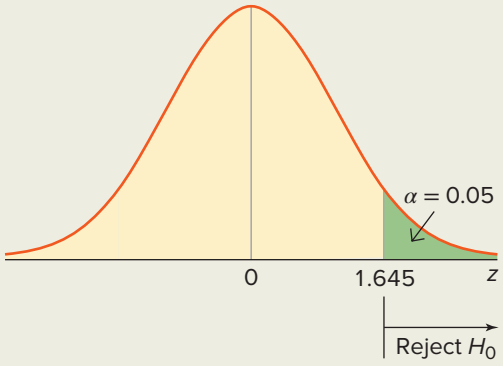
We always use sample evidence and the chosen significance level α to conduct hypothesis tests. The p-value approach makes the comparison in terms of probabilities. As discussed in Section 9.2, the value of the test statistic is used to compute the p-value, which is then compared with α in order to arrive at a decision. Most statistical software packages report p-values, so the p-value approach to hypothesis testing tends to be favored by most researchers and practitioners. The critical value approach, on the other hand, makes the comparison directly in terms of the value of the test statistic. This approach is particularly useful when a computer is unavailable and all calculations must be done manually. Both approaches, however, always lead to the same conclusion.

In Section 9.2, we used the p-value approach to validate a sociologist's claim that the mean retirement age in the United States is greater than 67 at the 5% significance level. In a random sample of 25 retirees, the average retirement age was 71. It was also assumed that the retirement age is normally distributed with a population standard deviation of 9 years. With the critical value approach, we still specify the competing hypotheses and calculate the value of the test statistic as we did with the p-value approach. In the retirement age example, the competing hypotheses are H_0: $\mu \leq 67$ versus H_A: $\mu > 67$ and the value of the test statistic is $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{71 - 67}{9/\sqrt{25}} = 2.22$.

The critical value approach specifies a range of values, also called the rejection region, such that if the value of the test statistic falls into the rejection region, then we reject the null hypothesis. The critical value is a point that separates the rejection region from the nonrejection region. Once again we need to make distinctions between the three types of competing hypotheses. For a right-tailed test, the critical value is z_α, where $P(Z \geq z_\alpha) = \alpha$. The resulting rejection region includes values greater than z_α.

With α known, we can easily find the corresponding z_α from the z table. In the retirement age example with $\alpha = 0.05$, we evaluate $P(Z \geq z_\alpha) = 0.05$ to derive the critical value as $z_\alpha = z_{0.05} = 1.645$. Figure A9.1 shows the critical value as well as the corresponding rejection region for the test.

FIGURE A9.1
The critical value for a right-tailed test with $\alpha = 0.05$



As shown in Figure A9.1, the decision rule is to reject H_0 if $z > 1.645$. Since the value of the test statistic, $z = 2.22$, exceeds the critical value, $z_\alpha = 1.645$, we reject the null hypothesis and conclude that the mean age is greater than 67. Thus, we confirm the conclusion reached with the p-value approach.

We would like to stress that we always arrive at the same conclusion whether we use the p-value approach or the critical value approach. If z falls in the rejection region, then the p-value must be less than α. Similarly, if z does not fall in the rejection region, then the p-value must be greater than α. Figure A9.2 shows the equivalence of the two results in the retirement age example for a right-tailed test.

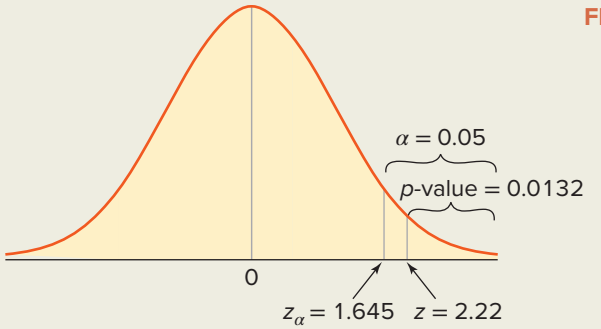


FIGURE A9.2 Equivalent conclusions resulting from the *p*-value and the critical value approaches

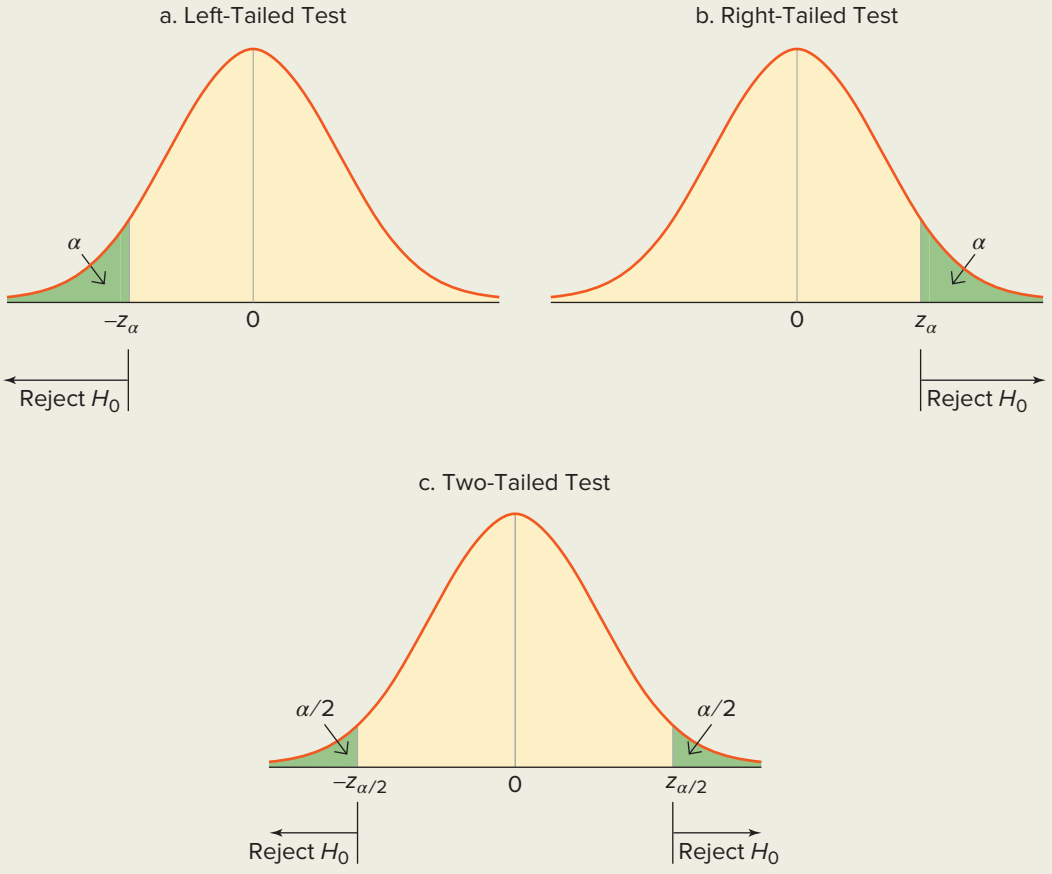
We reject the null hypothesis because the *p*-value = 0.0132 is less than $\alpha = 0.05$ or, equivalently, because $z = 2.22$ is greater than $z_\alpha = 1.645$.

The retirement age example uses a right-tailed test to calculate the critical value as z_α. Given the symmetry of the z distribution around zero, the critical value for a left-tailed test is simply $-z_\alpha$. For a two-tailed test, we split the significance level in half to determine *two* critical values, $-z_{\alpha/2}$ and $z_{\alpha/2}$, where $P(Z \geq z_{\alpha/2}) = \alpha/2$.

For a given α, Figure A9.3 shows the three different scenarios of determining the critical value(s) depending on the specification of the competing hypotheses. For the illustration, we assume that the test statistic follows the z distribution.

Figure A9.3a shows a negative critical value for a left-tailed test where we reject the null hypothesis if $z < -z_\alpha$. Similarly, Figure A9.3b shows a positive critical value for a right-tailed test where we reject the null hypothesis if $z > z_\alpha$. There are two critical values for a two-tailed test, where we reject the null hypothesis when $z < -z_{\alpha/2}$ or when $z > z_{\alpha/2}$ (see Figure A9.3c).

FIGURE A9.3 Critical values for one- and two-tailed tests



We now summarize the general procedure for implementing the critical value approach.

THE FOUR-STEP PROCEDURE USING THE CRITICAL VALUE APPROACH

Step 1. Specify the null and the alternative hypothesis. This step is the same as in the *p*-value approach.

Step 2. Specify the significance level and find the critical value(s). We first specify α. The critical value(s) is a point that separates the rejection region from the nonrejection region. If the test statistic follows the z distribution, then, for a given α, we find the critical value(s) as

- z_α where $P(Z \geq z_\alpha) = \alpha$ for a right-tailed test,
- $-z_\alpha$ where $P(Z \geq z_\alpha) = \alpha$ for a left-tailed test, or
- $-z_{\alpha/2}$ and $z_{\alpha/2}$ where $P(Z \geq z_{\alpha/2}) = \alpha/2$ for a two-tailed test.

Z and z_α are replaced with T_{df} and t_{df} if the test statistic follows the t_{df} distribution with degrees of freedom, $df = n - 1$.

Step 3. Calculate the value of the test statistic. We calculate the value of the test statistic by converting the estimate of the relevant population parameter into its corresponding standardized value, either z or t_{df}.

Step 4. State the conclusion and interpret the results. The decision rule is to reject the null hypothesis if the test statistic falls in the rejection region. We interpret the results in the context of the problem.

APPENDIX 9.2 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, JMP, and R. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands. When importing data into R, use the menu-driven option: File > Import Dataset > From Excel.

Minitab

Testing μ, σ Known

FILE
Debit_Spending

A. (Replicating Example 9.9). From the menu, choose **Stat > Basic Statistics > 1-Sample Z**.

B. Select **One or more samples, each in a column** and select Spending. Enter 500 for **Known standard deviation**. Select **Perform hypothesis test** and enter 8000 after **Hypothesized mean**. Choose **Options**. After **Alternative hypothesis**, select "Mean < hypothesized mean."

Testing μ, σ Unknown

FILE
Study_Hours

A. (Replicating Example 9.11) From the menu, choose **Stat > Basic Statistics > 1-Sample t**.

B. Select **One or more samples, each in a column** and select Hours. Select **Perform hypothesis test** and enter 14 after **Hypothesized mean**. Choose **Options**. After **Alternative hypothesis**, select "Mean ≠ hypothesized mean."

Testing p

A. (Replicating Example 9.12) From the menu, choose **Stat > Basic Statistics > 1-Proportion**.

B. Choose **Summarized data** and then enter 67 after **Number of events** and 180 after **Number of trials**. Select **Perform hypothesis test** and enter 0.40 for **Hypothesized proportion**. Choose **Options**. After **Alternative hypothesis**, select "Proportion < hypothesized proportion" and after **Method** select "Normal approximation."

SPSS

Testing μ, σ Unknown

A. (Replicating Example 9.11). From the menu, choose **Analyze > Compare Means > One-Sample T-Test**. Under **Test Variable(s)**, select Hours. After **Test Value**, enter 14.

FILE *Study_Hours*

JMP

Testing μ, σ Known

A. (Replicating Example 9.9). From the menu, choose **Analyze > Distribution**.

FILE *Debit_Spending*

B. Under **Select Columns**, select Spending, and then under **Cast Selected Columns into Roles**, select **Y, Columns**.

C. Click on the red triangle in the output window beside Spending. Choose **Test Mean**. After **Specify Hypothesized Mean**, enter 8000, and after **Enter true standard deviation to do z-test rather than t-test**, enter 500.

Testing μ, σ Unknown

A. (Replicating Example 9.11). From the menu, choose **Analyze > Distribution**.

B. Under **Select Columns**, select Hours, and then under **Cast Selected Columns into Roles**, select **Y, Columns**.

FILE *Study_Hours*

C. Click on the red triangle in the output window beside Hours. Choose **Test Mean**. After **Specify Hypothesized Mean**, enter 14.

R

Testing μ σ, Known

A. (Replicating Example 9.9) First calculate the value of the test statistic, labeled Teststat. Enter:

FILE *Debit_Spending*

```
> Teststat <- (mean(Debit_Spending$'Spending') - 8000)/(500/sqrt(20))
> List Teststat
```

And R returns: −1.200769.

B. Use the **pnorm** function to find the p-value. In order to find the p-value for this left-tailed—that is, $P(Z \leq -1.200769)$—enter:

```
> pnorm(Teststat, 0, 1, lower.tail=TRUE)
```

Testing μ, σ Unknown

A. (Replicating Example 9.11) First calculate the value of the test statistic, labeled Teststat. Enter:

FILE *Study_Hours*

```
> Teststat <- (mean(Study_Hours$'Hours') - 14)/(sd(Study_Hours$'Hours')/
  sqrt(35))
> List Teststat
```

And R returns: 1.944356.

B. Use the **pt** function to find the p-value. In order to find the p-value for this two-tailed test—that is, $2P(T_{34} \geq 1.944356)$— enter:

```
> 2*pt(Teststat, 34, lower.tail=FALSE).
```

10 Comparisons Involving Means

Learning Objectives

After reading this chapter you should be able to:

LO **10.1** Make inferences about the difference between two population means based on independent sampling.

LO **10.2** Make inferences about the mean difference based on matched-pairs sampling.

LO **10.3** Discuss features of the *F* distribution.

LO **10.4** Conduct and evaluate a one-way ANOVA test.

In the preceding two chapters, we used estimation and hypothesis testing to analyze a single parameter, such as the population mean and the population proportion. In this chapter, we extend our discussion from the analysis of a single population to the comparison of two or more population means. We first analyze differences between two population means. For instance, an economist may be interested in analyzing the salary difference between male and female employees. Similarly, a marketing researcher might want to compare the operating lives of two popular brands of batteries. In these examples, we use independent sampling for the analysis. We also consider the mean difference of two populations based on matched-pairs sampling. An example would be a consumer group activist wanting to analyze the mean weight of customers before and after they enroll in a new diet program. Finally, we use analysis of variance (ANOVA) to test for differences between three or more population means. For instance, we may want to determine whether all brands of small hybrid cars have the same average miles per gallon. ANOVA tests are based on a new distribution called the *F* distribution.



©John Smock/SIPA/Newscom

Introductory Case

Effectiveness of Mandatory Caloric Postings

The federal health care law enacted in March 2010 requires chain restaurants with 20 locations or more to post caloric information on their menus. The government wants calorie listings posted to make it easier for consumers to select healthier options. New York City pioneered the requirement of caloric information on menus in 2008, but research has shown mixed results on whether this requirement has prompted consumers to select healthier foods (*The Wall Street Journal*, August 31, 2010). Molly Hosler, a nutritionist in San Mateo, California, would like to study the effects of a recent local menu ordinance requiring caloric postings. She obtains transaction data for 40 Starbucks cardholders around the time that San Mateo implemented the ordinance. For each cardholder, drink and food calories were recorded prior to the ordinance and then after the ordinance. Table 10.1 shows a portion of the data.

TABLE 10.1 Average Caloric Intake Before and After Menu-Labeling Ordinance

Customer	Drink Calories		Food Calories	
	Before	After	Before	After
1	141	142	395	378
2	137	140	404	392
⋮	⋮	⋮	⋮	⋮
40	147	141	406	400

FILE
Drink_Calories
Food_Calories

Molly wants to use the sample information to

1. Determine whether the average calories of purchased drinks declined after the passage of the ordinance.
2. Determine whether the average calories of purchased food declined after the passage of the ordinance.
3. Assess the implications of caloric postings for Starbucks and other chains.

A synopsis of this case is provided at the end of Section 10.2.

LO **10.1**

Make inferences about the difference between two population means based on independent sampling.

10.1 INFERENCE CONCERNING THE DIFFERENCE BETWEEN TWO MEANS

In this section, we consider statistical inference about the difference between two population means based on **independent random samples**. Independent random samples are samples that are completely unrelated to one another. Consider the example where we are interested in the difference between male and female salaries. For one sample, we collect data from the male population, while for the other sample we collect data from the female population. The two samples are considered to be independent because the selection of one is in no way influenced by the selection of the other. Similarly, in a comparison of battery lives between Brand A and Brand B, one sample comes from the Brand A population, while the other sample comes from the Brand B population. Again, both samples can be considered to be drawn independently.

INDEPENDENT RANDOM SAMPLES

Two (or more) random samples are considered independent if the process that generates one sample is completely separate from the process that generates the other sample. The samples are clearly delineated.

Confidence Interval for $\mu_1 - \mu_2$

As discussed earlier, we use sample statistics to estimate the population parameter of interest. For example, the sample mean $\overline{X}$ is the point estimator for the population mean μ. In a similar vein, the difference between the two sample means $\overline{X}_1 - \overline{X}_2$ is a point estimator for the difference between two population means $\mu_1 - \mu_2$, where μ_1 is the mean of the first population and μ_2 is the mean of the second population. The estimate is found by taking the difference of the sample means $\bar{x}_1$ and $\bar{x}_2$ computed from two independent random samples with n_1 and n_2 observations, respectively.

Let's first discuss the sampling distribution of $\overline{X}_1 - \overline{X}_2$. As in the case of a single population mean, this estimator is unbiased; that is, $E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2$. Moreover, recall that the statistical inference regarding the population mean μ is based on the condition that the sample mean $\overline{X}$ is normally distributed. Similarly, for statistical inference regarding $\mu_1 - \mu_2$, it is imperative that the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is normal. Therefore, if we assume that the two sample means are derived from two independent and normally distributed populations, then $\overline{X}_1 - \overline{X}_2$ is also normally distributed. If the underlying populations cannot be assumed to be normally distributed, then by the central limit theorem, the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is approximately normal only if both sample sizes are sufficiently large—that is, $n_1 \geq 30$ and $n_2 \geq 30$.

As in the case of a single population mean, we consider two scenarios. If we know the variances of the two populations σ_1^2 and σ_2^2 (or the standard deviations σ_1 and σ_2), we use the z distribution for the statistical inference. A more common case is to use the t_{df} distribution, where the sample variances s_1^2 and s_2^2 are used in place of the unknown population variances. When σ_1^2 and σ_2^2 are not known, we will examine two cases: (a) they can be assumed equal ($\sigma_1^2 = \sigma_2^2$) or (b) they cannot be assumed equal ($\sigma_1^2 \neq \sigma_2^2$).

The confidence interval for the difference in means is based on the same procedure outlined in Chapter 8. In particular, the formula for the confidence interval will follow the standard format given by point estimate ± margin of error.

We use sample data to calculate the point estimate for $\mu_1 - \mu_2$ as the difference between the two sample means $\bar{x}_1 - \bar{x}_2$. The margin of error equals the standard error $se(\overline{X}_1 - \overline{X}_2)$ multiplied by $z_{\alpha/2}$ or $t_{\alpha/2,df}$, depending on whether or not the population variances are known.

CONFIDENCE INTERVAL FOR $\mu_1 - \mu_2$

A $100(1-\alpha)\%$ confidence interval for the difference between two population means $\mu_1 - \mu_2$ is given by

1. $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, if the population variances, σ_1^2 and σ_2^2, are known.
2. $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,df}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, if σ_1^2 and σ_2^2 are unknown but assumed equal. A pooled estimate of the common variance is $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$, where s_1^2 and s_2^2 are the corresponding sample variances and the degrees of freedom $df = n_1 + n_2 - 2$.
3. $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, if σ_1^2 and σ_2^2 are unknown and cannot be assumed equal. The degrees of freedom $df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$. Since the resultant value for df is rarely an integer, we generally round the value down to obtain the appropriate t value from the t table. Software packages use various rounding rules when reporting the resultant value for df.

These formulas are valid only if $\overline{X}_1 - \overline{X}_2$ (approximately) follows a normal distribution.

Note that in the case when we construct a confidence interval for $\mu_1 - \mu_2$ where σ_1^2 and σ_2^2 are unknown but assumed equal, we calculate a pooled estimate of the common variance s_p^2. In other words, because the two populations are assumed to have the same population variance, the two sample variances s_1^2 and s_2^2 are simply two separate estimates of this population variance. We estimate the population variance by a *weighted* average of s_1^2 and s_2^2, where the weights applied are their respective degrees of freedom relative to the total number of degrees of freedom. In the case when σ_1^2 and σ_2^2 are unknown and cannot be assumed equal, we cannot calculate a pooled estimate of the population variance.

EXAMPLE 10.1

A consumer advocate analyzes the nicotine content in two brands of cigarettes. A sample of 20 cigarettes of Brand A resulted in an average nicotine content of 1.68 milligrams with a standard deviation of 0.22 milligram; 25 cigarettes of Brand B yielded an average nicotine content of 1.95 milligrams with a standard deviation of 0.24 milligram.

Brand A	Brand B
$\bar{x}_1 = 1.68$	$\bar{x}_2 = 1.95$
$s_1 = 0.22$	$s_2 = 0.24$
$n_1 = 20$	$n_2 = 25$

Construct the 95% confidence interval for the difference between the two population means. Nicotine content is assumed to be normally distributed. In addition, the population variances are unknown but assumed equal.

SOLUTION: We wish to construct a confidence interval for $\mu_1 - \mu_2$ where μ_1 is the mean nicotine level for Brand A and μ_2 is the mean nicotine level for Brand B. Since the population variances are unknown but assumed equal, we use the formula

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,df}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

We calculate the point estimate $\bar{x}_1 - \bar{x}_2 = 1.68 - 1.95 = -0.27$. In order to find $t_{\alpha/2,df}$, we determine $df = n_1 + n_2 - 2 = 20 + 25 - 2 = 43$. For the 95% confidence interval ($\alpha = 0.05$), we reference the t table to find $t_{0.025,43} = 2.017$.

We then calculate the pooled estimate of the population variance as

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(20 - 1)(0.22)^2 + (25 - 1)(0.24)^2}{20 + 25 - 2} = 0.0535.$$

Inserting the appropriate values into the formula, we have

$$-0.27 \pm 2.017\sqrt{0.0535\left(\frac{1}{20} + \frac{1}{25}\right)} = -0.27 \pm 0.14.$$

In other words, the 95% confidence interval for the difference between the two means ranges from −0.41 to −0.13. Shortly, we will use this interval to conduct a two-tailed hypothesis test.

Hypothesis Test for $\mu_1 - \mu_2$

As always, when specifying the competing hypotheses, it is important to (1) identify the relevant population parameter, (2) determine whether a one- or a two-tailed test is appropriate, and (3) include some form of the equality sign in the null hypothesis and use the alternative hypothesis to establish a claim. In order to conduct a hypothesis test concerning the parameter $\mu_1 - \mu_2$, the competing hypotheses will take one of the following general forms:

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
H_0: $\mu_1 - \mu_2 = d_0$	H_0: $\mu_1 - \mu_2 \leq d_0$	H_0: $\mu_1 - \mu_2 \geq d_0$
H_A: $\mu_1 - \mu_2 \neq d_0$	H_A: $\mu_1 - \mu_2 > d_0$	H_A: $\mu_1 - \mu_2 < d_0$

In most applications, the hypothesized difference d_0 between two population means μ_1 and μ_2 is zero. In this scenario, a two-tailed test determines whether the two means differ from one another, a right-tailed test determines whether μ_1 is greater than μ_2, and a left-tailed test determines whether μ_1 is less than μ_2.

We can also construct hypotheses where the hypothesized difference d_0 is a value other than zero. For example, if we wish to determine if the mean return of an emerging market fund is more than two percentage points higher than that of a developed market fund, the resulting hypotheses are H_0: $\mu_1 - \mu_2 \leq 2$ versus H_A: $\mu_1 - \mu_2 > 2$.

EXAMPLE 10.2

Revisit Example 10.1.

a. Specify the competing hypotheses in order to determine whether the average nicotine levels differ between Brand A and Brand B.

b. Using the 95% confidence interval, what is the conclusion to the test?

SOLUTION:

a. We want to determine if the average nicotine levels differ between the two brands, or $\mu_1 \neq \mu_2$, so we formulate a two-tailed hypothesis test as

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 - \mu_2 \neq 0$$

b. In Example 10.1, we calculated the 95% confidence interval for the difference between the two means as -0.27 ± 0.14 or, equivalently, the confidence

interval ranges from −0.41 to −0.13. This interval does not contain zero, the value hypothesized under the null hypothesis. This information allows us to reject H_0; the sample data support the conclusion that average nicotine levels between the two brands differ at the 5% significance level.

While it is true that we can use confidence intervals to conduct two-tailed hypothesis tests, the four-step procedure outlined in Chapter 9 can be implemented to conduct one- or two-tailed hypothesis tests. (It is possible to adjust the confidence interval to accommodate a one-tailed test, but we do not discuss this modification.) The only real change in the process is the specification of the test statistic. We use the point estimate $\bar{x}_1 - \bar{x}_2$ to derive the value of the test statistic z or t_{df} by dividing $(\bar{x}_1 - \bar{x}_2) - d_0$ by the standard error of the estimator $se(\overline{X}_1 - \overline{X}_2)$.

TEST STATISTIC FOR TESTING $\mu_1 - \mu_2$

The value of the test statistic for a hypothesis test concerning the difference between two population means, $\mu_1 - \mu_2$, is computed using one of the following three formulas:

1. If σ_1^2 and σ_2^2 are known, then the value of the test statistic is computed as
$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$
2. If σ_1^2 and σ_2^2 are unknown but assumed equal, then the value of the test statistic is computed as $t_{df} = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, where $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$ and $df = n_1 + n_2 - 2$.
3. If σ_1^2 and σ_2^2 are unknown and cannot be assumed equal, then the value of the test statistic is computed as $t_{df} = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$, where $df = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$. For df, we generally round the value down; software packages use various rounding rules when reporting the resultant value for df.

These formulas are valid only if $\overline{X}_1 - \overline{X}_2$ (approximately) follows a normal distribution.

EXAMPLE 10.3

An economist claims that average weekly food expenditure for households in City 1 is more than the average weekly food expenditure for households in City 2. She surveys 35 households in City 1 and obtains an average weekly food expenditure of $164. A sample of 30 households in City 2 yields an average weekly food expenditure of $159. Prior studies suggest that the population standard deviation for City 1 and City 2 are $12.50 and $9.25, respectively.

City 1	City 2
$\bar{x}_1 = 164$	$\bar{x}_2 = 159$
$\sigma_1 = 12.50$	$\sigma_2 = 9.25$
$n_1 = 35$	$n_2 = 30$

a. Specify the competing hypotheses to test the economist's claim.

b. Calculate the value of the test statistic and the p-value.

c. At the 5% significance level, is the economist's claim supported by the data?

SOLUTION:

a. The relevant parameter of interest is $\mu_1 - \mu_2$, where μ_1 is the mean weekly food expenditure for City 1 and μ_2 is the mean weekly food expenditure for City 2. The economist wishes to determine if the mean weekly food expenditure in City 1 is more than that of City 2; that is, $\mu_1 > \mu_2$. This is an example of a right-tailed test where the appropriate hypotheses are

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_A: \mu_1 - \mu_2 > 0$$

b. Since the population standard deviations are known, we compute the value of the test statistic as

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(164 - 159) - 0}{\sqrt{\frac{(12.50)^2}{35} + \frac{(9.25)^2}{30}}} = 1.85.$$

The p-value of the right-tailed test is computed as p-value $= P(Z \geq 1.85) = 0.0322$.

c. We reject the null hypothesis since the p-value of 0.0322 is less than the chosen $\alpha = 0.05$. Therefore, at the 5% significance level, the economist concludes that average weekly food expenditure in City 1 is more than that of City 2.

Using Excel for Testing Hypotheses about $\mu_1 - \mu_2$

Excel provides several options that simplify the steps when conducting a hypothesis test about $\mu_1 - \mu_2$. If we are only provided with summary statistics, then the best way to calculate the value of the test statistic and the p-value would be to use methods analogous to those outlined in Chapter 9. However, when given raw sample data, it is possible to use a single function in Excel's Data Analysis Toolpak that generates all the necessary information. Consider the following example with raw sample data.

EXAMPLE 10.4

Table 10.2 shows annual return data for 10 firms in the gold industry and 10 firms in the oil industry. Can we conclude at the 5% significance level that the average returns in the two industries differ? Here we assume that the sample data are drawn independently from normally distributed populations. Since the variance is a common measure of risk when analyzing financial returns, we cannot assume that the risk from investing in the gold industry is the same as the risk from investing in the oil industry.

TABLE 10.2 Annual Returns (in percent)

FILE *Gold_Oil*

Gold	Oil
6	−3
15	15
19	28
26	18
2	32
16	31
31	15
14	12
15	10
16	15

SOLUTION: We let μ_1 denote the mean return for the gold industry and μ_2 denote the mean return for the oil industry. Since we wish to test whether the mean returns differ, we set up the null and alternative hypotheses as

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 - \mu_2 \neq 0$$

Given that we are testing the difference between two means when the population variances are unknown and not equal, we need to calculate $t_{df} = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$.

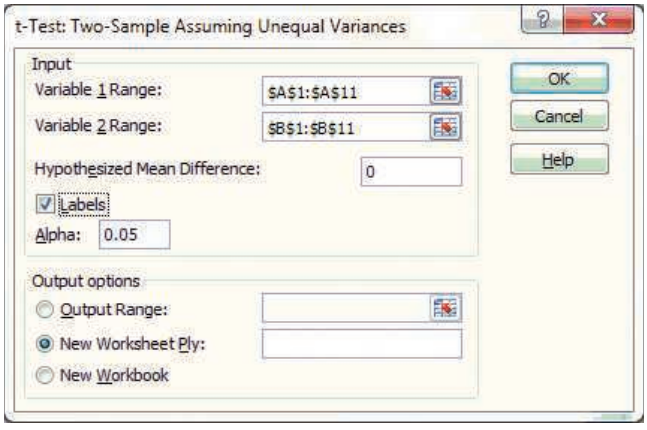
Recall that the calculation for the degrees of freedom for the corresponding test statistic is rather involved. Fortunately, Excel provides the degrees of freedom, the value of the test statistic, and the *p*-value.

a. Open the ***Gold_Oil*** data file.

b. Choose **Data > Data Analysis > t-Test: Two-Sample Assuming Unequal Variances > OK**. (Note: Excel provides two other options when we want to test the difference between two population means from independent samples and we have access to the raw data. If the population variances are known, we use the option **z-Test: Two-Sample for Means**. If the population variances are unknown but assumed equal, we use the option **t-Test: Two-Sample Assuming Equal Variances**.)

c. See Figure 10.1. In the dialog box, choose *Variable 1 Range* and select the Gold data. Then, choose *Variable 2 Range* and select the Oil data. Enter a *Hypothesized Mean Difference* of 0 since $d_0 = 0$, check the *Labels* box if you include Gold and Oil as headings, and enter an α value of 0.05 since the test is conducted at the 5% significance level. Click **OK**.

FIGURE 10.1 Excel's dialog box for *t*-test with unequal variances



Source: Microsoft Excel

Table 10.3 shows the Excel output.

The value of the test statistic and the p-value for this two-tailed test are −0.3023 and 0.7661, respectively (see these values in boldface in Table 10.3). At the 5% significance level, we cannot reject H_0 since the p-value is greater than 0.05. While average returns in the oil industry seem to slightly outperform average returns in the gold industry ($\bar{x}_2 = 17.3 > 16.0 = \bar{x}_1$), the difference is not statistically significant.

TABLE 10.3 Excel's Output for t-Test concerning $\mu_1 - \mu_2$

	Gold	Oil
Mean	16	17.3
Variance	70.6667	114.2333
Observations	10	10
Hypothesized Mean Difference	0	
Df	17	
t Stat	**−0.3023**	
P(T ≤ t) one-tail	0.3830	
t Critical one-tail	1.7396	
P(T ≤ t) two-tail	**0.7661**	
t Critical two-tail	**2.1098**	

Given the information in Table 10.3, it is also possible to calculate the corresponding 95% confidence interval for $\mu_1 - \mu_2$. Recall that when we estimate the difference between two population means when the population variances are unknown and cannot be assumed equal, we use $t_{\alpha/2,df}$ in its construction. Since we entered 0.05 for the significance level, the value for $t_{\alpha/2,df} = t_{0.025,17} = 2.1098$ (see this value in boldface in Table 10.3). We then calculate:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,df}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (16.0 - 17.3) \pm 2.1098\sqrt{\frac{70.6667}{10} + \frac{114.2333}{10}}$$
$$= -1.3 \pm 9.07.$$

That is, the 95% confidence interval for the difference between the two means ranges from −10.37 to 7.77. We note that this interval contains zero, the value hypothesized under the null hypothesis. Using the 95% confidence interval, we again cannot support the conclusion that the population mean returns differ at the 5% significance level.

EXERCISES 10.1

Mechanics

1. Consider the following data drawn independently from normally distributed populations:

$$\bar{x}_1 = 25.7 \quad \bar{x}_2 = 30.6$$
$$\sigma_1^2 = 98.2 \quad \sigma_2^2 = 87.4$$
$$n_1 = 20 \quad n_2 = 25$$

   a. Construct the 95% confidence interval for the difference between the population means.

   b. Specify the competing hypotheses in order to determine whether or not the population means differ.

   c. Using the confidence interval from part a, can you reject the null hypothesis? Explain.

2. Consider the following data drawn independently from normally distributed populations:

$$\bar{x}_1 = -10.5 \quad \bar{x}_2 = -16.8$$
$$s_1^2 = 7.9 \quad s_2^2 = 9.3$$
$$n_1 = 15 \quad n_2 = 20$$

   a. Construct the 95% confidence interval for the difference between the population means. Assume that the population variances are equal.

b. Specify the competing hypotheses in order to determine whether or not the population means differ.

c. Using the confidence interval from part a, can you reject the null hypothesis? Explain.

3. Consider the following competing hypotheses and accompanying sample data drawn independently from normally distributed populations.

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 - \mu_2 \neq 0$$

$\bar{x}_1 = 57$	$\bar{x}_2 = 63$
$\sigma_1 = 11.5$	$\sigma_2 = 15.2$
$n_1 = 20$	$n_2 = 20$

Test whether the population means differ at the 5% significance level.

4. Consider the following competing hypotheses and accompanying sample data. The two populations are known to be normally distributed.

$$H_0: \mu_1 - \mu_2 \leq 0$$
$$H_A: \mu_1 - \mu_2 > 0$$

$\bar{x}_1 = 20.2$	$\bar{x}_2 = 17.5$
$s_1 = 2.5$	$s_2 = 4.4$
$n_1 = 10$	$n_2 = 12$

a. Implement the test at the 5% significance level under the assumption that the population variances are equal.

b. Repeat the analysis at the 10% significance level.

5. Consider the following competing hypotheses and accompanying sample data drawn independently from normally distributed populations.

$$H_0: \mu_1 - \mu_2 \geq 0$$
$$H_A: \mu_1 - \mu_2 < 0$$

$\bar{x}_1 = 249$	$\bar{x}_2 = 262$
$s_1 = 35$	$s_2 = 23$
$n_1 = 10$	$n_2 = 10$

a. Implement the test at the 5% significance level under the assumption that the population variances are equal.

b. Implement the test at the 5% significance level under the assumption that the population variances are not equal.

6. Consider the following competing hypotheses and accompanying sample data.

$$H_0: \mu_1 - \mu_2 = 5$$
$$H_A: \mu_1 - \mu_2 \neq 5$$

$\bar{x}_1 = 57$	$\bar{x}_2 = 43$
$s_1 = 21.5$	$s_2 = 15.2$
$n_1 = 22$	$n_2 = 18$

Assume that the populations are normally distributed with equal variances.

a. Calculate the value of the test statistic and the *p*-value.

b. At the 5% significance level, can you conclude that the difference between the two means differs from 5?

7. Consider the following sample data drawn independently from normally distributed populations with equal population variances.

Sample 1	Sample 2
12.1	8.9
9.5	10.9
7.3	11.2
10.2	10.6
8.9	9.8
9.8	9.8
7.2	11.2
10.2	12.1

a. Construct the relevant hypotheses to test if the mean of the second population is greater than the mean of the first population.

b. Implement the test at the 1% significance level.

c. Implement the test at the 10% significance level.

8. Consider the following sample data drawn independently from normally distributed populations with unequal population variances.

Sample 1	Sample 2
88	98
110	114
102	118
96	128
74	102
120	110

a. Construct the relevant hypothesis to test if the means of the two populations differ.

b. What is the value of the test statistic and the *p*-value?

c. At the 10% significance level, do the two population means differ?

Applications

9. According to a Health of Boston report, female residents in Boston have a higher average life expectancy as compared to male residents (*The Boston Globe*, August 16, 2010). You collect the following sample data to verify the results of the report. You also use the historical (population) standard deviation of 8.2 years for females and 8.6 years for males.

Female	Male
$\bar{x}_1 = 81.1$	$\bar{x}_2 = 74.8$
$n_1 = 32$	$n_2 = 32$

a. Set up the hypotheses to test whether the average life expectancy of female Bostonians is higher than that of male Bostonians.

b. Calculate the value of the test statistic and the *p*-value.
c. At the 10% significance level, can we conclude that female Bostonians live longer than male Bostonians?

10. A joint project of the U.S. Census Bureau and the National Science Foundation shows that people with a bachelor's degree who transferred from a community college earn less than those who start at a four-year school (*USA TODAY*, March 17, 2009). Previous studies referred to this occurrence as a "community college penalty." Lucille Barnes wonders if a similar pattern applies to her university. The accompanying table shows the average salary of 100 graduates with an associate degree and the average salary of 100 graduates with no associate degree. Lucille believes that the population standard deviation is $4,400 for graduates with an associate degree and $1,500 for graduates with no associate degree.

Bachelor's Degree with Associate Degree	Bachelor's Degree with No Associate Degree
$\bar{x}_1 = 52,000$	$\bar{x}_2 = 54,700$
$n_1 = 100$	$n_2 = 100$

a. Set up the hypotheses to test if the report's conclusion also applies to Lucille's university.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 5% significance level, can we conclude that there is a "community college penalty" at Lucille's university?

11. The Chartered Financial Analyst (CFA) designation is fast becoming a requirement for serious investment professionals. It is an attractive alternative to getting an MBA for students wanting a career in investment. A student of finance is curious to know if a CFA designation is a more lucrative option than an MBA. He collects data on 38 recent CFAs with a mean salary of $138,000 and a standard deviation of $34,000. A sample of 80 MBAs results in a mean salary of $130,000 with a standard deviation of $46,000.
a. Specify the hypotheses to test whether a CFA designation is more lucrative than an MBA.
b. Calculate the value of the test statistic and the *p*-value. Do not assume that the population variances are equal.
c. At the 5% significance level, is a CFA designation more lucrative than an MBA?

12. An entrepreneur owns some land that he wishes to develop. He identifies two development options: build condominiums or build apartment buildings. Accordingly, he reviews public records and derives the following summary measures concerning annual profitability based on a random sample of 30 for each such local business venture. For the analysis, he uses a historical (population) standard deviation of $22,500 for condominiums and $20,000 for apartment buildings.

Condominiums	Apartment Buildings
$\bar{x}_1 = 244,200$	$\bar{x}_2 = 235,800$
$n_1 = 30$	$n_2 = 30$

a. Set up the hypotheses to test whether the mean profitability differs between condominiums and apartment buildings.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 5% significance level, what is the conclusion to the test? What if the significance level is 10%?

13. David Anderson has been working as a lecturer at Michigan State University for the last three years. He teaches two large sections of introductory accounting every semester. While he uses the same lecture notes in both sections, his students in the first section outperform those in the second section. He believes that students in the first section not only tend to get higher scores, they also tend to have lower variability in scores. David decides to carry out a formal test to validate his hunch regarding the difference in average scores. In a random sample of 18 students in the first section, he computes a mean and a standard deviation of 77.4 and 10.8, respectively. In the second section, a random sample of 14 students results in a mean of 74.1 and a standard deviation of 12.2.
a. Construct the null and the alternative hypotheses to test David's hunch.
b. Compute the value of the test statistic. What assumption regarding the populations is necessary to implement this step?
c. Implement the test at $\alpha = 0.01$ and interpret your results.

14. A design engineer at Sperling Manufacturing, a supplier of high-quality ball bearings, claims a new machining process can result in a higher daily output rate. Accordingly, the production group is conducting an experiment to determine if this claim can be substantiated. The mean and the standard deviation of bearings in a sample of 8 days' output using the new process equal 2,613.63 and 90.78, respectively. A similar sample of 10 days' output using the old process yields the mean and the standard deviation of 2,485.10 and 148.22, respectively.
a. Set up the hypotheses to test whether the mean output rate of the new process exceeds that of the old process. Assume normally distributed populations and equal population variances for each process.
b. Compute the value of the test statistic and the *p*-value.
c. At the 5% significance level, what is the conclusion of the experiment?
d. At the 1% significance level, what is the conclusion of the experiment?

15. A phone manufacturer wants to compete in the touch screen phone market. Management understands that the leading product has a less than desirable battery life. They aim to compete with a new touch screen phone that is guaranteed to have a battery life more than two hours longer than the

leading product. A recent sample of 120 units of the leading product provides a mean battery life of 5 hours and 40 minutes with a standard deviation of 30 minutes. A similar analysis of 100 units of the new product results in a mean battery life of 8 hours and 5 minutes and a standard deviation of 55 minutes. It is not reasonable to assume that the population variances of the two products are equal.

a. Set up the hypotheses to test if the new product has a battery life more than two hours longer than the leading product.
b. Implement the test at the 5% significance level.

16. In May 2008, CNN reported that sports utility vehicles (SUVs) are plunging toward the "endangered" list. Due to soaring oil prices and environmental concerns, consumers are replacing gas-guzzling vehicles with fuel-efficient smaller cars. As a result, there has been a big drop in the demand for new as well as used SUVs. A sales manager of a used car dealership believes that it takes an average of 30 days longer to sell an SUV as compared to a small car. In the last two months, he sold 18 SUVs that took an average of 95 days to sell with a standard deviation of 32 days. He also sold 38 small cars with an average of 48 days to sell and a standard deviation of 24 days.
    a. Construct the null and the alternative hypotheses to contradict the manager's claim.
    b. Compute the value of the test statistic and the *p*-value. Assume that the populations are normally distributed and that the variability of selling time for the SUVs and the small cars is the same.
    c. Implement the test at $\alpha = 0.10$ and interpret your results.
17. FILE ***Refrigerator_Longevity.*** A consumer advocate researches the length of life between two brands of refrigerators, Brand A and Brand B. He collects data (measured in years) on the longevity of 40 refrigerators for Brand A and repeats the sampling for Brand B.
    a. Specify the competing hypotheses to test whether the average length of life differs between the two brands.
    b. Calculate the value of the test statistic and the *p*-value. Assume that $\sigma_A^2 = 4.4$ and $\sigma_B^2 = 5.2$.
    c. At the 5% significance level, what is the conclusion?
18. FILE ***Website_Searches.*** The "See Me" marketing agency wants to determine if time of day for a television advertisement influences website searches for a product. They have extracted the number of website searches occurring during a one-hour period after an advertisement was aired for a random sample of 30 day and 30 evening advertisements. A portion of the data is shown in the accompanying table.

Day Searches	Evening Searches
96670	118379
97855	111005
⋮	⋮
95103	114721

    a. Set up the hypotheses to test whether the mean number of website searches differs between the day and evening advertisements.
    b. Calculate the value of the test statistic and the *p*-value. Assume that the population variances are equal.
    c. At the 5% significance level, what is the conclusion?
19. FILE ***Different_Diets.*** According to a study published in the *New England Journal of Medicine*, overweight people on low-carbohydrate and Mediterranean diets lost more weight and got greater cardiovascular benefits than people on a conventional low-fat diet (*The Boston Globe*, July 17, 2008). A nutritionist wishes to verify these results and documents the weight loss (in pounds) of 30 dieters on the low-carbohydrate and Mediterranean diets and 30 dieters on the low-fat diet.
    a. Set up the hypotheses to test the claim that the mean weight loss for those on low-carbohydrate or Mediterranean diets is greater than the mean weight loss for those on a conventional low-fat diet.
    b. Calculate the value of the test statistic and the *p*-value. Assume that the population variances are equal.
    c. At the 5% significance level, can the nutritionist conclude that people on low-carbohydrate or Mediterranean diets lose more weight than people on a conventional low-fat diet?
20. FILE ***Tractor_Times.*** The production department at Greenside Corporation, a manufacturer of lawn equipment, has devised a new manual assembly method for its lawn tractors. Now it wishes to determine if it is reasonable to conclude that the mean assembly time of the new method is less than the old method. Accordingly, they have randomly sampled assembly times (in minutes) from 40 tractors using the old method and 32 tractors using the new method. A portion of the data is shown in the accompanying table.

Old Method	New Method
32	30
36	32
⋮	⋮

    a. Set up the hypotheses to test the claim that the mean assembly time using the new method is less than the old method.
    b. Calculate the value of the test statistic and the *p*-value. Assume that the population variances are not equal.
    c. At the 5% significance level, what is the conclusion? What if the significance level is 10%?
21. FILE ***Nicknames.*** Baseball has always been a favorite pastime in America and is rife with statistics and theories. In a paper, researchers showed that major league players who have nicknames live an average of 2½ years longer than those without them (*The Wall Street Journal*, July 16, 2009). You do not believe in this result and decide to collect data on the

lifespan of 30 baseball players along with a nickname variable that equals 1 if the player had a nickname and 0 otherwise. A portion of the data is shown in the accompanying table.

Years	Nickname
74	1
62	1
⋮	⋮
64	0

a. Create two subsamples consisting of players with and without nicknames. Calculate the average longevity for each subsample.
b. Specify the hypotheses to contradict the claim made by the researchers.
c. Calculate the value of the test statistic and the *p*-value. Assume that the population variances are equal.
d. What is the conclusion of the test using a 5% level of significance?

22. FILE ***Starting_Salaries.*** Recent evidence suggests that graduating from college during bad economic times can impact the graduate's earning power for a long time (*Financial Times*, June 1, 2012). An associate dean at a prestigious college wants to determine if the starting salary of his college graduates has declined from 2008 to 2010. He expects the variance of the salaries to be different between these two years. A portion of the data is shown in the accompanying table.

Salary 2008	Salary 2010
35000	34000
56000	62000
⋮	⋮
47000	54000

At the 5% significance level, determine if the mean starting salary has decreased from 2008 to 2010.

23. FILE ***Spending_Gender.*** Researchers at the Wharton School of Business have found that men and women shop for different reasons. While women enjoy the shopping experience, men are on a mission to get the job done. Men do not shop as frequently, but when they do, they make big purchases like expensive electronics. The accompanying table shows a portion of the amount spent (in $) over the weekend by 40 men and 60 women at a local mall.

Spending by Men	Spending by Women
85	90
102	79
⋮	⋮

At the 1% significance level, determine if the mean amount spent by men is more than that by women. Assume that the population variances are equal.

LO **10.2**

Make inferences about the mean difference based on matched-pairs sampling.

10.2 INFERENCE CONCERNING MEAN DIFFERENCES

One of the crucial assumptions in Section 10.1 concerning differences between two population means is that the samples are drawn independently. As mentioned earlier, two samples are independent if the selection of one is not influenced by the selection of the other. When we want to conduct tests on two population means based on samples that we believe are not independent, we need to employ a different methodology.

A common case of dependent sampling, commonly referred to as **matched-pairs sampling**, is when the samples are paired or matched in some way. Such samples are useful in evaluating strategies because the comparison is made between "apples" and "apples." For instance, an effective way to assess the benefits of a new medical treatment is by evaluating the same patients before and after the treatment. If, however, one group of people is given the treatment and another group is not, then it is not clear if the observed differences are due to the treatment or due to other important differences between the groups.

For matched-pairs sampling, the parameter of interest is referred to as the mean difference μ_D where $D = X_1 - X_2$, and the random variables X_1 and X_2 are matched in a pair. The statistical inference regarding μ_D is based on the estimator $\overline{D}$, representing the sample mean difference. It requires that $X_1 - X_2$ is normally distributed or that the sample size is sufficiently large ($n \geq 30$).

Recognizing a Matched-Pairs Experiment

It is important to be able to determine whether a particular experiment uses independent or matched-pairs sampling. In general, two types of matched-pairs sampling occur:

1. The first type of matched-pairs sample is characterized by a measurement, an intervention of some type, and then another measurement. We generally refer to these experiments as "before" and "after" studies. For example, an operation manager of a production facility wants to determine whether a new workstation layout improves productivity at her plant. She first measures output of employees before the layout change. Then she measures output of the same employees after the change. Another classic before-and-after example concerns weight loss of clients at a diet center. In these examples, the same individual gets sampled before and after the experiment.
2. The second type of matched-pairs sample is characterized by a pairing of observations, where it is not the same individual who gets sampled twice. Suppose an agronomist wishes to switch to an organic fertilizer but is unsure what the effects might be on his crop yield. It is important to the agronomist that the yields be similar. He matches 20 adjacent plots of land using the nonorganic fertilizer on one half of the plot and the organic fertilizer on the other.

In order to recognize a matched-pairs experiment, we watch for a natural pairing between one observation in the first sample and one observation in the second sample. If a natural pairing exists, then the experiment involves matched samples.

Confidence Interval for μ_D

When constructing a confidence interval for the mean difference μ_D, we follow the same general format of point estimate ± margin of error.

CONFIDENCE INTERVAL FOR μ_D

A $100(1-\alpha)\%$ confidence interval for the mean difference μ_D is given by

$$\overline{d} \pm t_{\alpha/2,df} s_D/\sqrt{n},$$

where $\overline{d}$ and s_D are the mean and the standard deviation, respectively, of the n sample differences and $df = n - 1$. This formula is valid only if $\overline{D}$ (approximately) follows a normal distribution.

In the next example, the values for $\overline{d}$ and s_D are explicitly given; we will outline the calculations when we discuss hypothesis testing.

EXAMPLE 10.5

A manager is interested in improving productivity at a plant by changing the layout of the workstation. For each of 10 workers, she measures the time it takes to complete a task before the change and again after the change. She calculates the following summary statistics for the sample difference: $\overline{d} = 8.5, s_D = 11.38$, and $n = 10$. Construct the 95% confidence interval for the mean difference, assuming that the productivity variable, before minus after, is normally distributed.

SOLUTION: In order to construct the 95% confidence interval for the mean difference, we use $\overline{d} \pm t_{\alpha/2,df} s_D/\sqrt{n}$. With $df = n - 1 = 10 - 1 = 9$ and $\alpha = 0.05$, we find $t_{\alpha/2,df} = t_{0.025,9} = 2.262$. Plugging the relevant values into the formula, we calculate $8.5 \pm 2.262(11.38/\sqrt{10}) = 8.5 \pm 8.14$. That is, the 95% confidence interval for the mean difference ranges from 0.36 to 16.64. This represents a fairly wide interval, caused by the high standard deviation s_D of the 10 sample differences.

Hypothesis Test for μ_D

As before, we generally want to test whether the mean difference μ_D differs from, is greater than, or is less than a given hypothesized mean difference d_0, or:

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
H_0: $\mu_D = d_0$	H_0: $\mu_D \leq d_0$	H_0: $\mu_D \geq d_0$
H_A: $\mu_D \neq d_0$	H_A: $\mu_D > d_0$	H_A: $\mu_D < d_0$

In practice, the competing hypotheses tend to be based on $d_0 = 0$. For example, when testing if the mean difference differs from zero, we use a two-tailed test with the competing hypotheses defined as H_0: $\mu_D = 0$ versus H_A: $\mu_D \neq 0$. If, on the other hand, we wish to determine whether or not the mean difference differs by some amount, say by 5 units, we set $d_0 = 5$ and define the competing hypotheses as H_0: $\mu_D = 5$ versus H_A: $\mu_D \neq 5$. One-tailed tests are defined similarly.

EXAMPLE 10.6

Using the information from Example 10.5, can the manager conclude at the 5% significance level that there has been a change in productivity since the adoption of the new workstation?

SOLUTION: In order to determine whether or not there has been a change in the mean difference, we formulate the null and the alternative hypotheses as

$$H_0: \mu_D = 0$$
$$H_A: \mu_D \neq 0$$

In Example 10.5, we found that the 95% confidence interval for the mean difference ranges from 0.36 to 16.64. Although the interval is very wide, the entire range is above the hypothesized value of zero. Therefore, at the 5% significance level the sample data suggest that the mean difference differs from zero. In other words, there has been a change in productivity due to the different layout in the workstation.

We now examine the four-step procedure to conduct one- or two-tailed hypothesis tests concerning the mean difference. We again convert the sample mean difference into its corresponding t_{df} statistic by dividing the difference between the sample mean difference and the hypothesized mean difference by the standard error of the estimator $se(\overline{D})$.

TEST STATISTIC FOR TESTING μ_D

The value of the test statistic for a hypothesis test concerning the population mean difference μ_D is computed as $t_{df} = \frac{\overline{d} - d_0}{s_D/\sqrt{n}}$, where $df = n - 1$, $\overline{d}$ and s_D are the mean and the standard deviation, respectively, of the n sample differences and d_0 is the hypothesized mean difference. This formula is valid only if $\overline{D}$ (approximately) follows a normal distribution.

EXAMPLE 10.7

FILE
Drink_Calories

Let's revisit the chapter's introductory case. Recall that a local ordinance requires chain restaurants to post caloric information on their menus. A nutritionist wants to examine whether average drink calories declined at Starbucks after the passage of the ordinance. The nutritionist obtains transaction data for 40 Starbucks cardholders

and records each cardholder's drink calories prior to the ordinance and then after the ordinance. A portion of the data is shown in Table 10.4. Can she conclude at the 5% significance level that the ordinance reduced average drink calories?

SOLUTION: We first note that this is a matched-pairs experiment; specifically, it conforms to a "before" and "after" type of study. Moreover, we want to find out whether average drink calories consumed prior to the ordinance are greater than average drink calories consumed after passage of the ordinance. Thus, we want to test if the mean difference μ_D is greater than zero, where $D = X_1 - X_2$, X_1 denotes drink calories before the ordinance, and X_2 denotes drink calories after the ordinance for a randomly selected Starbuck's customer. We specify the competing hypotheses as

$$H_0: \mu_D \leq 0$$
$$H_A: \mu_D > 0$$

The normality condition for the test is satisfied since the sample size $n \geq 30$. The value of the test statistic is calculated as $t_{df} = \frac{\overline{d} - d_0}{s_D/\sqrt{n}}$ where d_0 equals 0. In order to determine $\overline{d}$ and s_D, we first calculate the difference d_i for each i-th customer. For instance, customer 1 consumes 141 calories prior to the ordinance and 142 calories after the ordinance, for a difference of $d_1 = 141 - 142 = -1$. The differences for a portion of the other customers appear in the fourth column of Table 10.4.

TABLE 10.4 Data and Calculations for Example 10.7, $n = 40$

	Drink Calories			
Customer	Before	After	d_i	$(d_i - \overline{d})^2$
1	141	142	−1	$(-1 - 2.1)^2 = 9.61$
2	137	140	−3	$(-3 - 2.1)^2 = 26.01$
⋮	⋮	⋮	⋮	⋮
40	147	141	6	$(6 - 2.1)^2 = 15.21$
			$\sum d_i = 84$	$\sum (d_i - \overline{d})^2 = 2593.60$

We obtain the sample mean as

$$\overline{d} = \frac{\Sigma d_i}{n} = \frac{84}{40} = 2.10$$

Similarly, in the last column of Table 10.4, we square the differences between d_i and $\overline{d}$. Summing these squared differences yields the numerator in the formula for the sample variance s_D^2. The denominator is simply $n - 1$, so

$$s_D^2 = \frac{\sum (d_i - \overline{d})^2}{n - 1} = \frac{2{,}593.60}{40 - 1} = 66.5026.$$

As usual, the standard deviation is the positive square root of the sample variance—that is, $s_D = \sqrt{66.5026} = 8.1549$. We compute the value of the t_{df} test statistic with $df = n - 1 = 40 - 1 = 39$ as

$$t_{39} = \frac{\overline{d} - d_0}{s_D/\sqrt{n}} = \frac{2.10 - 0}{8.1549/\sqrt{40}} = 1.629.$$

Given a right-tailed hypothesis test with $df = 39$, we can use the t table to approximate the p-value $= P(T_{39} \geq 1.629)$, as $0.05 < p\text{-value} < 0.10$. Using Excel, we find that the exact p-value $= 0.0557$. Since the p-value > 0.05, we do not reject H_0. At the 5% significance level, we cannot conclude that the posting of nutritional information decreases average drink calories.

We should note that once we have calculated the mean difference and the standard deviation of the mean difference, the hypothesis test essentially reduces to a one-sample *t*-test for the population mean.

Using Excel for Testing Hypotheses about μ_D

Excel provides several options that simplify the steps when conducting a hypothesis test about μ_D. If we are only provided with summary statistics, then the best way to calculate the value of the test statistic and the *p*-value would be to use methods analogous to those outlined in Chapter 9. However, when given raw sample data, it is possible to use a single function in Excel's Data Analysis Toolpak that generates all the necessary information. Consider the following example.

EXAMPLE 10.8

FILE
Food_Calories

The nutritionist from Example 10.7 also wants to use the data from the 40 Starbucks cardholders in order to determine if the posting of caloric information has reduced average food calories. This test is also conducted at the 5% significance level.

SOLUTION: We set up the same competing hypotheses as in Example 10.7 since we want to know if average food calories were greater before the ordinance as compared to after the ordinance.

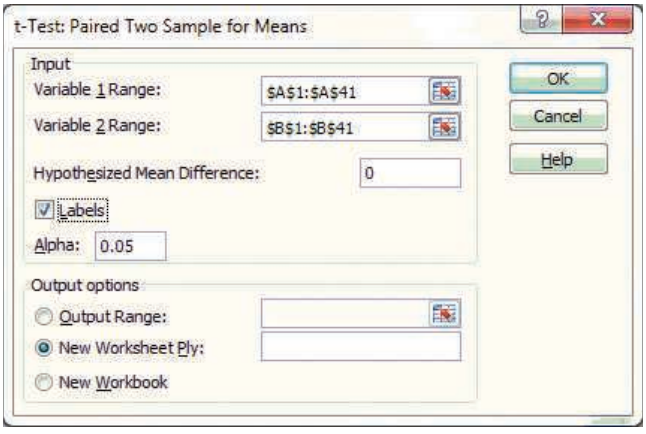
$$H_0: \mu_D \leq 0$$
$$H_A: \mu_D > 0$$

a. Open the ***Food_Calories*** data file.

b. Choose **Data > Data Analysis > *t*-Test: Paired Two Sample for Means > OK**.

c. See Figure 10.2. In the dialog box, choose *Variable 1 Range* and select the data in the Before column. Choose *Variable 2 Range* and select the data in the After column. Enter a *Hypothesized Mean Difference* of 0 since $d_0 = 0$, check the *Labels* box if you include Before and After as headings, and enter an α value of 0.05 since the test is conducted at the 5% significance level. Click **OK**.

FIGURE 10.2 Excel's dialog box for *t*-test with paired sample



Source: Microsoft Excel

Table 10.5 shows the Excel output. The value of the test statistic and the p-value for this one-tailed test are 6.7795 and 2.15E-08, respectively (see these values in boldface in Table 10.5). We can reject H_0 because the p-value < 0.05. Thus, at the 5% significance level, we can conclude that average food calories have declined after the passage of the ordinance.

TABLE 10.5 Excel's Output for t-Test concerning μ_D

	Before	After
Mean	400.275	391.475
Variance	49.9481	42.3583
Observations	40	40
Pearson Correlation	0.27080	
Hypothesized Mean Difference	0	
Df	39	
t Stat	**6.7795**	
P(T ≤ t) one-tail	**2.15E-08**	
t Critical one-tail	1.6849	
P(T ≤ t) two-tail	4.31E-08	
t Critical two-tail	2.0227	

Note: Although Excel calculates the p-value correctly, the expression it uses to denote the p-value is not always correct. In this example with a positive value for the test statistic, the expression should be "P(T ≥ t) one-tail" rather than "P(T ≤ t) one-tail."

SYNOPSIS OF INTRODUCTORY CASE



©Chris Hondros/Getty images News/Getty Images

In an effort to make it easier for consumers to select healthier options, the government wants chain restaurants to post caloric information on their menus. A nutritionist studies the effects of a recent local menu ordinance requiring caloric postings at a Starbucks in San Mateo, California. She obtains transaction data for 40 Starbucks cardholders and records each cardholder's drink and food calories prior to the ordinance and then after the ordinance. Two hypothesis tests are conducted. The first test examines whether average drink calories are less since the passage of the ordinance. After conducting a test on the mean difference at the 5% significance level, the nutritionist infers that the ordinance did not prompt customers to reduce their consumption of drink calories. The second test investigates whether average food calories are less since the passage of the ordinance. At the 5% significance level, the sample data suggest that customers have reduced their consumption of food calories since the passage of the ordinance. In sum, while the government is trying to ensure that customers process the calorie information as they are ordering, the results are consistent with research that has shown mixed results on whether mandatory caloric postings are prompting customers to select healthier foods.

EXERCISES 10.2

Mechanics

24. A sample of 20 paired observations generates the following data: $\overline{d} = 1.3$ and $s_D^2 = 2.6$. Assume a normal distribution.
    a. Construct the 90% confidence interval for the mean difference μ_D.
    b. Using the confidence interval, test whether the mean difference differs from zero. Explain.

25. The following table contains information on matched sample values whose differences are normally distributed.

Number	Sample 1	Sample 2
1	18	21
2	12	11
3	21	23
4	22	20
5	16	20
6	14	17
7	17	17
8	18	22

    a. Construct the 95% confidence interval for the mean difference μ_D.
    b. Specify the competing hypotheses in order to test whether the mean difference differs from zero.
    c. Using the confidence interval from part a, are you able to reject H_0? Explain.

26. Consider the following competing hypotheses and accompanying results from a matched-pairs sample:

$$H_0: \mu_D \geq 0;\ H_A: \mu_D < 0$$
$$\overline{d} = -2.8,\ s_D = 5.7,\ n = 12$$

    a. Calculate the value of the test statistic and the p-value, assuming that the sample difference is normally distributed.
    b. At the 5% significance level, what is the conclusion to the hypothesis test?

27. Consider the following competing hypotheses and accompanying results from a matched-pairs sample:

$$H_0: \mu_D \leq 2;\ H_A: \mu_D > 2$$
$$\overline{d} = 5.6,\ s_D = 6.2,\ n = 10$$

    a. Calculate the value of the test statistic and the p-value, assuming that the sample difference is normally distributed.
    b. Use the 1% significance level to make a conclusion.

28. A sample of 35 paired observations generates the following results: $\overline{d} = 1.2$ and $s_D = 3.8$.
    a. Specify the appropriate hypotheses to test if the mean difference is greater than zero.
    b. Calculate the value of the test statistic and the p-value.
    c. At the 5% significance level, can you conclude that the mean difference is greater than zero? Explain.

29. Consider the following matched-pairs sample that represents observations before and after an experiment. Assume that the sample differences are normally distributed.

Before	2.5	1.8	1.4	–2.9	1.2	–1.9	–3.1	2.5
After	2.9	3.1	3.9	–1.8	0.2	0.6	–2.5	2.9

    a. Construct the competing hypotheses to determine if the experiment increases the magnitude of the observations.
    b. Implement the test at the 5% significance level.
    c. Do the results change if we implement the test at the 1% significance level?

Applications

30. A manager of an industrial plant asserts that workers on average do not complete a job using Method A in the same amount of time as they would using Method B. Seven workers are randomly selected. Each worker's completion time (in minutes) is recorded by the use of Method A and Method B.

Worker	Method A	Method B
1	15	16
2	21	25
3	16	18
4	18	22
5	19	23
6	22	20
7	20	20

    a. Specify the null and alternative hypotheses to test the manager's assertion.
    b. Assuming that the completion time difference is normally distributed, calculate the value of the test statistic.
    c. Find the p-value.
    d. At the 10% significance level, is the manager's assertion supported by the data?

31. A diet center claims that it has the most effective weight loss program in the region. Its advertisements say, "Participants in our program lose more than 5 pounds within a month." Six clients of this program are weighed on the first day of the diet and then one month later.

Client	Weight on First Day of Diet	Weight One Month Later
1	158	151
2	205	200
3	170	169
4	189	179
5	149	144
6	135	129

    a. Specify the null and alternative hypotheses that test the diet center's claim.
    b. Assuming that weight loss is normally distributed, calculate the value of the test statistic.
    c. Find the p-value.
    d. At the 5% significance level, do the data support the diet center's claim?

32. A bank employs two appraisers. When approving borrowers for mortgages, it is imperative that the appraisers value the same types of properties consistently. To make sure that this is the case, the bank examines six properties (in $1,000s) that the appraisers had valued recently.

Property	Value from Appraiser 1	Value from Appraiser 2
1	235	239
2	195	190
3	264	271
4	315	310
5	435	437
6	515	525

a. Specify the competing hypotheses that determine whether there is any difference between the values estimated by Appraiser 1 and Appraiser 2.
b. Assuming that the value difference is normally distributed, calculate the value of the test statistic.
c. Find the *p*-value.
d. At the 5% significance level, is there sufficient evidence to conclude that the appraisers are inconsistent in their estimates? Explain.

33. The quality department at ElectroTech is examining which of two microscope brands (Brand A or Brand B) to purchase. They have hired someone to inspect six circuit boards using both microscopes. Below are the results in terms of the number of defects (e.g., solder voids, misaligned components) found using each microscope.

Circuit Board	Defects with Brand A	Defects with Brand B
1	12	14
2	8	9
3	16	16
4	14	12
5	9	8
6	13	15

a. Specify the null and alternative hypotheses to test for differences in the defects found between the microscope brands.
b. Assuming that the difference in defects is normally distributed, calculate the value of the test statistic and the *p*-value.
c. At the 5% significance level, is there a difference between the microscope brands?

34. A computer technology firm wishes to check whether the speed of a new processor exceeds that of an existing processor when used in one of its popular laptop computer models. Accordingly, it measures the time required (in seconds) to complete seven common tasks on two otherwise identical computers, one with the new processor and one with the existing processor. The time required is as follows:

Task	New Processor	Existing Processor
1	1.47	1.68
2	2.59	2.99
3	5.21	5.69
4	3.49	3.75
5	3.99	4.25
6	3.10	2.99
7	5.75	6.19

a. Specify the null and alternative hypotheses to test whether the time required for the new processor is less than the existing processor.
b. Assuming that the difference in time is normally distributed, calculate the value of the test statistic and the *p*-value.
c. At the 5% significance level, is the new processor faster than the old processor?

35. FILE ***Mock_SAT.*** A report criticizes SAT-test-preparation providers for promising big score gains without any hard data to back up such claims (*The Wall Street Journal*, May 20, 2009). Suppose eight college-bound students take a mock SAT, complete a three-month test-prep course, and then take the real SAT.

Student	Mock SAT	Real SAT
1	1830	1840
2	1760	1800
3	2000	2010
4	2150	2190
5	1630	1620
6	1840	1960
7	1930	1890
8	1710	1780

a. Specify the competing hypotheses that determine whether completion of the test-prep course increases a student's score on the real SAT.
b. Calculate the value of the test statistic and the *p*-value. Assume that the SAT scores difference is normally distributed.
c. At the 5% significance level, do the sample data support the test-prep providers' claims?

36. FILE ***Insurance_Premiums.*** The marketing department at Insure-Me, a large insurance company, wants to advertise that customers can save, on average, more than $100 on their annual automotive insurance policies (relative to their closest competitor) by switching their policies to Insure-Me. However, to avoid potential litigation for false advertising, they select a

random sample of 50 policyholders and compare their premiums to those of their closest competitor. A portion of the data is presented in the following table.

Policyholder	Competitor's Premium	"Insure-Me" Premium
1	958	1086
2	1034	366
⋮	⋮	⋮
50	1161	964

a. Specify the competing hypotheses to determine whether the mean difference between the competitor's premium and *Insure-Me's* premium is over $100.
b. Calculate the value of the test statistic and the *p*-value.
c. What is the conclusion at the 5% significance level? What is the conclusion at the 10% significance level?

37. FILE ***Electronic_Utilities.*** The following table shows the annual returns (in percent) for Fidelity's Select Electronic and Select Utilities mutual funds for the years 2001 through 2009.

Year	Electronic	Utilities
2001	–14.23	–21.89
2002	–50.54	–30.40
2003	71.89	26.42
2004	–9.81	24.22
2005	15.75	9.36
2006	0.30	30.08
2007	4.67	18.13
2008	–49.87	–36.00
2009	84.99	14.39

Source: www.finance.yahoo.com

a. Set up the hypotheses to test the claim that the mean return for the Electronic mutual fund differs from the mean return for the Utilities mutual fund.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 5% significance level, do the mean returns differ?

38. FILE ***Labor_Costs.*** The labor quotation department at Excabar, a large manufacturing company, wants to verify the accuracy of their labor bidding process (estimated cost per unit versus actual cost per unit). They have randomly chosen 35 product quotations that subsequently were successful (meaning the company won the contract for the product). A portion of the data is shown in the accompanying table.

Product	Estimated	Actual
1	13.90	12.90
2	18.80	15.80
⋮	⋮	⋮
35	17.80	14.80

a. Specify the competing hypotheses to determine whether there is a difference between the estimated cost and the actual cost.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 1% significance level, what is the conclusion?

39. FILE ***Smoking_Weight.*** It is fairly common for people to put on weight when they quit smoking. While a small weight gain is normal, excessive weight gain can create new health concerns that erode the benefits of not smoking. The accompanying table shows a portion of the weight data for 50 women before quitting and six months after quitting.

Weight before Quitting	Weight after Quitting
140	155
144	142
⋮	⋮
135	147

a. Construct and interpret the 95% confidence interval for the mean gain in weight.
b. Use the confidence interval to determine if the mean gain in weight differs from 5 pounds.

40. FILE ***Shift.*** When faced with a power hitter, many baseball teams utilize a defensive shift. A shift usually involves putting three infielders on one side of second base against pull hitters. Many believe that a power hitter's batting average is lower when he faces a shift defense as compared to when he faces a standard defense. Consider the following batting averages of 10 power hitters over the 2010 and 2011 seasons when they faced a shift defense versus when they faced a standard defense.

Player	Shift	Standard
Jack Cust	0.239	0.270
Adam Dunn	0.189	0.230
Prince Fielder	0.150	0.263
Adrian Gonzalez	0.186	0.251
Ryan Howard	0.177	0.317
Brian McCann	0.321	0.250
David Ortiz	0.245	0.232
Carlos Pena	0.243	0.191
Mark Teixeira	0.168	0.182
Jim Thome	0.211	0.205

Source: *The Fielding Bible-Volume III*, March 2012

a. Specify the competing hypotheses to determine whether the use of the defensive shift lowers a power hitter's batting average.
b. Calculate the value of the test statistic and the *p*-value. Assume that the batting average difference is normally distributed.
c. At the 5% significance level, is the defensive shift effective in lowering a power hitter's batting average?

10.3 INFERENCE CONCERNING DIFFERENCES AMONG MANY MEANS

We use an **analysis of variance (ANOVA) test** to determine if differences exist between the means of three or more populations under independent sampling. The ANOVA test is actually a generalization of the two-sample t test with equal but unknown variances discussed in Section 10.1. It is based on a new distribution, called the ***F* distribution**.[1] We will first discuss the characteristics of this important distribution before getting into the details of the ANOVA test.

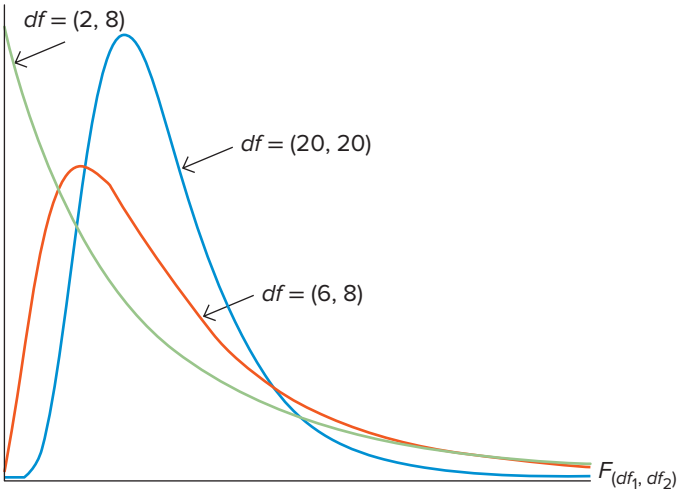
The *F* Distribution

LO 10.3

Discuss features of the *F* distribution.

Like the t_{df} distribution, the F distribution is characterized by a family of distributions; however, each distribution depends on *two* degrees of freedom: the numerator degrees of freedom df_1 and the denominator degrees of freedom df_2. It is common to refer to it as the $F_{(df_1,df_2)}$ distribution. The $F_{(df_1,df_2)}$ distribution is positively skewed with values ranging from zero to infinity but becomes increasingly symmetric as df_1 and df_2 increase.

FIGURE 10.3 The $F_{(df_1,df_2)}$ distribution with various degrees of freedom



From Figure 10.3 we note that all $F_{(df_1,df_2)}$ distributions are positively skewed, where skewness depends on degrees of freedom, df_1 and df_2. As df_1 and df_2 grow larger, the $F_{(df_1,df_2)}$ distribution becomes less skewed and approaches the normal distribution. For instance, $F_{(20,20)}$ is relatively less skewed and more bell-shaped as compared to $F_{(2,8)}$ or $F_{(6,8)}$.

SUMMARY OF THE $F_{(df_1,df_2)}$ DISTRIBUTION

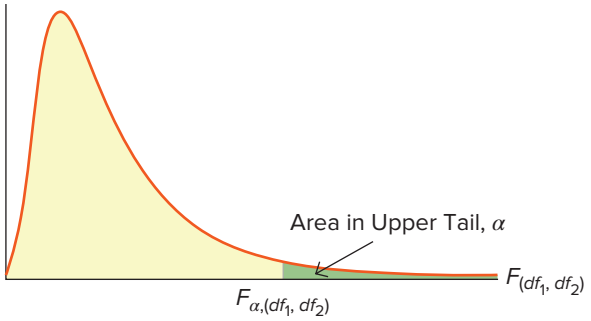
- The $F_{(df_1,df_2)}$ distribution is characterized by a family of distributions, where each distribution depends on two degrees of freedom, df_1 and df_2.
- The values of the $F_{(df_1,df_2)}$ distribution range from zero to infinity.
- The $F_{(df_1,df_2)}$ distribution is positively skewed, where the extent of skewness depends on df_1 and df_2. As df_1 and df_2 grow larger, the $F_{(df_1,df_2)}$ distribution approaches the normal distribution.

Finding $F_{(df_1,df_2)}$ Values and Probabilities

As with the t_{df} distribution, we use the notation $F_{\alpha,(df_1,df_2)}$ to represent a value such that the area in the upper (right) tail of the distribution is α. In other words, $P(F_{(df_1,df_2)} \geq F_{\alpha,(df_1,df_2)}) = \alpha$. Figure 10.4 illustrates this notation.

[1]The F distribution is named in honor of Sir Ronald Fisher, who discovered the distribution in 1922.

FIGURE 10.4 Graphical depiction of $P(F_{(df_1,df_2)} \geq F_{\alpha,(df_1,df_2)}) = \alpha$



A portion of the upper tail areas α and the corresponding $F_{(df_1,df_2)}$ values are given in Table 10.6. (Table 4 of Appendix A provides a more complete table.) Consider the degrees of freedom given by $df_1 = 6$ and $df_2 = 8$. With $df_1 = 6$ (read from the top row) and $df_2 = 8$ (read from the first column), we can easily determine the area in the upper tail as $P(F_{(6,8)} \geq 3.58) = 0.05$ and $P(F_{(6,8)} \geq 6.37) = 0.01$. The F table lists probabilities corresponding to a limited number of values in the upper tail of the distribution. For instance, the exact probability $P(F_{(6,8)} \geq 3.92)$ cannot be determined from the table, and we have to rely on approximate values. All we can say is the area to the right of 3.92 is between 0.025 and 0.05. Shortly, we will use Excel to find exact probabilities.

TABLE 10.6 Portion of the *F* Table

df_2	Area in Upper Tail, α	df_1 = 6	7	8
6	0.10	3.05	3.01	2.98
	0.05	4.28	4.21	4.15
	0.025	5.82	5.70	5.60
	0.01	8.47	8.26	8.10
7	0.10	2.83	2.78	2.75
	0.05	3.87	3.79	3.73
	0.025	5.12	4.99	4.90
	0.01	7.19	6.99	6.84
8	0.10	2.67	2.62	2.59
	0.05	**3.58**	3.50	3.44
	0.025	4.65	4.53	4.43
	0.01	**6.37**	6.18	6.03

One-Way ANOVA Test

LO **10.4**

Conduct and evaluate a one-way ANOVA test.

A **one-way ANOVA test** compares population means based on one categorical variable or factor. In general, it is used for testing c population means under the following assumptions:

1. The populations are normally distributed.
2. The population variances are unknown but assumed equal.
3. The samples are selected independently.

We will discuss a one-way ANOVA test through an example. Sean Cox, a research analyst at an environmental organization, believes that an upswing in the use of public transportation has taken place due to environmental concerns, the volatility of gas

prices, and the general economic climate. He is pleased with a recent study, which highlights the average annual cost savings when commuters use public transportation (*Boston Globe*, May 8, 2009). Sean wants to determine whether there are differences in mean cost savings among cities. He collects a representative sample of public transit riders in the top four cost-savings cities: Boston, New York, San Francisco, and Chicago. Table 10.7 shows each public transit rider's annual cost savings by city.

TABLE 10.7 Annual Cost Savings (in $) from Using Public Transportation

Boston	New York	San Francisco	Chicago
12500	12450	11800	10595
12640	12500	11745	10740
12600	12595	11700	10850
12625	12605	11800	10725
12745	12650	11700	10740
	12620	11575	
	12560		
	12700		

FILE
Public_Transportation

Since he wants to determine whether some differences exist in the mean cost savings of using public transportation by city (the categorical variable), he formulates the following competing hypotheses:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

H_A: Not all population means are equal.

Note that H_A does not require that all means must differ from one another. In principle, the sample data may support the rejection of H_0 in favor of H_A even if only two means differ.

When conducting the equality of means test, you might be tempted to set up a series of hypothesis tests, comparing μ_1 and μ_2, then μ_1 and μ_3, and so on, and then use the two-sample t test with equal variances discussed in Section 10.1. However, such an approach is not only cumbersome, but also flawed. In this example, where we evaluate the equality of four means, we would have to compare six combinations of two means at a time. Also, by conducting numerous pairwise comparisons, we inflate the risk of the Type I error α; that is, we increase the risk of incorrectly rejecting the null hypothesis. In other words, if we conduct all six pairwise tests at the 5% level of significance, the resulting significance level for the overall test will be greater than 5%.

Fortunately, the ANOVA technique avoids this problem by providing one test that simultaneously evaluates the equality of several means. In the public transportation example, if the four population means are equal, we would expect the resulting sample means, $\bar{x}_1, \bar{x}_2, \bar{x}_3$, and $\bar{x}_4$, to be relatively close to one another. Figure 10.5a illustrates the

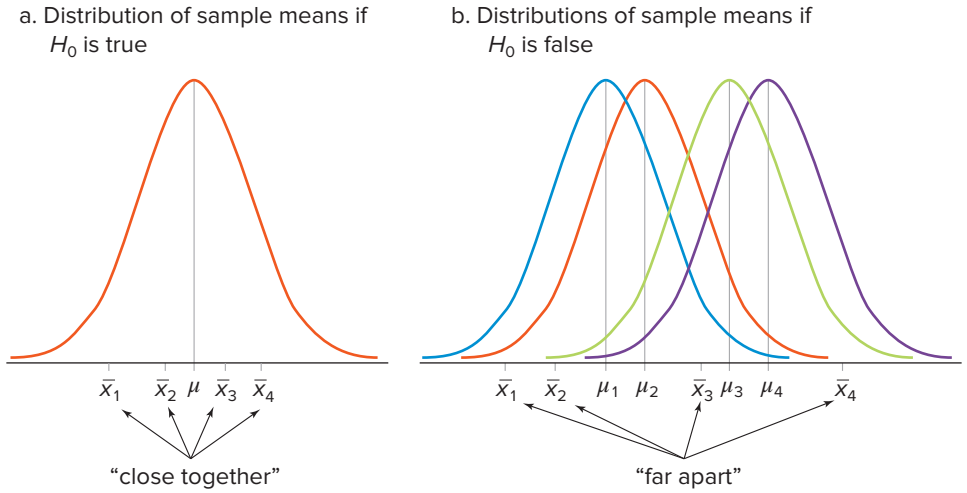


FIGURE 10.5
The logic of ANOVA

distribution of the sample means if H_0 is true. Here, the relatively small variability in the sample means can be explained by chance. What if the population means differ? Figure 10.5b shows the distributions of the sample means if the sample data do not support H_0. In this scenario, the sample means are relatively far apart since each sample mean is calculated from a population with a different mean. The resulting variability in the sample means cannot be explained by chance alone.

The term *treatments* is often used to identify the c populations being examined. The practice of referring to different populations as different treatments is due to the fact that many ANOVA applications were originally developed in connection with agricultural experiments where different fertilizers were regarded as different treatments applied to soil.

In order to determine if significant differences exist between some of the population means, we develop two independent estimates of the common population variance σ^2. One estimate of σ^2 can be attributed to the variability *between* the sample means. It is referred to as **between-treatments variance**. The other estimate of σ^2 can be attributed to the variability of the data *within* each sample; that is, the variability due to chance. It is referred to as **within-treatments variance**.

If the two independent estimates of σ^2 are relatively close together, then it is likely that the variability of the sample means can be explained by chance and the null hypothesis of equal population means is not rejected. However, if the between-treatments variance is significantly greater than the within-treatments variance, then the null hypothesis of equal population means is rejected. This is equivalent to concluding that the ratio of between-treatments variance to within-treatments variance is significantly greater than one. We will come back to this ratio shortly.

Between-Treatments Estimate of σ^2: *MSTR*

The between-treatments variance is based on a weighted sum of squared differences between the sample means and the overall mean of the data set, referred to as the **grand mean** and denoted as $\bar{\bar{x}}$. We compute the grand mean by summing all observations in the data set and dividing by the total number of observations.

Each squared difference of a sample mean from the grand mean $(\bar{x}_i - \bar{\bar{x}})^2$ is multiplied by the respective sample size for each treatment n_i. After summing the weighted squared differences, we arrive at a value called the **sum of squares due to treatments** or *SSTR*. When we divide *SSTR* by its degrees of freedom $c - 1$, we obtain the mean square for treatments; or equivalently, the between-treatments estimate of σ^2, which we denote by *MSTR*.

CALCULATIONS FOR *MSTR*

- The grand mean: $\bar{\bar{x}} = \dfrac{\sum_{i=1}^{c}\sum_{j=1}^{n_i} x_{ij}}{n_T}$,
- The sum of squares due to treatments: $SSTR = \sum_{i=1}^{c} n_i(\bar{x}_i - \bar{\bar{x}})^2$, and
- The between-treatments estimate of σ^2: $MSTR = \dfrac{SSTR}{c-1}$,

where c is the number of populations (treatments), $\bar{x}_i$ and n_i are the sample mean and the sample size of the ith sample, respectively, and n_T is the total sample size.

Referring back to the public transportation example, and Table 10.7, we first find the sample mean $\bar{x}_i$ and the sample size n_i, for each city. For Boston, New York, San Francisco, and Chicago, the sample means are 12,622, 12,585, 11,720, and 10,730, respectively. The

corresponding sample sizes are 5, 8, 6, and 5, respectively. The calculations for $\bar{\bar{x}}$, *SSTR*, and *MSTR* are as follows:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n_j} x_{ij}}{n_T} = \frac{12{,}500 + 12{,}640 + \cdots + 10{,}740}{24} = 11{,}990.$$

$$\begin{aligned} SSTR = \sum_{i=1}^{c} n_i(\bar{x}_i - \bar{\bar{x}})^2 &= 5(12{,}622 - 11{,}990)^2 + 8(12{,}585 - 11{,}990)^2 \\ &\quad + 6(11{,}720 - 11{,}990)^2 + 5(10{,}730 - 11{,}990)^2 \\ &= 13{,}204{,}720. \end{aligned}$$

$$MSTR = \frac{SSTR}{c-1} = \frac{13{,}204{,}720}{4-1} = 4{,}401{,}573.3333.$$

Within-Treatments Estimate of σ^2: *MSE*

We just calculated a value of *MSTR* equal to 4,401,573.3333. Is this value of *MSTR* large enough to indicate that the population means differ? To answer this question, we compare *MSTR* to the variability that we expect due to chance. We first calculate the sum of squares due to error, or equivalently, the **error sum of squares**, denoted as *SSE*. *SSE* provides a measure of the degree of variability that exists even if all population means are the same. We calculate *SSE* as a weighted sum of the sample variances of each treatment. When we divide *SSE* by its degrees of freedom $n_T - c$, we arrive at the **mean square error** or, equivalently, the within-treatments estimate of σ^2, which we denote by *MSE*.

CALCULATIONS FOR *MSE*

- The error sum of squares: $SSE = \sum_{i=1}^{c}(n_i - 1)s_i^2$, and
- The within-treatments estimate of σ^2: $MSE = \frac{SSE}{n_T - c}$,

where c is the number of populations (treatments), s_i^2 and n_i are the sample variance and the sample size of the ith sample, respectively, and n_T is the total sample size.

Here we first calculate the sample standard deviations for Boston, New York, San Francisco, and Chicago as 87.7924, 80.4008, 83.9643, and 90.6228, respectively. The values of *SSE* and *MSE* for the public transportation example are calculated as follows:

$$\begin{aligned} SSE &= \sum_{i=1}^{c}(n_i - 1)s_i^2 \\ &= (5-1)(87.7924)^2 + (8-1)(80.4008)^2 + (6-1)(83.9643)^2 \\ &\quad + (5-1)(90.6228)^2 = 144{,}180. \end{aligned}$$

$$MSE = \frac{SSE}{n_T - c} = \frac{144{,}180}{24-4} = 7{,}209.$$

As mentioned earlier, if the ratio of the between-treatments variance to the within-treatments variance is significantly greater than one, then this finding provides evidence for rejecting the null hypothesis of equal population means. Equivalently, if this ratio is not significantly greater than one, then we are not able to reject the null hypothesis in favor of the alternative hypothesis. We use this ratio to develop the test statistic for a one-way ANOVA test.

TEST STATISTIC FOR A ONE-WAY ANOVA TEST

The value of the test statistic for testing whether differences exist between the population means is computed as

$$F_{(df_1,df_2)} = \frac{MSTR}{MSE},$$

where $df_1 = c - 1$, $df_2 = n_T - c$, and n_T is the total sample size; $MSTR$ is the between-treatments variance and MSE is the within-treatments variance.

The values for $MSTR$ and MSE are based on independent samples drawn from c normally distributed populations with a common variance. ANOVA tests are always implemented as right-tailed tests.

We are now in a position to conduct a four-step hypothesis test at the 5% significance level for the public transportation example.

Step 1. Specify the null and the alternative hypothesis. For completeness, we repeat the competing hypotheses to determine whether average cost savings from using public transportation differ between the four cities:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_A: Not all population means are equal.

Step 2. Specify the significance level. We conduct the hypothesis test at the 5% significance level, so $\alpha = 0.05$.

Step 3. Calculate the value of the test statistic and the *p*-value. Given $MSTR = 4{,}401{,}573.3333$, $MSE = 7{,}209$, $df_1 = c - 1 = 4 - 1 = 3$ and $df_2 = n_T - c = 24 - 4 = 20$, we compute the value of the test statistic as

$$F_{(df_1,df_2)} = F_{(3,20)} = \frac{MSTR}{MSE} = \frac{4{,}401{,}573.3333}{7{,}209} = 610.566.$$

Since the ANOVA test is a right-tailed test, we calculate the p-value as $P(F_{3,20} \geq 610.566)$. We show a portion of the F table from Appendix A in Table 10.8.

TABLE 10.8 Portion of the F table

df_2	Area in Upper Tail	df_1		
		1	2	3
20	0.10	2.97	2.59	2.38
	0.05	4.35	3.49	3.10
	0.025	5.87	4.46	3.86
	0.01	8.10	5.85	4.94

For $df_1 = 3$ and $df_2 = 20$, we see that 610.566 is much greater than 4.94, implying that the p-value is less than 0.01. In order to find the exact p-value, we can use Excel's F.DIST.RT($F_{(df_1,df_2)}$, df_1, df_2) function where $F_{(df_1,df_2)}$ is the value of the test statistic, df_1 is degrees of freedom in the numerator, and df_2 is degrees of freedom in the denominator. In this example, we enter "=F.DIST.RT(610.566, 3, 20)" and Excel returns 7.956E-20 = 0 (approximately).

Step 4. State the conclusion and interpret the results. Since the p-value is less than 0.05, we reject H_0. Therefore, at the 5% significance level, we conclude that average cost savings from using public transportation differ between the four cities.

It is important to note that if we reject the null hypothesis, we can only conclude that not all population means are equal. The one-way ANOVA test does not allow us to infer which individual means differ. Therefore, even though the sample mean is the highest for Boston, we cannot conclude that Boston leads other cities in the amount that commuters

save by taking public transportation. Further analysis of the difference between paired population means is beyond the scope of this text.

The One-Way ANOVA Table

Most software packages summarize the ANOVA calculations in a table. The general format of the ANOVA table is presented in Table 10.9.

TABLE 10.9 General Format of a One-Way ANOVA Table

Source of Variation	*SS*	*df*	*MS*	*F*	*p*-value
Between Groups	*SSTR*	$c - 1$	*MSTR*	$F_{(df_1,df_2)} = \frac{MSTR}{MSE}$	$P\left(F_{(df_1,df_2)} \geq \frac{MSTR}{MSE}\right)$
Within Groups	*SSE*	$n_T - c$	*MSE*		
Total	*SST*	$n_T - 1$			

We should also note that **total sum of squares SST** is equal to the sum of the squared differences of each observation from the grand mean. This is equivalent to summing *SSTR* and *SSE;* that is, $SST = SSTR + SSE$.

Using Excel to Construct a One-Way ANOVA Table

As seen in the public transportation example, the calculations involved for conducting an ANOVA test are quite involved. Fortunately, given raw data, we can follow simple steps to obtain the necessary information outlined in Table 10.9.

EXAMPLE 10.9

FILE
Public_Transportation

Use Excel to obtain the ANOVA table for the public transportation example.

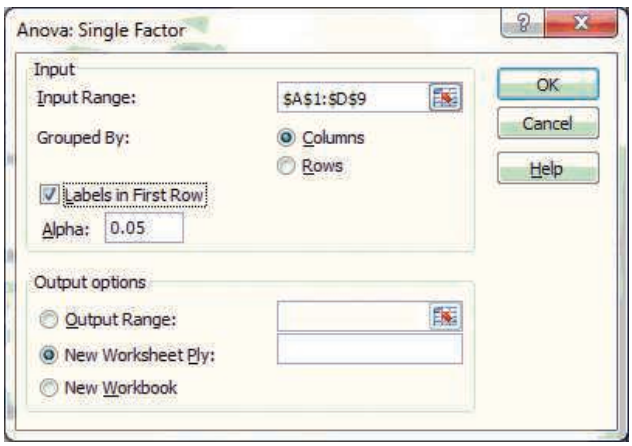
SOLUTION:

a. Open the ***Public_Transportation*** data file.

b. From the menu, choose **Data > Data Analysis > ANOVA: Single Factor**.

c. See Figure 10.6. In the *ANOVA: Single Factor* dialog box, choose the box next to *Input range*, and then select all the data, including the city names. Check the box in front of *Labels in First Row*. Click **OK**.

FIGURE 10.6
Excel's ANOVA: Single Factor dialog box



Source: Microsoft Excel

Table 10.10 shows the Excel-produced ANOVA table for the public transportation example. You should verify that all of your calculations match the values produced by Excel. Excel also shows a statistic called "*F* crit" which would be useful if we

were using the critical-value approach to conduct a hypothesis test. Since we use the *p*-value approach, we can ignore this statistic.

We find that the value of the test statistic, $F_{(3,20)} = 610.566$, and corresponding *p*-value of 0 (approximately) match our calculations (these values are in boldface in Table 10.10). As concluded before, average cost savings from using public transportation are not the same for each city.

TABLE 10.10 Excel-Produced ANOVA Table for Public Transportation Example

SUMMARY					
Groups	**Count**	**Sum**	**Average**	**Variance**	
Boston	5	63110	12622	7707.5	
New York	8	100680	12585	6464.3	
San Francisco	6	70320	11720	7050	
Chicago	5	53650	10730	8212.5	
ANOVA					
Source of Variation	***SS***	***df***	***MS***	***F***	***p*-value**
Between Groups	13204720	3	4401573	**610.566**	**7.96E-20**
Within Groups	144180	20	7209		
Total	13348900	23			

EXERCISES 10.3

Mechanics

41. A random sample of five observations from three normally distributed populations produced the following data:

Treatments		
A	**B**	**C**
22	20	19
25	25	22
27	21	24
24	26	21
22	23	19
$\bar{x}_A = 24$	$\bar{x}_B = 23$	$\bar{x}_C = 21$
$s_A^2 = 4.5$	$s_B^2 = 6.5$	$s_C^2 = 4.5$

a. Calculate the grand mean.
b. Calculate *SSTR* and *MSTR*.
c. Calculate *SSE* and *MSE*.
d. Specify the competing hypotheses in order to determine whether some differences exist between the population means.
e. Calculate the value of the $F_{(df_1,df_2)}$ test statistic and the *p*-value.
f. At the 5% significance level, what is the conclusion to the test?

42. Random sampling from four normally distributed populations produced the following data:

Treatments			
A	**B**	**C**	**D**
−11	−8	−8	−12
−13	−13	−13	−13
−10	−15	−8	−15
	−12	−13	
		−10	

a. Calculate the grand mean.
b. Calculate *SSTR* and *MSTR*.
c. Calculate *SSE* and *MSE*.
d. Specify the competing hypotheses in order to determine whether some differences exist between the population means.
e. Calculate the value of the $F_{(df_1,df_2)}$ test statistic and the *p*-value.
f. At the 10% significance level, what is the conclusion to the test?

43. Given the following information obtained from three normally distributed populations, construct an ANOVA table and perform an ANOVA test of mean differences at the 1% significance level.

$SSTR = 220.7$; $SSE = 2252.2$; $c = 3$; $n_1 = n_2 = n_3 = 8$

44. Given the following information obtained from four normally distributed populations, construct an ANOVA table and perform an ANOVA test of mean differences at the 5% significance level.
$SST = 70.47$; $SSTR = 11.34$; $c = 4$; $n_1 = n_2 = n_3 = n_4 = 15$

45. An analysis of variance experiment produced a portion of the accompanying ANOVA table.

Source of Variation	SS	df	MS	F	p-value
Between Groups	25.08	3	?	?	0.000
Within Groups	92.64	76	?		
Total	117.72	79			

a. Specify the competing hypotheses in order to determine whether some differences exist between the population means.
b. Fill in the missing statistics in the ANOVA table.
c. At the 5% significance level, what is the conclusion to the test?

46. An analysis of variance experiment produced a portion of the following ANOVA table.

Source of Variation	SS	df	MS	F	p-value
Between Groups		5	?	?	?
Within Groups	4321.11	54	?		
Total	4869.48	59			

a. Specify the competing hypotheses in order to determine whether some differences exist between the population means.
b. Fill in the missing statistics in the ANOVA table.
c. At the 10% significance level, what is the conclusion to the test?

Applications

47. Asian residents in Boston have the highest average life expectancy of any racial or ethnic group—a decade longer than black residents (*The Boston Globe*, August 16, 2010). Suppose sample results indicative of the overall results are as follows.

Asian	Black	Latino	White
$\bar{x}_1 = 83.7$ years	$\bar{x}_2 = 73.5$ years	$\bar{x}_3 = 80.6$ years	$\bar{x}_4 = 79.0$ years
$s_1^2 = 26.3$	$s_2^2 = 27.5$	$s_3^2 = 28.2$	$s_4^2 = 24.8$
$n_1 = 20$	$n_2 = 20$	$n_3 = 20$	$n_4 = 20$

a. Specify the competing hypotheses to test whether there are some differences in average life expectancies between the four ethnic groups.
b. Construct an ANOVA table. Assume life expectancies are normally distributed.
c. At the 5% significance level, what is the conclusion to the test?

48. FILE ***Detergent.*** A well-known conglomerate claims that its detergent "whitens and brightens better than all the rest." In order to compare the cleansing action of the top three brands of detergents, 24 swatches of white cloth were soiled with red wine and grass stains and then washed in front-loading machines with the respective detergents. The following whiteness readings were obtained:

Detergent		
1	2	3
84	78	87
79	74	80
87	81	91
85	86	77
94	86	78
89	89	79
89	69	77
83	79	78

a. Specify the competing hypotheses to test whether there are some differences in the average whitening effectiveness of the three detergents.
b. At the 5% significance level, what is the conclusion to the test? Assume whiteness readings are normally distributed.

49. A survey by Genworth Financial Inc., a financial-services company, concludes that the cost of long-term care in the United States varies significantly, depending on where an individual lives (*The Wall Street Journal*, May 16, 2009). An economist collects data from the five states with the highest annual costs (Alaska, Massachusetts, New Jersey, Rhode Island, and Connecticut), in order to determine if his sample data are consistent with the survey's conclusions. The economist provides the following portion of an ANOVA table:

Source of Variation	SS	df	MS	F	p-value
Between Groups	635.0542	4	?	?	?
Within Groups	253.2192	20	?		
Total	888.2734	24			

a. Specify the competing hypotheses to test whether some differences exist in the mean long-term care costs in these five states.
b. Complete the ANOVA table. Assume that long-term care costs are normally distributed.
c. At the 5% significance level, do mean costs differ?

50. FILE ***Sports.*** An online survey by the Sporting Goods Manufacturers Association, a trade group of sports retailers and marketers, claimed that household income of recreational athletes varies by sport (*The Wall Street Journal*, August 10, 2009). In order to verify this claim, an economist samples five sports enthusiasts participating in each of four different recreational sports and obtains each enthusiast's income (in $1,000s), as shown in the accompanying table.

Snorkeling	Sailing	Boardsailing/ Windsurfing	Bowling
90.9	87.6	75.9	79.3
86.0	95.0	75.6	75.8
93.6	94.6	83.1	79.6
98.8	87.2	74.4	78.5
98.4	82.5	80.5	73.2

a. Specify the competing hypotheses in order to test the association's claim.
b. Do some average incomes differ depending on the recreational sport? Explain. Assume incomes are normally distributed.

51. The following output summarizes the results of an analysis of variance experiment in which the treatments were three different hybrid cars and the variable measured was the miles per gallon (mpg) obtained while driving the same route. Assume mpg is normally distributed.

Source of Variation	*SS*	*df*	*MS*	*F*	*p*-value
Between Groups	1034.51	2	517.26	19.86	4.49E-07
Within Groups	1302.41	50	26.05		
Total	2336.92	52			

At the 5% significance level, can we conclude that average mpg differs between the hybrids? Explain.

52. Do energy costs vary dramatically depending on where you live in the United States? Annual energy costs are collected from 25 households in four regions in the United States. A portion of the ANOVA table is shown.

Source of Variation	*SS*	*df*	*MS*	*F*	*p*-value
Between Groups	7531769	3	?	?	7.13E-24
Within Groups	3492385	96	?		
Total	11024154	99			

a. Complete the ANOVA table. Assume energy costs are normally distributed.
b. At the 1% significance level, can we conclude that average annual energy costs vary by region?

53. FILE ***Buggies.*** Wenton Powersports produces dune buggies. They have three assembly lines, "Razor," "Blazer," and "Tracer," named after the particular dune buggy models produced on those lines. Each assembly line was originally designed using the same target production rate. However, over the years, various changes have been made to the lines. Accordingly, management wishes to determine whether the assembly lines are still operating at the same average hourly production rate. Production data (in dune buggies/hour) for the last eight hours are as follows.

Razor	Blazer	Tracer
11	10	9
10	8	9
8	11	10
10	9	9
9	11	8
9	10	7
13	11	8
11	8	9

a. Specify the competing hypotheses to test whether there are some differences in the mean production rates across the three assembly lines.
b. At the 5% significance level, what is the conclusion to the test? What about the 10% significance level? Assume production rates are normally distributed.

54. FILE ***Fill_Volumes.*** In the carbonated beverage industry, dispensing pressure can be an important factor in achieving accurate fill volumes. Too little pressure can slow down the dispensing process. Too much pressure can create excess "fizz" and, thus, inaccurate fill volumes. Accordingly, a leading beverage manufacturer wants to conduct an experiment at three different pressure settings to determine if differences

exist in the mean fill volumes. Forty bottles with a target fill volume of 12 ounces were filled at each pressure setting, and the resulting fill volumes (in ounces) were recorded. A portion of the data is shown in the accompanying table.

Low Pressure (60 psi)	Medium Pressure (80 psi)	High Pressure (100 psi)
12.00	12.00	11.56
11.97	11.87	11.55
⋮	⋮	⋮
12.00	12.14	11.80

a. Specify the competing hypotheses to test whether there are differences in the mean fill volumes across the three pressure settings.

b. At the 5% significance level, what is the conclusion to the test? What about the 1% significance level?

55. FILE ***Exam_Scores.*** A statistics instructor wonders whether significant differences exist in her students' average exam scores in her three different sections. She randomly selects the scores from 10 students in each section. A portion of the data is shown in the accompanying table. Assume exam scores are normally distributed.

Section 1	Section 2	Section 3
85	91	74
68	84	69
⋮	⋮	⋮
74	75	73

Do these data provide enough evidence at the 5% significance level to indicate that there are some differences in average exam scores among these three sections?

56. FILE ***Patronage.*** The accompanying table shows a portion of the number of customers that frequent a restaurant on weekend days over the past 52 weeks.

Fridays	Saturdays	Sundays
391	450	389
362	456	343
⋮	⋮	⋮
443	441	376

At the 5% significance level, can we conclude that the average number of customers that frequent the restaurant differs by weekend day?

57. FILE ***Nike_Revenues.*** The accompanying table shows a portion of quarterly data on Nike's revenue (in $ millions) for the fiscal years 2001 through 2010. Data for Nike's fiscal year refer to the time period from June 1 through May 31. Assume revenue is normally distributed.

Year	Quarters Ended			
	August 31	November 30	February 28	May 31
2001	2637	2199	2170	2483
2002	2614	2337	2260	2682
⋮	⋮	⋮	⋮	⋮
2010	4799	4406	4733	5077

Source: Annual Reports for Nike, Inc.

Use a one-way ANOVA test to determine if the data provide enough evidence at the 5% significance level to indicate that there are quarterly differences in Nike's average revenue.

58. FILE ***Field_Score.*** A human resource specialist wants to determine whether the average job satisfaction score (on a scale of 0 to 100) differs depending on a person's field of employment. She collects scores from 30 employees in three different fields. A portion of the data is shown in the accompanying table.

Field 1	Field 2	Field 3
80	76	81
76	73	77
⋮	⋮	⋮
79	67	80

At the 10% significance level, can we conclude that the average job satisfaction differs by field?

WRITING WITH STATISTICS



©Aaron Kohr/Getty Images

The Texas Transportation Institute, one of the finest higher-education-affiliated transportation research agencies in the nation, recently published its highly anticipated *2009 Annual Urban Mobility Report* (July 8, 2009). The study finds that the average U.S. driver languished in rush-hour traffic for 36.1 hours, as compared to 12 hours in 1982 when the records begin. This congestion also wasted approximately 2.81 billion gallons in fuel, or roughly three weeks' worth of gas per traveler. John Farnham, a research analyst at an environmental firm, is stunned by some of the report's

conclusions. John is asked to conduct an independent study in order to see if differences exist in congestion depending on the city where the traveler drives. He selects 25 travelers from each of the five cities that suffered from the worst congestion. He asks each traveler to approximate the time spent in traffic (in hours) over the last calendar year. Table 10.11 shows a portion of his sample results.

FILE
Congestion

TABLE 10.11 Annual Hours of Delay per Traveler in Five Cities

Los Angeles	Washington, DC	Atlanta	Houston	San Francisco
71	64	60	58	57
60	64	58	56	56
⋮	⋮	⋮	⋮	⋮
68	57	57	59	56

John wants to use the sample information to:

1. Determine whether significant differences exist in congestion, depending on the city where the traveler drives. Assume delay times are normally distributed.
2. Interpret the results.

Sample Report—Evaluating Traffic Congestion by City

Does traffic congestion vary by city? *The 2009 Annual Urban Mobility Report* found that traffic congestion, measured by annual hours of delay per traveler, was the worst in Los Angeles; followed by Washington, DC; Atlanta; Houston; and then San Francisco. An independent survey was conducted to verify some of the findings. Twenty-five travelers in each of these cities were asked how many hours they wasted in traffic over the past calendar year. Table 10.A reports the summary statistics. The sample data indicate that Los Angeles residents waste the most time sitting in traffic with an average of 69.2 hours per year. Washington, DC, residents rank a close second, spending an average of 62 hours per year in traffic. Residents in Atlanta, Houston, and San Francisco spend on average 57.0, 56.5, and 55.6 hours per year in traffic.

TABLE 10.A Summary Statistics

Los Angeles	Washington, DC	Atlanta	Houston	San Francisco
$\bar{x}_1 = 69.24$	$\bar{x}_2 = 61.96$	$\bar{x}_3 = 57.00$	$\bar{x}_4 = 56.52$	$\bar{x}_5 = 55.56$
$s_1 = 4.60$	$s_2 = 4.74$	$s_3 = 4.81$	$s_4 = 5.37$	$s_5 = 3.66$
$n_1 = 25$	$n_2 = 25$	$n_3 = 25$	$n_4 = 25$	$n_5 = 25$

A one-way ANOVA test is conducted to determine if significant differences exist in the average number of hours spent in traffic in these five worst-congested cities. The value of the test statistic is $F_{(4,120)} = 37.25$ with a p-value of approximately zero. Therefore, at the 5% level of significance, we reject the null hypothesis of equal means and conclude that traffic congestion does vary by city.

Although the above test allows us to conclude that congestion varies by city, it does not allow us to compare congestion between any two cities. For instance, we cannot conclude that Los Angeles has significantly higher average delays per traveler than San Francisco. Further analysis of the difference between paired population means is advised.

CONCEPTUAL REVIEW

LO 10.1 **Make inferences about the difference between two population means based on independent sampling.**

Independent samples are samples that are completely unrelated to one another. A $100(1 - \alpha)\%$ confidence interval for the difference between two population means $\mu_1 - \mu_2$, based on independent samples, is

- $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, if σ_1^2 and σ_2^2 are known.
- $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,df}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, if σ_1^2 and σ_2^2 are unknown but assumed equal.
  The pooled sample variance is $s_p^2 = \frac{(n_1 - 1)\,s_1^2 + (n_2 - 1)\,s_2^2}{n_1 + n_2 - 2}$, and $df = n_1 + n_2 - 2$.
- $(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,df}\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}$, if σ_1^2 and σ_2^2 are unknown and cannot be assumed equal; also, $df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$.

When conducting hypothesis tests about the difference between two means $\mu_1 - \mu_2$, based on independent samples, the value of the test statistic is

- $z = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$, if σ_1^2 and σ_2^2 are known.
- $t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, if σ_1^2 and σ_2^2 are unknown but assumed equal.
- $t_{df} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$, if σ_1^2 and σ_2^2 are unknown and cannot be assumed equal.

Here, d_0 is the hypothesized difference between μ_1 and μ_2 and the degrees of freedom for the last two tests are the same as the ones defined for the corresponding confidence intervals. The formulas for estimation and testing are valid only if $\bar{X}_1 - \bar{X}_2$ (approximately) follows a normal distribution.

LO 10.2 Make inferences about the mean difference based on matched-pairs sampling.

A common case of dependent sampling, commonly referred to as **matched-pairs sampling**, is when the samples are paired or matched in some way.

For matched-pairs sampling, the population parameter of interest is referred to as the mean difference μ_D where $D = X_1 - X_2$, and the random variables X_1 and X_2 are matched in a pair. A $100(1 - \alpha)\%$ confidence interval for the mean difference μ_D, based on a matched-pairs sample, is given by $\bar{d} \pm t_{\alpha/2,df}s_D/\sqrt{n}$, where $\bar{d}$ and s_D are the mean and the standard deviation, respectively, of the n sample differences, D, and $df = n - 1$. When conducting a hypothesis test about μ_D the value of the test statistic is calculated as $t_{df} = \frac{\bar{d} - d_0}{s_D/\sqrt{n}}$, where d_0 is a hypothesized mean difference and $df = n - 1$.

LO 10.3 Discuss features of the *F* distribution.

The ***F* distribution** is characterized by a family of distributions, where each distribution depends on two degrees of freedom: the numerator degrees of freedom df_1 and the denominator degrees of freedom df_2. It is common to refer to it as the $F_{(df_1,df_2)}$ distribution. The distribution is positively skewed with values ranging from zero to infinity, but it becomes increasingly symmetric as df_1 and df_2 increase.

LO 10.4 Conduct and evaluate a one-way ANOVA test.

A **one-way analysis of variance (ANOVA) test** is used to determine if differences exist between three or more population means. This test examines the amount of variability *between* the samples relative to the amount of variability *within* the samples.

The value of the test statistic for testing for differences between the c population means is calculated as $F_{(df_1,df_2)} = MSTR/MSE$, where $MSTR$ is the mean square for treatments, MSE is the mean square error, $df_1 = c - 1$, $df_2 = n_T - c$, and n_T is the total sample size. The values for $MSTR$ and MSE are based on independent samples drawn from c normally distributed populations with a common variance. An ANOVA test is always specified as a right-tailed test.

Exercises

59. A study has found that, on average, 6- to 12-year-old children are spending less time on household chores today compared to 1981 levels (*The Wall Street Journal*, August 27, 2008). Suppose two samples representative of the study's results report the following summary statistics for the two periods:

1981 Levels	2008 Levels
$\bar{x}_1 = 30$ minutes	$\bar{x}_2 = 24$ minutes
$s_1 = 4.2$ minutes	$s_1 = 3.9$ minutes
$n_1 = 30$	$n_2 = 30$

a. Specify the competing hypotheses to test the study's claim that children today spend less time on household chores as compared to children in 1981.
b. Calculate the value of the test statistic assuming that the unknown population variances are equal.
c. Find the *p*-value.
d. At the 5% significance level, do the data support the study's claim? Explain.

60. Do men really spend more money on St. Patrick's Day as compared to women? A survey found that men spend an average of $43.87 while women spend an average of $29.54 (*USA Today*, March 17, 2009). Assume that these data were based on a sample of 100 men and 100 women and the population standard deviations of spending for men and women are $32 and $25, respectively.
a. Specify the competing hypotheses to determine whether men spend more money on St. Patrick's Day as compared to women.
b. Calculate the value of the test statistic.
c. Find the *p*-value.
d. At the 1% significance level, do men spend more money on St. Patrick's Day as compared to women? Explain.

61. FILE ***Balanced_European.*** The accompanying table shows annual return data from 2001–2009 for Vanguard's Balanced Index and European Stock Index mutual funds.

Year	Balanced	European Stock
2001	–3.02	–20.30
2002	–9.52	–17.95
2003	19.87	38.70
2004	9.33	20.86
2005	4.65	9.26
2006	11.02	33.42
2007	6.16	13.82
2008	–22.21	–44.73
2009	20.05	31.91

Source: www.finance.yahoo.com

a. Set up the hypotheses to test whether the mean returns of the two funds differ. (*Hint:* This is a matched-pairs comparison.)
b. Calculate the value of the test statistic and the *p*-value. Assume that the return difference is normally distributed.
c. At the 5% significance level, what is the conclusion?

62. FILE ***Cholesterol_Levels.*** It is well documented that cholesterol over 200 is a risk factor in developing heart disease for both men and women (www.livestrong.com, January 11, 2011). Younger men are known to have higher cholesterol levels than younger women; however, beyond age 55, women are more likely to have higher cholesterol levels. A recent college graduate working at a local blood lab has access to the cholesterol data of 50 men and 50 women in the 20–40 age group. The accompanying table shows a portion of the data.

Men	Women
181	178
199	193
⋮	⋮
190	182

At the 1% significance level, determine if there are any differences in the mean cholesterol levels for men and women in the age group. It is fair to assume that the population variances for men and women are equal.

63. A farmer is concerned that a change in fertilizer to an organic variant might change his crop yield. He subdivides 6 lots and uses the old fertilizer on one half of each lot and the new fertilizer on the other half. The following table shows the results.

Lot	Crop Yield Using Old Fertilizer	Crop Yield Using New Fertilizer
1	10	12
2	11	10
3	10	13
4	9	9
5	12	11
6	11	12

a. Specify the competing hypotheses that determine whether there is any difference between the average crop yields from the use of the different fertilizers.
b. Assuming that differences in crop yields are normally distributed, calculate the value of the test statistic.
c. Find the *p*-value.
d. Is there sufficient evidence to conclude that the crop yields are different? Should the farmer be concerned?

64. FILE ***Pregnancy_Weight.*** It is important for women to gain the right amount of weight during pregnancy by eating a healthy, balanced diet. It is recommended that a woman of average weight before pregnancy should gain 25 to 35 pounds during pregnancy (www.babycenter.com, August 2016). The accompanying table shows a portion of the weight data for 40 women before and after pregnancy.

Weight before Pregnancy	Weight after Pregnancy
114	168
107	161
⋮	⋮
136	157

a. At the 5% level of significance, determine if the mean weight gain of women due to pregnancy is more than 30 pounds.
b. At the 5% level of significance, determine if the mean weight gain of women due to pregnancy is more than 35 pounds.

65. FILE ***SAT_Writing.*** The SAT is required of most students applying for college admission in the United States. This standardized test has gone through many revisions over the years. In 2005, a new writing section was introduced that includes a direct writing measure in the form of an essay. People argue that female students generally do worse on math tests but better on writing tests. Therefore, the new section may help reduce the usual male lead on the overall average SAT score (*The Washington Post*, August 30, 2006). Consider the following scores on the writing component of the test for eight male and eight female students.

Males	620	570	540	580	590	580	480	620
Females	660	590	540	560	610	590	610	650

a. Construct the null and the alternative hypotheses to test if females outscore males on writing tests.
b. Assuming that the difference in scores is normally distributed, calculate the value of the test statistic and the *p*-value. Do not assume that the population variances are equal.
c. Implement the test at $\alpha = 0.01$ and interpret your results.

66. FILE ***Safety_Program.*** An engineer wants to determine the effectiveness of a safety program. He collects annual loss of hours due to accidents in 12 plants before and after the program was put into operation.

Plant	Before	After	Plant	Before	After
1	100	98	7	88	90
2	90	88	8	75	70
3	94	90	9	65	62
4	85	86	10	58	60
5	70	67	11	67	60
6	83	80	12	104	98

a. Specify the competing hypotheses that determine whether the safety program was effective.
b. Calculate the value of the test statistic and the *p*-value. Assume that the hours difference is normally distributed.
c. At the 5% significance level, is there sufficient evidence to conclude that the safety program was effective? Explain.

67. FILE ***Battery_Times.*** Electrobat, a battery manufacturer, is investigating how storage temperature affects the performance of one of

its popular deep-cell battery models used in recreational vehicles. Samples of 30 fully charged batteries were subjected to a light load under each of four different storage temperature levels. The hours until deep discharge (meaning ≤ 20% of charge remaining) were measured. A portion of the data is shown in the accompanying table.

0 degrees F	30 degrees F	60 degrees F	90 degrees F
3	6	12	12
5	8	13	15
⋮	⋮	⋮	⋮
4	9	9	15

At the 5% significance level, can you conclude that mean discharge times differ across the four storage temperature levels? What about the 1% significance level?

68. FILE ***Transportation.*** A government agency wants to determine whether the average salaries of four kinds of transportation operators differ. The accompanying table shows the salaries (in $1,000s) for a random sample of five employees in each of the four categories. Assume that salaries are normally distributed.

Engineer	Truck Driver	Bus Driver	Limousine Driver
54.7	40.5	32.4	26.8
53.2	42.7	31.2	27.1
55.1	41.6	30.9	28.3
54.3	40.9	31.8	27.9
51.5	39.2	29.8	29.9

a. Specify the competing hypotheses in order to determine whether the average salaries of the transportation operators differ.
b. At the 5% significance level, what is the conclusion to the test?

69. FILE ***Foodco.*** The Marketing Manager at Foodco, a large grocery store, wants to determine if store display location influences sales of a particular grocery item. He instructs employees to rotate the display location of that item every week and then tallies the weekly sales at each location over a 24-week period (8 weeks per location). The following sales (in $) were obtained. Assume that sales are normally distributed.

Front of store	Center of store	Side aisle
947	858	1096
1106	780	1047
1143	786	910
1162	816	823
967	800	919
956	770	924
1057	876	1091
996	802	1027

a. Specify the competing hypotheses to test whether there are some differences in the mean weekly sales across the three store display locations.
b. At the 5% significance level, what is the conclusion to the test?

70. FILE ***SAT_Ethnicity.*** The manager of an SAT review program wonders whether average SAT scores differ depending on the ethnicity of the test taker. Thirty test scores for four ethnicities are collected. A portion of the data is shown in the accompanying table.

White	Black	Asian-American	Mexican-American
1587	1300	1660	1366
1562	1255	1576	1531
⋮	⋮	⋮	⋮
1500	1284	1584	1358

At the 5% significance level, can we conclude that the average SAT scores differ by ethnicity?

71. FILE ***Concrete_Mixing.*** Compressive strength of concrete is affected by several factors, including composition (sand, cement, etc.), mixer type (batch vs. continuous), and curing procedure. Accordingly, a concrete company is conducting an experiment to determine how mixing technique affects the resulting compressive strength. Four potential mixing techniques have been identified. Subsequently, samples of 20 specimens have been subjected to each mixing technique, and the resulting compressive strengths (in pounds per square inch, psi) were measured. A portion of the data is shown in the accompanying table. Assume that compressive strengths are normally distributed.

Technique 1	Technique 2	Technique 3	Technique 4
2972	2794	2732	2977
2818	3162	2905	2986
⋮	⋮	⋮	⋮
2665	2837	3073	3081

a. Specify the competing hypotheses to test whether there are some differences in the mean compressive strengths across the four mixing techniques.
b. At the 5% significance level, what is the conclusion to the test? What about the 1% significance level?

72. FILE ***Plywood.*** An engineer wants to determine whether the average strength of plywood boards (in pounds per square inch, psi) differs depending on the type of glue used. For three types of glue, she measures the strength of 20 plywood boards. A portion of the data is shown in the accompanying table.

Glue 1	Glue 2	Glue 3
38	41	42
34	38	38
⋮	⋮	⋮
38	49	50

At the 5% significance level, can she conclude that the average strength of the plywood boards differs by the type of glue used? Assume that the strength of plywood boards is normally distributed.

73. FILE ***Route.*** An employee of a small software company in Minneapolis bikes to work during the summer months. He can travel to work using one of three routes and wonders whether the average commute times (in minutes) differ between the three routes. He obtains the following data after traveling each route for one week.

Route 1	29	30	33	30	32
Route 2	27	32	28	30	29
Route 3	25	27	24	29	26

Determine at the 1% significance level whether the average commute times differ between the three routes. Assume that commute times are normally distributed.

74. FILE ***PEratios.*** An economist wants to determine whether average price/earnings (P/E) ratios differ for firms in three industries. Independent samples of five firms in each industry show the following results:

Industry A	12.19	12.44	7.28	9.96	10.51
Industry B	14.34	17.80	9.32	14.90	9.41
Industry C	26.38	24.75	16.88	16.87	16.70

At the 5% significance level, determine whether average P/E ratios differ in the three industries. Assume that P/E ratios are normally distributed.

75. Before the Great Recession, job-creating cities in the Sunbelt, like Las Vegas, Phoenix, and Orlando saw their populations, income levels, and housing prices surge. Las Vegas, however, offered something that often eluded these other cities: upward mobility for the working class. For example, hard-working hotel maids were able to prosper during the boom times. According to the Bureau of Labor Statistics, the average hourly rate for hotel maids was $14.25 in Las Vegas, versus $9.25 in Phoenix and $8.84 in Orlando (*The Wall Street Journal*, July 20, 2009). Suppose the following ANOVA table was produced from a sample of hourly wages of 25 hotel maids in each city.

Source of Variation	*SS*	*df*	*MS*	*F*	*p*-value
Between Groups	430.87	2	215.44	202.90	2.58E-30
Within Groups	76.44	72	1.06		
Total	507.31	74			

At the 5% significance level, do mean hourly rates for hotel maids differ between the three cities? Assume that hourly wages are normally distributed.

76. The marketing department for an upscale retail catalog company wants to determine if there are differences in the mean customer purchase amounts across the available purchase sources (Internet, phone, or mail-in). Accordingly, samples were taken for 20 random orders for each purchase source. The following ANOVA table was produced. Assume purchase amounts are normally distributed.

Source of Variation	*SS*	*df*	*MS*	*F*	*p*-value
Between Groups	12715.23	2	6357.62	0.433	0.651
Within Groups	836704.70	57	14679.03		
Total	849419.93	59			

At the 5% significance level, can we conclude that the mean purchase amount is different across the three purchase sources?

77. An accounting professor wants to know if students perform the same on the departmental final exam irrespective of the accounting section they attend. She randomly selects the exam scores of 20 students from three sections. A portion of the output from conducting a one-way ANOVA test is shown in the accompanying table. Assume exam scores are normally distributed.

Source of Variation	*SS*	*df*	*MS*	*F*	*p*-value
Between Groups	57.39	2	*MSTR* = ?	$F_{2,57}$ = ?	0.346
Within Groups	*SSE* = ?	57	*MSE* = ?		
Total	1570.19	59			

a. Find the missing values in the ANOVA table.
b. At the 5% significance level, can you conclude that average grades differ in the accounting sections?

78. FILE ***Website.*** A data analyst for an online store wonders whether average customer visits to the store's website vary by day of the week. He collects daily unique visits to the website for a 12-week period; a portion of the data is shown in the accompanying table. At the 5% significance level, what conclusion can the data analyst make? Assume that customer visits are normally distributed.

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
3088	2203	2331	1977	2036	2272	3065
2953	2288	2450	1912	2097	2205	2922
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3047	2233	2395	1913	2175	2279	2994

CASE STUDIES

CASE STUDY 10.1 Chad Perrone is a financial analyst in Boston studying the annual return data for the health and information technology industries. He randomly samples 20 firms in each industry and notes each firm's annual return. A portion of the data is shown in the accompanying table.

Data for Case Study 10.1 Annual Returns (in percent) for Firms in Health and Information Technology Industries

FILE
Health_Info

Health	Information Technology
10.29	4.77
32.17	1.14
⋮	⋮
13.21	22.61

In a report, use the sample information to

1. Provide descriptive statistics and comment on the reward and risk in each industry.
2. Determine whether the average returns in each industry differ at the 5% significance level. Assume that annual returns are normally distributed and that the population variances are not equal.

CASE STUDY 10.2 The Speedo LZR Racer Suit is a high-end, body-length swimsuit that was launched on February 13, 2008. When 17 world records fell at the December 2008 European Short Course Championships in Croatia, many believed a modification in the rules surrounding swimsuits was necessary. The FINA Congress, the international governing board for swimming, banned the LZR Racer and all other body-length swimsuits from competition, effective January 2010. In a statement to the public, FINA defended its position with the following statement: "FINA wishes to recall the main and core principle that swimming is a sport essentially based on the physical performance of the athlete" (*BBC Sport*, March 14, 2009).

Luke Johnson, a freelance journalist, wonders if the decision made by FINA has statistical backing. He conducts an experiment with the local university's Division I swim team. He times 10 of the swimmers swimming the 50-meter breaststroke in his/her bathing suit and then retests them while wearing the LZR Racer. A portion of the results is shown in the accompanying table.

Data for Case Study 10.2 50-Meter Breaststroke Times (in seconds)

FILE
LZR_Racer

Swimmer	Time in Bathing Suit	Time in LZR Racer
1	27.64	27.45
2	27.97	28.06
⋮	⋮	⋮
10	38.08	37.93

In a report, use the sample information to

1. Determine whether the LZR Racer improves swimmers' times at the 5% significance level. Assume that the time difference is normally distributed.
2. Comment on whether the data appear to support FINA's decision.

CASE STUDY 10.3 Lisa Grattan, a financial analyst for a small investment firm, collects annual stock return data for 10 firms in the energy industry, 13 firms in the retail industry, and 16 firms in the utilities industry. A portion of the data is shown in the accompanying table.

Data for Case Study 10.3 Annual Stock Returns (in %)

Energy	Retail	Utilities
12.5	6.6	3.5
8.2	7.4	6.4
⋮	⋮	⋮
6.9	7.9	4.3

FILE *Industry_Returns*

In a report, use the sample information to

1. Provide descriptive statistics and comment on the reward and risk in each industry.
2. Determine whether significant differences exist in the annual returns for the three industries at the 5% significance level. Assume that annual returns are normally distributed.

APPENDIX 10.1 Guidelines for Other Software Packages

The following section provides brief commands for specific software packages: Minitab, SPSS, JMP, and R. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands. When importing data into R, use the menu-driven option: File > Import Dataset > From Excel.

Minitab

Testing $\mu_1 - \mu_2$

A. (Replicating Example 10.4) From the menu, choose **Stat > Basic Statistics > 2-Sample t**. Choose **Each sample is in its own column**, and after **Sample 1**, select Gold and after **Sample 2**, select Oil.

FILE *Gold_Oil*

B. Choose **Options**. After **Alternative hypothesis**, select "Difference ≠ hypothesized difference."

Testing μ_D

A. (Replicating Example 10.8) From the menu, choose **Stat > Basic Statistics > Paired t. Each sample is in its own column**, and after **Sample 1**, select Before and after **Sample 2**, select After.

FILE *Food_Calories*

B. Choose **Options**. After **Alternative hypothesis**, select "Difference > hypothesized difference."

One-Way ANOVA

(Replicating Example 10.9) From the menu, choose **Stat > ANOVA > One-Way**.

FILE *Public_Transportation*

SPSS

Testing $\mu_1 - \mu_2$

FILE *Gold_Oil*

A. (Replicating Example 10.4) Pool all ***Gold_Oil*** data in one column and label Pooled. In adjacent column (labeled Group), denote all Gold values with 0 and all Oil values with 1.

B. From the menu, choose **Analyze > Compare Means > Independent-Samples T-Test**.

C. Select Pooled as **Test Variable(s)** and Group as **Grouping Variable**. Select **Define Groups**, and enter 0 for **Group 1** and 1 for **Group 2**.

Testing μ_D

FILE *Food_Calories*

A. (Replicating Example 10.8) From the menu, choose **Analyze > Compare Means > Paired-Samples T-Test**.

B. Select Before as **Variable1** and After as **Variable2**.

One-Way ANOVA

FILE *Public_Transportation*

A. (Replicating Example 10.9) Stack all cost values in one column and label Cost. In adjacent column (label City), denote all Boston costs with value 1, all New York costs with value 2, etc.

B. From the menu, choose **Analyze > Compare Means > One-Way ANOVA**.

C. Under **Dependent**, **List**, select Cost, and under **Factor**, select City.

JMP

Testing $\mu_1 - \mu_2$

FILE *Gold_Oil*

A. (Replicating Example 10.4) Pool all ***Gold_Oil*** data in one column and label it Pooled. In adjacent column (labeled Group and read as nominal data), denote all Gold values with 0 and all Oil values with 1.

B. From the menu, choose **Analyze > Fit Y by X**.

C. Select Pooled as **Y**, **Response** and Group as **X**, **Factor**.

D. Click on the red triangle next to the header that reads **Oneway Analysis of Column 1 by Column 2** and select **t-test** (to use a pooled variance, select **Means/Anova/ Pooled t**).

Testing μ_D

FILE *Food_Calories*

A. (Replicating Example 10.8) From the menu, choose **Analyze > Matched Pairs**.

B. Choose Before and After as **Y**, **Paired Response**.

One-Way ANOVA

FILE *Public_Transportation*

A. (Replicating Example 10.9) In order to arrange the data, follow the SPSS instructions for One-Way ANOVA, step A.

B. From the menu, select **Analyze > Fit Y by X**.

C. Under **Select Columns**, select Pooled, then under **Cast Selected Columns into Roles**, select **Y**, **Columns**. Under **Select Columns**, select Group, then under **Cast Selected Columns into Roles**, select **X**, **Factor**.

D. Click on the red triangle next to **Oneway Analysis of Pooled by Group** and select **Means/Anova**.

R

Testing $\mu_1 - \mu_2$

(Replicating Example 10.4) Use the **t.test** function. For options within the **t.test** function use *alternative* to denote the specification of the alternative hypothesis (denoted as "two.sided" for a two-tailed test, "less" for a left-tailed test, and "greater" for a right-tailed test), *mu* to denote the value of the hypothesized difference, *paired* to indicate if it is a matched-pairs sample, *var.equal* to indicate if the variances are assumed equal, and *conf.level* to specify the confidence level. Enter:

FILE
Gold_Oil

```
> t.test(Gold_Oil$'Gold', Gold_Oil$'Oil', alternative="two.sided",
  mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

Testing μ_D

(Replicating Example 10.8) Use the **t.test** function. Refer to the instructions for testing $\mu_1 - \mu_2$ for specifics about the **t.test** function. Enter:

FILE
Food_Calories

```
> t.test(Food_Calories$'Before', Food_Calories$'After',
  alternative="greater", mu = 0, paired = TRUE)
```

One-Way ANOVA

A. (Replicating Example 10.9) Install and load the *reshape2* package. This package is necessary in order to reconfigure the data frame. Enter:

FILE
Public_Transportation

```
> install.packages("reshape2")
> library(reshape2)
```

B. Stack the city names and cost values using the **melt** function. Label the reconfigured data frame as Stacked. (R displays a warning, which you can ignore.) For clarity, use the **colnames** function to label the columns in Stacked as City and Cost, respectively. Enter:

```
> Stacked <- melt(Public_Transportation)
> colnames(Stacked) <- c("City", "Cost")
```

(You can make sure that your data has been properly reconfigured by entering Stacked at the prompt sign.)

C. Use the **aov** function, which creates an analysis of variance model object; label this object as Transit. Within the **aov** function, first specify the quantitative variable of interest as a function of the categorical factor(s) or treatment(s). Enter:

```
> Transit <- aov(Cost ~ City, data = Stacked)
```

D. To obtain the ANOVA table, use the **anova** function with the object created in Step C. Enter:

```
> anova(Transit)
```

11 Comparisons Involving Proportions

Learning Objectives

After reading this chapter you should be able to:

LO **11.1** **Make inferences about the difference between two population proportions based on independent sampling.**

LO **11.2** **Discuss features of the χ^2 distribution.**

LO **11.3** **Conduct a goodness-of-fit test for a multinomial experiment.**

LO **11.4** **Conduct a test of independence.**

In Chapter 10, we used quantitative data to make inferences regarding the means of two or more populations. In this chapter we focus on qualitative data. We first compare the difference between two population proportions. For instance, marketing executives and advertisers are often interested in the different preferences between males and females when determining where to target advertising dollars. We then introduce the χ^2 (chi-square) distribution to develop statistical tests that compare observed data with what we would expect from a population with a specific distribution. Generally, chi-square tests are used to assess two types of comparison. First, a *goodness-of-fit test* is commonly used with a frequency distribution representing sample data of a qualitative variable. For instance, we may want to substantiate a claim that market shares in the automotive industry have changed dramatically over the past 10 years. Whereas a goodness-of-fit test focuses on a single qualitative variable, a *test for independence* is used to compare two qualitative variables. For example, we may want to determine whether a person's gender influences his/her purchase of a product.



©RaymondAsiaPhotography/Alamy Stock Photo

Introductory Case

Sportswear Brands

In the introductory case to Chapter 4, Annabel Gonzalez, chief retail analyst at a marketing firm, studies the relationship between the brand name of compression garments in the sport-apparel industry and the age of the customer. Specifically, she wants to know whether the age of the customer influences the brand name purchased.

Her initial feeling is that the Under Armour brand attracts a younger customer, whereas the more established companies, Nike and Adidas, draw an older clientele. She believes this information is relevant to advertisers and retailers in the sporting-goods industry, as well as to some in the financial community. Suppose she collects data on 600 recent purchases in the compression-gear market. For ease of exposition, the contingency table (cross-classified by age and brand name) from Chapter 4 is reproduced here as Table 11.1.

TABLE 11.1 Purchases of Compression Garments Based on Age and Brand Name

	Brand Name		
Age Group	Under Armour	Nike	Adidas
Under 35 years	174	132	90
35 years or older	54	72	78

Annabel wants to use the above sample information to

1. Determine whether the two variables (Age Group and Brand Name) are related at the 5% significance level.
2. Discuss how Under Armour can use the findings from the test in its marketing campaigns.

A synopsis of this case will be provided at the end of Section 11.3.

LO 11.1

Make inferences about the difference between two population proportions based on independent sampling.

11.1 INFERENCE CONCERNING THE DIFFERENCE BETWEEN TWO PROPORTIONS

In the preceding chapter, we focused on quantitative data, where we compared means of two or more populations. Now we turn our attention to qualitative data, where we provide statistical inference concerning the difference between two population proportions. This technique has many practical applications. For instance, an investor may want to determine if the bankruptcy rate is the same in the technology and construction industries. The resulting analysis will help determine the relative risk of investing in these two industries. Or perhaps a marketing executive maintains that the proportion of women who buy a firm's product is greater than the proportion of men who buy the product. If this claim is supported by the data, it provides information as to where the firm should advertise. In another case, a consumer advocacy group may state that the proportion of young adults who carry health insurance is less than the proportion of older adults. Health and government officials might be particularly interested in this type of information. All of these examples deal with comparing two population proportions. Our parameter of interest is $p_1 - p_2$, where p_1 and p_2 denote the proportions in the first and second populations, respectively. The estimator for the difference between two population proportions is $\overline{P}_1 - \overline{P}_2$.

Confidence Interval for $p_1 - p_2$

Since the population proportions p_1 and p_2 are unknown, we estimate them by $\bar{p}_1$ and $\bar{p}_2$, respectively. The first sample proportion is computed as $\bar{p}_1 = x_1/n_1$ where x_1 denotes the number of successes in n_1 observations drawn from population 1. Similarly, $\bar{p}_2 = x_2/n_2$ is the sample proportion derived from population 2 where x_2 is the number of successes in n_2 observations drawn from population 2. The difference $\bar{p}_1 - \bar{p}_2$ is a point estimate of $p_1 - p_2$. Recall from Chapter 7 that the standard errors for the estimators $\overline{P}_1$ and $\overline{P}_2$ are $se(\overline{P}_1) = \sqrt{\frac{p_1(1-p_1)}{n_1}}$ and $se(\overline{P}_2) = \sqrt{\frac{p_2(1-p_2)}{n_2}}$, respectively. Therefore, for two independently drawn samples, the standard error, $se(\overline{P}_1 - \overline{P}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$. Since p_1 and p_2 are unknown, we estimate the standard error by $\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$. Finally, when both n_1 and n_2 are sufficiently large, the sampling distribution of $\overline{P}_1 - \overline{P}_2$ can be approximated by the normal distribution. We construct a confidence interval for the difference between two population proportions using the following formula.

CONFIDENCE INTERVAL FOR $p_1 - p_2$

A $100(1 - \alpha)\%$ confidence interval for the difference between two population proportions $p_1 - p_2$ is given by:

$$(\bar{p}_1 - \bar{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}.$$

As noted, the above formula is valid only when the two samples are sufficiently large; the general guideline is that n_1p_1, $n_1(1 - p_1)$, n_2p_2, and $n_2(1 - p_2)$ must all be greater than or equal to 5, where p_1 and p_2 are evaluated at $\bar{p}_1$ and $\bar{p}_2$, respectively.

EXAMPLE 11.1

Despite his inexperience, candidate A appears to have gained support among the electorate. Three months ago, in a survey of 120 registered voters, 55 said that they would vote for Candidate A. Today, 41 registered voters in a sample of 80 said that they would vote for Candidate A. Construct the 95% confidence interval for the difference between the two population proportions.

SOLUTION: Let p_1 and p_2 represent the population proportion of the electorate who support the candidate today and three months ago, respectively. In order to calculate the 95% confidence interval for $p_1 - p_2$, we use the formula $(\bar{p}_1 - \bar{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$. We compute the sample proportions as

$$\bar{p}_1 = x_1/n_1 = 41/80 = 0.5125 \quad \text{and} \quad \bar{p}_2 = x_2/n_2 = 55/120 = 0.4583.$$

Note that the normality condition is satisfied because $n_1\bar{p}_1$, $n_1(1-\bar{p}_1)$, $n_2\bar{p}_2$, and $n_2(1-\bar{p}_2)$ all exceed 5. For the 95% confidence interval, we use the z table to find $z_{\alpha/2} = 1.96$. Substituting the values into the formula, we find

$$(0.5125 - 0.4583) \pm 1.96\sqrt{\frac{0.5125(1-0.5125)}{80} + \frac{0.4583(1-0.4583)}{120}}$$
$$= 0.0542 \pm 0.1412 \text{ or } [-0.0870, 0.1954].$$

With 95% confidence, we can report that the percentage change of support for the candidate is between −8.70% and 19.54%.

Hypothesis Test for $p_1 - p_2$

The null and alternative hypotheses for testing the difference between two population proportions under independent sampling will take one of the following forms:

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
H_0: $p_1 - p_2 = d_0$	H_0: $p_1 - p_2 \le d_0$	H_0: $p_1 - p_2 \ge d_0$
H_A: $p_1 - p_2 \ne d_0$	H_A: $p_1 - p_2 > d_0$	H_A: $p_1 - p_2 < d_0$

We use the symbol d_0 to denote a given hypothesized difference between the unknown population proportions p_1 and p_2. In most cases, d_0 is set to zero. For example, when testing if the population proportions differ—that is, if $p_1 \ne p_2$—we use a two-tailed test, with the competing hypotheses defined as H_0: $p_1 - p_2 = 0$ versus H_A: $p_1 - p_2 \ne 0$. If, on the other hand, we wish to determine whether or not the proportions differ by some amount, say 0.20, we set $d_0 = 0.20$ and define the competing hypotheses as H_0: $p_1 - p_2 = 0.20$ versus H_A: $p_1 - p_2 \ne 0.20$. One-tailed tests are defined similarly.

EXAMPLE 11.2

Let's revisit Example 11.1. Specify the competing hypotheses in order to determine whether the proportion of those who favor Candidate A has changed over the three-month period. Using the 95% confidence interval, what is the conclusion to the test? Explain.

SOLUTION: In essence, we would like to determine whether $p_1 \ne p_2$, where p_1 and p_2 represent the population proportion of the electorate who support the candidate today and three months ago, respectively. We formulate the competing hypotheses as

$$H_0: p_1 - p_2 = 0$$
$$H_A: p_1 - p_2 \ne 0$$

In the previous example, we constructed the 95% confidence interval for the difference between the population proportions as [−0.0870, 0.1954]. We note that the interval contains zero, the value hypothesized under the null hypothesis. Therefore, we are unable to reject the null hypothesis. In other words, we cannot conclude at the 5% significance level that the support for candidate A has changed.

We now introduce the standard four-step procedure for conducting one- or two-tailed hypothesis tests concerning the difference between two proportions $p_1 - p_2$. We transform its estimate $\bar{p}_1 - \bar{p}_2$ into its corresponding z statistic by first subtracting the hypothesized difference d_0 from this estimate, and then dividing by the standard error of the estimator $se(\bar{P}_1 - \bar{P}_2)$. When we developed the confidence interval for $p_1 - p_2$, we assumed $se(\bar{P}_1 - \bar{P}_2) = \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$. However, if d_0 is zero—that is, H_0: $p_1 = p_2$—both $\bar{p}_1$ and $\bar{p}_2$ are essentially the estimates of the same unknown population proportion. For this reason, the standard error can be improved upon by computing the pooled estimate $\bar{p} = (x_1 + x_2)/(n_1 + n_2)$ for the unknown population proportion, which is now based on a larger sample.

TEST STATISTIC FOR TESTING $p_1 - p_2$

The value of the test statistic for a hypothesis test concerning the difference between two proportions $p_1 - p_2$ is computed using one of the following two formulas:

1. If the hypothesized difference d_0 is zero, then the value of the test statistic is

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where $\bar{p}_1 = \frac{x_1}{n_1}$, $\bar{p}_2 = \frac{x_2}{n_2}$, and $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$.

2. If the hypothesized difference d_0 is not zero, then the value of the test statistic is

$$z = \frac{(\bar{p}_1 - \bar{p}_2) - d_0}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}}.$$

As in the case of the confidence interval, the above formulas are valid only when the two samples are sufficiently large.

EXAMPLE 11.3

Research by analysts and retailers claims gender differences when it comes to online shopping (*The Wall Street Journal*, March 13, 2008). A survey revealed that 5,400 of 6,000 men said they "regularly" or "occasionally" make purchases online, compared with 8,600 of 10,000 women surveyed. At the 5% significance level, test whether the proportion of all men who regularly or occasionally make purchases online is greater than the proportion of all women.

SOLUTION: Let p_1 and p_2 denote the population proportions of men and of women who make online purchases, respectively. We wish to test whether the proportion of men who make purchases online is greater than the proportion of women; that is, $p_1 - p_2 > 0$. Therefore, we construct the competing hypotheses as

$$H_0: p_1 - p_2 \leq 0$$
$$H_A: p_1 - p_2 > 0$$

Since the hypothesized difference is zero, or $d_0 = 0$, we compute the value of the test statistic as $z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$. We first compute the sample proportions

$\bar{p}_1 = x_1/n_1 = 5{,}400/6{,}000 = 0.90$ and $\bar{p}_2 = x_2/n_2 = 8{,}600/10{,}000 = 0.86$. The normality condition is satisfied since $n_1\bar{p}_1$, $n_1(1-\bar{p}_1)$, $n_2\bar{p}_2$, and $n_2(1-\bar{p}_2)$ all exceed 5. Next we calculate $\bar{p} = \frac{x_1+x_2}{n_1+n_2} = \frac{5{,}400+8{,}600}{6{,}000+10{,}000} = 0.875$. Thus,

$$z = \frac{(0.90-0.86)}{\sqrt{0.875(1-0.875)\left(\frac{1}{6{,}000}+\frac{1}{10{,}000}\right)}} = 7.41.$$

The p-value, computed as $P(Z \geq 7.41)$, is approximately zero. Since the p-value $< \alpha = 0.05$, we reject H_0. At the 5% significance level, we conclude that the proportion of men who shop online either regularly or occasionally is greater than the proportion of women. Our results are consistent with the recent decision by so many retailers to redesign their websites to attract male customers.

In Example 11.4, we consider a case where the hypothesized value d_0 does not equal 0.

EXAMPLE 11.4

While we expect relatively expensive wines to have more desirable characteristics than relatively inexpensive wines, people are often confused in their assessment of the quality of wine in a blind test (*The Telegraph*, April 30, 2015). In a recent experiment at a local winery, the same wine is served to two groups of people but with different price information. In the first group, 60 people are told that they are tasting a $25 wine, of which 48 like the wine. In the second group, only 20 of 50 people like the wine when they are told that it is a $10 wine. The experiment is conducted to determine if the proportion of people who like the wine in the first group is more than 20 percentage points higher than in the second group. Conduct this test at the 5% significance level.

SOLUTION: Let p_1 and p_2 denote the proportions of people who like the wine in groups 1 and 2, respectively. We want to test if the proportion of people who like the wine in the first group is more than 20 percentage points higher than in the second group. Thus, we construct the competing hypotheses as

$$H_0: p_1 - p_2 \leq 0.20$$
$$H_A: p_1 - p_2 > 0.20$$

We first compute the sample proportions as $\bar{p}_1 = x_1/n_1 = 48/60 = 0.80$ and $\bar{p}_2 = x_2/n_2 = 20/50 = 0.40$ and note that the normality condition is satisfied since $n_1\bar{p}_1$, $n_1(1-\bar{p}_1)$, $n_2\bar{p}_2$, and $n_2(1-\bar{p}_2)$ all exceed 5.

Since $d_0 = 0.20$, the value of the test statistic is computed as

$$z = \frac{(\bar{p}_1-\bar{p}_2)-d_0}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1}+\frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}} = \frac{(0.80-0.40)-0.20}{\sqrt{\frac{0.80(1-0.80)}{60}+\frac{0.40(1-0.40)}{50}}} = 2.31.$$

For this right-tailed test, we compute the p-value as $P(Z \geq 2.31) = 0.0104$. Since the p-value $< \alpha = 0.05$, we reject the null hypothesis. At the 5% significance level, we conclude that the proportion of people who like the wine in the first group is more than 20 percentage points higher than in the second group. Overall, this result is consistent with scientific research, which has demonstrated the power of suggestion in wine tasting.

EXERCISES 11.1

Mechanics

1. Given $\bar{p}_1 = 0.85$, $n_1 = 400$, $\bar{p}_2 = 0.90$, $n_2 = 350$, construct the 90% confidence interval for the difference between the population proportions. Is there a difference between the population proportions at the 10% significance level? Explain.
2. Given $x_1 = 50$, $n_1 = 200$, $x_2 = 70$, $n_2 = 250$, construct the 95% confidence interval for the difference between the population proportions. Is there a difference between the population proportions at the 5% significance level? Explain.
3. Consider the following competing hypotheses and accompanying sample data.

$$H_0: p_1 - p_2 \geq 0$$
$$H_A: p_1 - p_2 < 0$$
$$x_1 = 250 \qquad x_2 = 275$$
$$n_1 = 400 \qquad n_2 = 400$$

   a. Calculate the value of the test statistic.
   b. Find the *p*-value.
   c. At the 5% significance level, what is the conclusion to the test? Is p_1 less than p_2?

4. Consider the following competing hypotheses and accompanying sample data.

$$H_0: p_1 - p_2 = 0$$
$$H_A: p_1 - p_2 \neq 0$$
$$x_1 = 100 \qquad x_2 = 172$$
$$n_1 = 250 \qquad n_2 = 400$$

   a. Calculate the value of the test statistic.
   b. Find the *p*-value.
   c. At the 5% significance level, what is the conclusion to the test? Do the population proportions differ?

5. Consider the following competing hypotheses and accompanying sample data.

$$H_0: p_1 - p_2 = 0$$
$$H_A: p_1 - p_2 \neq 0$$
$$x_1 = 300 \qquad x_2 = 325$$
$$n_1 = 600 \qquad n_2 = 500$$

   a. Calculate the value of the test statistic.
   b. Find the *p*-value.
   c. At the 5% significance level, what is the conclusion to the test? Do the population proportions differ?

6. Consider the following competing hypotheses and accompanying sample data.

$$H_0: p_1 - p_2 = 0.20$$
$$H_A: p_1 - p_2 \neq 0.20$$
$$x_1 = 150 \qquad x_2 = 130$$
$$n_1 = 250 \qquad n_2 = 400$$

   a. Calculate the value of the test statistic.
   b. Find the *p*-value.
   c. At the 5% significance level, what is the conclusion to the test? Can you conclude that the difference between the population proportions differs from 0.20?

Applications

7. A study claims that girls and boys do not do equally well on math tests taken from the 2nd to 11th grades (*Chicago Tribune*, July 25, 2008). Suppose in a representative sample, 344 of 430 girls and 369 of 450 boys score at proficient or advanced levels on a standardized math test.
   a. Construct the 95% confidence interval for the difference between the population proportions of girls and boys who score at proficient or advanced levels.
   b. Develop the appropriate null and alternative hypotheses to test whether the proportion of girls who score at proficient or advanced levels differs from the proportion of boys.
   c. At the 5% significance level, what is the conclusion to the test? Do the results support the study's claim?
8. Reducing scrap of 4-foot planks of hardwood is an important factor in reducing cost at a wood-flooring manufacturing company. Accordingly, engineers at Lumberworks are investigating a potential new cutting method involving lateral sawing that may reduce the scrap rate. To examine its viability, samples of planks were examined under the old and new methods. Sixty-two of the 500 planks were scrapped under the old method, whereas 36 of the 400 planks were scrapped under the new method.
   a. Construct the 95% confidence interval for the difference between the population scrap rates between the old and new methods.
   b. Specify the null and alternative hypotheses to test for differences in the population scrap rates between the old and new cutting methods.
   c. Using the results from part (a), can we conclude at the 5% significance level that the scrap rate of the new method is different from the old method?
9. According to a Pew report, 14.6% of newly married couples in 2008 reported that their spouse was of another race or ethnicity (*CNNLiving*, June 7, 2010). In a similar survey in 1980, only 6.8% of newlywed couples reported marrying outside their race or ethnicity. Suppose both of these surveys were conducted on 500 newly married couples.
   a. Specify the competing hypotheses to test the claim that there is an increase in the proportion of people who marry outside their race or ethnicity.
   b. Calculate the value of the test statistic and the *p*-value.
   c. At the 5% level of significance, what is the conclusion to the test?

10. Research by Harvard Medical School experts suggests that boys are more likely than girls to grow out of childhood asthma when they hit their teenage years (*BBC News*, August 15, 2008). Scientists followed over 1,000 children between the ages of 5 and 12, all of whom had mild to moderate asthma. By the age of 18, 27% of the boys and 14% of the girls had grown out of asthma. Suppose the analysis was based on 500 boys and 500 girls.
    a. Develop the hypotheses to test whether the proportion of boys who grow out of asthma in their teenage years is more than that of girls.
    b. Test the assertion in part (a) at the 5% significance level.
    c. A medical researcher has asserted that the proportion of boys who grow out of asthma in their teenage years is more than 0.10 than that of girls. Test this assertion at the 5% significance level.

11. More people are using social media to network, rather than phone calls or e-mails (U.S. *News & World Report*, October 20, 2010). From an employment perspective, jobseekers are no longer calling up friends for help with job placement, as they can now get help online. In a recent survey of 150 jobseekers, 67 said they used LinkedIn to search for jobs. A similar survey of 140 jobseekers, conducted three years ago, had found that 58 jobseekers had used LinkedIn for their job search. Is there sufficient evidence to suggest that more people are now using LinkedIn to search for jobs as compared to three years ago? Use a 5% level of significance for the analysis.

12. The director of housekeeping at *Elegante*, a luxury resort hotel with two locations (*Seaside* and *Oceanfront*), wants to evaluate housekeeping performance at those two locations. Random samples of 100 rooms were inspected at each location for defects (e.g., missing towels, missing soap, dirty floors or showers, dusty tables) after being cleaned. It was found that 21 of the rooms at *Seaside* had some housekeeping defect(s), and 28 rooms at *Oceanfront* had some housekeeping defect(s).
    a. Develop the hypotheses to test whether the proportion of housekeeping defects differs between the two hotel locations.
    b. Calculate the value of the test statistic and the *p*-value.
    c. Do the results suggest that the proportion of housekeeping defects differs between the two hotel locations at the 5% significance level?
    d. Construct the 95% confidence interval for the difference between the population housekeeping defect rates at the two hotel locations. How can this confidence interval be used to reach the same conclusion as in part (c)?

13. In an effort to make children's toys safer and more tamperproof, toy packaging has become cumbersome for parents to remove in many cases. Accordingly, the director of marketing at Toys4Tots, a large toy manufacturer, wants to evaluate the effectiveness of a new packaging design that engineers claim will reduce customer complaints by more than 10 percentage points. Customer satisfaction surveys were sent to 250 parents who registered toys packaged under the old design and 250 parents who registered toys packaged under the new design. Of these, 85 parents expressed dissatisfaction with packaging of the old design, and 40 parents expressed dissatisfaction with packaging of the new design.
    a. Specify the null and alternative hypotheses to test whether customer complaints have been reduced by more than 10 percentage points under the new packaging design.
    b. Calculate the value of the test statistic and the *p*-value.
    c. At the 5% significance level, do the results support the engineers' claim?
    d. At the 1% significance level, do the results support the engineers' claim?

14. According to a report by the Centers for Disease Control and Prevention (CDC), 36.5% of American adults are obese (*Adult Obesity Facts*, August 29, 2017). Among ethnic groups in general, African American women are more overweight than Caucasian women, but African American men are less obese than Caucasian men. Sarah Weber, a recent college graduate, is curious to determine if the same pattern also exists in her hometown on the West Coast. She randomly selects 220 African American adults and 300 Caucasian adults for the analysis. The following table contains the sample information.

Race	Gender	Obese	Not Obese
African Americans	Males	36	94
	Females	35	55
Caucasians	Males	62	118
	Females	31	89

    a. Test if the proportion of obese African American men is less than the proportion of obese Caucasian men at $\alpha = 0.05$.
    b. Test if the proportion of obese African American women is more than the proportion of obese Caucasian women at $\alpha = 0.05$.
    c. Test if the proportion of obese African American adults differs from the proportion of obese Caucasian adults at the 5% significance level.

15. Only 26% of psychology majors are satisfied with their career paths as compared to 50% of accounting majors (*The Wall Street Journal*, October 11, 2010). Suppose these results were obtained from a survey of 300 psychology majors and 350 accounting majors.
    a. Develop the null and alternative hypotheses to test whether the proportion of accounting majors satisfied with their career paths is higher than that of psychology majors by more than 20 percentage points.
    b. Calculate the value of the test statistic and the *p*-value.
    c. At the 5% significance level, what is the conclusion?

16. Due to late delivery problems with an existing supplier, the director of procurement at ElectroTech began to place orders for electrical switches with a new supplier as part of a "dual-source" (two-supplier) strategy. Now she wants to revert

to a "single-source" (i.e., one supplier) strategy to simplify purchasing activities. She wishes to conduct a test to infer whether the new supplier will continue to outperform the old supplier. Based on recent sample data, she found that 27 of 150 orders placed with the old supplier arrived late, whereas 6 of 75 orders placed with the new supplier arrived late.

a. Specify the null and alternative hypotheses to test for whether the proportion of late deliveries with the new supplier is less than that of the old supplier.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 5% significance level, what is the conclusion?

17. A report suggests that business majors spend the least amount of time on course work than all other college students (*The New York Times*, November 17, 2011). A provost of a university decides to conduct a survey where students are asked if they study hard, defined as spending at least 20 hours per week on course work. Of 120 business majors included in the survey, 20 said that they studied hard, as compared to 48 out of 150 nonbusiness majors who said that they studied hard. At the 5% significance level, can we conclude that the proportion of business majors who study hard is less than that of nonmajors? Provide the details.

18. It has generally been believed that it is not feasible for men and women to be just friends (*The New York Times*, April 12, 2012). Others argue that this belief may not be true anymore since gone are the days when men worked and women stayed at home and the only way they could get together was for romance. In a recent survey, 200 heterosexual college students were asked if it was feasible for male and female students to be just friends. Thirty-two percent of females and 57% of males reported that it was not feasible for men and women to be just friends. Suppose the study consisted of 100 female and 100 male students. At the 5% significance level, can we conclude that there is a greater than 10 percentage point difference between the proportion of male and female students with this view? Provide the details.

LO **11.2**

Discuss features of the χ^2 distribution.

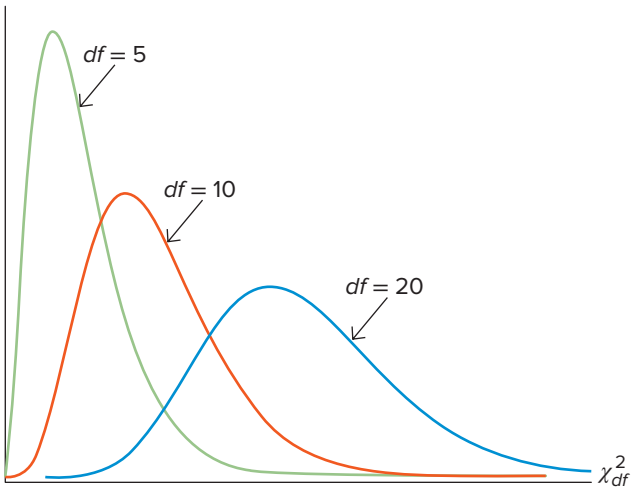
11.2 GOODNESS-OF-FIT TEST FOR A MULTINOMIAL EXPERIMENT

There are many instances where we may want to make inferences about the relative sizes of more than two population proportions. For instance, in a heavily concentrated industry consisting of four firms, we may want to determine whether each firm has an equal market share. Or, in a political contest, we may want to determine whether Candidates A, B, and C will receive 70%, 20%, and 10% of the vote, respectively. These tests are based on a new distribution called the χ^2 **(chi-square)** distribution.

The χ^2 Distribution

Like the *t* distribution, the χ^2 distribution is characterized by a family of distributions, where each distribution depends on its particular degrees of freedom *df*. It is common, therefore, to refer to it as the χ^2_{df} distribution. The χ^2_{df} distribution is positively skewed, where the extent of skewness depends on the degrees of freedom. As the *df* grow larger, the χ^2_{df} distribution approaches the normal distribution. For instance, in Figure 11.1, the χ^2_{20} distribution resembles the shape of the normal distribution.

FIGURE 11.1
The χ^2_{df} distribution with various degrees of freedom



SUMMARY OF THE χ^2_{df} DISTRIBUTION

- The χ^2_{df} distribution is characterized by a family of distributions, where each distribution depends on its particular degrees of freedom *df*.
- The values of the χ^2_{df} distribution range from zero to infinity.
- The χ^2_{df} distribution is positively skewed, and the extent of skewness depends on the *df*. As the *df* grow larger, the χ^2_{df} distribution approaches the normal distribution.

Finding χ^2_{df} Values and Probabilities

For a random variable that follows the χ^2_{df} distribution, we use the notation $\chi^2_{\alpha,df}$ to represent a value such that the area in the upper (right) tail of the distribution is α. In other words, $P(\chi^2_{df} \geq \chi^2_{\alpha,df}) = \alpha$. Figure 11.2 illustrates this notation.

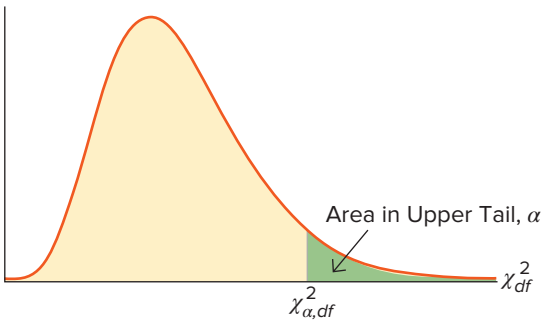


FIGURE 11.2
Graphical depiction of $P(\chi^2_{df} \geq \chi^2_{\alpha,df}) = \alpha$

A portion of the upper tail areas and the corresponding values for the χ^2_{df} distributions are given in Table 11.2. (Table 3 of Appendix A provides a more complete table.) Suppose we want to find the value of $\chi^2_{\alpha,df}$ with $\alpha = 0.05$ and $df = 10$; that is, $\chi^2_{0.05,10}$. Using Table 11.2, we look at the first column labeled *df* and find the value 10. We then continue along this row until we reach the column 0.050. Here we see the value $\chi^2_{0.05,10} = 18.307$ such that $P(\chi^2_{10} \geq 18.307) = 0.05$.

TABLE 11.2 Portion of the χ^2 table

	α									
df	**0.995**	**0.990**	**0.975**	**0.950**	**0.900**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	2.156	2.558	3.247	3.940	4.865	15.987	**18.307**	20.483	23.209	25.188
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	67.328	70.065	74.222	77.929	82.358	118.342	124.342	129.561	135.807	140.170

LO 11.3
Conduct a goodness-of-fit test for a multinomial experiment.

We will now examine whether two or more population proportions equal each other or some predetermined (hypothesized) set of values. Before conducting this test, we must first ensure that the random experiment satisfies the conditions of a **multinomial experiment**, which is simply a generalization of the Bernoulli process first introduced in Chapter 5.

Recall that a Bernoulli process, also referred to as a binomial experiment, is a series of *n* independent and identical trials of an experiment, where each trial has only two possible outcomes, conventionally labeled "success" and "failure." For the binomial experiment, we generally denote the probability of success as *p* and the probability of failure as $1 - p$. Alternatively, we could let p_1 and p_2 represent these probabilities, where $p_1 + p_2 = 1$. In a multinomial experiment, the number of outcomes per trial is *k* where $k \geq 2$.

A MULTINOMIAL EXPERIMENT

A multinomial experiment consists of a series of n independent and identical trials, such that for each trial:

- There are k possible outcomes called categories.
- The probability p_i associated with the ith category remains the same.
- The sum of the probabilities is one; that is, $p_1 + p_2 + \cdots + p_k = 1$.

Note that when $k = 2$, the multinomial experiment specializes to a binomial experiment.

Numerous experiments fit the conditions of a multinomial experiment. For instance,

- As compared to the previous day, a stockbroker records whether the price of a stock rises, falls, or stays the same. This example has three possible categories ($k = 3$).
- A customer rates service at a restaurant as excellent, good, fair, or poor ($k = 4$).
- The admissions office records which of the six business concentrations a student picks ($k = 6$).

When setting up the competing hypotheses for a multinomial experiment, we have essentially two choices. We can set all population proportions equal to the same specific value or, equivalently, equal to one another. For instance, if we want to test on the basis of sample data whether the proportion of voters who favor four different candidates is not the same, the competing hypotheses would take the following form:

H_0: $p_1 = p_2 = p_3 = p_4 = 0.25$

H_A: Not all population proportions are equal to 0.25.

Note that the hypothesized value under the null hypothesis is 0.25 because the population proportions must sum to one. We can also set each population proportion equal to a different predetermined (hypothesized) value. Suppose we want to contest the prediction that 40% of the voters favor Candidate 1, 30% favor Candidate 2, 20% favor Candidate 3, and 10% favor Candidate 4. The competing hypotheses are formulated as

H_0: $p_1 = 0.40$, $p_2 = 0.30$, $p_3 = 0.20$, and $p_4 = 0.10$

H_A: Not all population proportions equal their hypothesized values.

When conducting a test, we take a random sample and determine the extent to which the sample proportions deviate from the hypothesized population proportions. For this reason, this type of test is called a **goodness-of-fit test**. Under the usual assumption that the null hypothesis is true, we derive the expected frequencies of the categories in a multinomial experiment and compare them with observed frequencies. The objective is to determine whether we can reject the null hypothesis in favor of the alternative hypothesis. To see how to conduct a goodness-of-fit test, consider the following example.

One year ago, the management at a restaurant chain surveyed its patrons to determine whether changes should be made to the menu. One question on the survey asked patrons to rate the quality of the restaurant's entrées. The percentages of the patrons responding Excellent, Good, Fair, or Poor are listed in the following table:

Excellent	Good	Fair	Poor
15%	30%	45%	10%

Based on responses to the overall survey, management decided to revamp the menu. Recently, the same question concerning the quality of entrées was asked of a random sample of 250 patrons. Their responses are shown in the following table:

Excellent	Good	Fair	Poor
46	83	105	16

At the 5% significance level, we want to determine whether there has been any change in the population proportions calculated one year ago.

Since we want to determine whether the responses of the 250 patrons are inconsistent with the earlier proportions, we let the earlier population proportions denote the hypothesized proportions for the test. Thus, we use p_1, p_2, p_3, and p_4 to denote the population proportions of those that responded Excellent, Good, Fair, or Poor, respectively, and construct the following competing hypotheses.

H_0: $p_1 = 0.15$, $p_2 = 0.30$, $p_3 = 0.45$, and $p_4 = 0.10$

H_A: Not all population proportions equal their hypothesized values.

The first step in calculating the value of the test statistic is to calculate the expected frequency for each category. That is, we need to estimate the frequencies that we would expect to get if the null hypothesis is true. In general, in order to calculate the expected frequency e_i for category i, we multiply the sample size n by the respective hypothesized value of the population proportion p_i. For example, consider the category Excellent. If H_0 is true, then we expect that 15% ($p_1 = 0.15$) of 250 patrons will find the quality of entrées to be excellent. Therefore, the expected frequency of Excellent responses is 37.5 ($= 250 \times 0.15$), whereas the corresponding observed frequency is 46. Expected frequencies for other responses are found similarly. Ultimately, when computing the value of the test statistic, we compare these expected frequencies to the frequencies we actually observe. The test statistic follows the χ^2_{df} distribution.

TEST STATISTIC FOR THE GOODNESS-OF-FIT TEST

For a multinomial experiment with k categories, the value of the test statistic is calculated as

$$\chi^2_{df} = \Sigma \frac{(o_i - e_i)^2}{e_i},$$

where $df = k - 1$, o_i is the observed frequency for category i, $e_i = np_i$ is the expected frequency for category i, and n is the number of observations.

Note: The test is valid when the expected frequency for each category is five or more.

Table 11.3 shows the expected frequency e_i for each category. The condition that each expected frequency e_i must equal five or more is satisfied here. As we will see shortly, sometimes it is necessary to combine data from two or more categories to achieve this result.

TABLE 11.3 Calculation of Expected Frequency for Restaurant Example

	Hypothesized Proportion, p_i	Expected Frequency, $e_i = np_i$
Excellent	0.15	$250 \times 0.15 = 37.5$
Good	0.30	$250 \times 0.30 = 75.0$
Fair	0.45	$250 \times 0.45 = 112.5$
Poor	0.10	$250 \times 0.10 = 25.0$
		$\Sigma e_i = 250$

As a check on the calculations, the sum of the expected frequencies Σe_i must equal the sample size n, which in this example equals 250. Once the expected frequencies are estimated, we are ready to calculate the value of the test statistic.

The χ^2_{df} statistic measures how much the observed frequencies deviate from the expected frequencies. In particular, χ^2_{df} is computed as the sum of the standardized squared deviations. The smallest value that χ^2_{df} can assume is zero—this occurs when each observed frequency equals its expected frequency. Rejection of the null hypothesis occurs when χ^2_{df} is significantly greater than zero. As a result, these tests of hypotheses regarding multiple population proportions (p_1, p_2, p_3, . . .) are always implemented as

right-tailed tests. However, since the alternative hypothesis states that not all population proportions equal their hypothesized values, rejection of the null hypothesis does not indicate which proportions differ from these values.

In this example, there are four categories ($k = 4$), so $df = k - 1 = 3$. The value of the test statistic is calculated as

$$\chi^2_{df} = \chi^2_3 = \Sigma \frac{(o_i - e_i)^2}{e_i}$$
$$= \frac{(46 - 37.5)^2}{37.5} + \frac{(83 - 75)^2}{75} + \frac{(105 - 112.5)^2}{112.5} + \frac{(16 - 25)^2}{25}$$
$$= 1.927 + 0.853 + 0.500 + 3.240 = 6.520.$$

Since a goodness-of-fit test is a right-tailed test, we calculate the p-value as $P(\chi^2_3 \geq 6.520)$. We show a portion of the χ^2 table from Appendix A in Table 11.4.

For $df = 3$, we see that 6.520 lies between the values 6.251 and 7.815, implying that the p-value is between 0.05 and 0.10. In order to find the exact p-value, we can use Excel's CHISQ.DIST.RT(χ^2_{df}, df) function where χ^2_{df} is the value of the test statistic and df are the respective degrees of freedom. In this example, we enter '=CHISQ.DIST.RT(6.520, 3)' and Excel returns 0.089. Since the p-value is greater than 0.05, we do not reject H_0. We cannot conclude that the proportions differ from the ones from one year ago at the 5% significance level. Management may find this news disappointing in that the goal of the menu change was to improve customer satisfaction. Responses to other questions on the survey may shed more light on whether the goals of the menu change met or fell short of expectations.

TABLE 11.4 Portion of the χ^2 table

	α									
df	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838

EXAMPLE 11.5

FILE *Share*

Table 11.5 lists the market shares in 2010 of the five firms that manufacture a particular product. A marketing analyst wonders whether the market shares have changed since 2010. He surveys 200 customers. The last column of Table 11.5 shows the number of customers who recently purchased the product at each firm.

TABLE 11.5 Market Share of Five Firms

Firm	Market Share in 2010	Number of Recent Customers
1	0.40	70
2	0.32	60
3	0.24	54
4	0.02	10
5	0.02	6

a. Specify the competing hypotheses to test whether the market shares have changed since 2010.

b. Calculate the value of the test statistic.

c. Use $a = 0.05$ to determine if the market shares have changed since 2010.

SOLUTION:

a. Let p_i denote the market share for the ith firm. In order to test whether the market shares have changed since 2010, we *initially* set up the competing hypotheses as

$$H_0: p_1 = 0.40,\ p_2 = 0.32,\ p_3 = 0.24,\ p_4 = 0.02, \text{ and } p_5 = 0.02$$
$$H_A: \text{Not all market shares equal their hypothesized values.}$$

b. The value of the test statistic is calculated as $\chi^2_{df} = \Sigma\frac{(o_i - e_i)^2}{e_i}$. The last column of Table 11.5 shows each firm's observed frequency o_i, so before applying the formula, we first calculate each firm's expected frequency e_i:

$$\left.\begin{aligned} e_1 &= 200 \times 0.40 = 80 \\ e_2 &= 200 \times 0.32 = 64 \\ e_3 &= 200 \times 0.24 = 48 \\ e_4 &= 200 \times 0.02 = 4 \\ e_5 &= 200 \times 0.02 = 4 \end{aligned}\right\} 8$$

We note that the expected frequencies for firms 4 and 5 are less than five. The test is valid so long as the expected frequencies in each category are five or more. In order to achieve this result, we combine the expected frequencies for firms 4 and 5 to obtain a combined frequency of eight ($e_4 + e_5 = 8$). We could have made other combinations, say e_4 with e_1 and e_5 with e_2, but we preferred to maintain a category for the less dominant firms. After making this combination, we now respecify the competing hypotheses as

$$H_0: p_1 = 0.40,\ p_2 = 0.32,\ p_3 = 0.24, \text{ and } p_4 = 0.04$$
$$H_A: \text{Not all market shares equal their hypothesized values.}$$

With $df = k - 1 = 3$, we calculate the value of the test statistic as

$$\chi^2_3 = \Sigma\frac{(o_i - e_i)^2}{e_i} = \frac{(70 - 80)^2}{80} + \frac{(60 - 64)^2}{64} + \frac{(54 - 48)^2}{48} + \frac{(16 - 8)^2}{8}$$
$$= 1.250 + 0.250 + 0.750 + 8.000 = 10.250.$$

c. We calculate the p-value as $P(\chi^2_3 \geq 10.250)$. From Table 11.4, we see that 10.520 lies between the values 9.348 and 11.345, implying that the p-value is between 0.01 and 0.025. Using Excel's CHISQ.DIST.RT function, we find that the exact p-value is 0.017. Since the p-value is less than 0.05, we reject H_0. At the 5% significance level, we conclude that some market shares have changed.

As mentioned earlier, one limitation of this type of chi-square test is that we cannot tell which proportions differ from their hypothesized values. However, given the divergence between the observed and expected frequencies for the less dominant firms, it appears that they may be making some headway in this industry. Further testing can be conducted to see if this is the case.

EXERCISES 11.2

Mechanics

19. Consider a multinomial experiment with $n = 250$ and $k = 4$. The null hypothesis to be tested is $H_0: p_1 = p_2 = p_3 = p_4 = 0.25$. The observed frequencies resulting from the experiment are

Category	1	2	3	4
Frequency	70	42	72	66

a. Specify the alternative hypothesis.
b. Calculate the value of the test statistic and the p-value.
c. At the 5% significance level, what is the conclusion to the hypothesis test?

20. Consider a multinomial experiment with $n = 400$ and $k = 3$. The null hypothesis is $H_0: p_1 = 0.60,\ p_2 = 0.25$, and $p_3 = 0.15$. The observed frequencies resulting from the experiment are

Category	1	2	3
Frequency	250	94	56

a. Specify the alternative hypothesis.
b. Calculate the value of the test statistic and the p-value.
c. At the 5% significance level, what is the conclusion to the hypothesis test?

21. A multinomial experiment produced the following results:

Category	1	2	3	4	5
Frequency	57	63	70	55	55

Can we conclude at the 1% significance level that not all population proportions are equal to 0.20?

22. A multinomial experiment produced the following results:

Category	1	2	3
Frequency	128	87	185

At the 1% significance level, can we reject H_0: $p_1 = 0.30$, $p_2 = 0.20$, and $p_3 = 0.50$?

Applications

23. You suspect that an unscrupulous employee at a casino has tampered with a die; that is, he is using a loaded die. In order to test this claim, you roll the die 200 times and obtain the following frequencies:

Category	1	2	3	4	5	6
Frequency	40	35	33	30	33	29

a. Specify the null and alternative hypotheses in order to test your claim.
b. Calculate the value of the test statistic and the p-value.
c. At the 10% significance level, can you conclude that the die is loaded?

24. A study conducted in September and October of 2010 found that fewer than half of employers who hired new college graduates last academic year plan to definitely do so again (*The Wall Street Journal*, November 29, 2010). Suppose the hiring intentions of the respondents were as follows:

Definitely Hire	Likely to Hire	Hire Uncertain	Will not Hire
37%	17%	28%	18%

Six months later, a sample of 500 employers were asked their hiring intentions and gave the following responses:

Definitely Hire	Likely to Hire	Hire Uncertain	Will not Hire
170	100	120	110

a. Specify the competing hypotheses to test whether the proportions from the initial study have changed.
b. Calculate the value of the test statistic and the p-value.
c. At the 5% significance level, what is the conclusion to the hypothesis test? Interpret your results.

25. A rent-to-own (RTO) agreement appeals to low-income and financially distressed consumers. It allows immediate access to merchandise, and by making all payments, the consumer acquires the merchandise. At the same time, goods can be returned at any point without penalty. Suppose a recent study documents that 65% of RTO contracts are returned, 30% are purchased, and the remaining 5% default. In order to test the validity of this claim, an RTO researcher looks at the transaction data of 420 RTO contracts, of which 283 are returned, 109 are purchased, and the rest defaulted.
a. Set up the competing hypothesis to test whether the return, purchase, and default probabilities of RTO contracts differ from 0.65, 0.30, and 0.05, respectively.
b. Compute the value of the test statistic.
c. Conduct the test at the 5% level of significance, and interpret the test results.

26. Despite Zimbabwe's shattered economy, with endemic poverty and widespread political strife and repression, thousands of people from overseas still head there every year (*BBC News*, August 27, 2008). Main attractions include the magnificent Victoria Falls, the ruins of Great Zimbabwe, and herds of roaming wildlife. A tourism director claims that Zimbabwe visitors are equally represented by Europe, North America, and the rest of the world. Records show that of the 380 tourists who recently visited Zimbabwe, 148 were from Europe, 106 were from North America, and 126 were from the rest of the world.
a. A recent visitor to Zimbabwe believes that the tourism director's claim is wrong. Set up the competing hypotheses to test the visitor's belief.
b. Conduct the test at the 5% significance level. Do the sample data support the visitor's belief?

27. In 2003, *The World Wealth Report* first started publishing market shares of global millionaires (*The Wall Street Journal*, June 25, 2008). At this time, the distribution of the world's people worth $1 million or more was as follows:

Region	Millionaires
Europe	35.7%
North America	31.4%
Asia Pacific	22.9%
Latin America	4.3%
Middle East	4.3%
Africa	1.4%

Source: *The Wealth Report*, 2003.

A recent sample of 500 global millionaires produces the following results:

Region	Number of Millionaires
Europe	153
North America	163
Asia Pacific	139
Latin America	20
Middle East	20
Africa	5

a. Test whether the distribution of millionaires today is different from the distribution in 2003 at $\alpha = 0.05$.
b. Would the conclusion change if we tested it at $\alpha = 0.10$?

28. An Associated Press/GfK Poll shows that 38% of American drivers favor U.S. cars, while 33% prefer Asian brands, with the remaining 29% going for other foreign cars (www.msnbc.com, April 21, 2010). A researcher wonders whether the preferences for cars have changed since the Associated Press/GfK Poll. He surveys 200 Americans and finds that the number of respondents in the survey who prefer American, Asian, and other foreign cars are 66, 70, and 64, respectively. At the 5% significance level, can the researcher conclude that preferences have changed since the Associated Press/GfK Poll?

29. The quality department at an electronics company has noted that, historically, 92% of the units of a specific product pass a test operation, 6% fail the test but are able to be repaired, and 2% fail the test and need to be scrapped. Due to recent process improvements, the quality department would like to test if the rates have changed. A recent sample of 500 parts revealed that 475 parts passed the test, 18 parts failed the test but were repairable, and 7 parts failed the test and were scrapped.
    a. State the null and alternative hypotheses to test if the current proportions are different than the historical proportions.
    b. Calculate the value of the test statistic and the *p*-value.
    c. At the 5% significance level, what is your conclusion? Would your conclusion change at the 1% significance level?

30. An agricultural grain company processes and packages various grains purchased from farmers. A high-volume conveyor line contains four chutes at the end, each of which is designed to receive and dispense equal proportions of grain into bags. Each bag is then stamped with a date code and the number of the chute from which it came. If the chute output proportions are not relatively equal, then a bottleneck effect is created upstream and the conveyor cannot function at peak output. Recently, a series of repairs and modifications have led management to question whether the grains still are being equally distributed among the chutes. Packaging records from 800 bags yesterday indicate that 220 bags came from Chute 1, 188 bags from Chute 2, 218 bags from Chute 3, and 174 bags from Chute 4.
    a. State the null and alternative hypotheses to test if the proportion of bags filled by any of the chutes is different from 0.25.
    b. Calculate the value of the test statistic and the *p*-value.
    c. What is your conclusion at the 10% significance level? Would your conclusion change at the 5% significance level?

11.3 CHI-SQUARE TEST FOR INDEPENDENCE

LO **11.4**
Conduct a test of independence.

Recall from Chapter 4 that a contingency table is a useful tool when we want to examine or compare two qualitative variables defined on the same population.

CONTINGENCY TABLE

A contingency table generally shows frequencies for two qualitative (categorical) variables, x and y, where each cell represents a mutually exclusive combination of the pair of x and y values.

In this section, we use the data in a contingency table to conduct a hypothesis test that determines whether the two qualitative variables depend upon one another. Whereas a goodness-of-fit test examines a single qualitative variable, a **test for independence**—also called a **chi-square test of a contingency table**—assesses the relationship between

two qualitative variables. Many examples of the use of this test arise, especially in marketing, biomedical research, and courts of law. For instance, a retailer may be trying to determine whether there is a relationship between the age of its clientele and where it chooses to advertise. Doctors might want to investigate whether or not losing weight through stomach surgery can extend the lives of severely obese patients. Or one party in a discrimination lawsuit may be trying to show that one's gender and the likelihood of promotion are related. All of these examples lend themselves to applications of the hypothesis test discussed in this section.

In the introductory case study, we are presented with a contingency table cross-classified by the variables Age Group and Brand Name. Specifically, we want to determine whether or not the age of a customer influences his/her decision to buy a garment from Under Armour, Nike, or Adidas. We will conduct this test at the 5% significance level.

In general, the competing hypotheses for a statistical test for independence are formulated such that rejecting the null hypothesis leads to the conclusion that the two qualitative variables are dependent. Formally,

H_0: The two qualitative variables are independent.

H_A: The two qualitative variables are dependent.

Since the criteria upon which we classify the data in the introductory case are Age Group and Brand Name, we write the competing hypotheses as

H_0: Age Group and Brand Name are independent.

H_A: Age Group and Brand Name are dependent.

Table 11.6 reproduces Table 11.1 of the introductory case. The variable Age Group has two possible categories: (1) Under 35 years and (2) 35 years or older. The variable Brand Name has three possible categories: (1) Under Armour, (2) Nike, and (3) Adidas. Each cell in this table represents an observed frequency o_{ij}, where the subscript ij refers to the ith row and the jth column. Thus, o_{13} refers to the cell in the first row and the third column. Here, $o_{13} = 90$, or, equivalently, 90 customers under 35 years of age purchased an Adidas product.

TABLE 11.6 Purchases of Compression Garments Based on Age and Brand Name

Age Group	Brand Name		
	Under Armour	Nike	Adidas
Under 35 years	174	132	90
35 years or older	54	72	78

We will use the independence assumption postulated under the null hypothesis to derive an expected frequency for each cell from the sample data. In other words, we first estimate values as if no relationship exists between the age of a customer and the brand name of the clothing purchased. Then we will compare these expected frequencies with the observed values to compute the value of the test statistic.

Calculating Expected Frequencies

For ease of exposition, we let events A_1 and A_2 represent "Under 35 years" and "35 years or older," respectively; events B_1, B_2, and B_3 stand for Under Armour, Nike, and Adidas, respectively. We then sum the frequencies for each column and row. For instance, the sum of the frequencies for Event A_1 is 396; this is obtained by summing the values in row A_1: 174, 132, and 90. Totals for the other rows and columns are shown in Table 11.7.

TABLE 11.7 Row and Column Totals

Age Group	Brand Name			Row Total
	B_1	B_2	B_3	
A_1	e_{11}	e_{12}	e_{13}	396
A_2	e_{21}	e_{22}	e_{23}	204
Column Total	228	204	168	600

Our goal is to calculate the expected frequency e_{ij} for each cell, where again the subscript ij refers to the ith row and the jth column. Thus, e_{13} refers to the cell in the first row and the third column, or the expected number of customers who are under 35 years of age and purchase an Adidas product.

Before we can arrive at the expected frequencies, we first calculate marginal row probabilities (the proportion of people under 35 years of age and those 35 years old or older) and marginal column probabilities (the proportion of people purchasing from each brand name). We calculate a marginal row (column) probability by dividing the row (column) sum by the total sample size:

Marginal Row Probabilities:

$$P(A_1) = \frac{396}{600} \quad \text{and} \quad P(A_2) = \frac{204}{600}$$

Marginal Column Probabilities:

$$P(B_1) = \frac{228}{600}, \quad P(B_2) = \frac{204}{600}, \quad \text{and } P(B_3) = \frac{168}{600}$$

We can now calculate each cell probability by applying the multiplication rule for independent events from Chapter 4. That is, if two events are independent, say events A_1 and B_1 (our assumption under the null hypothesis), then their joint probability is

$$P(A_1 \cap B_1) = P(A_1)P(B_1) = \left(\frac{396}{600}\right)\left(\frac{228}{600}\right) = 0.2508.$$

Multiplying this joint probability by the sample size yields the expected frequency for e_{11}; that is, the expected number of customers who are under 35 years of age and purchase an Under Armour product is

$$e_{11} = 600(0.2508) = 150.48.$$

CALCULATING EXPECTED FREQUENCIES FOR A TEST FOR INDEPENDENCE

We use the following general formula to calculate the expected frequencies for each cell in a contingency table:

$$e_{ij} = \frac{(\text{Row } i \text{ total})(\text{Column } j \text{ total})}{\text{Sample Size}},$$

where e_{ij} is the expected frequency for each cell in a contingency table, and the subscript ij refers to the ith row and the jth column.

Applying the formula, we calculate all expected frequencies as

$$e_{11} = \frac{(396)(228)}{600} = 150.48 \quad e_{12} = \frac{(396)(204)}{600} = 134.64 \quad e_{13} = \frac{(396)(168)}{600} = 110.88$$

$$e_{21} = \frac{(204)(228)}{600} = 77.52 \quad e_{22} = \frac{(204)(204)}{600} = 69.36 \quad e_{23} = \frac{(204)(168)}{600} = 57.12$$

Table 11.8 shows the expected frequency e_{ij} for each cell. In order to satisfy subsequent assumptions, each expected frequency e_{ij} *must equal five or more*. This condition is satisfied here. As we saw in Example 11.5, it may be necessary to combine two or more rows or columns to achieve this result in other applications.

TABLE 11.8 Expected Frequencies for Contingency Table

Age Group	Brand Name			Row Total
	B_1	B_2	B_3	
A_1	150.48	134.64	110.88	396
A_2	77.52	69.36	57.12	204
Column Total	228	204	168	600

When conducting a test for independence, we calculate the value of the chi-square test statistic χ^2_{df}. Analogous to the discussion in Section 11.2, χ^2_{df} measures how much the observed frequencies deviate from the expected frequencies. The smallest value that χ^2_{df} can assume is zero—this occurs when each observed frequency equals its expected frequency. Thus, a test for independence is also implemented as a *right-tailed test*.

TEST STATISTIC FOR A TEST FOR INDEPENDENCE

For a test for independence applied to a contingency table with r rows and c columns, the value of the test statistic is calculated as

$$\chi^2_{df} = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

where $df = (r - 1)(c - 1)$, o_{ij} is the observed frequency for row i and column j, and e_{ij} is the expected frequency for row i and column j.

Note: This test is valid when the expected frequency for each cell is five or more.

With two rows and three columns in the contingency table, degrees of freedom are calculated as $df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$. We apply the formula to compute the value of the test statistic as

$$\chi^2_2 = \frac{(174 - 150.48)^2}{150.48} + \frac{(132 - 134.64)^2}{134.64} + \frac{(90 - 110.88)^2}{110.88}$$
$$+ \frac{(54 - 77.52)^2}{77.52} + \frac{(72 - 69.36)^2}{69.36} + \frac{(78 - 57.12)^2}{57.12}$$
$$= 3.6762 + 0.0518 + 3.9319 + 7.1361 + 0.1005 + 7.6326 = 22.529.$$

For $df = 2$, we calculate the p-value as $P(\chi^2_2 \geq 22.529)$. From Table 11.4, we see that 22.529 is greater than 10.597, implying that the p-value is less than 0.005. As discussed in Section 11.2, we can use Excel's CHISQ.DIST.RT function to find the exact p-value as $1.282 \times 10^{-5} = 0$ (approximately). Since the p-value is less than 0.05, we reject H_0. At the 5% significance level, we conclude that the two qualitative variables are dependent; that is, there is a relationship between the age of a customer and the brand name purchased.

SYNOPSIS OF INTRODUCTORY CASE



©Gallo Images/Alamy Stock Photo

Under Armour pioneered clothing in the compression-gear market. Compression garments are meant to keep moisture away from a wearer's body during athletic activities in warm and cool weather. Wicking moisture is a secondary characteristic of compression gear. Wicking materials were in widespread use before Under Armour and are used in most noncompression athletic wear. The characteristic that defines compression gear is that it is tight to help compress muscles, which supposedly helps them work better, avoid injury, and recover faster. Under Armour has experienced exponential growth since the firm went public in November 2005 (*USA Today*, June 16, 2010); however, Nike and Adidas have aggressively entered the compression-gear market as well. An analysis was conducted to examine whether the age of the customer matters when making a purchase in the compression-gear market. This information is relevant not only for Under Armour and how the firm may focus its advertising efforts, but also to competitors and retailers in this market. Data were collected on 600 recent purchases in the compression-gear market; the data were then cross-classified by age group and brand name. A test for independence was conducted at the 5% significance level. The results suggest that a customer's age and the brand name purchased are related to one another. Given that age influences the brand name purchased, it is not surprising that Under Armour signed NFL quarterback Tom Brady (http://cnbc.com, October 6, 2010) to endorse its products, a move likely to attract a younger customer. Brady had spent most of his career with Nike before breaking away to go with Under Armour.

EXAMPLE 11.6

In general, Latinos and Caucasians use social media networks equally, but there are some differences in their preferences for specific social media sites (www.pewresearch.org, February 5, 2015). In particular, Instagram is more popular among Latinos while Pinterest is more popular among Caucasians. Kate Dawson, a junior in college, decides to test if similar differences exist among students on her campus. She collects data on 400 students cross-classified by Race (Latinos versus Caucasians) and Social Media Preference (Instagram versus Pinterest). The results are shown in Table 11.9. At the 10% significance level, determine whether the sample data support racial differences in social media preferences.

TABLE 11.9 Social Media Preference by Race

Race	Social Media Preference		Row Total
	Instagram	Pinterest	
Latinos	50	60	110
Caucasians	120	170	290
Column Total	170	230	400

SOLUTION: In order to determine whether social media preference depends on race, we specify the competing hypotheses as

H_0: Race and Social Media Preference are independent.

H_A: Race and Social Media Preference are dependent.

The value of the test statistic is calculated as $\chi^2_{df} = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$. Table 11.9 provides each cell's observed frequency o_{ij}, so before applying the formula, we first calculate each cell's expected frequency e_{ij}:

$$e_{11} = \frac{(110)(170)}{400} = 46.75 \qquad e_{12} = \frac{(110)(230)}{400} = 63.25$$

$$e_{21} = \frac{(290)(170)}{400} = 123.25 \qquad e_{22} = \frac{(290)(230)}{400} = 166.75$$

With two rows and two columns in the contingency table, the degrees of freedom are calculated as $df = (r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$. The value of the test statistic is calculated as

$$\chi^2_1 = \frac{(50 - 46.75)^2}{46.75} + \frac{(60 - 63.25)^2}{63.25} + \frac{(120 - 123.25)^2}{123.25} + \frac{(170 - 166.75)^2}{166.75}$$

$$= 0.2259 + 0.1670 + 0.0857 + 0.0633 = 0.542.$$

Using Excel's CHISQ.DIST.RT function, we find that the exact p-value, $P(\chi^2_1 \geq 0.542)$, is equal to 0.462. Since the p-value is greater than 0.10, we do not reject H_0. At the 10% significance level, the sample data do not support racial differences in social media preferences.

EXERCISES 11.3

Mechanics

31. Given the following contingency table, conduct a test for independence at the 5% significance level.

Variable B	Variable A	
	1	2
1	23	47
2	32	53

32. Given the following contingency table, conduct a test for independence at the 1% significance level.

Variable B	Variable A			
	1	2	3	4
1	120	112	100	110
2	127	115	120	124
3	118	115	110	124

Applications

33. According to an online survey by Harris Interactive for job site CareerBuilder.com (InformationWeek.com, September 27, 2007), more than half of IT workers say they have fallen asleep at work. The same is also true for government workers. Assume that the following contingency table is representative of the survey results.

Slept on the Job?	Job Category	
	IT Professional	Government Professional
Yes	155	256
No	145	144

a. Specify the competing hypotheses to determine whether sleeping on the job is associated with job category.
b. Calculate the value of the test statistic.
c. Find the p-value.
d. At the 5% significance level, can you conclude that sleeping on the job depends on job category?

34. A market researcher for an automobile company suspects differences in preferred color between male and female buyers. Advertisements targeted to different groups should take such differences into account, if they exist. The researcher examines the most recent sales information of a particular car that comes in three colors.

Color	Sex of Automobile Buyer	
	Male	Female
Silver	470	280
Black	535	285
Red	495	350

a. Specify the competing hypotheses to determine whether color preference depends on the automobile buyer's sex.

b. Calculate the value of the test statistic and the *p*-value.
c. At the 1% significance level, does your conclusion suggest that the company should target advertisements differently for males versus females? Explain.

35. The following sample data reflect shipments received by a large firm from three different vendors and the quality of those shipments.

Vendor	Defective	Acceptable
1	14	112
2	10	70
3	22	150

a. Specify the competing hypotheses to determine whether quality is associated with the source of the shipments.
b. Conduct the test at the 1% significance level.
c. Should the firm be concerned about the source of the shipments? Explain.

36. A marketing agency would like to determine if there is a relationship between union membership and type of vehicle owned (domestic or foreign brand). The goal is to develop targeted advertising campaigns for particular vehicle brands likely to appeal to specific groups of customers. A survey of 500 potential customers revealed the following results.

	Union Member	Not Union Member
Domestic brand	133	147
Foreign brand	67	153

a. Specify the competing hypotheses to determine whether vehicle brand (domestic, foreign) is associated with union membership.
b. Conduct the test at the 10% significance level. What is your conclusion?
c. Is the conclusion reached in part (b) sensitive to the choice of the significance level?

37. The quality manager believes there may be a relationship between the experience level of an inspector and whether a product passes or fails inspection. Inspection records were reviewed for 630 units of a particular product, and the number of units which passed and failed inspection was determined based on three inspector experience levels. The results are shown in the following table.

	Experience Level		
Decision	Low (< 2 years)	Medium (2–8 years)	High (> 8 years)
Pass	152	287	103
Fail	16	46	26

a. Specify the competing hypotheses to determine whether the inspector pass/fail decision depends on experience level.
b. Calculate the value of the test statistic.
c. Find the *p*-value.
d. At the 5% significance level, what is your conclusion? Does your conclusion change at the 1% significance level?

38. According to a 2008 survey by the Pew Research Center, people in China are highly satisfied with their roaring economy and the direction of their nation (*USA Today*, July 22, 2008). Eighty-six percent of those who were surveyed expressed positive views of the way China is progressing and described the economic situation as good. A political analyst wants to know if this optimism among the Chinese depends on age. In an independent survey of 280 Chinese residents, the respondents are asked how happy they are with the direction that their country is taking. Their responses are tabulated in the following table.

Age	Very Happy	Somewhat Happy	Not Happy
20 up to 40	23	50	18
40 up to 60	51	38	16
60 and older	19	45	20

a. Set up the hypotheses to test the claim that optimism regarding China's direction depends on the age of the respondent.
b. Calculate the value of the test statistic.
c. Find the *p*-value.
d. At the 1% level of significance, can we infer that optimism among the Chinese is dependent on age?

39. A study by the Massachusetts Community & Banking Council found that blacks, and, to a lesser extent, Latinos, remain largely unable to borrow money at the same interest rate as whites (*The Boston Globe*, February 28, 2008). The following contingency table shows representative data for the city of Boston, cross-classified by race and type of interest rate received:

	Type of Interest Rate on Loan	
Race	High Interest Rate	Low Interest Rate
Black	553	480
Latino	265	324
White	491	3,701

At the 5% significance level, do the data indicate that the interest rate received on a loan is dependent on race? Provide the details.

40. Founded in February 2004, Facebook is a social utility that helps people communicate with their friends and family. In a survey of 3,000 Facebook users, researchers looked at why Facebook users break up in a relationship (*The Wall Street Journal*, November 27–28, 2010).

Reasons for Breakup	Sex of Respondent	
	Percentage of Men	Percentage of Women
Nonapproval	3	4
Distance	21	16
Cheating	18	22
Lost Interest	28	26
Other	30	32

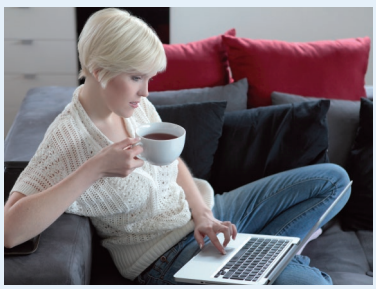
Source: Internal survey of 3,000 Facebook users.

Suppose the survey consisted of 1,800 men and 1,200 women. Use the data to determine whether the reasons for breakup depend on one's sex at the 1% significance level. Provide the details.

WRITING WITH STATISTICS



©JGI/Blend Images LLC



©STOCK4B-RF/Getty Images

The phenomenon of online dating has made it as likely for would-be couples to meet via e-mail or other virtual matchmaking services as through friends and family (*CNN*, February 6, 2012). Studies that have looked at gender differences in mate selection have found that women put greater emphasis on the race and financial stability of a partner, while men mostly look for physical attractiveness. Survey results reported in *USA Today* (February 2, 2012) showed that 13% of women and 8% of men want their partner to be of the same ethnic background. The same survey also reported that 36% of women and 13% of men would like to meet someone who makes as much money as they do.

Anka Wilder, working for a small matchmaking service in Cincinnati, Ohio, wants to know if a similar pattern also exists with her customers. She has access to the preferences of 160 women and 120 men customers. In this sample, she finds that 28 women and 12 men want their partner to be of the same ethnicity. Also, 50 women and 10 men want their partner to make as much money as they do.

Anka wants to use this sample information to:

1. Determine whether the proportion of women who want their partner to be of the same ethnic background is greater than that of men.
2. Determine whether the proportion of women who want their partner to make as much money as they do is more than 20 percentage points greater than that of men.

Sample Report—Online Dating Preferences

With the advent of the Internet, there has been a surge in online dating services that connect individuals with similar interests, religions, and cultural backgrounds for personal relationships. In 1992, when the Internet was still in its infancy, less than 1% of Americans met their partners through online dating services. By 2009, about 22% of heterosexual couples and 61% of same-sex couples reported meeting online (CNN, February 6, 2012). A survey suggested that a higher proportion of women than men would like to meet someone with a similar ethnic background. Also, the difference between the proportion of women and men who would like to meet someone who makes as much money as they do is greater than 20%.

A couple of hypothesis tests were performed to determine if similar gender differences existed for online dating customers in Cincinnati, Ohio. The sample consisted of responses from 160 women and 120 men. The summary of the test results is presented in Table 11.A.

Table 11.A Test Statistics and *p*-values for Hypothesis Tests

Hypotheses	Test Statistic	*p*-value
H_0: $p_1 - p_2 \le 0$ H_A: $p_1 - p_2 > 0$	$z = \dfrac{0.175 - 0.10}{\sqrt{0.1429(1 - 0.1429)\left(\frac{1}{160} + \frac{1}{120}\right)}} = 1.77$	0.0384
H_0: $p_1 - p_2 \le 0.20$ H_A: $p_1 - p_2 > 0.20$	$z = \dfrac{0.3125 - 0.0833 - 0.20}{\sqrt{\frac{0.3125(1 - 0.3125)}{160} + \frac{0.0833(1 - 0.0833)}{120}}} = 0.66$	0.2546

First, it was tested if the proportion of women, denoted p_1, who want their partner to be of the same ethnicity is greater than that of men, denoted p_2. It was found that 28 out of 160 women valued this trait, yielding a sample proportion of $\bar{p}_1 = 28/160 = 0.175$; a similar proportion for men was calculated as $\bar{p}_2 = 12/120 = 0.10$. The first row of Table 11.A shows the competing hypotheses, the value of the test statistic, and the *p*-value for this test. At the 5% significance level, the proportion of women who want the same ethnicity was greater than that of men. In the second test, p_1 and p_2 denoted the proportion of women and men, respectively, who would like their partner to make as much money as they do; here $\bar{p}_1 = 50/160 = 0.3125$ and $\bar{p}_2 = 10/120 = 0.0833$. The second row of Table 11.A shows the competing hypotheses, the value of the test statistic, and the *p*-value for this test. At the 5% significance level, the proportion of women who want their partner to make as much income as they do is not more than 20 percentage points greater than that of men. Online dating is a relatively new market and any such information is important for individuals looking for relationships as well as for service providers.

CONCEPTUAL REVIEW

LO 11.1 Make inferences about the difference between two population proportions based on independent sampling.

A $100(1 - \alpha)\%$ confidence interval for the difference between two population proportions $p_1 - p_2$ is given by $(\bar{p}_1 - \bar{p}_2) \pm z_{a/2}\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$. When conducting hypothesis tests about the difference between two proportions $p_1 - p_2$, the value of the test statistic is calculated as

- $z = \dfrac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$, if the hypothesized difference d_0 between p_1 and p_2 is zero. The pooled sample proportion is $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$.
- $z = \dfrac{(\bar{p}_1 - \bar{p}_2) - d_0}{\sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}}$, if the hypothesized difference d_0 between p_1 and p_2 is not zero.

LO 11.2 Discuss features of the χ^2 distribution.

The **χ^2 distribution** is characterized by a family of distributions, where each distribution depends on its particular degrees of freedom *df;* thus, it is common to refer to it as the χ^2_{df} distribution. It is positively skewed with values ranging from zero to infinity. As *df* grow larger, the χ^2_{df} distribution tends to the normal distribution.

LO 11.3 Conduct a goodness-of-fit test for a multinomial experiment.

A **multinomial experiment** consists of a series of n independent and identical trials such that on each trial there are k possible outcomes, called categories; the probability p_i associated with the ith category remains the same; and the sum of the probabilities is one.

A **goodness-of-fit test** is conducted to determine if the population proportions differ from some predetermined (hypothesized) values. The value of the test statistic is calculated as $\chi^2_{df} = \Sigma\frac{(o_i - e_i)^2}{e_i}$, where $df = k - 1$, o_i is the observed frequency for category i, $e_i = np_i$ is the expected frequency for category i, and n is the number of observations. The test is valid when the expected frequency for each category is five or more. This test is always implemented as a right-tailed test.

LO 11.4 Conduct a test of independence.

A goodness-of-fit test examines a single qualitative variable, whereas a **test of independence**, also called a **chi-square test of a contingency table**, analyzes the relationship between two qualitative variables defined on the same population. A contingency table shows frequencies for two qualitative variables, x and y, where each cell of the table represents a mutually exclusive combination of the pair of x and y values.

In order to determine whether or not the two variables are related, we again compare observed frequencies with expected frequencies. The expected frequency for each cell, e_{ij}, is calculated as $e_{ij} = \frac{(\text{Row } i \text{ total})(\text{Column } j \text{ total})}{\text{Sample Size}}$, where the subscript ij refers to the ith row and the jth column of the contingency table. The value of the chi-square test statistic is calculated as $\chi^2_{df} = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$, where o_{ij} is the observed frequency. The degrees of freedom are calculated as $(r - 1)(c - 1)$, where r and c refer to the number of rows and columns, respectively, in the contingency table. The test for independence is always implemented as a right-tailed test and is valid when the expected frequency for each cell is five or more.

ADDITIONAL EXERCISES AND CASE STUDIES

Exercises

41. A Health of Boston report suggests that 14% of female residents suffer from asthma as opposed to 6% of males (*The Boston Globe*, August 16, 2010). Suppose 250 females and 200 males responded to the study.
    a. Develop the appropriate null and alternative hypotheses to test whether the proportion of females suffering from asthma is greater than the proportion of males.
    b. Calculate the value of the test statistic and the p-value.
    c. At the 5% significance level, what is the conclusion? Do the data suggest that females suffer more from asthma than males?

42. Fresh numbers from the U.S. Department of Transportation suggest that fewer flights in the United States arrive on time than before. The explanations offered for the lackluster performance are understaffed airlines, a high volume of travelers, and overtaxed air traffic control. A transportation analyst is interested in comparing the performance at two major international airports, namely Kennedy International (JFK) in New York and O'Hare International in Chicago. She finds that 70% of the flights were on time at JFK compared with 63% at O'Hare. Suppose these proportions were based on 200 flights at each of these two airports. The analyst believes that the proportion of on-time flights at JFK is more than 5 percentage points higher than that of O'Hare.
    a. Develop the competing hypotheses to test the transportation analyst's belief.
    b. Calculate the value of the test statistic and the p-value.
    c. At the 5% significance level, do the data support the transportation analyst's belief? Explain.

43. Depression engulfs millions of Americans every day. A federal study reported that 10.9% of adults

aged 18–24 identified with some level of depression versus 6.8% of adults aged 65 or older (*The Boston Globe*, October 18, 2010). Suppose 1,000 young adults (18–24 years old) and 1,000 older adults (65 years old and older) responded to the study.

a. Develop the appropriate null and alternative hypotheses to test whether the proportion of young adults suffering from depression is greater than the proportion of older adults suffering from depression.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 5% significance level, do the sample data suggest that young adults suffer more from depression than older adults?

44. The following table lists the market shares of the four firms in a particular industry in 2010 and total sales (in $ billions) for each firm in 2011.

Firm	Market Share in 2010	Total Sales in 2011
1	0.40	200
2	0.30	180
3	0.20	100
4	0.10	70

a. Specify the competing hypotheses to test whether the market shares in 2010 are not valid in 2011.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 1% significance level, do the sample data suggest that the market shares changed from 2010 to 2011?

45. A study suggests that airlines have increased restrictions on cheap fares by raising overnight requirements (*The Wall Street Journal*, August 19, 2008). This would force business travelers to pay more for their flights, since they tend to need the most flexibility and want to be home on weekends. Eight months ago, the overnight stay requirements were as follows:

One night	Two nights	Three nights	Saturday night
37%	17%	28%	18%

A recent sample of 644 flights found the following restrictions:

One night	Two nights	Three nights	Saturday night
117	137	298	92

a. Specify the competing hypotheses to test whether the proportions cited by the study have changed.
b. Calculate the value of the test statistic.
c. At the 5% significance level, what is the conclusion to the hypothesis test? Interpret your results.

46. Although founded only in 2004, Facebook has nearly 2 billion active users, of which 50% log on to Facebook on any given day. In a survey by Facebook, young users were asked about their preference for delivering the news about breaking up a relationship (*The Wall Street Journal*, November 27–28, 2010). One of the shocking results was that only 47% of users preferred to break the news in person. A researcher decides to verify the survey results of Facebook by taking her own sample of 200 young Facebook users. The preference percentages from Facebook and the researcher's survey are presented in the following table.

Delivery Method	Facebook Results	Researcher's Results
In Person	47%	55%
Phone	30%	25%
E-mail	4%	6%
Facebook	5%	5%
Instant Message	14%	9%

At the 5% level of significance, test if the researcher's results are inconsistent with the survey results conducted by Facebook. Provide the details.

47. A local TV station claims that 60% of people support Candidate A, 30% support Candidate B, and 10% support Candidate C. A survey of 500 registered voters is taken. The accompanying table indicates how they are likely to vote.

Candidate A	Candidate B	Candidate C
350	125	25

a. Specify the competing hypotheses to test whether the TV station's claim can be rejected by the data.
b. Test the hypothesis at the 1% significance level.

48. A study in the *Journal of the American Medical Association* (February 20, 2008) found that patients who go into cardiac arrest while in the hospital are more likely to die if it happens after 11 pm. The study investigated 58,593 cardiac arrests during the day or evening. Of those, 11,604 survived to leave the hospital. There were 28,155 cardiac arrests during the shift that began at 11 pm, commonly referred to as the graveyard shift. Of those, 4,139

survived for discharge. The following contingency table summarizes the results of the study:

Shift	Survived for Discharge	Did Not Survive for Discharge
Day or Evening Shift	11,604	46,989
Graveyard Shift	4,139	24,016

a. Specify the competing hypotheses to determine whether a patient's survival depends on the time at which he/she experiences cardiac arrest.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 1% significance level, is the timing of when a cardiac arrest occurs dependent on whether or not the patient survives for discharge? Given your answer, what type of recommendations might you give to hospitals?

49. An analyst is trying to determine whether the prices of certain stocks on the NASDAQ depend on the industry to which they belong. She examines four industries and, within each industry, categorizes each stock according to its price (high-priced, average-priced, low-priced).

Stock Price	Industry			
	I	II	III	IV
High	16	8	10	14
Average	18	16	10	12
Low	7	8	4	9

a. Specify the competing hypotheses to determine whether stock price depends on the industry.
b. Calculate the value of the test statistic and the *p*-value.
c. At the 1% significance level, what can the analyst conclude?

50. The operations manager at ElectroTech, an electronics manufacturing company, believes that workers on particular shifts may be more likely to phone in "sick" than those on other shifts. To test this belief, she has compiled the following table containing frequencies based on work shift and days absent over the past year.

	First Shift	Second Shift	Third Shift
0–2 days absent	44	20	10
3–6 days absent	38	25	12
7–10 days absent	14	9	13
11 or more days absent	4	6	5

a. Specify the competing hypotheses to determine whether days absent depend on work shift.
b. Calculate the value of the test statistic and the *p*-value.
c. What is your conclusion at the 5% significance level? What about the 1% significance level?

51. A poll asked 3,228 Americans aged 16 to 21 whether they are likely to serve in the U.S. military. The following table, cross-classified by a person's sex and race, reports those who responded that are likely or very likely to serve in the active-duty military.

Sex	Race		
	Hispanic	Black	White
Male	1,098	678	549
Female	484	355	64

Source: Defense Human Resources Activity telephone poll of 3,228 Americans conducted October through December 2005.

a. State the competing hypotheses to test whether a person's sex and race are dependent when making a choice to serve in the military.
b. Conduct the test at the 5% significance level.

52. Given a shaky economy and high heating costs, more and more households are struggling to pay utility bills (*The Wall Street Journal*, February 14, 2008). Particularly hard hit are households with homes heated with propane or heating oil. Many of these households are spending twice as much to stay warm this winter compared to those who heat with natural gas or electricity. A representative sample of 500 households was taken to investigate if the type of heating influences whether or not a household is delinquent in paying its utility bill. The following table reports the results.

Delinquent in Payment?	Type of Heating			
	Natural Gas	Electricity	Heating Oil	Propane
Yes	50	20	15	10
No	240	130	20	15

At the 5% significance level, test whether the type of heating influences a household's delinquency in payment. Interpret your results.

53. An automotive parts company has been besieged with poor publicity over the past few years due to several highly publicized product recalls that have tarnished its public image. This has prompted a series of quality improvement initiatives. Currently, the marketing manager would like to determine if these initiatives have been successful in changing public perception about the company. The accompanying table shows the results of two

surveys, each of 600 randomly selected. Survey 1 was conducted prior to the quality initiatives. Survey 2 was conducted after the quality initiatives were implemented and publicized.

	Public Perception		
	Negative	Neutral	Positive
Survey 1	324	180	96
Survey 2	246	146	208

a. State the appropriate null and alternative hypotheses to test if the public perception has changed since the quality initiatives have been implemented.
b. Make a conclusion at the 1% significance level.

54. Color coding is often used in manufacturing operations to display production status or to identify/prioritize materials. For example, suppose "green" status indicates that an assembly line is operating normally, "yellow" indicates it is down waiting on personnel for set up or repair, "blue" indicates it is down waiting on materials to be delivered, and "red" indicates an emergency condition. Management has set realistic goals whereby the assembly line should be operating normally 80% of the time, waiting on personnel 9% of the time, waiting on materials 9% of the time, and in an emergency condition 2% of the time. Based on 250 recent status records, the status was green 185 times, yellow 24 times, blue 32 times, and red 9 times.
    a. State the appropriate null and alternative hypotheses to test if the proportions of assembly line statuses differ from the goals set by management.
    b. Calculate the value of the test statistic and the *p*-value.
    c. Are management's goals being met at $\alpha = 0.05$? Will your conclusion change at $\alpha = 0.01$?

55. Many parents have turned to St. John's wort, an herbal remedy, to treat their children with attention deficit hyperactivity disorder (ADHD). *The Journal of the American Medical Association* (June 11, 2008) published an article that explored the herb's effectiveness. Children with ADHD were randomly assigned to take either St. John's wort capsules or placebos. The accompanying contingency table broadly reflects the results found in the study.

Treatment	Effect on ADHD	
	No Change in ADHD	Improvement in ADHD
St. John's wort	12	15
Placebo	14	13

At the 5% significance level, do the data indicate that there is a relationship between the type of treatment and the condition of children with ADHD?

56. The human resources department would like to consolidate the current set of retirement plan options offered to specific employee pay groups into a single plan for all pay groups (salaried, hourly, or piecework). A sample of 585 employees of various pay groups were asked which of three potential plans they preferred (A, B, or C). The results are shown in the accompanying table. The human resources department is hoping to conclude that the retirement plan preferred by the majority of employees is independent of pay group, in order to avoid the impression that the preferred plan may favor a particular group.

Preferred Plan	Employee Pay Group		
	Salaried	Hourly	Piecework
A	78	98	37
B	121	95	30
C	51	57	18

    a. Specify the competing hypotheses to determine whether the preferred retirement plan depends on employee pay group.
    b. Calculate the value of the test statistic and the *p*-value.
    c. What is your conclusion at the 10% significance level? What about the 5% significance level?

57. FILE ***Degrees***. The value of a college degree is greater than it has been in nearly half a century, at least when compared to the prospect of not getting a degree (www.pewresearch.org, January 28, 2014). Due to this fact, more and more people are obtaining college degrees, despite the soaring costs. The accompanying table shows the proportions of college degrees awarded in 2010 by colleges and universities, categorized by a graduate's race and ethnicity. The race and ethnicity of 500 recent graduates are recorded and shown in the last column of the table.

Race/Ethnicity	2010 Proportions	Recent Numbers
White	0.73	350
Black	0.10	50
Hispanic	0.09	60
Asian	0.08	40

Source: The 2010 proportions come from Table 300 in *Digest of Education Statistics*, 2011.

At the 5% significance level, test if the proportions have changed since 2010.

CASE STUDIES

CASE STUDY 11.1 According to a recent study, cell phones are the main medium for teenagers to stay connected with friends and family (*CNN*, March 19, 2012). These days text messaging followed by cell calling have become an integral part of life for teenagers. It is estimated that 90% of older teens (aged 14–17) and 60% of younger teens (aged 12–13) own a cell phone. Moreover, one in four teenagers who owns a cell phone uses a smartphone. Susan Alder works at an AT&T store in the campus town of Ames, Iowa. She has been tasked to determine if the patterns reported in this study apply to Ames teens. She surveys 120 older teens and 90 younger teens. In her sample, 100 older teens and 48 younger teens own a cell phone. She also finds that 26 older teens and 9 younger teens use a smartphone.

In a report use the above information to:

1. Determine at the 5% significance level whether the proportion of older teens who own a cell phone is more than 20 percentage points greater than that of younger teens.
2. Of the teens who own a cell phone, determine at the 5% significance level whether the proportion of older teens that use a smartphone is greater than that of younger teens.

CASE STUDY 11.2 A detailed study of Americans' religious beliefs and practices by the Pew Forum on Religion & Public Life revealed that religion is quite important in an individual's life (*Boston Globe* , June 24, 2008). The second column of the accompanying table reports the proportion of Americans who feel a certain way about religion. The study also concludes that Massachusetts residents are the least likely to say that they are religious. In order to test this claim, assume 400 randomly selected Massachusetts residents are asked about the importance of religion in his/her life. The results of this survey are shown in the last column of the accompanying table.

Data for Case Study 11.2 Importance of Religion, U.S. versus Massachusetts

Importance of Religion	U.S. Results	Responses of Massachusetts Residents
Very important	0.58	160
Somewhat important	0.25	140
Not too important	0.15	96
Don't know	0.02	4

In a report, use the sample information to:

1. Determine whether Massachusetts residents' religious beliefs differ from those based on the study of all Americans at a 5% significance level.
2. Discuss whether you would expect to find the same conclusions if you conducted a similar test for the state of Utah or states in the Southern Belt of the United States.

CASE STUDY 11.3 A University of Utah study examined 7,925 severely obese adults who had gastric bypass surgery and an identical number of people who did not have the surgery (*Boston Globe*, August 23, 2007). The study wanted to investigate whether losing weight through stomach surgery prolonged the lives of severely obese patients, thereby reducing their deaths from heart disease, cancer, and diabetes.

Over the course of the study, 534 of the participants died. Of those who died, the cause of death was classified as either a disease death (disease deaths include heart disease, cancer, and diabetes) or a nondisease death (nondisease deaths include suicide or accident). The following contingency table summarizes the study's findings:

Data for Case Study 11.3 Deaths Cross-Classified by Cause and Method of Losing Weight

	Method of Losing Weight	
Cause of Death	**No Surgery**	**Surgery**
Death from disease	285	150
Death from nondisease	36	63

In a report, use the sample information to:

1. Determine at the 5% significance level whether the cause of death is related to the method of losing weight.
2. Discuss how the findings of the test used in question 1 might be used by those in the health industry.

APPENDIX 11.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, JMP, and R. Where a data file is specified, copy and paste it into the relevant software spreadsheet prior to following the commands. When importing data into R, use the menu-driven option: File > Import Dataset > From Excel.

MINITAB

Testing $p_1 - p_2$

A. (Replicating Example 11.3) From the menu choose **Stat > Basic Statistics > 2 Proportions**. Choose **Summarized data**. Under **Sample 1**, enter 5400 for **Number of events** and 6000 for **Number of trials**. Under **Sample 2**, enter 8600 for **Number of events** and 10000 for **Number of trials**.

B. Choose **Options**. After **Alternative hypothesis**, select "Difference > hypothesized difference." After **Test method**, select "Use the pooled estimate for the proportion."

Goodness-of-Fit Test

A. (Replicating Example 11.5) From the menu, choose **Stat > Tables > Chi-Square Goodness-of-Fit Test (One Variable)**.

FILE *Share*

B. Choose **Observed counts** and then select Number. Under **Test**, select **Proportions specified by historical counts**, and then select Share. Choose **Results** and select **Display test results**.

Test of Independence

FILE *Social_Media*

A. (Replicating Example 11.6) From the menu, choose **Stat > Tables > Cross Tabulation and Chi-Square**.

B. Select "Summarized data in a two-way table." Under **Columns containing the table**, select Instagram and Pinterest. Choose Chi-Square and select **Chi-square test**.

SPSS

Goodness-of-Fit Test

A. (Replicating Example 11.5) From the menu, choose **Data > Weight Cases.** Select **Weight cases by**, and under **Frequency Variable**, select Number.

FILE *Share*

B. Select **Analyze > Nonparametric Tests > Legacy Dialogs > Chi-square.**

C. Under **Test Variable List**, select Firm. Under **Expected Values**, select **Values**, and **Add** 0.40, 0.32, 0.24, and 0.04.

Test of Independence

FILE
Social_Media

A. (Replicating Example 11.6) In order to conduct this test in SPSS, the data need to be reconfigured. Label Columns 1, 2, and 3 as "Race," "Media," and "Frequency," respectively. In the first row, enter Latinos, Instagram, 50; in the second row, enter Latinos, Pinterest, 60; in the third row enter Caucasians, Instagram, 120; and in the last row enter Caucasians, Pinterest, 160.

B. From the menu, choose **Data > Weight Cases**. Select **Weight cases by**, and under **Frequency Variable**, select Frequency.

C. From the menu, select **Analyze > Descriptive Statistics > Crosstabs**.

D. Under **Rows**, select Race, and under **Columns**, select Media. Choose **Statistics**, check **Chi-square**.

JMP

Test of Independence

FILE
Social_Media

A. (Replicating Example 11.6) In order to conduct this test in JMP, the data need to be reconfigured. Follow Step A under the SPSS instructions for Test of Independence.

B. From the menu, select **Analyze > Fit Y by X**.

C. Under **Select Columns**, select Race, and then under **Cast Selected Columns into Roles**, select **Y, Response**. Under **Select Columns**, select Media and then under **Cast Selected Columns into Roles**, select **X, Factor**. Under **Select Columns**, select Frequency, and then under **Cast Selected Columns into Roles**, select **Freq**.

R

Goodness-of-Fit Test

FILE
Share

(Replicating Example 11.5) Use the **chisq.test** function to calculate the value of the test statistic and the *p*-value. For options within the **chisq.test** function, use *p* to indicate the location of the hypothesized proportions.

```
> chisq.test(Share$'Customers', p=Share$'Share')
```

Test of Independence

FILE
Social Media

(Replicating Example 11.6) Use the **chisq.test** function to calculate the value of the test statistic and the *p*-value. Within the **chisq.test** function, indicate that the relevant data are in columns 2 and 3 of the data frame; also set the option *correct* to FALSE which will turn off the continuity correction factor.

```
> chisq.test(Social_Media[ , 2:3], correct=FALSE)
```

12 Basics of Regression Analysis

Learning Objectives

After reading this chapter you should be able to:

LO **12.1** Estimate and interpret a simple linear regression model.

LO **12.2** Estimate and interpret a multiple linear regression model.

LO **12.3** Interpret goodness-of-fit measures.

LO **12.4** Conduct a test of individual significance.

LO **12.5** Conduct a test of joint significance.

LO **12.6** Address common violations of the OLS assumptions.

Regression analysis is one of the most widely used statistical techniques in business, engineering, and the social sciences. It is commonly used to predict and/or describe changes in a variable of interest, called the response variable, on the basis of several input variables, called the explanatory variables. For example, we may want to predict a firm's sales based on its various marketing campaigns or predict the selling price of a house based on its size and location. In this chapter we explore the procedure for estimating a linear regression model using the method of ordinary least squares (OLS). We then examine a number of goodness-of-fit measures and conduct hypothesis tests in order to assess which explanatory variables matter most for making predictions. Finally, we examine the importance of the assumptions on the statistical properties of the OLS estimator, as well as the validity of the testing procedures. We address common violations to the model assumptions, the consequences when these assumptions are violated, and offer some remedial measures.



©Blend Images-JGI/Jamie Grill/Band X Pictures/Getty Images

Introductory Case

Consumer Debt Payments

A study of 26 metropolitan areas found that American consumers are making average monthly debt payments of $983 (Experian.com, November 11, 2010). However, it turns out that the actual amount a consumer pays depends a great deal on where the consumer lives. For instance, Washington, D.C. residents pay the most ($1,285 per month), while Pittsburgh residents pay the least ($763 per month). Madelyn Davis, an economist at a large bank, believes that income differences between cities are associated with the disparate debt payments. For example, the Washington, D.C. area's high incomes have likely contributed to its placement on the list. She is unsure about the relationship between the unemployment rate and consumer debt payments. On the one hand, higher unemployment rates may reduce consumer debt payments, as consumers forgo making major purchases such as large appliances and cars. On the other hand, higher unemployment rates may raise consumer debt payments as consumers struggle to pay their bills. In order to analyze the relationship between consumer debt payments, income, and the unemployment rate, Madelyn gathers data on average consumer debt (Debt in $), the annual median household income (Income in $1,000s), and the monthly unemployment rate (Unemployment in %) from the same 26 metropolitan areas used in the debt payment study. Table 12.1 shows a portion of the data.

TABLE 12.1 Average Consumer Debt, Median Income, and the Unemployment Rate, 2010–2011

Metropolitan Area	Debt	Income	Unemployment
Washington, D.C.	1285	103.50	6.3
Seattle	1135	81.70	8.5
⋮	⋮	⋮	⋮
Pittsburgh	763	63.00	8.3

Source: Experian.com collected average monthly consumer debt payments in August 2010 and published the data in November 2010; eFannieMae.com reports 2010–2011 Area Median Household Incomes; bls.com gives monthly unemployment rates for August 2010.

FILE
Debt_payments

Madelyn would like to use the sample information in Table 12.1 to

1. Use regression analysis to make predictions for debt payments for given values of income and the unemployment rate.
2. Use various goodness-of-fit measures to determine the regression model that best fits the data.
3. Determine the significance of income and the unemployment rate at the 5% significance level.

A synopsis of this case is provided at the end of Section 12.4.

LO 12.1

Estimate and interpret a simple linear regression model.

12.1 THE SIMPLE LINEAR REGRESSION MODEL

Regression analysis is one of the most important statistical methodologies used in business, engineering, and the social sciences. It is used to examine the relationship between two or more variables. In the introductory case, Madelyn is interested in examining how income and the unemployment rate might be related to debt payments. In another scenario, we may want to predict a firm's sales based on its advertising; estimate an individual's salary based on education and years of experience; predict the selling price of a house on the basis of its size and location; or describe auto sales with respect to consumer income, interest rates, and price discounts. In all of these examples, we can use regression analysis to describe the relationship between the variables of interest.

Regression analysis allows us to analyze the linear relationship between the target variable, called the **response variable**, and other variables, called the **explanatory variables**. Consequently, we use information on the explanatory variables to predict and/or describe changes in the response variable. Alternative names for the explanatory variables are independent variables, predictor variables, control variables, or regressors, while the response variable is often referred to as the dependent variable, the explained variable, the predicted variable, or the regressand. It is important to note that regression models appear to search for causality when they basically detect correlation. Causality can only be established through randomized experiments and advanced models, which is outside the scope of this text.

No matter the response variable that we choose to examine, we cannot expect to predict its exact value. If the value of the response variable is uniquely determined by the values of the explanatory variables, we say that the relationship between the variables is **deterministic**. This is often the case in the physical sciences. For example, momentum p is the product of the mass m and velocity v of an object, that is, $p = mv$. In most fields of research, however, we tend to find that the relationship between the explanatory variables and the response variable is **stochastic**, due to the omission of relevant factors (sometimes not measurable) that influence the response variable. For instance, debt payments are likely to be influenced by housing costs—a variable that is not included in the introductory case. Similarly, when trying to predict an individual's salary, the individual's natural ability is often omitted since it is extremely difficult, if not impossible, to quantify.

DETERMINISTIC VERSUS STOCHASTIC RELATIONSHIPS

The relationship between the response variable and the explanatory variables is deterministic if the value of the response variable is uniquely determined by the explanatory variables; otherwise, the relationship is stochastic.

Our objective is to develop a mathematical model that captures the relationship between the response variable y and the k explanatory variables $x_1, x_2, \ldots, x_k$. The model must also account for the stochastic nature of the relationship. In order to develop a linear regression model, we start with a deterministic component that approximates the relationship we want to model, and then add a random term to it, making the relationship stochastic.

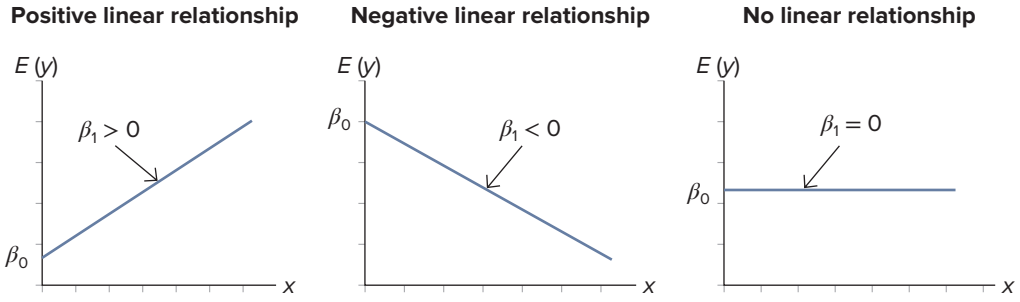
In this section we focus on the **simple linear regression model**, which uses one explanatory variable, denoted x_1, to explain the variability in the response variable, denoted y. For ease of exposition when discussing the simple linear regression model, we often drop the subscript on the explanatory variable and refer to it solely as x. In the next section we extend the simple linear regression model to the **multiple linear regression model**, which uses more than one explanatory variable to explain the variability in the response variable.

A fundamental assumption underlying the simple linear regression model is that the expected value of y lies on a straight line, denoted by $\beta_0 + \beta_1 x$, where β_0 and β_1 (the Greek letters read as betas) are the unknown intercept and slope parameters, respectively. (You have actually seen this relationship before, but you just used different notation. Recall the equation for a line: $y = \text{mx} + \text{b}$, where b and m are the intercept and the slope, respectively, of the line.)

The expression $\beta_0 + \beta_1 x$ is the deterministic component of the simple linear regression model, which can be thought of as the expected value of y for a given value of x. In other

words, conditional on x, $E(y) = \beta_0 + \beta_1 x$. The slope parameter β_1 determines whether the linear relationship between x and $E(y)$ is positive ($\beta_1 > 0$) or negative ($\beta_1 < 0$); $\beta_1 = 0$ indicates that there is no linear relationship between x and $E(y)$. Figure 12.1 shows the deterministic portion of the simple linear regression model for various values of the intercept β_0 and the slope β_1 parameters.

FIGURE 12.1 Various examples of a simple linear regression model



As noted earlier, the observed value y may differ from the expected value $E(y)$. Therefore, we add a random error term ε (the Greek letter read as epsilon) to develop a simple linear regression model.

THE SIMPLE LINEAR REGRESSION MODEL

The simple linear regression model is defined as

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

where y is the response variable, x is the explanatory variable, and ε is the random error term. The coefficients β_0 and β_1 are the unknown parameters to be estimated.

The population parameters β_0 and β_1 used in the simple linear regression model are unknown, and, therefore, must be estimated. As always, we use sample data to estimate the population parameters of interest. Here sample data consist of n pairs of observations on y and x.

Let b_0 and b_1 represent the estimates of β_0 and β_1, respectively. We form the **sample regression equation** as $\hat{y} = b_0 + b_1 x$, where $\hat{y}$ (read as y-hat) is the predicted value of the response variable given a specified value of the explanatory variable x. For a given value of x, the observed and the predicted values of the response variable are likely to be different since many factors besides x influence y. We refer to the difference between the observed and the predicted values of y, that is $y - \hat{y}$, as the **residual** e.

THE SAMPLE REGRESSION EQUATION FOR THE SIMPLE LINEAR REGRESSION MODEL

The sample regression equation for the simple linear regression model is denoted as

$$\hat{y} = b_0 + b_1 x,$$

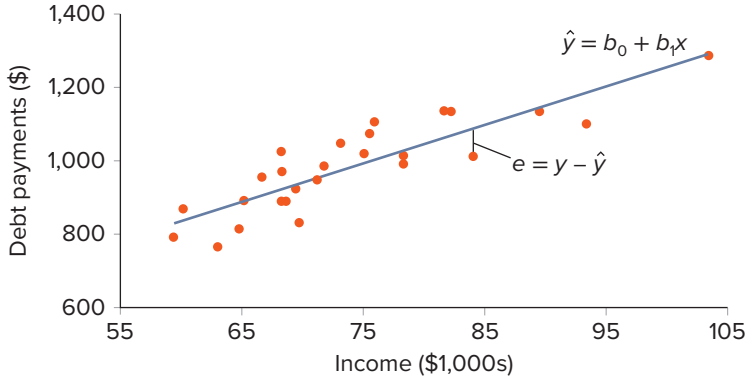
where b_0 and b_1 are the estimates of β_0 and β_1, respectively.

The difference between the observed value and the predicted value of y represents the residual e—that is, $e = y - \hat{y}$.

Before estimating a simple linear regression model, it is useful to visualize the relationship between y and x by constructing a scatterplot. Here, we explicitly place y on the vertical axis and x on the horizontal axis, implying that x is used to explain the variation in y. In Figure 12.2 we use the data from the introductory case to show a scatterplot of debt payments against income. We then superimpose a linear trendline through the points on the scatterplot.

The superimposed line in Figure 12.2 is the sample regression equation, $\hat{y} = b_0 + b_1 x$, where y and x represent debt payments and income, respectively. The upward slope of the line suggests that as income increases, the predicted debt payments also increase. Also, the vertical distance between any data point on the scatterplot (y) and the corresponding point on the line ($\hat{y}$) represents the residual, $e = y - \hat{y}$.

FIGURE 12.2 Scatterplot with a superimposed trend line



Determining the Sample Regression Equation

A common approach to fitting a line to the scatterplot is the **method of least squares**, also referred to as **ordinary least squares (OLS)**. In other words, we use OLS to estimate the parameters β_0 and β_1. OLS estimators have many desirable properties if certain assumptions hold. (These assumptions are discussed in Section 12.5.) The OLS method chooses the line whereby the **error sum of squares**, ***SSE***, is minimized, where $SSE = \Sigma(y_i - \hat{y}_i)^2 = \Sigma e_i^2$. *SSE* is the sum of the squared difference between each observed value y and its predicted value $\hat{y}$ or, equivalently, the sum of the squared distances from the regression equation. Thus, using this distance measure, we say that the OLS method produces the straight line that is "closest" to the data. In the context of Figure 12.2, the superimposed line has been estimated by OLS.

Using calculus, equations have been developed for b_0 and b_1 that satisfy the OLS criterion. These equations, or formulas, are as follows.

CALCULATING THE REGRESSION COEFFICIENTS b_1 AND b_0

The slope b_1 and the intercept b_0 of the sample regression equation are calculated as

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \text{ and}$$

$$b_0 = \bar{y} - b_1 \bar{x}.$$

Fortunately, virtually every statistical software package produces values for b_1 and b_0. So we can focus on interpreting these regression coefficients rather than performing the grueling calculations. The slope estimate b_1 represents the change in $\hat{y}$ when x increases by one unit. As we will see in the following example, it is not always possible to provide an economic interpretation of the intercept estimate b_0; mathematically, however, it represents the predicted value of $\hat{y}$ when x has a value of zero.

EXAMPLE 12.1

FILE
Debt_Payments

Using the data from Table 12.1, let debt payments represent the response variable and income represent the explanatory variable in a simple linear regression model.

a. What is the sample regression equation?

b. Interpret b_1.

c. Interpret b_0.

d. Predict debt payments if income is $80,000.

SOLUTION: Table 12.2 shows the Excel-produced output from estimating the model: Debt $= \beta_0 + \beta_1$ Income $+ \varepsilon$, or simply, $y = \beta_0 + \beta_1 x + \varepsilon$, where y and x represent debt payments and income, respectively. We will provide Excel instructions for obtaining this output shortly.

TABLE 12.2 Excel-Produced Regression Results for Example 12.1

Regression Statistics						
Multiple R	0.8675					
R Square	0.7526					
Adjusted R Square	0.7423					
Standard Error	63.2606					
Observations	26					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	292136.9086	292136.9086	73.000	9.66E-09	
Residual	24	96045.5529	4001.8980			
Total	25	388182.4615				
	Coefficients	*Standard Error*	*t Stat*	*p-Value*	*Lower 95%*	*Upper 95%*
Intercept	**210.2977**	91.3387	2.302	0.030	21.78	398.81
Income	**10.4411**	1.2220	8.544	9.66E-09	7.92	12.96

a. As Table 12.2 shows, Excel produces quite a bit of information. In order to formulate the sample regression equation, we need estimates for β_0 and β_1, which are found at the bottom of the table (see values in boldface). We will address the remaining information in Sections 12.3 and 12.4. We find that $b_0 = 210.2977$ and $b_1 = 10.4411$. Thus, the sample regression equation is $\hat{y} = 210.2977 + 10.4411x$; that is, $\widehat{\text{Debt}} = 210.2977 + 10.4411\text{Income}$.

b. The estimated slope coefficient of 10.4411 suggests a positive relationship between income and debt payments. If median household income increases by $1,000 (since income is measured in $1,000s), then we predict consumer debt payments to increase by b_1—that is, by $10.44.

c. The estimated intercept coefficient of 210.2977 suggests that if income equals zero, then predicted debt payments are $210.30. In this particular application, this conclusion makes some sense, since a household with no income still needs to make debt payments for any credit card use, automobile loans, and so on. However, we should be careful about predicting y when we use a value for x that is not included in the sample range of x. In the ***Debt_Payments*** data file, the lowest and highest values for income (in $1,000s) are 59.40 and 103.50, respectively; plus, the scatterplot suggests that a line fits the data well within this range of the explanatory variable. Unless we assume that income and debt payments will maintain the same linear relationship at income values less than 59.40 and more than 103.50, we should refrain from making predictions based on values of the explanatory variable outside the sample range.

d. Recall that income is measured in $1,000s. So if we are predicting debt payments for an income of $80,000, then we input the value of 80 for Income in the sample regression equation. Thus, we find

$$\widehat{\text{Debt}} = 210.2977 + 10.4411 \times 80 = 1{,}045.59.$$

That is, debt payments are predicted to be $1,045.59.

Using Excel

Constructing a Scatterplot with Trendline

We replicate Figure 12.2.

FILE
Debt_Payments

A. Open the ***Debt_Payments*** data file. (You can disregard the data in the Unemployment column for this application.)

B. When constructing a scatterplot, Excel, by default, chooses the first column as the explanatory variable (x) and the second column as the response variable (y). Since the data file shows Debt (y) in the first column and Income (x) in the second column, the easiest solution is to copy and paste the Debt column to the right of the Income column. Then simultaneously select the data for Income and Debt and, from the menu, choose **Insert > Scatter**. Select the graph on the top left. (If you are having trouble finding this option after selecting **Insert**, look for the graph with data points above **Charts**.)

C. Right-click on the scatter points, choose **Add Trendline**, and then choose **Linear**.

D. Formatting (regarding axis titles, gridlines, etc.) can be done by selecting **Format > Add Chart Element** from the menu.

Estimating a Simple Linear Regression Model

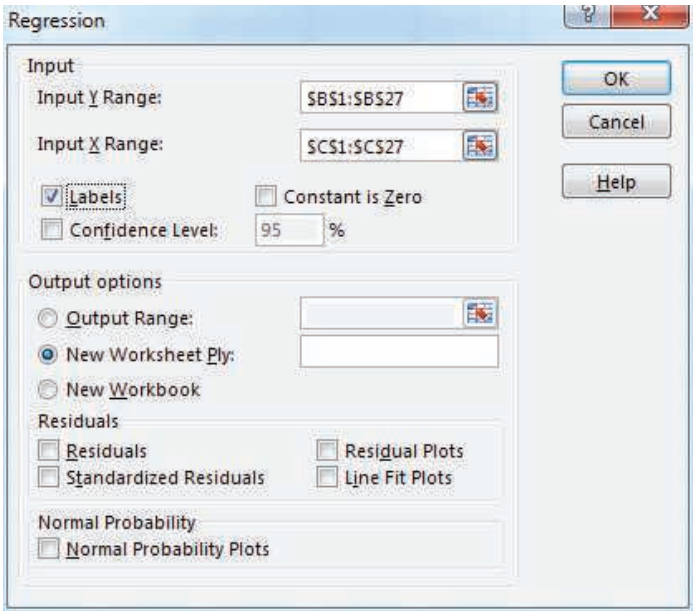
We replicate Table 12.2 from Example 12.1.

A. Open the ***Debt_Payments*** data file.

B. From the menu, choose **Data > Data Analysis > Regression**.

C. See Figure 12.3. In the *Regression* dialog box, click on the box next to *Input Y Range*, and then select the Debt data, including its heading. For *Input X Range*, select the Income data, including its heading. Check *Labels*. Click **OK**. Your results should be identical to Table 12.2.

FIGURE 12.3 Excel's Regression dialog box



Source: Microsoft Excel

EXERCISES 12.1

Mechanics

1. In a simple linear regression, the following sample regression coefficients were estimated: $b_0 = -1.5$ and $b_1 = 3.4$.
   a. Formulate the sample regression equation.
   b. Predict y if x equals 5.
2. In a simple linear regression, the following sample regression equation is obtained:
   $$\hat{y} = 15 + 2.5x.$$
   a. Predict y if x equals 10.
   b. What happens to this prediction if x doubles in value to 20?
3. In a simple linear regression, the following sample regression equation is obtained:
   $$\hat{y} = 436 - 17x.$$
   a. Interpret the slope coefficient.
   b. Predict y if x equals -15.
4. Thirty observations were used to estimate $y = \beta_0 + \beta_1 x + \varepsilon$. A portion of the results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	41.82	8.58	4.87	3.93E-05
x	0.49	0.10	4.81	4.65E-05

   a. What is the estimate for β_1? Interpret this value.
   b. What is the sample regression equation?
   c. If $x = 30$, what is $\hat{y}$?
5. Twenty-four observations were used to estimate $y = \beta_0 + \beta_1 x + \varepsilon$. A portion of the regression results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	2.25	2.36	0.95	0.3515
x	–0.16	0.30	–0.53	0.6017

   a. What is the estimate for β_1? Interpret this value.
   b. What is the sample regression equation?
   c. What is the predicted value for y if $x = 2$? If $x = -2$?

Applications

6. If a firm spends more on advertising, is it likely to increase sales? Data on annual sales (in $100,000s) and advertising expenditures (in $10,000s) were collected for 20 firms in order to estimate the model Sales $= \beta_0 + \beta_1$ Advertising $+ \varepsilon$. A portion of the regression results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	–7.42	1.46	–5.09	7.66E-05
Advertising	0.42	0.05	8.70	7.26E-08

   a. Is the sign on the slope as expected? Explain.
   b. What is the sample regression equation?
   c. Predict the sales for a firm that spends $500,000 annually on advertising.
7. The owner of several used-car dealerships believes that the selling price of a used car can best be predicted using the car's age. He uses data on the recent selling price (in $) and age of 20 used sedans to estimate Price $= \beta_0 + \beta_1$Age $+ \varepsilon$. A portion of the regression results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	21187.94	733.42	28.89	1.56E-16
Age	–1208.25	128.95	–9.37	2.41E-08

   a. What is the estimate for β_1? Interpret this value.
   b. What is the sample regression equation?
   c. Predict the selling price of a 5-year-old sedan.
8. FILE ***GPA.*** The director of graduate admissions at a large university is analyzing the relationship between scores on the math portion of the Graduate Record Examination (GRE) and subsequent performance in graduate school, as measured by a student's grade point average (GPA). She uses a sample of 24 students who graduated within the past five years. A portion of the data is as follows:

GPA	GRE
3.0	700
3.5	720
⋮	⋮
3.5	780

   a. Find the sample regression equation for the model: GPA $= \beta_0 + \beta_1$GRE $+ \varepsilon$.
   b. What is a student's predicted GPA if he/she scored 710 on the math portion of the GRE?
9. FILE ***Education.*** A social scientist would like to analyze the relationship between educational attainment (in years of higher education) and annual salary (in $1,000s). He collects data on 20 individuals. A portion of the data is as follows:

Salary	Education
40	3
53	4
⋮	⋮
38	0

   a. Find the sample regression equation for the model: Salary $= \beta_0 + \beta_1$Education $+ \varepsilon$.
   b. Interpret the coefficient for Education.
   c. What is the predicted salary for an individual who completed 7 years of higher education?

10. FILE ***Consumption_Function.*** The consumption function, first developed by John Maynard Keynes, captures one of the key relationships in economics. It expresses consumption as a function of disposable income, where disposable income is income after taxes. The accompanying table shows a portion of average U.S. annual consumption (in $) and disposable income (in $) for the years 1985–2006.

	Consumption	Disposable Income
1985	23490	22887
1986	23866	23172
⋮	⋮	⋮
2006	48398	58101

Source: *The Statistical Abstract of the United States.*

a. Estimate the model: Consumption $= \beta_0 + \beta_1$Disposable Income $+ \varepsilon$.
b. In this model, the slope coefficient is called the marginal propensity to consume. Interpret its meaning.
c. What is predicted consumption if disposable income is $57,000?

11. FILE ***MLB_Pitchers.*** The following table lists a portion of Major League Baseball's (MLB's) leading pitchers, each pitcher's salary (In $ millions), and earned run average (ERA) for 2008.

	Salary	ERA
J. Santana	17.0	2.53
C. Lee	4.0	2.54
⋮	⋮	⋮
C. Hamels	0.5	3.09

Source: www.ESPN.com.

a. Estimate the model: Salary $= \beta_0 + \beta_1$ERA $+ \varepsilon$ and interpret the coefficient of ERA.
b. Use the estimated model to predict the salary for each player, given his ERA. For example, use the sample regression equation to predict the salary for J. Santana with ERA $= 2.53$.
c. Derive the corresponding residuals and explain why the residuals might be so high.

12. FILE ***Happiness_Age.*** Refer to the accompanying data file on happiness and age for this exercise.
a. Estimate a simple linear regression model with Happiness as the response variable and Age as the explanatory variable.
b. Use the sample regression equation to predict Happiness when Age equals 25, 50, and 75.
c. Construct a scatterplot of Happiness against Age. Discuss why your predictions might not be accurate.

13. FILE ***Property_Taxes.*** The accompanying table shows a portion of data that refers to the property taxes owed by a homeowner (in $) and the size of the home (in square feet) in an affluent suburb 30 miles outside New York City.

Property Taxes	Size
21928	2449
17339	2479
⋮	⋮
29235	2864

a. Estimate the sample regression equation that enables us to predict property taxes on the basis of the size of the home.
b. Interpret the slope coefficient.
c. Predict the property taxes for a 1,500-square-foot home.

14. FILE ***Test_Scores.*** The accompanying table shows a portion of the scores that 32 students obtained on the final and the midterm in a course in statistics.

Final	Midterm
86	78
94	97
⋮	⋮
91	47

a. Estimate the sample regression equation that enables us to predict a student's final score on the basis of his/her midterm score.
b. Predict the final score of a student who received an 80 on the midterm.

15. FILE ***Fertilizer.*** A horticulturist is studying the relationship between tomato plant height and fertilizer amount. Thirty tomato plants grown in similar conditions were subjected to various amounts of fertilizer (in ounces) over a four-month period, and then their heights (in inches) were measured. A portion of the data is shown in the accompanying table.

Height	Fertilizer
20.4	1.9
49.2	5.0
⋮	⋮
46.4	3.1

a. Estimate the model: Height $= \beta_0 + \beta_1$Fertilizer $+ \varepsilon$.
b. Interpret the coefficient of Fertilizer. Does the y-intercept make practical sense?
c. Use the estimated model to predict, after four months, the height of a tomato plant which received 3.0 ounces of fertilizer.

16. FILE ***Dexterity.*** Finger dexterity, the ability to make precisely coordinated finger movements to grasp or assemble very small objects, is important in jewelry making. Thus, the manufacturing manager at Gemco, a manufacturer of high-quality watches, wants to develop a regression model to predict the productivity (in watches per shift) of new employees based on dexterity. He has subjected a sample of 20 current employees to the O'Connor dexterity test in which the time required to place 3 pins in each of 100 small holes using tweezers is measured in seconds. A portion of the data is shown in the accompanying table.

Watches	Time
23	513
19	608
⋮	⋮
20	437

a. Estimate the model: Watches $= \beta_0 + \beta_1$Time $+ \varepsilon$.
b. Interpret the coefficient of Time.
c. Explain why the y-intercept makes no practical sense in this particular problem.
d. Suppose a new employee takes 550 seconds on the dexterity test. How many watches per shift is she expected to produce?

12.2 THE MULTIPLE LINEAR REGRESSION MODEL

LO **12.2**

Estimate and interpret a multiple linear regression model.

The simple linear regression model allows us to analyze the linear relationship between one explanatory variable and the response variable. However, by restricting the number of explanatory variables to one, we sometimes reduce the potential usefulness of the model. In Section 12.5, we will discuss how the OLS estimates can be quite misleading when important explanatory variables are excluded. A **multiple linear regression model** allows us to analyze the linear relationship between the response variable and two or more explanatory variables. The choice of the explanatory variables is based on economic theory, intuition, and/or prior research. The multiple linear regression model is a straightforward extension of the simple linear regression model.

THE MULTIPLE LINEAR REGRESSION MODEL

The multiple linear regression model is defined as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

where y is the response variable, $x_1, x_2, \ldots, x_k$ are the k explanatory variables, and ε is the random error term. The coefficients $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters to be estimated.

As in the case of the simple linear regression model, we apply the OLS method that minimizes SSE, where $SSE = \Sigma(y_i - \hat{y}_i)^2 = \Sigma e_i^2$.

THE SAMPLE REGRESSION EQUATION FOR THE MULTIPLE LINEAR REGRESSION MODEL

The sample regression equation for the multiple linear regression model is denoted as

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k,$$

where $b_0, b_1, \ldots, b_k$ are the estimates of $\beta_0, \beta_1, \ldots, \beta_k$.

The difference between the observed value and the predicted value of y represents the residual e—that is, $e = y - \hat{y}$.

For each explanatory variable x_j ($j = 1, \ldots, k$), the corresponding slope coefficient b_j is the estimate of β_j. We slightly modify the interpretation of the slope coefficients in the context of a multiple linear regression model. Here b_j measures the change in the predicted value of the response variable $\hat{y}$ given a unit increase in the associated explanatory variable x_j, *holding all other explanatory variables constant*. In other words, it represents the partial influence of x_j on $\hat{y}$.

EXAMPLE 12.2

FILE
Debt_Payments

In this example we analyze how debt payments may be influenced jointly by income and the unemployment rate.

a. Given the data from Table 12.1, estimate the multiple linear regression model with debt payments as the response variable and income and the unemployment rate as the explanatory variables.

b. Interpret the regression coefficients.

c. Predict debt payments if income is $80,000 and the unemployment rate is 7.5%.

SOLUTION:

a. Table 12.3 shows the Excel-produced output from estimating the model: Debt $= \beta_0 + \beta_1$Income $+ \beta_2$Unemployment $+ \varepsilon$. We will provide Excel instructions for obtaining this output shortly.

TABLE 12.3 Excel-Produced Regression Output for Example 12.2

Regression Statistics						
Multiple R		0.8676				
R Square		0.7527				
Adjusted R Square		0.7312				
Standard Error		64.6098				
Observations		26				
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	292170.7719	146085.3860	34.995	1.05E-07	
Residual	23	96011.6896	4174.4213			
Total	25	388182.4615				
	Coefficients	*Standard Error*	*t Stat*	*p-Value*	*Lower 95%*	*Upper 95%*
Intercept	**198.9956**	156.3619	1.273	0.216	−124.46	522.45
Income	**10.5122**	1.4765	7.112	2.98E-07	7.46	13.57
Unemployment	**0.6186**	6.8679	0.090	0.929	−13.59	14.83

Using the boldface estimates from Table 12.3, $b_0 = 198.9956$, $b_1 = 10.5122$, and $b_2 = 0.6186$, we derive the sample regression equation as

$$\widehat{\text{Debt}} = 198.9956 + 10.5122\text{Income} + 0.6186\text{Unemployment}.$$

b. The regression coefficient of Income is 10.5122. Since income is measured in $1,000s, the model suggests that if income increases by $1,000, then debt payments are predicted to increase by $10.51, holding the unemployment rate constant. Similarly, the regression coefficient of Unemployment is 0.6186, implying that a 1 percentage point increase in the unemployment rate leads to a predicted increase in debt payments of $0.62, holding income constant. It seems that the predicted impact of Unemployment, with Income held

constant, is rather small. In fact, the influence of the unemployment rate is not even statistically significant at any reasonable level; we will discuss such tests of significance in Section 12.4.

c. If income is \$80,000 and the unemployment rate is 7.5%, we find

$$\widehat{\text{Debt}} = 198.9956 + 10.5122 \times 80 + 0.6186 \times 7.5 = 1{,}044.61.$$

That is, debt payments are predicted to be \$1,044.61.

Using Excel to Estimate a Multiple Linear Regression Model

When estimating a multiple linear regression model in Excel as compared to a simple linear regression, there are very minor modifications. For these reasons, we are brief.

FILE
Debt_Payments

A. Open the ***Debt_Payments*** data file.

B. From the menu, choose **Data > Data Analysis > Regression**.

C. In the *Regression* dialog box, click on the box next to *Input Y Range*, then select the Debt data, including its heading. For *Input X Range*, *simultaneously* select the Income and the Unemployment data (including both headings). Check *Labels*. Click **OK**. Your results should be identical to Table 12.3.

EXERCISES 12.2

Mechanics

17. In a multiple regression, the following sample regression coefficients were estimated: $b_0 = -1.5$, $b_1 = 3.4$, and $b_2 = -9.2$.
    a. Formulate the sample regression equation.
    b. Predict y if x_1 equals 10 and if x_2 equals 5.

18. In a multiple regression, the following sample regression equation is obtained:
$$\hat{y} = -8 + 2.6x_1 - 47.2x_2.$$
    a. Predict y if x_1 equals 40 and x_2 equals −10.
    b. Interpret the slope coefficient of x_2.

19. In a multiple regression, the following sample regression equation is obtained:
$$\hat{y} = 152 + 12.9x_1 + 2.7x_2.$$
    a. Predict y if x_1 equals 20 and x_2 equals 35.
    b. Interpret the slope coefficient of x_1.

20. Thirty observations were used to estimate $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. A portion of the regression results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	21.97	2.98	7.37	6.31E-08
x_1	30.00	2.23	13.44	1.75E-13
x_2	−1.88	0.27	−6.96	1.75E-07

    a. What is the estimate for β_1? Interpret this value.
    b. What is the sample regression equation?
    c. If $x_1 = 30$ and $x_2 = 20$, what is $\hat{y}$?

21. Forty observations were used to estimate $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. A portion of the regression results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	13.83	2.42	5.71	1.56E-06
x_1	−2.53	0.15	−16.87	5.84E-19
x_2	0.29	0.06	4.83	2.38E-05

    a. What is the estimate for β_1? Interpret this value.
    b. What is the sample regression equation?
    c. What is the predicted value for y if $x_1 = -9$ and $x_2 = 25$?

Applications

22. On the first day of class, an economics professor administers a test to gauge the math preparedness of her students. She believes that the performance on this math test and the number of hours studied per week on the course are the primary factors that predict a student's score on the final exam. Using data from her class of 60 students, she estimates $\text{Final} = \beta_0 + \beta_1 \text{Math} + \beta_2 \text{Hours} + \varepsilon$.

A portion of the regression results is shown in the following table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	40.55	3.37	12.03	2.83E-17
Math	0.25	0.04	6.06	1.14E-07
Hours	4.85	0.57	8.53	9.06E-12

a. What is the slope coefficient of Hours?
b. What is the sample regression equation?
c. What is the predicted final exam score for a student who has a math score of 70 and studies 4 hours per week?

23. Using data from 50 workers, a researcher estimates Wage $= \beta_0 + \beta_1$Education $+ \beta_2$Experience $+ \beta_3$Age $+ \varepsilon$, where Wage is the hourly wage rate and Education, Experience, and Age are the years of higher education, the years of experience, and the age of the worker, respectively. A portion of the regression results is shown in the following table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	7.87	4.09	1.93	0.0603
Education	1.44	0.34	4.24	0.0001
Experience	0.45	0.14	3.16	0.0028
Age	−0.01	0.08	−0.14	0.8920

a. Interpret the estimates for β_1 and β_2.
b. What is the sample regression equation?
c. Predict the hourly wage rate for a 30-year-old worker with 4 years of higher education and 3 years of experience.

24. A sociologist believes that the crime rate in an area is significantly influenced by the area's poverty rate and median income. Specifically, she hypothesizes crime will increase with poverty and decrease with income. She collects data on the crime rate (crimes per 100,000 residents), the poverty rate (in %), and the median income (in $1,000s) from 41 New England cities. A portion of the regression results is shown in the following table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	−301.62	549.71	−0.55	0.5864
Poverty	53.16	14.22	3.74	0.0006
Income	4.95	8.26	0.60	0.5526

a. Are the signs as expected on the slope coefficients?
b. Interpret the slope coefficient for Poverty.
c. Predict the crime rate in an area with a poverty rate of 20% and a median income of $50,000.

25. Osteoporosis is a degenerative disease that primarily affects women over the age of 60. A research analyst wants to forecast sales of StrongBones, a prescription drug for treating this debilitating disease. She uses the model Sales $= \beta_0 + \beta_1$Population $+ \beta_2$Income $+ \varepsilon$, where Sales refers to the sales of StrongBones (in $1,000,000s), Population is the number of women over the age of 60 (in millions), and Income is the average income of women over the age of 60 (in $1,000s). She collects data on 38 cities across the United States and obtains the following regression results:

	Coefficients	Standard Error	t Stat	p-value
Intercept	10.35	4.02	2.57	0.0199
Population	8.47	2.71	3.12	0.0062
Income	7.62	6.63	1.15	0.2661

a. What is the sample regression equation?
b. Interpret the slope coefficients.
c. Predict sales if a city has 1.5 million women over the age of 60 and their average income is $44,000.

26. FILE ***Engine_Overhaul.*** The maintenance manager at a trucking company wants to build a regression model to forecast the time (in years) until the first engine overhaul based on four explanatory variables: (1) annual miles driven (in 1,000s of miles), (2) average load weight (in tons), (3) average driving speed (in mph), and (4) oil change interval (in 1,000s of miles). Based on driver logs and onboard computers, data have been obtained for a sample of 25 trucks. A portion of the data is shown in the accompanying table.

Time until First Engine Overhaul	Annual Miles Driven	Average Load Weight	Average Driving Speed	Oil Change Interval
7.9	42.8	19	46	15
0.9	98.5	25	46	29
⋮	⋮	⋮	⋮	⋮
6.1	61.2	24	58	19

a. For each explanatory variable, discuss whether it is likely to have a positive or negative influence on time until the first engine overhaul.
b. Estimate the regression model (use all four explanatory variables).
c. Based on part (a), are the signs of the regression coefficients logical?
d. Predict the time before the first engine overhaul for a particular truck driven 60,000 miles per year with an average load of 22 tons, an average driving speed of 57 mph, and 18,000 miles between oil changes.

27. FILE ***Arlington_Homes.*** A realtor in Arlington, Massachusetts, is analyzing the relationship between the sale price of a home (Price in $), its square footage (Sqft), the number of bedrooms (Beds), and the number of bathrooms (Baths). She collects data on 36 sales in Arlington in the first quarter of 2009 for the analysis. A portion of the data is shown in the accompanying table.

Price	Sqft	Beds	Baths
840000	2768	4	3.5
822000	2500	4	2.5
⋮	⋮	⋮	⋮
307500	850	1	1

Source: http://Newenglandmoves.com.

a. Estimate the model: Price $= \beta_0 + \beta_1$Sqft $+ \beta_2$Beds $+ \beta_3$Baths $+ \varepsilon$.
b. Interpret the slope coefficients.
c. Predict the price of a 2,500-square-foot home with three bedrooms and two bathrooms.

28. FILE ***Electricity_Cost.*** The facility manager at a pharmaceutical company wants to build a regression model to forecast monthly electricity cost. Three main variables are thought to dictate electricity cost (in $): (1) average outdoor temperature (Avg Temp in °F), (2) working days per month, and (3) tons of product produced. A portion of the past year's monthly data is shown in the accompanying table.

Cost	Avg Temp	Work Days	Tons Produced
24100	26	24	80
23700	32	21	73
⋮	⋮	⋮	⋮
26000	39	22	69

a. For each explanatory variable, discuss whether it is likely to have a positive or negative influence on monthly electricity cost.
b. Estimate the regression model.
c. Are the signs of the regression coefficients as expected? If not, speculate as to why this could be the case.
d. What is the predicted electricity cost in a month during which the average outdoor temperature is 65°, there are 23 working days, and 76 tons are produced?

29. FILE ***MCAS.*** Education reform is one of the most hotly debated subjects on both state and national policy makers' list of socioeconomic topics. Consider a linear regression model that relates school expenditures and family background to student performance in Massachusetts using 224 school districts. The response variable is the mean score on the MCAS (Massachusetts Comprehensive Assessment System) exam given in May 1998 to 10th graders. Four explanatory variables are used: (1) STR is the student-to-teacher ratio in %, (2) TSAL is the average teacher's salary in $1,000s, (3) INC is the median household income in $1,000s, and (4) SGL is the percentage of single-parent households. A portion of the data is shown in the accompanying table.

Score	STR	TSAL	INC	SGL
227.00	19.00	44.01	48.89	4.70
230.67	17.90	40.17	43.91	4.60
⋮	⋮	⋮	⋮	⋮
230.67	19.20	44.79	47.64	5.10

Source: Massachusetts Department of Education and the Census of Population and Housing.

a. For each explanatory variable, discuss whether it is likely to have a positive or negative influence on Score.
b. Find the sample regression equation. Are the signs of the slope coefficients as expected?
c. What is the predicted score if STR = 18, TSAL = 50, INC = 60, SGL = 5?
d. What is the predicted score if everything else is the same as in part (c) except INC = 80?

30. FILE ***Quarterback_Salaries.*** American football is the highest paying sport on a per-game basis. The quarterback, considered the most important player on the team, is appropriately compensated. A sports statistician wants to use 2009 data to estimate a multiple linear regression model that links the quarterback's salary (in $ millions) with his pass completion percentage (PCT), total touchdowns scored (TD), and his age. A portion of the data is shown in the accompanying table.

Name	Salary	PCT	TD	Age
Philip Rivers	25.5566	65.2	28	27
Jay Cutler	22.0441	60.5	27	26
⋮	⋮	⋮	⋮	⋮
Tony Romo	0.6260	63.1	26	29

Source: *USA Today* database for salaries; http://NFL.com for other data.

a. Estimate the model: Salary $= \beta_0 + \beta_1$PCT $+ \beta_2$TD $+ \beta_3$Age $+ \varepsilon$.
b. Are you surprised by the estimated coefficients?
c. Drew Brees earned 12.9895 million dollars in 2009. According to the model, what is his predicted salary if PCT = 70.6, TD = 34, and Age = 30?
d. Tom Brady earned 8.0073 million dollars in 2009. According to the model, what is his predicted salary if PCT = 65.7, TD = 28, and Age = 32?
e. Compute and interpret the residual salary for Drew Brees and Tom Brady.

31. FILE ***Car_Prices.*** The accompanying table shows a portion of data consisting of the selling price, the age, and the mileage for 20 used sedans.

Selling Price	Age	Mileage
13590	6	61485
13775	6	54344
⋮	⋮	⋮
11988	8	42408

a. Estimate the sample regression equation that enables us to predict the price of a sedan on the basis of its age and mileage.
b. Interpret the slope coefficient of Age.
c. Predict the selling price of a five-year-old sedan with 65,000 miles.

32. FILE ***AnnArbor_Rental.*** The accompanying table shows a portion of data consisting of the rent, the number of bedrooms, the number of bathrooms, and the square footage for 40 apartments in the college town of Ann Arbor, Michigan.

Rent	Bed	Bath	Sqft
645	1	1	500
675	1	1	648
⋮	⋮	⋮	⋮
2400	3	2.5	2700

a. Determine the sample regression equation that enables us to predict the rent of an Ann Arbor apartment on the basis of the number of bedrooms, the number of bathrooms, and the square footage.
b. Interpret the slope coefficient of Bath.
c. Predict the rent for a 1,500-square-foot apartment with 2 bedrooms and 1 bathroom.

LO **12.3**

Interpret goodness-of-fit measures.

12.3 GOODNESS-OF-FIT MEASURES

So far we have focused on the estimation and the interpretation of the linear regression models. By simply observing the sample regression equation, we cannot assess how well the explanatory variables explain the variation in the response variable. We rely on several objective "goodness-of-fit" measures that summarize how well the sample regression equation fits the data. If each predicted value $\hat{y}$ is equal to its observed values y, then we have a perfect fit. Since that almost never happens, we evaluate the models on a relative basis.

In the introductory case study, we were interested in analyzing consumer debt payments. In Section 12.1 and Section 12.2, we estimated the following two linear regression models:

$$\text{Model 1: Debt} = \beta_0 + \beta_1 \text{Income} + \varepsilon$$
$$\text{Model 2: Debt} = \beta_0 + \beta_1 \text{Income} + \beta_2 \text{Unemployment} + \varepsilon$$

For ease of exposition, we use the same notation to refer to the coefficients in Models 1 and 2. We note, however, that these coefficients and their estimates have a different meaning depending on which model we are referencing.

If you had to choose one of these models to predict debt payments, which model would you choose? It may be that by using more explanatory variables, you can better describe the response variable. However, for a given sample, *more* is not always better. In order to select the preferred model, we examine several goodness-of-fit measures: the **standard error of the estimate**, the **coefficient of determination**, and the **adjusted coefficient of determination**. We first discuss these measures in general, and then determine whether Model 1 or Model 2 is the preferred model.

The Standard Error of the Estimate

We first describe goodness-of-fit measures in the context of a simple linear regression model, so we use Model 1 for exposition. Figure 12.4 reproduces the scatterplot of debt payments against income, as well as the superimposed sample regression line. Recall that the residual e represents the difference between an observed value and the predicted value

of the response variable—that is, $e = y - \hat{y}$. If all the data points had fallen on the line, then each residual would be zero; in other words, there would be no dispersion between the observed and the predicted values. Since in practice we rarely, if ever, obtain this result, we evaluate models on the basis of the relative magnitude of the residuals. The sample regression equation provides a good fit when the dispersion of the residuals is relatively small.

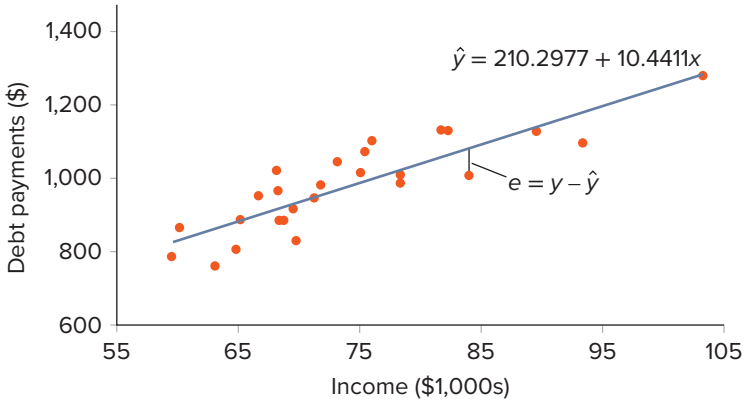


FIGURE 12.4
Scatterplot of debt payments *y* against income *x*

A numerical measure that gauges dispersion from the sample regression equation is the sample variance of the residual, denoted s_e^2. This measure is defined as the average squared difference between y_i and $\hat{y}_i$. The numerator of the formula is the error sum of squares, $SSE = \Sigma(y_i - \hat{y}_i)^2 = \Sigma e_i^2$. Dividing SSE by its respective degrees of freedom $n - k - 1$ yields s_e^2. Recall that k denotes the number of explanatory variables in the linear regression model; thus, for a simple linear regression model, k equals one. Instead of s_e^2, we generally report the standard deviation of the residual, denoted s_e, more commonly referred to as the standard error of the estimate. As usual, s_e is the positive square root of s_e^2. The less the dispersion, the smaller the s_e, which typically implies that the model provides a good fit for the sample data.

THE STANDARD ERROR OF THE ESTIMATE

The standard error of the estimate s_e is calculated as

$$s_e = \sqrt{\frac{SSE}{n - k - 1}},$$

where SSE is the error sum of squares. Theoretically, s_e can assume any value between zero and infinity, $0 \leq s_e < \infty$. The closer s_e is to zero, the better the estimated model fits the sample data.

For a given sample size n, increasing the number k of the explanatory variables reduces both the numerator (SSE) and the denominator ($n - k - 1$) in the formula for s_e. The net effect, shown by the value of s_e, allows us to determine if the added explanatory variables improve the fit of the model.

Virtually all statistical software packages report s_e. Excel reports s_e in the *Regression Statistics* portion of the regression output and refers to it as Standard Error.

The Coefficient of Determination, R^2

Like the standard error of the estimate, the coefficient of determination, commonly referred to as R^2, evaluates how well the sample regression equation fits the data. In particular,

R^2 quantifies the sample variation in the response variable y that is explained by the sample regression equation. It is computed as the ratio of the explained variation of the response variable to its total variation. For example, if $R^2 = 0.72$, we say that 72% of the sample variation in the response variable is explained by the sample regression equation. Other factors, which have not been included in the model, account for the remaining 28% of the sample variation.

We use analysis of variance (ANOVA) in the context of the linear regression model, to derive R^2. We denote the total variation in y as $\Sigma(y_i - \bar{y})^2$, which is the numerator in the formula for the variance of y. This value, called the **total sum of squares**, ***SST***, can be broken down into two components: explained variation and unexplained variation. Figure 12.5 illustrates the decomposition of the total variation in y into its two components for a simple linear regression model.

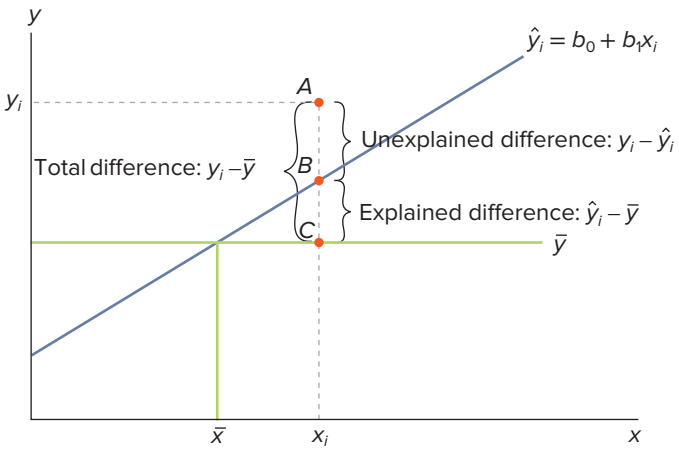


FIGURE 12.5 Total, explained, and unexplained differences

For ease of exposition, we show a scatterplot with all the points removed except one (point A). Point A refers to the observation (x_i, y_i). The blue line represents the estimated regression equation based on the entire sample data; the horizontal and vertical green lines represent the sample means $\bar{y}$ and $\bar{x}$, respectively. The vertical distance between the data point A and $\bar{y}$ (point C) is the difference $y_i - \bar{y}$ (distance AC). For each data point, we square these differences and then find their sum—this amounts to $SST = \Sigma(y_i - \bar{y})^2$. As mentioned above, SST is a measure of the total variation in y.

Now, we focus on the distance between the predicted value $\hat{y}_i$ (point B) and $\bar{y}$; that is, the explained difference (distance BC). It is called "explained" because the difference between $\hat{y}_i$ and $\bar{y}$ can be explained by the difference between x_i and $\bar{x}$. Squaring all such differences and summing them yields the **regression sum of squares**, ***SSR***, where $SSR = \Sigma(\hat{y}_i - \bar{y})^2$. SSR is a measure of the explained variation in y.

The distance between the particular observation and its predicted value (distance AB) is the unexplained difference. This is the portion that remains unexplained; it is due to random error or chance. Squaring all such differences and summing them yields the familiar error sum of squares, $SSE = \Sigma(y_i - \hat{y}_i)^2$. SSE is a measure of the unexplained variation in y.

Thus, the total variation in y can be decomposed into explained and unexplained variation as follows:

$$SST = SSR + SSE.$$

Dividing both sides by SST and rearranging yields:

$$\frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Each side of the above equation shows two equivalent formulas for the coefficient of determination R^2; that is, $R^2 = SSR/SST$ or $R^2 = 1 - SSE/SST$. The value of R^2 falls

between zero and one, $0 \le R^2 \le 1$. The closer R^2 is to one, the stronger the fit; the closer it is to zero, the weaker the fit.

THE COEFFICIENT OF DETERMINATION, R^2

The coefficient of determination, R^2, is the proportion of the sample variation in the response variable that is explained by the sample regression equation. It is computed as

$$R^2 = \frac{SSR}{SST}, \text{ or equivalently, } R^2 = 1 - \frac{SSE}{SST},$$

where $SSR = \Sigma(\hat{y}_i - \bar{y})^2$, $SSE = \Sigma(y_i - \hat{y}_i)^2$, and $SST = \Sigma(y_i - \bar{y})^2$.

The value of R^2 falls between zero and one; the closer the value is to one, the better the fit.

Most statistical packages, including Excel, report the coefficient of determination. Excel reports R^2 in the *Regression Statistics* portion of the regression output and refers to it as R Square. Excel also reports another interesting statistic referred to as **Multiple *R***. Multiple R is simply the sample correlation coefficient between the response variable y and its predicted value $\hat{y}$ which, using notation from Section 3.7, implies that Multiple $R = r_{y\hat{y}}$. Note that R^2 is the square of Multiple R—that is, $R^2 = r_{y\hat{y}}^2$.

In general, the objective in adding another explanatory variable to a linear regression model is to increase the model's usefulness. It turns out that we cannot use R^2 for model comparison when the competing models do not include the same number of explanatory variables. This occurs because R^2 never decreases as we add more explanatory variables to the model. A popular goodness-of-fit measure in such situations is to choose the model that has the highest adjusted R^2 value.

The Adjusted R^2

Since R^2 never decreases as we add more explanatory variables to the linear regression model, it is possible to increase its value unintentionally by including a group of explanatory variables that may have no economic or intuitive foundation in the linear regression model. This is true especially when the number of explanatory variables k is large relative to the sample size n. In order to avoid the possibility of R^2 creating a false impression, virtually all software packages, including Excel, include adjusted R^2. Unlike R^2, adjusted R^2 explicitly accounts for the number of explanatory variables k. It is common to use adjusted R^2 for model selection because it imposes a penalty for any additional explanatory variable that is included in the analysis.

ADJUSTED R^2

The adjusted coefficient of determination is calculated as

$$\text{Adjusted } R^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right).$$

Adjusted R^2 is used to compare competing linear regression models with different numbers of explanatory variables; the higher the adjusted R^2, the better the model.

If SSE is substantially greater than zero and k is large compared to n, then adjusted R^2 will differ substantially from R^2. Adjusted R^2 may be negative if the correlation between the response variable and the explanatory variables is sufficiently low.

We would also like to point out that both the standard error of the estimate and the adjusted R^2 are useful for comparing the linear regression models with different numbers of explanatory variables. Adjusted R^2, however, is the more commonly used criterion for model selection.

EXAMPLE 12.3

Table 12.4 provides goodness-of-fit measures from estimating Model 1 and Model 2:

$$\text{Model 1: Debt} = \beta_0 + \beta_1 \text{Income} + \varepsilon$$
$$\text{Model 2: Debt} = \beta_0 + \beta_1 \text{Income} + \beta_2 \text{Unemployment} + \varepsilon$$

TABLE 12.4 Goodness-of-Fit Measures for Model 1 and Model 2

	Model 1	Model 2
Multiple R	0.8675	0.8676
R Square	0.7526	0.7527
Adjusted R Square	0.7423	0.7312
Standard Error	63.2606	64.6098

a. Based on two goodness-of-fit measures, which model is the preferred model.

b. Interpret the coefficient of determination for the preferred model.

c. What percentage of the sample variation in debt payments is unexplained by the preferred model?

SOLUTION:

a. Model 1 has the lower standard error of the estimate ($63.2606 < 64.6098$), as well as the higher adjusted R^2($0.7423 > 0.7312$). Thus, Model 1 is the preferred model. Note that we cannot use the coefficient of determination R^2 to compare the two models since the models have different numbers of explanatory variables.

b. The coefficient of determination R^2 for Model 1 is 0.7526 which means that 75.26% of the sample variation in debt payments is explained by the regression model.

c. If 75.26% of the sample variation in debt payments is explained by Model 1, then 24.74% of the variation is unexplained by the regression equation.

EXERCISES 12.3

Mechanics

33. In a simple linear regression based on 30 observations, it is found that $SSE = 2{,}540$ and $SST = 13{,}870$.
    a. Calculate the standard error of the estimate s_e.
    b. Calculate the coefficient of determination R^2.

34. In a multiple regression with two explanatory variables, and 50 observations, it is found that $SSE = 35$ and $SST = 90$.
    a. Calculate the standard error of the estimate s_e.
    b. Calculate the coefficient of determination R^2.

35. In a multiple regression with four explanatory variables and 100 observations, it is found that $SSR = 4.75$ and $SST = 7.62$.
    a. Calculate the standard error of the estimate s_e.
    b. Calculate the coefficient of determination R^2.
    c. Calculate adjusted R^2.

36. The accompanying table lists goodness-of-fit measures that were obtained when estimating the following two models:

$$\text{Model 1: } y = \beta_0 + \beta_1 x_1 + \varepsilon$$
$$\text{Model 2: } y = \beta_0 + \beta_1 x_2 + \varepsilon$$

	Model 1	Model 2
R^2	0.459	0.496
Adjusted R^2	0.445	0.483
s_e	104.914	101.274

Which model provides a better fit for y? Justify your response with two goodness-of-fit measures.

37. The accompanying table lists goodness-of-fit measures that were obtained when estimating the following two models:

Model 1: $y = \beta_0 + \beta_1 x_1 + \varepsilon$
Model 2: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

	Model 1	Model 2
R^2	0.751	0.752
Adjusted R^2	0.748	0.747
s_e	13.652	13.694

Which model provides a better fit for y? Justify your response with two goodness-of-fit measures.

38. The accompanying table lists goodness-of-fit measures that were obtained when estimating the following two models:

Model 1: $y = \beta_0 + \beta_1 x_1 + \varepsilon$
Model 2: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

	Model 1	Model 2
R^2	0.804	0.828
Adjusted R^2	0.801	0.819
s_e	17.746	16.924

Which model provides a better fit for y? Justify your response with two goodness-of-fit measures.

39. The accompanying table lists goodness-of-fit measures that were obtained when estimating the following two models:

Model 1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
Model 2: $y = \beta_0 + \beta_1 x_3 + \beta_2 x_4 + \varepsilon$

	Model 1	Model 2
R^2	0.640	0.610
Adjusted R^2	0.627	0.597
s_e	34.706	36.103

Which model provides a better fit for y? Justify your response with two goodness-of-fit measures.

Applications

40. An analyst estimates the sales of a firm as a function of its advertising expenditures using the model: Sales $= \beta_0 + \beta_1$Advertising $+ \varepsilon$. Using 20 observations, he finds that $SSR = 199.93$ and $SST = 240.92$.
    a. What proportion of the sample variation in sales is explained by advertising expenditures?
    b. What proportion of the sample variation in sales is unexplained by advertising expenditures?

41. FILE ***Test_Scores.*** The accompanying data file shows the midterm and final scores for 32 students in a statistics course.
    a. Estimate a student's final score as a function of his/her midterm score.
    b. Find the standard error of the estimate.
    c. Find and interpret the coefficient of determination.

42. The director of college admissions at a local university is trying to determine whether a student's high school GPA or SAT score is a better predictor of the student's subsequent college GPA. She formulates two models:

Model 1. College GPA $= \beta_0 + \beta_1$High School GPA $+ \varepsilon$
Model 2. College GPA $= \beta_0 + \beta_1$SAT Score $+ \varepsilon$

She estimates these models and obtains the following goodness-of-fit measures.

	Model 1	Model 2
R^2	0.5595	0.5322
Adjusted R^2	0.5573	0.5298
s_e	40.3684	41.6007

Which model provides a better fit for y? Justify your response with two goodness-of-fit measures.

43. FILE ***Property_Taxes.*** The accompanying data file shows the property taxes and square footage for 20 homes in an affluent suburb 30 miles outside New York City.
    a. Estimate a home's property taxes as a linear function of the size of the home (measured by its square footage).
    b. What proportion of the sample variation in property taxes is explained by the home's size?
    c. What proportion of the sample variation in property taxes is unexplained by the home's size?

44. FILE ***Car_Prices.*** The accompanying data file shows the selling price of a used sedan, its age, and its mileage. Estimate two models:

Model 1: Price $= \beta_0 + \beta_1$Age $+ \varepsilon$
Model 2: Price $= \beta_0 + \beta_1$Age $+ \beta_2$Mileage $+ \varepsilon$

Which model provides a better fit for y? Justify your response with two goodness-of-fit measures.

45. For a sample of 41 New England cities, a sociologist studies the crime rate in each city (crimes per 100,000 residents) as a function of its poverty rate (in %) and its median income (in $1,000s). He finds that $SSE = 4{,}182{,}663$ and $SST = 7{,}732{,}451$.
    a. Calculate the standard error of the estimate.
    b. What proportion of the sample variation in crime rate is explained by the variability in the explanatory variables? What proportion is unexplained?

46. A financial analyst uses the following model to estimate a firm's stock return: Return $= \beta_0 + \beta_1$P/E $+ \beta_2$P/S $+ \varepsilon$, where P/E is a firm's price-to-earnings ratio and P/S is a firm's price-to-sales ratio. For a sample of 30 firms, she finds that $SSE = 4{,}402.786$ and $SST = 5{,}321.532$.
    a. Calculate the standard error of the estimate.
    b. Calculate and interpret the coefficient of determination.
    c. Calculate the adjusted R^2.

47. FILE ***Football.*** Is it defense or offense that wins football games? Consider the following portion of data, which includes a team's winning record (Win in %), the average number of yards gained, and the average number of yards allowed during the 2009 NFL season.

Team	Win	Yards Gained	Yards Allowed
Arizona Cardinals	62.50	344.40	346.40
Atlanta Falcons	56.30	340.40	348.90
⋮	⋮	⋮	⋮
Washington Redskins	25.00	312.50	319.70

Source: NFL website.

    a. Compare two simple linear regression models, where Model 1 predicts the winning percentage based on Yards Gained and Model 2 uses Yards Allowed.
    b. Estimate a multiple linear regression model, Model 3, that applies both Yards Gained and Yards Allowed to forecast the winning percentage. Is this model an improvement over the other two models? Explain.

48. FILE ***Executive_Compensation.*** Executive compensation has risen dramatically beyond the rising levels of an average worker's wage over the years. The government is even considering a cap on high-flying salaries for executives (*The New York Times*, February 9, 2009). Consider the following portion of data which links total compensation (in $ millions) of the 455 highest-paid CEOs in 2006 with three measures: industry-adjusted return on assets (ROA), industry-adjusted stock return (Return) and the firm's size (Total Assets in $ millions).

Compensation	ROA	Return	Total Assets
16.58	2.53	−0.15	20917.5
26.92	1.27	−0.57	32659.5
⋮	⋮	⋮	⋮
2.3	0.45	0.75	44875.0

Source: SEC website and Compustat.

    a. Estimate three simple linear regression models that use Compensation as the response variable with ROA, Return, or Total Assets as the explanatory variable. Which model do you select? Explain.
    b. Estimate multiple linear regression models that use various combinations of two, or all three, explanatory variables. Which model do you select? Explain.

LO **12.4**

Conduct a test of individual significance.

12.4 TESTS OF SIGNIFICANCE

In this section, we continue our assessment of the linear regression model by turning our attention to hypothesis tests about the unknown parameters (coefficients) $\beta_0, \beta_1, \ldots, \beta_k$. In particular, we test for the individual and joint significance of the regression coefficients to determine whether there is evidence of a linear relationship between the response and the explanatory variables. We note that for the tests to be valid, the OLS estimators $b_0, b_1, \ldots, b_k$ must be normally distributed. This condition is satisfied if the random error term ε is normally distributed. If we cannot assume the normality of ε, then the tests are valid only for large sample sizes. We will discuss the underlying assumptions of the linear regression model in Section 12.5.

Tests of Individual Significance

Consider the following multiple regression model, which links the response variable y with the k explanatory variables, $x_1, x_2, \ldots, x_k$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon.$$

If any slope coefficient β_j equals zero, then the corresponding x_j basically drops out of the above equation, implying that x_j does not influence y. In other words, if β_j equals zero, then there is no linear relationship between x_j and y. Conversely, if β_j does not equal zero, then x_j influences y.

Following the hypothesis testing methodology introduced in earlier chapters, we want to test whether the population slope coefficient β_j is different from, greater than, or less than β_{j0} where β_{j0} is the hypothesized value of β_j. That is, the competing hypotheses take one of the following forms:

Two-Tailed Test	Right-Tailed Test	Left-Tailed Test
H_0: $\beta_j = \beta_{j0}$	H_0: $\beta_j \leq \beta_{j0}$	H_0: $\beta_j \geq \beta_{j0}$
H_A: $\beta_j \neq \beta_{j0}$	H_A: $\beta_j > \beta_{j0}$	H_A: $\beta_j < \beta_{j0}$

When testing whether x_j significantly influences y, we set $\beta_{j0} = 0$ and specify a two-tailed test as H_0: $\beta_j = 0$ and H_A: $\beta_j \neq 0$. We could easily specify one-tailed competing hypotheses for a positive linear relationship (H_0: $\beta_j \leq 0$ and H_A: $\beta_j > 0$) or a negative linear relationship (H_0: $\beta_j \geq 0$ and H_A: $\beta_j < 0$).

Although tests of significance are commonly based on $\beta_{j0} = 0$, in some situations we might wish to determine whether the slope coefficient differs from a nonzero value. For instance, if we are analyzing the relationship between a student's exam score on the basis of hours studied, we may want to determine if an extra hour of review before the exam will increase a student's score by more than 5 points. Here, we formulate the hypotheses as H_0: $\beta_j \leq 5$ and H_A: $\beta_j > 5$, where $\beta_{j0} = 5$.

Finally, although in most applications we are interested in conducting hypothesis tests on the slope coefficient(s), there are instances where we may also be interested in testing the intercept β_0. The testing framework for the intercept remains the same; that is, if we want to test whether the intercept differs from zero, we specify the competing hypotheses as H_0: $\beta_0 = 0$ and H_A: $\beta_0 \neq 0$.

As in all hypothesis tests, the next essential piece of information is how we define the appropriate test statistic.

TEST STATISTIC FOR THE TEST OF INDIVIDUAL SIGNIFICANCE

The value of the test statistic for a test of individual significance is calculated as

$$t_{df} = \frac{b_j - \beta_{j0}}{se(b_j)},$$

where $df = n - k - 1$, b_j is the estimate for β_j, $se(b_j)$ is the estimated standard error of b_j, and β_{j0} is the hypothesized value of β_j. If $\beta_{j0} = 0$, the value of the test statistic reduces to $t_{df} = \frac{b_j}{se(b_j)}$.

EXAMPLE 12.4

Let's revisit Model 1, Debt $= \beta_0 + \beta_1$Income $+ \varepsilon$, estimated with the sample data in Table 12.1. We reproduce a portion of the regression results in Table 12.5. Conduct a hypothesis test to determine whether Income influences Debt at the 5% significance level.

TABLE 12.5 Portion of Regression Results for Model 1: Debt $= \beta_0 + \beta_1$Income $+ \varepsilon$

	Coefficients	Standard Error	t Stat	p-value
Intercept	10.2977	91.3387	2.302	0.030
Income	10.4411	1.2220	8.544	9.66E-09

SOLUTION: We set up the following competing hypotheses in order to determine whether Debt and Income have a linear relationship:

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

From Table 12.5, we find that $b_1 = 10.4411$ and $se(b_1) = 1.2220$. In addition, given $n = 26$ and $k = 1$, we find $df = n - k - 1 = 26 - 1 - 1 = 24$. So we calculate the value of the test statistic as $t_{24} = \frac{b_1 - \beta_{10}}{\text{se}(b_1)} = \frac{10.4411 - 0}{1.2220} = 8.544$. Note that this calculation is not necessary since virtually all statistical computer packages automatically provide the value of the test statistic and its associated p-value.

Since the p-value of $9.66 \times 10^{-9} \approx 0 < 0.05 = \alpha$, we reject H_0. At the 5% significance level, there is a linear relationship between Debt and Income; in other words, Income is significant in explaining Debt.

EXAMPLE 12.5

Let's revisit Model 2, Debt $= \beta_0 + \beta_1$Income $+ \beta_2$Unemployment $+ \varepsilon$, estimated with the sample data in Table 12.1. We reproduce a portion of the regression results in Table 12.6. At the 5% significance level, determine whether Unemployment is significant in explaining Debt.

TABLE 12.6 Portion of Regression Results for Model 2: *Debt* $= \beta_0 + \beta_1$Income $+ \beta_2$Unemployment $+ \varepsilon$

	Coefficients	*Standard Error*	*t Stat*	*p-value*
Intercept	198.9956	156.3619	1.273	0.216
Income	10.5122	1.4765	7.112	2.98E-07
Unemployment	0.6186	6.8679	0.090	0.929

SOLUTION: For testing whether Unemployment significantly influences Debt, we set up the following competing hypotheses:

$$H_0: \beta_2 = 0$$
$$H_A: \beta_2 \neq 0$$

From Table 12.6, we see that the p-value $= 0.929 > 0.05 = \alpha$; thus, we cannot reject H_0. At the 5% significance level, we cannot conclude that Unemployment is significant in explaining Debt.

This result is not surprising since the goodness-of-fit measures suggested that Model 1 with only one explanatory variable (Income) provided a better fit for Debt as compared to Model 2 with two explanatory variables (Income and Unemployment).

It is important to note that the computer-generated results are valid only in a standard case where a two-tailed test is implemented to determine whether a coefficient differs from zero. In Examples 12.4 and 12.5 we could use the computer-generated values of the test statistics as well as the corresponding p-values because they represented standard cases. For a one-tailed test with $\beta_{j0} = 0$, the value of the test statistic is valid, but the

p-value is not; in this case, the computer-generated p-value must be divided in half. For a one- or two-tailed test to determine if the regression coefficient differs from a nonzero value, both the computer-generated value of the test statistic and the p-value become invalid. These facts are summarized below.

COMPUTER-GENERATED TEST STATISTIC AND THE p-VALUE

Excel and virtually all other statistical packages report a value of the test statistic and its associated p-value for a two-tailed test that assesses whether the regression coefficient differs from zero.

- If we specify a one-tailed test with $\beta_{j0} = 0$, then we need to divide the computer-generated p-value in half.
- If we specify a one- or two-tailed test with $\beta_j \neq 0$, then we cannot use the value of the computer-generated test statistic and its p-value.

We would also like to point out that for a one-tailed test with $\beta_{j0} = 0$, there are rare instances when the computer generated p-value is invalid. This occurs when the sign of b_j (and the value of the accompanying test statistic) is not inconsistent with the null hypothesis. For example, for a right-tailed test, H_0: $\beta_j \leq 0$ and H_A: $\beta_j > 0$, the null hypothesis cannot be rejected if the estimate b_j (and the value of the accompanying test statistic t_{df}) is negative. Similarly, no further testing is necessary if $b_j > 0$ (and thus $t_{df} > 0$) for a left-tailed test.

A Test for a Nonzero Slope Coefficient

In Examples 12.4 and 12.5, the null hypothesis included a zero value for the slope coefficient; that is, $\beta_{j0} = 0$. We now motivate a test where the hypothesized value is not zero by using a renowned financial application—the **capital asset pricing model (CAPM)**.

Let R represent the return on a stock or portfolio of interest. Given the market return R_M and the risk-free return R_f, the CAPM expresses the risk-adjusted return of an asset, $R - R_f$, as a function of the risk-adjusted market return, $R_M - R_f$. It is common to use the return of the S&P 500 index for R_M and the return on a Treasury bill for R_f. For empirical estimation, we express the CAPM as

$$R - R_f = \alpha + \beta(R_M - R_f) + \varepsilon.$$

We can rewrite the model as $y = \alpha + \beta x + \varepsilon$, where $y = R - R_f$ and $x = R_M - R_f$. Note that this is essentially a simple linear regression model that uses α and β, in place of the usual β_0 and β_1, to represent the intercept and the slope coefficients, respectively. The slope coefficient β, called the stock's **beta**, measures how sensitive the stock's return is to changes in the level of the overall market. When β equals 1, any change in the market return leads to an identical change in the given stock return. A stock for which $\beta > 1$ is considered more "aggressive" or riskier than the market, whereas one for which $\beta < 1$ is considered "conservative" or less risky. We also give importance to the intercept coefficient α, called the stock's **alpha**. The CAPM theory predicts α to be zero, and thus a nonzero estimate indicates abnormal returns. Abnormal returns are positive when $\alpha > 0$ and negative when $\alpha < 0$.

EXAMPLE 12.6

Johnson & Johnson (J&J) was founded more than 120 years ago on the premise that doctors and nurses should use sterile products to treat people's wounds. Since that time, J&J products have become staples in most people's homes. Consider the CAPM where the J&J risk-adjusted stock return $R - R_f$ is used as the response

variable and the risk-adjusted market return $R_M - R_f$ is used as the explanatory variable. A portion of 60 months of data is shown in Table 12.7.

FILE
Johnson_Johnson

TABLE 12.7 Risk-Adjusted Stock Return of J&J and Market Return

Date	$R - R_f$	$R_M - R_f$
1/1/2006	−4.59	2.21
2/1/2006	0.39	−0.31
⋮	⋮	⋮
12/1/2010	0.48	2.15

Source: finance.yahoo.com and U.S. Treasury.

a. Since consumer staples comprise many of the products sold by J&J, its stock is often considered less risky; that is, people need these products whether the economy is good or bad. At the 5% significance level, is the beta coefficient less than one?

b. At the 5% significance level, are there abnormal returns? In other words, is the alpha coefficient significantly different from zero?

SOLUTION: Using the CAPM notation, we estimate the model, $R - R_f = \alpha + \beta(R_M - R_f) + \varepsilon$; the relevant portion of the regression output is presented in Table 12.8.

TABLE 12.8 Portion of CAPM Regression Results for J&J

	Coefficients	Standard Error	t Stat	p-Value
Intercept	0.2666	0.4051	0.658	0.513
$R_M - R_f$	0.5844	0.0803	7.276	0.000

a. The estimate for the beta coefficient is 0.5844 and its standard error is 0.0803. Interestingly, our estimate is identical to the beta reported in the popular press (www.dailyfinance.com, March 4, 2011). In order to determine whether the beta coefficient is significantly less than one, we formulate the hypotheses as

$$H_0: \beta \geq 1$$
$$H_A: \beta < 1$$

Given 60 data points, $df = n - k - 1 = 60 - 1 - 1 = 58$. We cannot use the test statistic value reported in Table 12.8, since the hypothesized value of β is not zero. We calculate the value of the test statistic as $t_{58} = \frac{b_1 - \beta_{10}}{\text{se}(b_1)} = \frac{0.5844 - 1}{0.0803} = -5.176$. We can use the t table to approximate the p-value, $P(T_{58} \leq -5.176)$, as a value that is less than 0.005. Using Excel's T.DIST.RT function and recognizing that $P(T_{58} \leq -5.176) = P(T_{58} \geq 5.176)$, we find that the exact p-value is 1.48×10^{-6}. Since the p-value $< \alpha = 0.05$, we reject H_0 and conclude that β is significantly less than one; that is, the return on J&J stock is less risky than the return on the market.

b. Abnormal returns exist when α is significantly different from zero. Thus, the competing hypotheses are H_0: $\alpha = 0$ versus H_A: $\alpha \neq 0$. Since it is a standard case, where the hypothesized value of the coefficient is zero, we can use the reported test statistic value of 0.658 with an associated p-value of 0.513. We cannot reject H_0 at any reasonable level of significance. Therefore, we cannot conclude that there are abnormal returns for J&J stock.

Test of Joint Significance

LO **12.5**

Conduct a test of joint significance.

So far, we have considered a test of individual significance. For instance, we used a *t*-test to determine whether a household's income has a statistically significant influence on debt payments. When we assess a multiple linear regression model, it is also important to conduct a **test of joint significance**. A test of joint significance is often regarded as a test of the overall usefulness of a regression. This test determines whether the explanatory variables $x_1, x_2, \ldots, x_k$ have a joint statistical influence on y.

In the null hypothesis of the test of joint significance, *all* of the slope coefficients are assumed zero. The competing hypotheses for a test of joint significance are specified as

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_A: \text{At least one } \beta_j \neq 0.$$

You might be tempted to implement this test by performing a series of tests of individual significance with the *t* statistic. However, such an option is not appropriate. The test of joint significance determines if at least one of the explanatory variables is significant. Therefore, it is not clear if one or all of the explanatory variables must be significant in order to document a joint significance. In addition, recall from the discussion of ANOVA in Chapter 10 that if we conduct many individual tests at (say) a 5% level of significance, the resulting significance level for the joint test will be greater than 5%.

Testing a series of individual hypotheses is not equivalent to testing the same hypotheses jointly.

To conduct the test of joint significance, we employ a right-tailed *F* test. (Recall that the $F_{(df_1,df_2)}$ distribution was used for hypothesis testing in Chapter 10.) The test statistic measures how well the regression equation explains the variability in the response variable. It is defined as the ratio of the **mean square regression (*MSR*)** to the **mean square error (*MSE*)** where $MSR = SSR/k$ and $MSE = SSE/(n - k - 1)$.

TEST STATISTIC FOR THE TEST OF JOINT SIGNIFICANCE

The value of the test statistic for a test of joint significance is calculated as

$$F_{(df_1,df_2)} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE},$$

where $df_1 = k$, $df_2 = n - k - 1$, *SSR* is the regression sum of squares, *SSE* is the error sum of squares, *MSR* is the mean square regression, and *MSE* is the mean square error.

In general, a large value of $F_{(df_1,df_2)}$ indicates that a large portion of the sample variation in y is explained by the regression model; thus, the model is useful. A small value of $F_{(df_1,df_2)}$ implies that a large portion of the sample variation in y remains unexplained. In fact, the test of joint significance is sometimes informally referred to as the test of the significance of R^2. Note that while the test of joint significance is important for a multiple regression model, it is redundant for a simple regression model. In fact, in a simple regression

model, the p-value of the F test is identical to that of the t-test; we advise you to verify this fact.

Most statistical computer packages, including Excel, produce an ANOVA table that decomposes the total variability of the response variable y into two components: (1) the variability explained by the regression and (2) the variability that is unexplained. In addition, the value for the $F_{(df_1,df_2)}$ test statistic and its p-value are also provided. Table 12.9 shows the general format of an ANOVA table. Excel explicitly provides an ANOVA table with its regression output, with the p-value reported under the heading *Significance F*.

TABLE 12.9 General Format of an ANOVA Table for Regression

ANOVA	df	SS	MS	F	Significance F
Regression	k	SSR	$MSR = \frac{SSR}{k}$	$F_{(df_1,df_2)} = \frac{MSR}{MSE}$	$P\left(F_{(df_1,df_2)} \geq \frac{MSR}{MSE}\right)$
Residual	$n - k - 1$	SSE	$MSE = \frac{SSE}{n - k - 1}$		
Total	$n - 1$	SST			

FILE
Debt_Payments

EXAMPLE 12.7

Let's revisit Model 2, Debt $= \beta_0 + \beta_1$Income $+ \beta_2$Unemployment $+ \varepsilon$, estimated with the sample data in Table 12.1. We reproduce the ANOVA portion of the regression results in Table 12.10. Conduct a test to determine if Income and Unemployment are jointly significant in explaining Debt at $\alpha = 0.05$.

TABLE 12.10 ANOVA Portion of Regression Results for Model 2: Debt $= \beta_0 + \beta_1$Income $+ \beta_2$Unemployment $+ \varepsilon$

ANOVA	df	SS	MS	F	Significance F
Regression	2	292170.7719	146085.3860	34.995	1.05E-07
Residual	23	96011.6896	4174.4213		
Total	25	388182.4615			

SOLUTION: When testing whether the explanatory variables are jointly significant in explaining Debt, we set up the following competing hypotheses:

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_A: \text{At least one } \beta_j \neq 0.$$

Given $n = 26$ and $k = 2$, we find that $df_1 = k = 2$ and $df_2 = n - k - 1 = 23$. From Table 12.10, we calculate the value of the test statistic as

$$F_{(2,23)} = \frac{MSR}{MSE} = \frac{146{,}085.3860}{4{,}174.4213} = 34.995.$$

Note that this is the same value that is reported by Excel. The p-value, $P(F_{(2,23)} \geq 34.995)$, is equal to $1.05 \times 10^{-7} \approx 0$. Since the p-value is less than $\alpha = 0.05$, we reject H_0. At the 5% significance level, Income and Unemployment are jointly significant in explaining Debt. The results of the individual significance tests, in Examples 12.4 and 12.5, suggested that only Income (not Unemployment) significantly influences Debt.

Reporting Regression Results

Regression results are often reported in a "user-friendly" table. Table 12.11 reports the regression results for the two models discussed in this chapter. The response variable is Debt and the explanatory variables are Income in Model 1, and Income and Unemployment in Model 2. If we were supplied with only this table, we would be able to compare these models, construct the sample regression equation of the chosen model, and perform a respectable assessment of the model with the statistics provided. Many tables contain a Notes section at the bottom explaining some of the notation. We choose to put the *p*-values in parentheses; however, some researchers place the standard errors of the coefficients or the values of the test statistics in parentheses. Whichever format is chosen, it must be made clear to the reader in the Notes section.

TABLE 12.11 Model Estimates for the Response Variable Debt

Variable	Model 1	Model 2
Intercept	210.2977* (0.030)	198.9956 (0.216)
Income	10.4411* (0.000)	10.5122* (0.000)
Unemployment	NA	0.6186 (0.929)
s_e	63.2606	64.6098
R^2	0.7526	0.7527
Adjusted R^2	0.7423	0.7312
F statistic (p-value)	NA	34.995* (0.000)

NOTES: Parameter estimates are in the top half of the table with the *p*-values in parentheses; NA denotes not applicable; * represents significance at the 5% level. The lower part of the table contains goodness-of-fit measures.

SYNOPSIS OF INTRODUCTORY CASE



©Andy Dean Photography/Alamy Stock Photo

A recent study shows substantial variability in consumer debt payments depending on where a consumer resides (Experian.com, November 11, 2010). A possible explanation is that a relationship exists between consumer debt payments and an area's median household income. In addition to income, the unemployment rate may also impact consumer debt payments. In order to substantiate these claims, relevant data on 26 metropolitan areas are collected.

Two linear regression models are estimated for the analysis. A simple linear regression model (Model 1), which uses consumer debt payments as the response variable and median household income as the explanatory variable, is estimated as $\widehat{\text{Debt}} = 210.30 + 10.44\text{Income}$. For every \$1,000 increase in median household income, consumer debt payments are predicted to increase by \$10.44. In an attempt to improve upon the prediction, a multiple linear regression model (Model 2) is proposed, where median household income and the unemployment rate are used as explanatory variables. The sample regression equation for Model 2 is $\widehat{\text{Debt}} = 199.00 + 10.51\text{Income} + 0.62\text{Unemployment}$. Given its slope coefficient of only 0.62, the economic impact of the unemployment rate on consumer debt payments, with median household income held fixed, seems extremely weak.

Goodness-of-fit measures confirm that Model 1 provides a better fit than Model 2. The standard error of the estimate is smaller for Model 1, suggesting less dispersion of the data from the sample regression equation. In addition, the adjusted R^2 is higher for Model 1, implying a better fit. Lastly, at the 5% significance level, median household income is significant in explaining consumer debt payments but unemployment is not. Using Model 1 and assuming that an area's median household income is \$80,000, consumer debt payments are predicted to be \$1,045.59.

EXERCISES 12.4

Mechanics

49. In a simple linear regression based on 30 observations, it is found that $b_1 = 3.25$ and $se(b_1) = 1.36$. Consider the hypotheses:

$$H_0: \beta_1 = 0 \text{ and } H_A: \beta_1 \neq 0.$$

a. Calculate the value of the test statistic.
b. Find the p-value.
c. At the 5% significance level, what is the conclusion? Is the explanatory variable statistically significant?

50. In a simple linear regression based on 25 observations, it is found that $b_1 = 0.5$ and $se(b_1) = 0.3$. Consider the hypotheses:

$$H_0: \beta_1 \leq 0 \text{ and } H_A: \beta_1 > 0.$$

a. Calculate the value of the test statistic and the p-value.
b. At the 5% significance level, what is the conclusion to the test?

51. In a simple linear regression based on 30 observations, it is found that $b_1 = 7.2$ and $se(b_1) = 1.8$. Consider the hypotheses:

$$H_0: \beta_1 \geq 10 \text{ and } H_A: \beta_1 < 10.$$

a. Calculate the value of the test statistic and the p-value.
b. At the 5% significance level, what is the conclusion to the test?

52. Consider the following regression results based on 20 observations.

	Coefficients	Standard Error	t Stat	p-value
Intercept	34.2123	4.5665	7.420	0.000
x_1	0.1223	0.1794	0.682	0.504

a. Specify the hypotheses to determine if the intercept differs from zero. Perform this test at the 5% significance level.
b. At the 5% significance level, does the slope differ from zero? Explain.

53. Consider the following regression results based on 40 observations.

	Coefficients	Standard Error	t Stat	p-value
Intercept	43.1802	12.6963	3.401	0.002
x_1	0.9178	0.9350	0.982	0.333

a. Specify the hypotheses to determine if the slope differs from minus one.
b. Calculate the value of the test statistic and the p-value.
c. At the 5% significance level, does the slope differ from minus one? Explain.

54. When estimating a multiple linear regression model based on 30 observations, the following results were obtained.

	Coefficients	Standard Error	t Stat	p-value
Intercept	152.27	119.70	1.272	0.214
x_1	12.91	2.68	4.817	5.06E-05
x_2	2.74	2.15	1.274	0.213

a. Specify the hypotheses to determine whether x_1 is linearly related to y. At the 5% significance level, are x_1 and y linearly related?
b. At the 5% significance level, is x_2 significant in explaining y? Explain.
c. At the 5% significance level, can you conclude that β_1 is less than 20? Show the relevant steps of the hypothesis test.

55. The following ANOVA table was obtained when estimating a multiple linear regression model.

ANOVA	df	SS	MS	F	Significance F
Regression	2	22016.75	11008.375		0.0228
Residual	17	39286.93	2310.996		
Total	19	61303.68			

a. How many explanatory variables were specified in the model? How many observations were used?
b. Specify the hypotheses to determine whether the explanatory variables are jointly significant.
c. Compute the value of the test statistic.
d. At the 5% significance level, what is the conclusion to the test? Explain.

Applications

56. A marketing manager analyzes the relationship between the annual sales of a firm (in \$100,000s) and its advertising expenditures (in \$10,000s). He collects data from 20 firms and estimates Sales $= \beta_0 + \beta_1$Advertising $+ \varepsilon$. A portion of the regression results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	–7.42	1.46	–5.082	7.66E-05
Advertising	0.42	0.05		1.21E-07

a. Specify the competing hypotheses in order to determine whether advertising expenditures and sales have a positive linear relationship.
b. Calculate the value of the test statistic.
c. At the 5% significance level, do advertising expenditures and sales have a positive linear relationship?

57. In order to examine the relationship between the selling price of a used car and its age, an analyst uses data from 20 recent transactions and estimates Price $= \beta_0 + \beta_1$Age $+ \varepsilon$. A portion of the regression results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	21187.94	733.42	28.889	1.56E-16
Age	–1208.25	128.95		2.41E-08

a. Specify the competing hypotheses in order to determine whether the selling price of a used car and its age are linearly related.

b. Calculate the value of the test statistic.
c. At the 5% significance level, is the age of a used car significant in explaining its selling price?
d. Conduct a hypothesis test at the 5% significance level in order to determine if β_1 differs from −1,000. Show all of the relevant steps.

58. A study on the evolution of mankind shows that, with a few exceptions, world-record holders in the 100-meter dash have progressively gotten bigger over time (*The Wall Street Journal*, July 22, 2009). The following table shows runners who have held the record, along with their record-holding times (in seconds) and heights (in inches):

Record Holder/Year	Time	Height
Eddie Tolan (1932)	10.30	67
Jesse Owens (1936)	10.20	70
Charles Greene (1968)	9.90	68
Eddie Hart (1972)	9.90	70
Carl Lewis (1991)	9.86	74
Asafa Powell (2007)	9.74	75
Usain Bolt (2008)	9.69	77

A portion of the regression results from estimating Time = $\beta_0 + \beta_1$Height + ε is:

	Coefficients	Standard Error	t Stat	p-value
Intercept	13.353	1.1714	11.399	9.1E-05
Height	−0.0477	0.0163		

a. Specify the sample regression equation.
b. Specify the hypotheses to determine whether Height is linearly related to Time.
c. Calculate the value of the test statistic.
d. At the 5% significance level, is Height statistically significant? Explain.

59. An economist examines the relationship between changes in short-term interest rates and long-term interest rates. He believes that changes in short-term rates are significant in explaining long-term interest rates. He estimates the model Dlong = $\beta_0 + \beta_1$Dshort + ε, where Dlong is the change in the long-term interest rate (10-year Treasury bill) and Dshort is the change in the short-term interest rate (3-month Treasury bill). Monthly data from January 2006 through December 2010 ($n = 60$) were obtained from the St. Louis Federal Reserve's website. A portion of the regression results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	−0.0038	0.0088	−0.427	0.671
Dshort	0.0473	0.0168	2.813	0.007

Use a 5% significance level to determine whether there is a linear relationship between Dshort and Dlong.

60. For a sample of 20 New England cities, a sociologist studies the crime rate in each city (crimes per 100,000 residents) as a function of its poverty rate (in %) and its median income (in $1,000s). A portion of the regression results is shown in the accompanying table.

ANOVA	df	SS	MS	F	Significance F
Regression	2	188246.8	94123.40	35.20	9.04E-07
Residual	17	45457.32	2673.96		
Total	19	233704.1			

	Coefficients	Standard Error	t Stat	p-value
Intercept	−301.7927	549.7135	−0.549	0.590
Poverty	53.1597	14.2198	3.738	0.002
Income	4.9472	8.2566	0.599	0.557

a. Specify the sample regression equation.
b. At the 5% significance level, show whether the poverty rate and the crime rate are linearly related.
c. At the 5% significance level, determine whether income influences the crime rate at the 5% significance level.
d. At the 5% significance level, are the poverty rate and income jointly significant in explaining the crime rate?

61. Akiko Hamaguchi is a manager at a small sushi restaurant in Phoenix, Arizona. Akiko is concerned that the weak economic environment has hampered foot traffic in her area, thus causing a dramatic decline in sales. In order to offset the decline in sales, she has pursued a strong advertising campaign. She believes advertising expenditures have a positive influence on sales. To support her claim, Akiko estimates the following linear regression model: Sales = $\beta_0 + \beta_1$Unemployment + β_2 Advertising + ε. A portion of the regression results is shown in the accompanying table.

ANOVA	df	SS	MS	F	Significance F
Regression	2	72.6374	36.3187	8.760	0.003
Residual	14	58.0438	4.1460		
Total	16	130.681			

	Coefficients	Standard Error	t Stat	p-value
Intercept	17.5060	3.9817	4.397	0.007
Unemployment	−0.6879	0.2997	−2.296	0.038
Advertising	0.0266	0.0068	3.932	0.002

a. At the 5% significance level, test whether the explanatory variables jointly influence sales.
b. At the 1% significance level, test whether the unemployment rate is negatively related with sales.
c. At the 1% significance level, test whether advertising expenditures are positively related with sales.

62. A researcher estimates the following model relating the return on a firm's stock as a function of its price-to-earnings ratio and

its price-to-sales ratio: Return $= \beta_0 + \beta_1$P/E $+ \beta_2$P/S $+ \varepsilon$. A portion of the regression results is shown in the accompanying table.

ANOVA	df	SS	MS	F	Significance F
Regression	2	918.746	459.3728	2.817	0.077
Residual	27	4402.786	163.0661		
Total	29	5321.532			

	Coefficients	Standard Error	t Stat	p-value
Intercept	−12.0243	7.886858	−1.525	0.139
P/E	0.1459	0.4322	0.338	0.738
P/S	5.4417	2.2926	2.374	0.025

a. Specify the sample regression equation.
b. At the 10% significant level, are P/E and P/S jointly significant? Show the relevant steps of the test.
c. Are both explanatory variables individually significant at the 10% significance level? Show the relevant steps of the test.

63. FILE ***Test_Scores.*** The accompanying data file shows midterm and final grades for 32 students. Estimate a student's final grade as a linear function of a student's midterm grade. At the 1% significance level, is a student's midterm grade significant in explaining a student's final grade? Show the relevant steps of the test.

64. FILE ***Property_Taxes.*** The accompanying data file shows the property taxes and the square footage for 20 homes in an affluent suburb 30 miles outside of New York City. Estimate a home's property taxes as a linear function of its square footage. At the 5% significance level, is square footage significant in explaining property taxes? Show the relevant steps of the test.

65. FILE ***Fertilizer.*** A horticulturist is studying the relationship between tomato plant height and fertilizer amount. Thirty tomato plants grown in similar conditions were subjected to various amounts of fertilizer (in ounces) over a four-month period, and then their heights (in inches) were measured.
a. Estimate: Height $= \beta_0 + \beta_1$Fertilizer $+ \varepsilon$.
b. At the 5% significance level, determine if an ounce of fertilizer increases height by more than 3 inches. Show the relevant steps of the test.

66. FILE ***Dexterity.*** Finger dexterity, the ability to make precisely coordinated finger movements to grasp or assemble very small objects, is important in jewelry making. Thus, the manufacturing manager at Gemco, a manufacturer of high-quality watches, wants to develop a regression model to predict the productivity, measured by watches per shift, of new employees based on the time required (in seconds) to place 3 pins in each of 100 small holes using tweezers. He has subjected a sample of 20 current employees to the O'Connor dexterity test in which the time required to place the pins and the number of watches produced per shift are measured.
a. Estimate: Watches $= \beta_0 + \beta_1$Time $+ \varepsilon$.
b. The manager claims that for every extra second taken on placing the pins, the number of watches produced decreases by more than 0.02. Test this claim at the 5% significance level. Show the relevant steps of the test.

67. FILE ***Engine_Overhaul.*** The maintenance manager at a trucking company wants to build a regression model to forecast the time until the first engine overhaul (Time in years) based on four explanatory variables: (1) annual miles driven (Miles in 1,000s), (2) average load weight (Load in tons), (3) average driving speed (Speed in mph), and (4) oil change interval (Oil in 1,000s miles). Based on driver logs and onboard computers, data have been obtained for a sample of 25 trucks.
a. Estimate the time until the first engine overhaul as a function of all four explanatory variables.
b. At the 10% significance level, are the explanatory variables jointly significant? Show the relevant steps of the test.
c. Are the explanatory variables individually significant at the 10% significance level? Show the relevant steps of the test.

68. FILE ***Electricity_Cost.*** The facility manager at a pharmaceutical company wants to build a regression model to forecast monthly electricity cost. Three main variables are thought to influence electricity cost: (1) average outdoor temperature (Temp in °F), (2) working days per month (Days), and (3) tons of product produced (Tons).
a. Estimate the regression model.
b. At the 10% significance level, are the explanatory variables jointly significant? Show the relevant steps of the test.
c. Are the explanatory variables individually significant at the 10% significance level? Show the relevant steps of the test.

69. FILE ***Caterpillar.*** Caterpillar, Inc. manufactures and sells heavy construction equipment worldwide. The performance of Caterpillar's stock is likely to be strongly influenced by the economy. For instance, during the subprime mortgage crisis, the value of Caterpillar's stock plunged dramatically. Monthly data for Caterpillar's risk-adjusted return ($R - R_f$) and the risk-adjusted market return ($R_M - R_f$) are collected for a five-year period ($n = 60$). A portion of the data is shown in the accompanying table.

Date	$R - R_f$	$R_M - R_f$
1/1/2006	17.66	2.21
2/1/2006	7.27	−0.31
⋮	⋮	⋮
11/1/2010	3.37	2.15

Source: http://finance.yahoo.com and U.S. Treasury.

a. Estimate the CAPM model for Caterpillar, Inc. Show the regression results in a well-formatted table.
b. At the 5% significance level, determine if investment in Caterpillar is riskier than the market (beta significantly greater than 1).
c. At the 5% significance level, is there evidence of abnormal returns?

70. FILE ***Arlington_Homes.*** A realtor examines the factors that influence the price of a house in Arlington, Massachusetts. He collects data on recent house sales (Price) and notes each house's square footage (Sqft) as well as its number of bedrooms (Beds) and number of bathrooms (Baths). A portion of the data is shown in the accompanying table.

Price	Sqft	Beds	Baths
840000	2768	4	3.5
822000	2500	4	2.5
⋮	⋮	⋮	⋮
307500	850	1	1

a. Estimate: Price $= \beta_0 + \beta_1$Sqft $+ \beta_2$Beds $+ \beta_3$Baths $+ \varepsilon$. Show the regression results in a well-formatted table.
b. At the 5% significance level, are the explanatory variables jointly significant in explaining Price?
c. At the 5% significance level, are all explanatory variables individually significant in explaining Price?

71. FILE ***Final_Test.*** On the first day of class, an economics professor administers a test to gauge the math preparedness of her students. She believes that the performance on this math test and the number of hours studied per week on the course are the primary factors that predict a student's score on the final exam. She collects data from 60 students, a portion of which is shown in the accompanying table.

Final	Math	Hours
94	92	5
74	90	3
⋮	⋮	⋮
63	64	2

a. Estimate the sample regression equation that enables us to predict a student's final exam score on the basis of his/her math score and the number of hours studied per week.
b. At the 5% significance level, are a student's math score and the number of hours studied per week jointly significant in explaining a student's final exam score?
c. At the 5% significance level, is each explanatory variable individually significant in explaining a student's final exam score?

12.5 MODEL ASSUMPTIONS AND COMMON VIOLATIONS

So far we have focused on the estimation and the assessment of linear regression models. It is important to understand that the statistical properties of the OLS estimator, as well as the validity of the testing procedures, depend on the assumptions of the classical linear regression model. In this section, we discuss these assumptions. We also address common violations to the assumptions, discuss the consequences when the assumptions are violated, and, where possible, offer some remedies.

REQUIRED ASSUMPTIONS OF REGRESSION ANALYSIS

1. The regression model given by $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$ is *linear in the parameters*, $\beta_0, \beta_1, \ldots, \beta_k$.
2. Conditional on $x_1, x_2, \ldots, x_k$, the error term has an *expected value of zero*, or $E(\varepsilon) = 0$. This implies that $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.
3. There is no exact linear relationship among the explanatory variables; or, in statistical terminology, there is *no perfect multicollinearity*.
4. Conditional on $x_1, x_2, \ldots, x_k$, the variability of the error term ε is the same for all observations; or, in statistical terminology, there is *no heteroskedasticity*. The assumption is violated if observations have a *changing variability*.

5. Conditional on $x_1, x_2, \ldots, x_k$, the error term ε is uncorrelated across observations; or, in statistical terminology, there is *no serial correlation*. The assumption is violated if *observations are correlated*.
6. The error term ε is not correlated with any of the explanatory variables $x_1, x_2, \ldots, x_k$; or, in statistical terminology, there is *no endogeneity*. In general, this assumption is violated if important *explanatory variables are excluded*.
7. Conditional on $x_1, x_2, \ldots, x_k$, the error term ε is *normally distributed*. This assumption allows us to construct interval estimates and conduct tests of significance. If ε is not normally distributed, the interval estimates and the hypothesis tests are valid only for large sample sizes.

Under the assumptions of the classical linear regression model, the OLS estimators have desirable properties. In particular, the OLS estimators of the regression coefficients β_j are unbiased; that is, $E(b_j) = \beta_j$. Moreover, among all linear unbiased estimators, they have minimum variations between samples. These desirable properties of the OLS estimators become compromised as one or more model assumptions are violated. Aside from coefficient estimates, the validity of the significance tests is also impacted by the assumptions. For certain violations, the estimated standard errors of the OLS estimators are inappropriate; in these cases it is not possible to make meaningful inferences from the t and the F test results.

The assumptions of the classical linear regression model are, for the most part, based on the error term ε. Since the residuals, or the observed error term, $e = y - \hat{y}$, contain useful information regarding ε, it is common to use the residuals to investigate the assumptions. In this section, we will rely on **residual plots** to detect some of the common violations to the assumptions. These graphical plots are easy to use and provide informal analysis of the estimated regression models. Formal tests are beyond the scope of this text.

RESIDUAL PLOTS

For the regression model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$, the residuals are computed as $e = y - \hat{y}$, where $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$. These residuals can be plotted sequentially or against an explanatory variable x_j to look for model inadequacies.

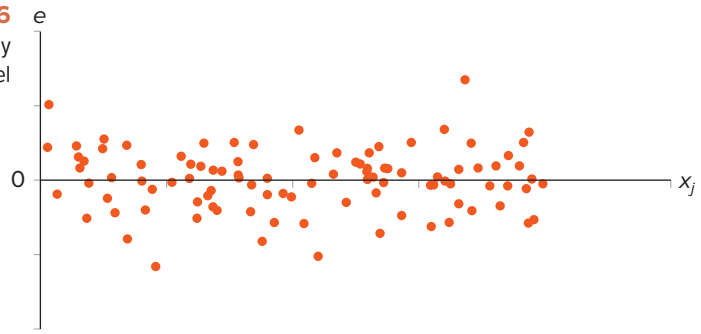
It is common to plot the residuals e on the vertical axis and the explanatory variable x_j, or the predicted values $\hat{y}$ on the horizontal axis. Such plots are useful for detecting departures from linearity as well as constant variability. If the regression is based on time series data, we can plot the residuals sequentially to detect if the observations are correlated.

Residual plots can also be used to detect outliers. Recall that outliers are observations that stand out from the rest of the data. For an outlier observation, the resulting residual will appear distinct in a plot; it will stand out from the rest. While outliers can greatly impact the estimates, it is not always clear what to do with them. As mentioned in Chapter 3, outliers may indicate bad data due to incorrectly recorded (or included) observations in the data set. In such cases, the relevant observation should be corrected or simply deleted. Alternatively, outliers may just be due to random variations, in which case the relevant observations should remain. In any event, residual plots help us identify potential outliers so that we can take corrective actions, if needed.

In Figure 12.6, we present a hypothetical residual plot when none of the assumptions has been violated. Note that all the points are randomly dispersed around the zero value

of the residuals. Also, there is no evidence of outliers since no residual stands out from the rest. Any discernible pattern of the residuals indicates that one or more assumptions have been violated.

FIGURE 12.6 Residual plot of a correctly specified model



We discuss how to obtain residual plots in Excel at the end of this section, but first we describe common violations of the assumptions and offer remedies.

Common Violation 1: Nonlinear Patterns

LO **12.6**

Address common violations of the OLS assumptions.

Linear regression models are often justified on the basis of their computational simplicity. The model $y = \beta_0 + \beta_1 x + \varepsilon$ implies that if x goes up by one unit, we expect y to change by β_1, irrespective of the value of x. However, in many applications, the relationship cannot be represented by a straight line and, therefore, must be captured by an appropriate curve. It is always good to rely on economic theory and intuition to determine if the linearity assumption is appropriate. We confirm our intuition by analyzing scatterplots or residual plots. The OLS estimates can be quite misleading if there are obvious nonlinear patterns in the data.

Detection

We can use residual plots to identify nonlinear patterns. Linearity is justified if the residuals are randomly dispersed across the values of an explanatory variable. A discernible trend in the residuals is indicative of nonlinear patterns.

EXAMPLE 12.8

A sociologist wishes to study the relationship between age and happiness. He interviews 24 individuals and collects data on age and happiness, measured on a scale from 0 to 100. A portion of the data is shown in Table 12.12. Examine the linearity assumption in the regression model, Happiness $= \beta_0 + \beta_1$Age $+ \varepsilon$.

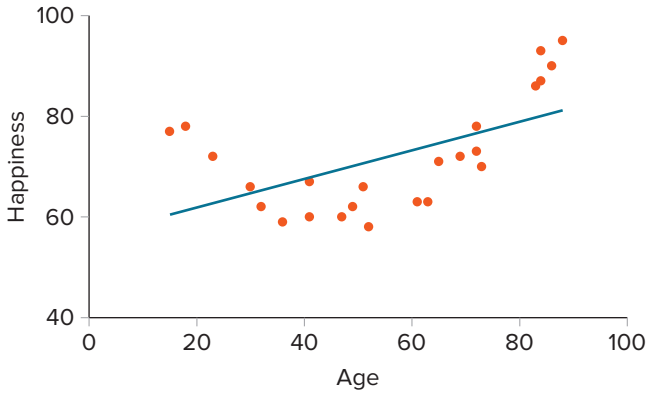
TABLE 12.12 Happiness and Age

FILE *Happiness_Age*

Happiness	Age
62	49
66	51
⋮	⋮
72	69

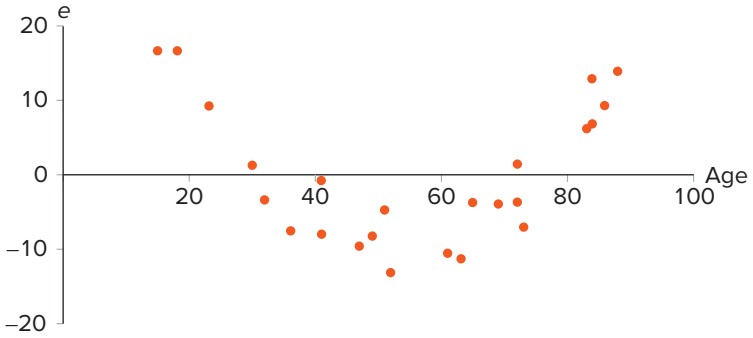
SOLUTION: We start the analysis with a scatterplot of Happiness against Age. Figure 12.7 shows the scatterplot and the superimposed trend line, which is based on the sample regression equation, $\widehat{\text{Happiness}} = 56.18 + 0.28$Age. It is fairly clear from Figure 12.7 that the linear regression model does not appropriately capture the relationship between Happiness and Age. In other words, it is misleading to conclude that a person's happiness increases by 0.28 units every year.

FIGURE 12.7 Scatterplot and the superimposed trendline (Example 12.8)



A residual plot, shown in Figure 12.8, further explores the linearity assumption of the regression model.

FIGURE 12.8 Residual plot against Age (Example 12.8)



The above residual plot shows that there is an obvious trend with the residuals decreasing until the age of 50 and steadily increasing thereafter. The linear regression model is inappropriate as it underestimates at lower and higher age levels and overestimates in the middle. This result is consistent with a report that shows that happiness initially decreases with age and then increases with age (*The Economist*, December 16, 2010).

Remedy

Linear regression models are often used as a first pass for most empirical work. In many instances, they provide a very good approximation for the actual relationship. However, if residual plots exhibit strong nonlinear patterns, the inferences made by a linear regression model can be quite misleading. In such instances, we should employ nonlinear regression methods based on simple transformations of the response and the explanatory variables; these methods are discussed in the next chapter.

Common Violation 2: Multicollinearity

Perfect multicollinearity exists when two or more explanatory variables have an exact linear relationship. Consider the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where y is bonus, x_1 is the number of cars sold, and x_2 is the number of cars remaining in the lot. If all car salesmen started with the same inventory, we have a case of *perfect* multicollinearity ($x_2 = \text{Constant} - x_1$). Perfect multicollinearity is easy to detect because the model cannot

be estimated. However, if x_2 represents the proportion of positive reviews from customers, we have *some* **multicollinearity** since the number of cars sold and the proportion of positive reviews are likely to be correlated. In most applications, some degree of correlation exists between the explanatory variables.

The problem with (non-perfect) multicollinearity is similar to that of small samples. Multicollinearity does not violate any of the assumptions; however, its presence results in imprecise estimates of the slope coefficients. In other words, multicollinearity makes it difficult to disentangle the separate influences of the explanatory variables on the response variable. If multicollinearity is severe, we may find insignificance of important explanatory variables; some coefficient estimates may even have wrong signs.

Detection

The detection methods for multicollinearity are mostly informal. The presence of a high R^2 coupled with individually insignificant explanatory variables can indicate multicollinearity. Sometimes researchers examine the correlations between the explanatory variables to detect severe multicollinearity. One such guideline suggests that multicollinearity is severe if the sample correlation coefficient between any two explanatory variables is more than 0.80 or less than −0.80. Seemingly wrong signs of the estimated regression coefficients may also indicate multicollinearity.

EXAMPLE 12.9

Examine the multicollinearity issue in a linear regression model that uses median home values (in $) as the response variable and median household incomes (in $), per capita incomes (in $), and the proportion of owner-occupied homes (in %) as the explanatory variables. A portion of 2010 data for all states in the United States is shown in Table 12.13.

TABLE 12.13 Home Values and Other Factors

State	Home Value	HH Income	Per Cap Inc	Pct Owner Occ
Alabama	117600	42081	22984	71.1
Alaska	229100	66521	30726	64.7
⋮	⋮	⋮	⋮	⋮
Wyoming	174000	53802	27860	70.2

Source: 2010 U.S. Census.

FILE *Home_Values*

SOLUTION: We estimate three models to examine the multicollinearity issue; Table 12.14 presents the regression results.

TABLE 12.14 Summary of Model Estimates (Example 12.9)

Variable	Model 1	Model 2	Model 3
Intercept	417,892.04* (0.001)	348,187.14* (0.002)	285,604.08 (0.083)
HH Income	9.04* (0.000)	7.74* (0.000)	NA
Per Cap Inc	−3.27 (0.309)	NA	13.21* (0.000)
Pct Owner Occ	−8,744.30* (0.000)	−8,027.90* (0.000)	−6,454.08* (0.001)
Adjusted R^2	0.8071	0.8069	0.6621

Notes: The table contains parameter estimates with *p*-values in parentheses; * represents significance at the 5% level. NA denotes not applicable. Adjusted R^2, reported in the last row, is used for model selection.

Model 1 uses all three explanatory variables to explain home values. Surprisingly, the per capita income variable has a negative estimated coefficient of −3.27 and, with a p-value of 0.31, is not even statistically significant at the 5% level. Multicollinearity might be the reason for this surprising result since household income and per capita income are likely to be correlated. We compute the sample correlation coefficient between these two variables as 0.8582, which suggests that multicollinearity is severe. We estimate two more models where one of these collinear variables is removed; Model 2 removes per capita income and Model 3 removes household income. Note that per capita income in Model 3 now exerts a positive and significant influence on home values. Between these two models, Model 2 is preferred to Model 3 because of its higher adjusted R^2 (0.8069 > 0.6621). The choice between Model 1 and Model 2 is unclear. In general, Model 1, with the highest adjusted R^2 value of 0.8071, is preferred if the sole purpose of the analysis is to make predictions. However, if the coefficient estimates need to be evaluated, then Model 2 may be the preferred choice.

Remedy

Inexperienced researchers tend to include too many explanatory variables in their quest not to omit anything important and in doing so may include redundant variables that essentially measure the same thing. When confronted with multicollinearity, a good remedy is to drop one of the collinear variables. The difficult part is to decide which of the collinear variables is redundant and, therefore, can safely be removed. Another option is to obtain more data, since the sample correlation may get weaker as we include more observations. Sometimes it helps to express the explanatory variables differently so that they are not collinear. At times, the best approach may be to *do nothing* when there is a justification to include all explanatory variables. This is especially so if the estimated model yields a high R^2, which implies that the estimated model is good for prediction.

Common Violation 3: Changing Variability

The assumption of constant variability of observations often breaks down in studies with cross-sectional data. Consider the model $y = \beta_0 + \beta_1 x + \varepsilon$, where y is a household's consumption expenditure and x is its disposable income. It may be unreasonable to assume that the variability of consumption is the same across a cross-section of household incomes. For example, we would expect higher-income households to have a higher variability in consumption as compared to lower-income households. Similarly, home prices tend to vary more as homes get larger, and sales tend to vary more as firm size increases.

In the presence of **changing variability**, the OLS estimators are still unbiased. However, the estimated standard errors of the OLS estimators are inappropriate. Consequently, we cannot put much faith in the standard t or F tests since they are based on these estimated standard errors.

Detection

We can use residual plots to gauge changing variability. The residuals are generally plotted against each explanatory variable x_j; for a multiple regression model, we can also plot them against the predicted value $\hat{y}$. There is no violation if the residuals are randomly dispersed across the values of x_j. On the other hand, there is a violation if the variability increases or decreases over the values of x_j.

EXAMPLE 12.10

Consider a simple regression model that relates a store's monthly sales (Sales in $1,000s) with its square footage (Sqft) for a chain of 40 convenience stores. A portion of the data used for the analysis is shown in Table 12.15. Estimate the model and use a residual plot to determine if the observations have a changing variability.

TABLE 12.15 Sales and Square Footage of Convenience Stores

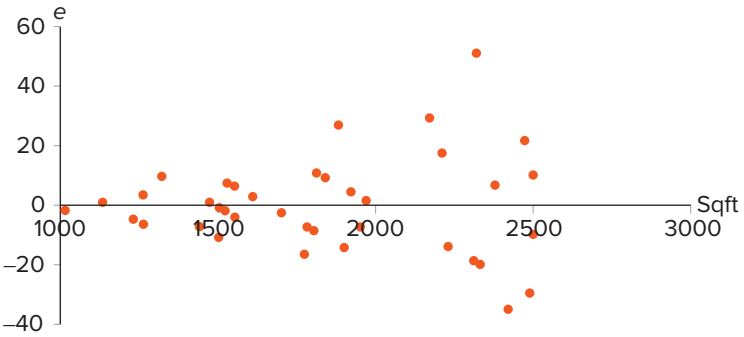
Sales	Sqft
140	1810
160	2500
⋮	⋮
110	1470

FILE
Convenience_Stores

SOLUTION: The sample regression is given by $\widehat{\text{Sales}} = 22.0795 + 0.0591\text{Sqft}$.

A residual plot of the estimated model is shown in Figure 12.9. Note that the residuals seem to fan out across the horizontal axis. Therefore, we conclude that changing variability is a likely problem in this application relating sales to square footage. This result is not surprising, since you would expect sales to vary more as square footage increases. For instance, a small convenience store is likely to include only bare essentials for which there is a fairly stable demand. A larger store, on the other hand, may include specialty items, resulting in more fluctuation in sales.

FIGURE 12.9 Residual plot against square footage (Example 12.10)



Remedy

As mentioned earlier, in the presence of changing variability, the OLS estimators are unbiased but their estimated standard errors are inappropriate. Therefore, OLS still provides reasonable coefficient estimates, but the t and the F tests are no longer valid. This has prompted some researchers to use the OLS estimates along with a correction for the standard errors, often referred to as robust standard errors. Unfortunately, the current version of Excel does not include a correction for the standard errors.

Common Violation 4: Correlated Observations

When obtaining the OLS estimators, we assume that the observations are uncorrelated. This assumption often breaks down in studies with time series data. Variables such as GDP, employment, and asset returns exhibit business cycles. As a consequence, successive observations are likely to be correlated.

In the presence of **correlated observations**, the OLS estimators are unbiased, but their estimated standard errors are inappropriate. Generally, these standard errors are distorted downward, making the model look better than it really is with a spuriously high R^2. Furthermore, the t and F tests may suggest that the explanatory variables are individually and jointly significant when this is not true.

Detection

We can plot the residuals sequentially over time to look for correlated observations. If there is no violation, then the residuals should show no pattern around the horizontal

EXAMPLE 12.11

Consider $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ where y represents sales (in \$1,000s) at a sushi restaurant and x_1 and x_2 represent advertising costs (AdsCost in \$) and the unemployment rate (Unemp in %), respectively. A portion of monthly data from January 2008 to June 2009 is given in Table 12.16. Inspect the behavior of the residuals to comment on serial correlation.

FILE
Sushi_Restaurant

TABLE 12.16 Sales, Advertising Costs, and Unemployment Data for Example 12.11

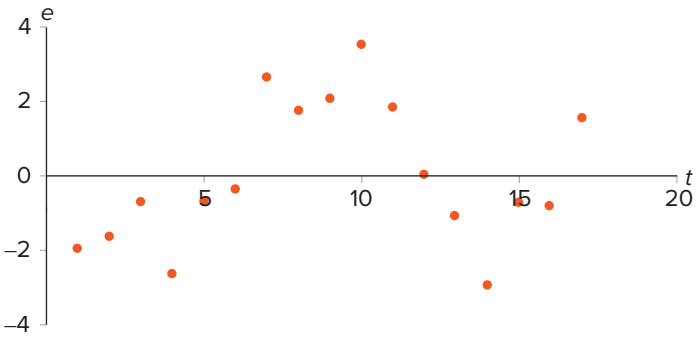
Month	Year	Sales	AdsCost	Unemp
January	2008	27.0	550	4.6
February	2008	24.2	425	4.3
⋮	⋮	⋮	⋮	⋮
May	2009	27.4	550	9.1

Source for the unemployment rate data: Development Department, State of California, June 2009.

SOLUTION: The sample regression is given by $\hat{y} = 17.5060 + 0.0266x_1 - 0.6879x_2$.

In order to detect serial correlation, we plot the residuals sequentially against time t, where t is given by 1, 2, . . ., 17 for the 17 months of time series data. Figure 12.10 shows a wavelike movement in the residuals over time, first clustering below the horizontal axis, then above the horizontal axis, and so on. Given this pattern around the horizontal axis, we conclude that the observations are correlated.

FIGURE 12.10 Scatterplot of residuals against time t



axis. A violation is indicated when a positive residual in one period is followed by positive residuals in the next few periods, followed by negative residuals for a few periods, then positive residuals, and so on. Although not as common, a violation is also indicated when a positive residual is followed by a negative residual, then a positive residual, and so on.

Remedy

As mentioned earlier, in the presence of correlated observations, the OLS estimators are unbiased but their standard errors are inappropriate and generally distorted downward, making the model look better than it really is. Therefore, OLS still provides reasonable coefficient estimates, but the t and the F tests are no longer valid. This has prompted some researchers to use the OLS estimates along with a correction for the standard errors, often referred to as robust standard errors. As in the case of changing variability, the current version of Excel does not include this correction.

Common Violation 5: Excluded Variables

Another crucial assumption in a linear regression model is that the error term is not correlated with the explanatory variables. In general, this assumption breaks down when important explanatory variables are excluded. If one or more of the relevant explanatory variables are excluded, then the resulting OLS estimators are biased. The extent of the bias depends on the degree of the correlation between the included and the excluded explanatory variables.

Suppose we want to estimate $y = \beta_0 + \beta_1 x + \varepsilon$, where y is salary and x is years of education. This model excludes innate ability, which is an important ingredient for salary. Since innate ability is omitted, it gets incorporated in the error term and the resulting error term is likely to be correlated with years of education. Now consider someone who is highly educated and also commands a high salary. The model will associate high salary with education, when, in fact, it may be the person's unobserved high level of innate ability that has raised both education and salary. In sum, this violation leads to unreliable coefficient estimates; some estimates may even have the wrong signs.

Remedy

It is important that we include all relevant explanatory variables in the regression model. An important first step before running a regression model is to compile a comprehensive list of potential explanatory variables. We can then build down to perhaps a smaller list of explanatory variables using the adjusted R^2 criterion. Sometimes, due to data limitations, we are unable to include all relevant variables. For example, innate ability may be an important explanatory variable for a model that explains salary, but we are unable to include it since innate ability is not observable. In such instances, we use a technique called the instrumental variable technique, which is outside the scope of this text.

Summary

Regression models are an integral part of business statistics. It takes practice to become an effective user of the regression methodology. We should think of regression modeling as an iterative process. We start with a clear understanding of what the regression model is supposed to do. We define the relevant response variable and compile a comprehensive list of potential explanatory variables. The emphasis should be to pick a model that makes economic and intuitive sense and avoid explanatory variables that more or less measure the same thing, thus causing multicollinearity.

We then apply this model to data and refine and improve its fit. Specifically, from the comprehensive list, we build down to perhaps a smaller list of explanatory variables using significance tests and goodness-of-fit measures such as the standard error of the estimate and the adjusted R^2. It is important that we explore residual plots to look for signs of changing variability and correlated observations in cross-sectional and time series studies, respectively. If we identify any of these two violations, we can still trust the coefficient estimates of the regression coefficients. However, we cannot place much faith in the standard t or F tests of significance unless we employ the necessary correction.

Using Excel to Construct Residual Plots

We replicate Figure 12.9.

FILE
Convenience_Stores

A. Open the ***Convenience_Stores*** data file.

B. From the menu, choose **Data > Data Analysis > Regression**.

C. For *Input Y Range*, select the Sales data, and for *Input X Range*, select the Sqft data.

D. Select *Residual Plots*.

E. Click **OK**. Formatting (regarding colors, axes, etc.) can be done by selecting **Format** from the menu.

We replicate Figure 12.10.

FILE
Sushi_Restaurant

A. Open the ***Sushi_Restaurant*** data file.

B. From the menu, choose **Data > Data Analysis > Regression**.

C. For *Input Y Range*, select the Sales data, and for *Input X Range*, simultaneously select the AdsCost and Unemp data.

D. Select *Residuals*. Click **OK**.

E. Given the regression output, select the residual data and choose **Insert > Scatter**; choose the option on the top left. (If you are having trouble finding this option after selecting **Insert**, look for the graph with data points above **Charts**.) Formatting (regarding colors, axes, etc.) can be done by selecting **Format** from the menu.

EXERCISES 12.5

Mechanics

72. A simple linear regression, $y = \beta_0 + \beta_1 x + \varepsilon$, is estimated with cross-sectional data. The resulting residuals e along with the values of the explanatory variable x are shown in the accompanying table.

x	4	5	6	7	8	9	10	11
e	14	4	−6	−16	−11	−1	7	9

a. Graph the residuals e against the values of the explanatory variable x and look for any discernible pattern.

b. Which assumption is being violated? Discuss its consequences and suggest a possible remedy.

73. Using 20 observations, the multiple regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ was estimated. A portion of the regression results is shown in the accompanying table.

a. At the 5% significance level, are the explanatory variables jointly significant?

b. At the 5% significance level, is each explanatory variable individually significant?

c. What is the likely problem with this model?

	df	SS	MS	F	Significance F	
Regression	2	2.12E+12	1.06E+12	56.556	3.07E-08	
Residual	17	3.19E+11	1.88E+10			
Total	19	2.44E+12				
	Coefficients	Standard Error	t Stat	p-value	Lower 95%	Upper 95%
Intercept	−987557	131583	−7.505	0.000	−1265173	−709941
x_1	29233	32653	0.895	0.383	−39660	98125
x_2	30283	32645	0.928	0.367	−38592	99158

74. A simple linear regression, $y = \beta_0 + \beta_1 x + \varepsilon$, is estimated with cross-sectional data. The resulting residuals e along with the values of the explanatory variable x are shown in the accompanying table.

x	1	2	5	7	10	14	15	20	24	30
e	−2	1	−3	2	4	−5	−6	8	11	−10

a. Graph the residuals e against the values of the explanatory variable x and look for any discernible pattern.

b. Which assumption is being violated? Discuss its consequences and suggest a possible remedy.

75. A simple linear regression, $y = \beta_0 + \beta_1 x + \varepsilon$, is estimated with time series data. The resulting residuals e and the time variable t are shown in the accompanying table.

t	1	2	3	4	5	6	7	8	9	10
e	−5	−4	−2	3	6	8	4	−5	−3	−2

a. Graph the residuals against time and look for any discernible pattern.

b. Which assumption is being violated? Discuss its consequences and suggest a possible remedy.

Applications

76. FILE ***Television.*** Numerous studies have shown that watching too much television hurts school grades. Others have argued that television is not necessarily a bad thing for children (*Mail Online*, July 18, 2009). Like books and stories, television not only entertains, it also exposes a child to new information about the world. While watching too much television is harmful, a little bit may actually help. Researcher Matt Castle gathers information on the grade point average (GPA) of 28 middle-school children and the number of hours of television they watched per week. Examine the linearity assumption in the regression model, $\text{GPA} = \beta_0 + \beta_1 \text{Hours} + \varepsilon$.

77. FILE ***Delivery.*** Quick2U, a delivery company, would like to standardize its delivery charge model for shipments (Charge in $) such that customers will better understand their delivery costs. Three explanatory variables are used: (1) distance (in miles), (2) shipment weight (in lbs), and (3) number of boxes. A sample of 30 recent deliveries is collected; a portion of the data is shown in the accompanying table.

Charge	Distance	Weight	Boxes
92.50	29	183	1
157.60	96	135	3
⋮	⋮	⋮	⋮
143.00	47	117	7

a. Estimate the model $\text{Charge} = \beta_0 + \beta_1 \text{Distance} + \beta_2 \text{Weight} + \beta_3 \text{Boxes} + \varepsilon$ and examine the joint and individual significance of the explanatory variables at the 1% level.

b. Is there any evidence of multicollinearity?

c. Graph the residuals against the predicted values and determine if there is any evidence of changing variability.

78. Consider the results of a survey where students were asked about their GPA and also to break down their typical 24-hour day into study, leisure (including work), and sleep. Consider the model $\text{GPA} = \beta_0 + \beta_1 \text{Study} + \beta_2 \text{Leisure} + \beta_3 \text{Sleep} + \varepsilon$.

a. What is wrong with this model?

b. Suggest a simple way to reformulate the model.

79. FILE ***AnnArbor_Rental.*** Consider the monthly rent (Rent in $) of a home in Ann Arbor, Michigan, as a function of the number of bedrooms (Beds), the number of bathrooms (Baths), and square footage (Sqft).

a. Estimate: $\text{Rent} = \beta_0 + \beta_1 \text{Beds} + \beta_2 \text{Baths} + \beta_3 \text{Sqft} + \varepsilon$.

b. Which of the explanatory variables might cause changing variability? Explain.

c. Use residual plots to verify your economic intuition.

80. FILE ***Work_Experience.*** Consider the accompanying data on salary (in $) and work experience (in years) for 100 employees in a marketing firm. Estimate: $\text{Salary} = \beta_0 + \beta_1 \text{Experience} + \varepsilon$.

a. Explain why you would be concerned about changing variability in this application.

b. Use a residual plot to confirm your economic intuition.

81. FILE ***Healthy_Living.*** Healthy living has always been an important goal for any society. In an ad campaign for Walt Disney, former First Lady Michelle Obama shows parents and children that eating well and exercising can also be fun (*USA Today*, September 30, 2010). Consider a regression model that conjectures that fruits and vegetables and regular exercising have a positive effect on health and smoking has a negative effect on health. The sample consists of the percentage of these variables observed in various states in the United States in 2009. A portion of the data is shown in the accompanying table.

State	Healthy	Fruits/Vegetables	Exercise	Smoke
AK	88.7	23.3	60.6	14.6
AL	78.3	20.3	41.0	16.4
⋮	⋮	⋮	⋮	⋮
WY	87.5	23.3	57.2	15.2

Source: Centers for Disease Control and Prevention.

a. Estimate: Healthy $= \beta_0 + \beta_1$Fruits/Vegetables $+ \beta_2$Exercise $+ \beta_3$Smoke $+ \varepsilon$.
b. Analyze the data to determine if multicollinearity and changing variability are present.

82. FILE ***Johnson_Johnson.*** A capital asset pricing model (CAPM) for Johnson & Johnson (J&J) was discussed in Example 12.6. The model uses the risk-adjusted stock return $R - R_f$ for J&J as the response variable and the risk-adjusted market return $R_M - R_f$ as the explanatory variable. Since serial correlation may occur with time series data, it is prudent to inspect the behavior of the residuals. Construct a scatterplot of the residuals against time to comment on correlated observations.

83. FILE ***Consumption_Quarterly.*** The consumption function is one of the key relationships in economics, where consumption y depends on disposable income x. Consider the quarterly data for these seasonally adjusted variables, measured in billions of dollars. A portion of the data is shown in the accompanying table.

Date	Consumption	Disposable Income
2006:01	9148.2	9705.2
2006:02	9266.6	9863.8
⋮	⋮	⋮
2010:04	10525.2	11514.7

Source: U.S. Department of Commerce.

a. Estimate: Consumption $= \beta_0 + \beta_1$ Disposable Income $+ \varepsilon$. Plot the residuals against time to determine if there is a possibility of correlated observations.
b. Discuss the consequences of correlated observations and suggest a possible remedy.

84. FILE ***Mowers.*** The marketing manager at Turfco, a lawn mower company, believes that monthly sales across all outlets (stores, online, etc.) are influenced by three key variables: (1) outdoor temperature (in °F), (2) advertising expenditures (in $1,000s), and (3) promotional discounts (in %). A portion of the monthly sales data for the past two years is shown in the accompanying table.

Sales	Temperature	Advertising	Discount
17235	33	15	5.0
19854	42	25	5.0
⋮	⋮	⋮	⋮
22571	44	21	5.0

a. Estimate: Sales $= \beta_0 + \beta_1$Temperature $+ \beta_2$Advertising $+ \beta_3$Discount $+ \varepsilon$, and test for the joint and individual significance of the explanatory variables at the 5% level.
b. Examine the data for evidence of multicollinearity. Provide two reasons why it might be best to do nothing about multicollinearity in this application.
c. Examine the residual plots for evidence of changing variability.

WRITING WITH STATISTICS



©Matthew Cavanaugh/EPA/REX/Shutterstock

Gavin Cann listens to two sports analysts quarrel over which statistic is a better predictor of a Major League Baseball team's winning proportion (Win). One argues that the team's batting average (BA) is a better predictor of a team's success since the team with the higher batting average has won approximately 75% of the World Series contests. The other insists that a team's pitching is clearly the main factor in determining a team's winning proportion—the lower a team's earned run average (ERA), the higher the team's winning proportion.

In order to determine if either of these claims is backed by the data, Gavin collects relevant information for the 14 American League (AL) and 16 National League (NL) teams during the regular season of 2010. A portion of the data is shown in Table 12.17.

TABLE 12.17 Winning Proportion, Batting Average, and Earned Run Average in Baseball

Team	League	Win	BA	ERA
Baltimore, Orioles	AL	0.407	0.259	4.59
Boston, Red Sox	AL	0.549	0.268	4.20
⋮	⋮	⋮	⋮	⋮
Washington, Nationals	NL	0.426	0.25	4.13

SOURCE: http://mlb.mlb.com.

FILE
Baseball

Gavin wants to use this sample information to:

1. Estimate three linear regression models where winning proportion is based on BA (Model 1), ERA (Model 2), and both BA and ERA (Model 3).
2. Use goodness-of-fit measures to determine which of the three models best fits the data.
3. Determine the individual and the joint significance of BA and ERA at the 5% significance level.

Sample Report—Analyzing the Winning Proportion in Baseball

Two sports analysts have conflicting views over how best to predict a Major League Baseball team's winning proportion (Win). One argues that the team's batting average (BA) is a better predictor of a team's success, while the other analyst insists that a team's pitching is the main factor as measured by the pitchers' earned run average (ERA). Three linear regression models are used to analyze a baseball team's winning proportion. The explanatory variables are BA in Model 1, ERA in Model 2, and both BA and ERA in Model 3. A priori, one expects that BA positively influences Win, whereas ERA negatively affects Win. The regression results for the three models are presented in Table 12.A.

TABLE 12.A Model Estimates for the Response Variable Win

Variable	Model 1	Model 2	Model 3
Intercept	−0.2731 (0.342)	0.9504* (0.000)	0.1269 (0.492)
Batting Average	3.0054* (0.011)	NA	3.2754* (0.000)
Earned Run Average	NA	−0.1105* (0.000)	−0.1153* (0.000)
s_e	0.0614	0.0505	0.0375
R^2	0.2112	0.4656	0.7156
Adjusted R^2	0.1830	0.4465	0.6945
F statistic (p-value)	NA	NA	33.9663* (0.000)

NOTES: Parameter estimates are in the top half of the table with the p-values in parentheses; NA denotes not applicable;
* represents significance at the 5% level. The lower part of the table contains goodness-of-fit measures.

If simply choosing between Models 1 and 2 where only one explanatory variable influences Win, then Model 2 with ERA as an explanatory variable appears to provide the better fit since it has a lower standard error of the estimate ($s_e = 0.0505$) and a higher coefficient of determination ($R^2 = 0.4656$). However, Model 3 uses both BA and ERA as explanatory variables and it has the lowest standard error of the estimate ($s_e = 0.0375$) and the highest adjusted R^2 with a value of 0.6945. Thus, Model 3 provides the best overall fit with a sample regression equation of $\widehat{\text{Win}} = 0.13 + 3.28\text{BA} - 0.12\text{ERA}$. As expected, the slope coefficient of BA is positive and the slope coefficient of ERA is negative.

Further testing of Model 3 reveals that the two explanatory variables are jointly as well as individually significant in explaining a team's winning proportion at the 5% significance level. It appears that neither sports analyst is totally right or totally wrong. With an $R^2 = 0.7156$, approximately 72% of the sample variability in the winning proportion is explained by the estimated Model 3. However, 28% of the variability in winning proportion remains unexplained. This is not entirely surprising, since other factors, besides a team's batting average and earned run average, influence a baseball team's winning proportion.

CONCEPTUAL REVIEW

LO 12.1 Estimate and interpret a simple linear regression model.

Regression analysis allows us to analyze the relationship between the target variable, called the **response variable**, and other variables, called the **explanatory variables**.

The **simple linear regression model** uses only one explanatory variable to predict and/or describe change in the response variable. The model is expressed as $y = \beta_0 + \beta_1 x + \varepsilon$, where y is the response variable, x is the explanatory variable, and ε is the random error term. The coefficients β_0 and β_1 are the unknown parameters to be estimated.

We apply the **ordinary least squares (OLS)** method to find a sample regression equation, $\hat{y} = b_0 + b_1 x$, where $\hat{y}$ is the predicted value of the response variable and b_0 and b_1 are the estimates of β_0 and β_1, respectively. The estimated slope coefficient b_1 represents the change in $\hat{y}$ when x increases by one unit.

LO 12.2 Estimate and interpret a multiple linear regression model.

The **multiple linear regression model** allows us to analyze the linear relationship between the response variable and two or more explanatory variables. It is defined as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$, where y is the response variable, $x_1, x_2, \ldots, x_k$ are the k explanatory variables, and ε is the random error term. The coefficients $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters to be estimated. We again use the OLS method to arrive at the following sample regression equation: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$, where $b_0, b_1, \ldots, b_k$ are the estimates of $\beta_0, \beta_1, \ldots, \beta_k$, respectively.

For each explanatory variable x_j ($j = 1, \ldots, k$), the corresponding slope coefficient b_j is the estimated regression coefficient. It measures the change in the predicted value of the response variable $\hat{y}$ given a unit increase in the associated explanatory variable x_j, *holding all other explanatory variables constant*. In other words, it represents the partial influence of x_j on $\hat{y}$.

LO 12.3 Interpret goodness-of-fit measures.

- The **standard error of the estimate** s_e is the standard deviation of the residual. It is calculated as as $s_e = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE}$, where SSE is the error sum of squares and MSE is the mean square error. Theoretically, s_e can assume any value between zero and infinity, $0 \le s_e < \infty$; the closer s_e is to zero, the better the model fits.
- The **coefficient of determination R^2** is the proportion of the variation in the response variable that is explained by the sample regression equation. It falls between 0 and 1; the closer the value is to 1, the better the model fits. For example, if $R^2 = 0.72$, we say that 72% of the sample variation in y is explained by the estimated model.

  We compute the coefficient of determination as $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$, where SSR is the regression sum of squares, SSE is the error sum of squares, and SST is the total sum of squares.
- **Adjusted R^2** adjusts R^2 by accounting for the sample size n and the number of explanatory variables k used in the regression. It is calculated as adjusted $R^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right)$. In comparing competing models with different numbers of explanatory variables, the preferred model will have the highest adjusted R^2.

LO 12.4 Conduct a test of individual significance.

A test of individual significance determines whether the explanatory variable x_j has an individual statistically significant influence on y. The value of the test statistic is calculated as $t_{df} = \frac{b_j - \beta_{j0}}{se(b_j)}$, where $df = n - k - 1$, b_j is the estimate for β_j, $se(b_j)$ is the standard

error of the OLS estimator b_j, and β_{j0} is the hypothesized value of β_j, If $\beta_{j0} = 0$, the value of the test statistic reduces to $t_{df} = \frac{b_j}{se(b_j)}$.

Excel reports a value of a test statistic and its associated p-value for a two-tailed test that assesses whether the regression coefficient differs from zero, that is, $\beta_j \neq 0$.

- If we specify a one-tailed test with $\beta_{j0} = 0$, then we need to divide the computer-generated p-value in half.
- If we specify a one- or two-tailed test with $\beta_{j0} \neq 0$, then we cannot use the value of the computer-generated test statistic and its p-value.

LO 12.5 Conduct a test of joint significance.

A **test of joint significance** determines whether the explanatory variables $x_1, x_2, \ldots, x_k$ in a multiple linear regression model have a joint statistically significant influence on y. The test statistic is calculated as $F_{(df_1, df_2)} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}$, where $df_1 = k$, $df_2 = n - k - 1$, MSR is the mean square regression and MSE is the mean square error.

Excel reports the value of the test statistic and its associated p-value.

LO 12.6 Address common violations of the OLS assumptions.

Under the assumptions of the classical linear regression model, OLS provides the best estimates. However, the desirable properties of the OLS estimators become compromised as one or more model assumptions are violated. In addition, for certain violations, it is not possible to make meaningful inferences from the t and F test results. Residual plots are used to identify some of these violations; they also help identify outliers. The model is adequate if the residuals are randomly dispersed around the zero value.

If **nonlinear patterns** exist in the data, yet we estimate a linear relationship between the response and the predictor variables, the resulting OLS estimates can be quite misleading. Often, a plot of the residuals against the predictor variable(s) will reveal whether or not a nonlinear relationship should be incorporated into the model.

Some degree of **multicollinearity** is present in most applications. We can drop one of the collinear variables if its omission can be justified. We can obtain more data, as that may weaken the correlation. Another option is to express the predictor variables differently. At times the best approach may be to do nothing, especially if the estimated model yields a high R^2.

The assumption of **constant variability** often breaks down in cross-sectional studies. The resulting OLS estimators are unbiased, but the standard errors of the OLS estimators are inappropriate, making the standard t or F tests invalid. Analysts often use the OLS estimates along with corrected standard errors, referred to as robust standard errors.

The assumption of **uncorrelated observations** often breaks down in time series studies. The resulting OLS estimators are unbiased but their standard errors are inappropriate. Analysts often use the OLS estimates along with corrected standard errors, referred to as robust standard errors.

It is important that the regression model incorporates all relevant predictor variables. In the case of **excluded variables**, the OLS estimators are generally biased.

85. In an attempt to determine whether or not a linear relationship exists between the price of a home (in $1,000s) and the number of days it takes to sell the home, a real estate agent collected data from recent sales in his city and estimated: Price $= \beta_0 + \beta_1$Days $+ \varepsilon$. A portion of the results is shown in the accompanying table.

	Coefficients	Standard Error	t Stat	p-value
Intercept	−491.27	156.94	−3.13	0.0203
Days	6.17	1.19	5.19	0.0020

a. What is the sample regression equation?
b. Predict the price of a home that has been on the market for 100 days.
c. Specify the competing hypotheses to determine whether Days is significant in explaining a house's price.
d. At the 5% significance level, what is the conclusion to the test? Explain.

86. FILE ***Happiness_Age.*** A sociologist wishes to study the relationship between an individual's age and his/her happiness. He interviews 24 individuals and collects data on his/her age and happiness score, measured on a scale from 0 to 100. A portion of the data is shown in the accompanying table.

Happiness	Age
62	49
66	51
⋮	⋮
72	69

Estimate: Happiness $= \beta_0 + \beta_1$Age $+ \varepsilon$.
a. What is the sample regression equation?
b. Predict the happiness score for a 45-year old person.
c. Interpret the coefficient of determination.
d. At the 1% significance level, does Age influence Happiness? Show the relevant steps of the hypothesis test.

87. FILE ***Yields.*** While the Federal Reserve controls short-term interest rates, long-term interest rates essentially depend on supply/demand dynamics, as well as longer-term interest rate expectations. The accompanying table shows a portion of annualized rates for 3-month Treasury yields and 10-year Treasury yields.

Year	3-Month	10-Year
2001	3.47	5.02
2002	1.63	4.61
⋮	⋮	⋮
2010	0.14	3.21

SOURCE: Federal Reserve Bank of Dallas.

a. Construct and interpret a scatterplot of the 10-year treasury yield against the 3-month yield.
b. Determine the sample regression equation that enables us to predict the 10-year yield on the basis of the 3-month yield.
c. Interpret the coefficient of determination.
d. At the 5% significance level, is the 3-month yield significant in explaining the 10-year yield?
e. Many wonder whether a change in the 3-month yield implies the same change in the 10-year yield. Verify this hypothesis at the 5% significance level.

88. FILE ***Home_Ownership.*** The homeownership rate in the United States was 67.4% in 2009. In order to determine if homeownership is linked with income, 2009 state level data on the homeownership rate (Ownership in %) and median household income (Income in $) were collected. A portion of the data is shown in the accompanying table.

State	Ownership	Income
Alabama	74.1	39980
Alaska	66.8	61604
⋮	⋮	⋮
Wyoming	73.8	52470

SOURCE: http://www.census.gov.

a. Estimate and interpret the model: Ownership $= \beta_0 + \beta_1$Income $+ \varepsilon$.
b. What is the standard error of the estimate?
c. Interpret the coefficient of determination.

89. FILE ***Quotations.*** The labor estimation group at Sturdy Electronics, a contract electronics manufacturer of printed circuit boards, wants to simplify the process it uses to quote production costs to potential customers. They have identified the primary drivers for production time (and thus production cost) as being the number of electronic parts that can be machine-installed and the number of parts that must be manually installed. Accordingly, they wish to develop a multiple regression model to predict production time, measured as minutes per board, using a random

sample of 25 recent product quotations. A portion of the data is shown in the accompanying table.

Production Time	Machine Parts	Manual Parts
9.1	275	14
10.8	446	12
⋮	⋮	⋮
15.5	618	16

a. What is the sample regression equation?
b. Predict production time for a circuit board with 475 machine-installed components and 16 manually installed components.
c. What proportion of the sample variability in production time is explained by the two explanatory variables?
d. At the 5% significance level, are the explanatory variables jointly significant? Are they individually significant?

90. FILE ***DOW_2010.*** A research analyst is trying to determine whether a firm's price-earnings (P/E) and price-sales (P/S) ratios can explain the firm's stock performance over the past year. Generally, a high P/E ratio suggests that investors are expecting higher earnings growth in the future compared to companies with a lower P/E ratio. Investors use the P/S ratio to determine how much they are paying for a dollar of the firm's sales rather than a dollar of its earnings (P/E ratio). In short, the higher the P/E ratio and the lower the P/S ratio, the more attractive the investment. The accompanying table shows the year-to-date (YTD in %) returns and the P/E and P/S ratios for a portion of the 30 firms included in the Dow Jones Industrial Average.

YTD return	P/E	P/S
4.4	14.37	2.41
−4.5	11.01	0.78
⋮	⋮	⋮
16.3	13.94	1.94

SOURCE: The 2010 returns (January 1, 2010–December 31, 2010) were obtained from *The Wall Street Journal*, January 3, 2010; the P/E ratios and the P/S ratios were obtained from http://finance.yahoo.com on January 20, 2011.

a. Estimate: Return $= \beta_0 + \beta_1 \text{P/E} + \beta_2 \text{P/S} + \varepsilon$. Are the signs on the coefficients as expected? Explain.
b. Interpret the slope coefficient of the P/S ratio.
c. What is the predicted return for a firm with a P/E ratio of 10 and a P/S ratio of 2?
d. What is the standard error of the estimate?
e. Interpret R^2.
f. At the 5% significance level, are the explanatory variables jointly significant?
g. At the 5% significance level, are the explanatory variables individually significant?

91. FILE ***SAT.*** A researcher studies the relationship between a test-taker's SAT score, family income (Income in $), and his/her grade point average (GPA). Data are collected from 24 students. A portion of the data is shown in the accompanying table.

SAT	Income	GPA
1651	47000	2.79
1581	34000	2.97
⋮	⋮	⋮
1940	113000	3.96

a. Estimate three models:
   (i) $\text{SAT} = \beta_0 + \beta_1 \text{Income} + \varepsilon$,
   (ii) $\text{SAT} = \beta_0 + \beta_1 \text{GPA} + \varepsilon$, and
   (iii) $\text{SAT} = \beta_0 + \beta_1 \text{Income} + \beta_2 \text{GPA} + \varepsilon$.
b. Use goodness-of-fit measures to select the best-fitting model.
c. Use the best-fitting model to predict SAT given the mean value of the explanatory variable(s).

92. FILE ***Startups.*** Many of today's leading companies, including Google, Microsoft, and Facebook, are based on technologies developed within universities. Lisa Fisher is a business school professor who would like to analyze university factors that enhance innovation. She collects data on 143 universities in 2008 where the response variable is the number of startups (Startups), which is used as a measure for innovation. The explanatory variables include the university's research expenditure (Research in $ millions), the number of patents issued (Patents), and the age of its technology transfer office in years (Duration). A portion of the data is shown in the accompanying table.

Startups	Research	Patents	Duration
1	145.52	8	23
1	237.52	16	23
⋮	⋮	⋮	⋮
1	154.38	3	9

SOURCE: Association of University Managers and National Science Foundation.

a. Estimate: Startups $= \beta_0 + \beta_1 \text{Research} + \beta_2 \text{Patents} + \beta_3 \text{Duration} + \varepsilon$.
b. Predict the number of startups for a university that spent $120 million on research, issued 8 patents, and has had a technology transfer office for 20 years.
c. How much more research expenditure is needed for the university to have an additional predicted startup, with everything else being the same?

93. FILE ***Hourly_Wage.*** A researcher interviews 50 employees of a large manufacturer and collects data on each worker's hourly wage (Wage in $), years of higher education (EDUC), experience (EXPER), and age (AGE).

a. Estimate: Wage $= \beta_0 + \beta_1$EDUC $+\beta_2$EXPER $+ \beta_3$AGE $+ \varepsilon$.
b. Are the signs as expected?
c. Interpret the coefficient of EDUC.
d. Interpret the coefficient of determination.
e. Predict the hourly wage of a 40-year-old employee who has 5 years of higher education and 8 years of experience.
f. At the 5% significance level, are the explanatory variables jointly significant?
g. At the 5% significance level, are the explanatory variables individually significant?

94. FILE ***AnnArbor_Rental.*** The accompanying data file shows the rent, the number of bedrooms, the number of bathrooms, and the square footage for 40 apartments in the college town of Ann Arbor, Michigan.
a. Determine the sample regression equation that enables us to predict the rent on the basis of the number of bedrooms, the number of bathrooms, and the square footage.
b. Interpret the coefficient of determination. What percent of the variation in the rent is unexplained by the sample regression equation?
c. At the 5% significance level, are the explanatory variables jointly significant?
d. At the 5% significance level, are the explanatory variables individually significant?

95. FILE ***Smoking.*** A nutritionist wants to understand the influence of income and healthy food on the incidence of smoking. He collects 2009 data on the percentage of smokers in each state in the U.S. and the corresponding median income (in $) and the percentage of the population that regularly eats fruits and vegetables. A portion of the data is shown in the accompanying table.

State	Smoke	Fruits/Vegetables	Median Income
AK	14.6	23.3	61604
AL	16.4	20.3	39980
⋮	⋮	⋮	⋮
WY	15.2	23.3	52470

Source: Centers for Disease Control and Prevention and U.S. Census Bureau.

a. Estimate: Smoke $= \beta_0 + \beta_1$Fruits/Vegetables $+ \beta_2$Median Income $+ \varepsilon$.
b. At the 5% level of significance, are the explanatory variables individually and jointly significant? Explain.
c. Use the sample correlation coefficients to evaluate the potential problem of multicollinearity.

96. FILE ***Turnover_Expense.*** George believes that the returns of mutual funds are influenced by annual turnover rates and annual expense ratios. In order to substantiate his claim, he randomly selects 20 mutual funds and collects data on each fund's five-year annual return (Return), its annual holding turnover rate (Turnover), and its annual expense ratio (Expense). All variables are measured in percentages. A portion of the data is shown in the accompanying table.

Return	Turnover	Expense
–7.32	16	1.27
10.14	17	0.64
⋮	⋮	⋮
8.94	18	1.18

a. Estimate: Return $= \beta_0 + \beta_1$Turnover $+ \beta_2$Expense $+ \varepsilon$. Conduct appropriate tests to verify George's theory at the 5% significance level.
b. Discuss the potential problems of multicollinearity and changing variability.

97. FILE ***Crime.*** A government researcher examines the factors that influence a city's crime rate. For 41 cities, she collects the crime rate (crimes per 100,000 residents), the poverty rate (in %), the median income (in $1,000s), the percent of residents younger than 18, and the percent of residents older than 65. A portion of the data is shown in the accompanying table.

Crime	Poverty	Income	Under 18	Over 65
710.6	3.8	58.422	18.3	23.4
1317.7	16.7	48.729	19.0	10.3
⋮	⋮	⋮	⋮	⋮
139.7	3.9	59.445	19.7	16

a. Estimate: Crime $= \beta_0 + \beta_1$Poverty $+ \beta_2$Income $+ \beta_3$Under 18 $+ \beta_4$Over 65 $+ \varepsilon$. Discuss the individual and joint significance of the explanatory variables at the 5% significance level.
b. Which explanatory variables are likely to be collinear? Find their sample correlation coefficients to confirm.

98. FILE ***PerCapita.*** Consider a regression model for per capita income, y. The explanatory variables consist of the percentage of the population in the U.S. that is (a) without a high school diploma, x_1, (b) foreign born, x_2, and (c) non-English speaking, x_3. A portion of the data is shown in the accompanying table.

State	Per capita Income	No High School	Foreign Born	No English
Alabama	22984	18.6	3.4	4.9
Alaska	30726	9.3	7.2	16.5
⋮	⋮	⋮	⋮	⋮
Wyoming	27860	8.7	3.1	6.7

Source: 2010 U.S. Census

a. Estimate the model, $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$, and test for the joint and individual significance of the explanatory variables at the 5% level.
b. What proportion of the sample variability in per capita income is explained by the explanatory variables?
c. Do you suspect multicollinearity in the model? Explain.

CASE STUDIES

CASE STUDY 12.1 Akiko Hamaguchi, the manager at a small sushi restaurant in Phoenix, Arizona, is concerned that the weak economic environment has hampered foot traffic in her area, thus causing a dramatic decline in sales. Her cousin in San Francisco, Hiroshi Sato, owns a similar restaurant, but he has seemed to prosper during these rough economic times. Hiroshi agrees that higher unemployment rates have likely forced some customers to dine out less frequently, but he maintains an aggressive marketing campaign to thwart this apparent trend. For instance, he advertises in local papers with valuable two-for-one coupons and promotes early-bird specials over the airwaves. Despite the fact that advertising increases overall costs, he believes that this campaign has positively affected sales at his restaurant. In order to support his claim, Hiroshi provides his restaurant's monthly sales (in $1,000s) and advertising costs (AdsCost in $), as well as the monthly unemployment rate (Unemp in %) from San Francisco County. A portion of the data is shown in the accompanying table.

Data for Case Study 12.1 Hiroshi's Sales, Advertising Costs, and Unemployment Data

FILE
Sushi_Restaurant

Month	Year	Sales	AdsCost	Unemp
January	2006	27.0	550	4.6
February	2008	24.2	425	4.3
⋮	⋮	⋮	⋮	⋮
May	2009	27.4	550	9.1

Source for unemployment rate data: Development Department, State of California, June 2009.

In a report, use the sample information to

1. Estimate a simple regression model, Sales $= \beta_0 + \beta_1$AdsCost $+ \varepsilon$, as well as a multiple regression model, Sales $= \beta_0 + \beta_1$AdsCost $+ \beta_2$Unemp $+ \varepsilon$.
2. Show that the multiple regression model is more appropriate for making predictions.
3. Make predictions for sales with an unemployment rate of 6% and advertising costs of $400 and $600.

CASE STUDY 12.2 Megan Hanson, a realtor in Brownsburg, Indiana, would like to use estimates from a multiple regression model to help prospective sellers determine a reasonable asking price for their homes. She believes that the following four factors influence the asking price (Price) of a house: (1) the square footage of the house (SQFT), (2) the number of bedrooms (Bed), (3) the number of bathrooms (Bath), and (4) the lot size (LTSZ in acres). She randomly collects online listings for 50 single-family homes. A portion of the data is presented in the accompanying table.

Data for Case Study 12.2 Real Estate Data for Brownsburg, Indiana

FILE
Indiana_Real Estate

Price	SQFT	Bed	Bath	LTSZ
399900	5026	4	4.5	0.3
375000	3200	4	3	5
⋮	⋮	⋮	⋮	⋮
102900	1938	3	1	0.1

Source: *Indianapolis Star,* February 27, 2008.

In a report, use the sample information to

1. Provide summary statistics on the asking price, square footage, the number of bedrooms, the number of bathrooms, and the lot size.

2. Estimate and interpret a multiple regression model where the asking price is the response variable and the other four factors are the explanatory variables.
3. Interpret the resulting coefficient of determination.
4. Conduct joint and individual significance tests at the 5% significance level.

CASE STUDY 12.3 Apple Inc. has established a unique reputation in the consumer electronics industry with its development of products such as the iPod, the iPhone, and the iPad. As of May 2010, Apple had surpassed Microsoft as the most valuable company in the world (*The New York Times*, May 26, 2010). Michael Gomez is a stock analyst and wonders if the return on Apple's stock is best modeled using the CAPM model. He collects five years of monthly data, a portion of which is shown in the accompanying table.

FILE
Apple

Data for Case Study 12.3 Apple Return Data, $n = 60$

Date	$R - R_f$	$R_M - R_f$
1/1/2006	4.70	2.21
2/1/2006	−9.65	−0.31
⋮	⋮	⋮
11/1/2010	1.68	2.15

Source: finance.yahoo.com and U.S. Treasury.

In a report, use the sample information to

1. Estimate and interpret CAPM: $R - R_f = \beta_0 + \beta_1(R_M - R_f) + \varepsilon$. Search for Apple's reported Beta on the Web and compare it with your estimate.
2. At the 5% significance level, is the stock return for Apple riskier than that of the market? At the 5% significance level, do abnormal returns exist? Explain.
3. Use a residual plot to analyze the potential problem of correlated observations.

CASE STUDY 12.4 According to a report by the government, new home construction fell to an 18-month low in October 2010 (CNNMoney.com, November 17, 2010). Housing starts, or the number of new homes being built, experienced an 11.7% drop in its seasonally adjusted annual rate. Urmil Singh works for a mortgage company in Madison, Wisconsin. She wants to better understand the quantitative relationship between housing starts (in 1,000s), the mortgage rate (in %), and the unemployment rate (in %). She gathers seasonally adjusted monthly data on these variables from 2006:01–2010:12. A portion of the data is shown in the accompanying table.

FILE
Housing_Starts

Data for Case Study 12.4 Housing Starts and Other Factors, $n = 60$

Date	Housing Starts	Mortgage	Unemployment
2006–01	2273	6.15	4.7
2006–02	2119	6.25	4.8
⋮	⋮	⋮	⋮
2010–12	520	4.71	9.4

Source: Census Bureau and Board of Governors.

In a report, use the sample information to

1. Estimate a multiple regression model for housing starts using the mortgage rate and the unemployment rate as the explanatory variables.
2. At the 5% significance level, evaluate the individual and joint significance of the explanatory variables.
3. Discuss the potential problems of multicollinearity and correlated observations in this time series data application.

APPENDIX 12.1 Guidelines for Other Software Packages

The following section provides brief commands for Minitab, SPSS, JMP, and R. Copy and paste the specified data file into the relevant software spreadsheet prior to following the commands. When importing data into R, use the menu-driven option: File > Import Dataset > From Excel.

Minitab

Simple Linear Regression

(Replicating Example 12.1) From the menu, choose **Stat > Regression > Regression > Fit Regression Model**. Select Debt for **Responses**, and select Income for **Continuous predictors**.

FILE *Debt_Payments*

Multiple Regression

(Replicating Example 12.2) From the menu, choose **Stat > Regression > Regression > Fit Regression Model**. Select Debt for **Responses**, and select Income and Unemployment for **Continuous predictors**.

FILE *Debt_Payments*

Residual Plots—Changing Variability

FILE *Convenience_Stores*

A. (Replicating Figure 12.9) From the menu, choose **Stat > Regression > Regression > Fit Regression Model**.

B. Next to **Response**, select Sales, and next to **Continuous predictors**, select Sqft. Choose **Graphs**. Under **Residuals Plots**, select **Individual plots**, and under **Residuals versus the variables**, select Sqft.

Residual Plots—Correlated Observations

FILE *Sushi_Restaurant*

A. (Replicating Figure 12.10) From the menu, choose **Stat > Regression > Regression > Fit Regression Model**.

B. Next to **Response**, select Sales, and next to **Continuous predictors**, select AdsCost and Unemp. Choose **Graphs**. Under **Residuals Plots**, select **Individual plots**, and then select **Residuals versus order**.

Assessing Multicollinearity with a Correlation Matrix

(Replicating Example 12.9) From the menu, choose **Stat > Basic Statistics > Correlation**. Under **Variables**, select HH Income and Per Cap Inc.

FILE *Home_Values*

SPSS

Simple Linear Regression

(Replicating Example 12.1) From the menu, choose **Analyze > Regression-Linear**. Select Debt as **Dependent**, and Income as **Independent(s)**.

FILE *Debt_Payments*

Multiple Regression

(Replicating Example 12.2) From the menu, choose **Analyze > Regression-Linear**. Select Debt as **Dependent**, and Income and Unemployment as **Independent(s)**.

FILE *Debt_Payments*

Residual Plots—Changing Variability

FILE *Convenience_Stores*

A. (Replicating Figure 12.9) From the menu, choose **Analyze > Regression > Linear**.

B. Under **Dependent**, select Sales, and under **Independent(s)**, select Sqft. Choose **Save**, and under **Residuals**, select **Unstandardized**.

C. From the menu, choose **Graphs > Legacy Dialogs > Scatter/Dot > Simple Scatter**.

D. Under **Y Axis**, select Unstandardized Residual, and under **X Axis**, select Sqft.

Residual Plots—Correlated Observations

FILE *Sushi_Restaurant*

A. (Replicating Figure 12.10) Add a column to data labeled "time" and number from 1 to 17.

B. From the menu, choose **Analyze > Regression > Linear**.

C. Under **Dependent**, select Sales, and under **Independent(s)**, select AdsCost and Unempl. Choose **Save**, and under **Residuals**, select **Unstandardized**.

D. From the menu, choose **Graphs > Legacy Dialogs > Scatter/Dot > Simple Scatter**.

E. Under **Y Axis**, select Unstandardized Residual, and under **X Axis**, select time.

Assessing Multicollinearity with a Correlation Matrix

FILE *Home_Values*

A. (Replicating Example 12.9) From the menu, choose **Analyze > Correlate > Bivariate**.

B. Under **Variables**, select HH Income and Per Cap Inc.

JMP

Simple Linear Regression

FILE *Debt_Payments*

A. (Replicating Example 12.1) From the menu, choose **Analyze > Fit Y by X**. Select Debt as **Y, Response**, and select Income as **X, Factor**.

B. Click on the red triangle next to the header that reads **Bivariate Fit of Debt by Income** and select **Fit line**.

Multiple Regression

FILE *Debt_Payments*

(Replicating Example 12.2) From the menu, choose **Analyze > Fit Model**. Under **Pick Role Variables**, select Debt, and under **Construct Model Effects**, select Income and Unemployment, and choose **Add**.

Residual Plots—Changing Variability

FILE *Convenience_Stores*

A. (Replicating Figure 12.9) From the menu, choose **Analyze > Fit Y by X**.

B. Under **Select Columns**, select Sales, and then under **Cast Selected Columns Into Roles**, select **Y, Response**. Under **Select Columns**, select Sqft, and then under **Cast Selected Columns Into Roles**, select **X, Factor**.

C. Click on the red triangle next to **Bivariate Fit Sales by Sqft**, and select **Fit Line**.

D. Click on the red triangle next to **Linear Fit**, and select **Plot Residuals**.

Residual Plots—Correlated Observations

FILE *Sushi_Restaurant*

A. (Replicating Figure 12.10) From the menu, choose **Analyze > Fit Model**.

B. Under **Select Columns**, select Sales, and then under **Pick Role Variables**, select **Y**. Under **Select Columns**, select AdsCost and Unemp, and then under **Construct Model Effects**, select **Add**.

C. Click on the red triangle next to **Response Sales**, select **Role Diagnostics > Plot Residual by Row**.

Assessing Multicollinearity with a Correlation Matrix

FILE *Home_Values*

(Replicating Example 12.9) From the menu, choose **Analyze > Multivariate Methods > Multivariate**. Under **Select Columns**, select HH Income and Per Cap Inc, and under **Cast Selected Columns into Roles**, select **Y, Columns**.

R

Simple Linear Regression

(Replicating Example 12.1) Use the **lm** function to create a linear model, labeled Simple; in R terminology this is also referred to as an object. In order to view the output, use the **summary** function. Enter:

FILE *Debt_Payments*

```
> Simple <- lm(Debt ~ Income, data = Debt_Payments)
> summary(Simple)
```

Multiple Regression

(Replicating Example 12.2) Use the **lm** function to create a linear model, labeled Multiple, and then use the **summary** function to view the output. Enter:

FILE *Debt_Payments*

```
> Multiple <- lm(Debt ~ Income + Unemployment, data = Debt_Payments)
> summary(Multiple)
```

Assessing Multicollinearity with a Correlation Matrix

(Replicating Example 12.9) Use the **cor** function to find all pairwise correlations for the quantitative variables in the data frame. Enter:

FILE *Home_Values*

```
> cor(Home_Values[,2:5])
```

Residual Plots – Changing Variability

A. (Replicating Figure 12.9) Use the **lm** function to create a linear model, labeled Simple. Enter:

FILE *Convenience_Stores*

```
> Simple <- lm(Sales~Sqft, data = Convenience_Stores)
```

B. Use the **resid** function to obtain the residuals from the model, labeled as Simple_Residuals. Enter:

```
> Simple_Residuals <- resid(Simple)
```

C. Use the **plot** function to create a scatterplot of the residuals against the explanatory variable, Sqft. For options within the function, use *ylab* and *xlab* to label the *y*-axis and the *x*-axis, respectively. Enter:

```
> plot(Simple_Residuals ~  Convenience_Stores$'Sqft', ylab ="e", xlab ="Sqft")
```

Residual Plots – Correlated Observations

A. (Replicating Figure 12.10) Use the **lm** function to create a linear model, labeled Multiple. Use the resid function to obtain the residuals, labeled as Multiple_Residuals. Enter:

FILE *Sushi_Restaurant*

```
> Multiple <- lm(Sales ~ AdsCost + Unemp, data=Sushi_Restaurant)
> Multiple_Residuals <- resid(Multiple)
```

B. Use the **seq** function to create a time variable, labeled as T, that has the same number of observations as Multiple_Residuals. Enter:

```
> T <- seq(from = 1, to = length(Multiple_Residuals))
```

C. Use the **plot** function to plot Multiple_Residuals against T. For options within the function, use *ylab* and *xlab* to label the *y*-axis and the *x*-axis, respectively. Finally, use the **abline** function to insert a line at the *x*-axis. Enter:

```
> plot(Multiple_Residuals ~  T, ylab = "e", xlab = "time")
> abline(h = 0)
```

13 More on Regression Analysis

Learning Objectives

After reading this chapter you should be able to:

LO **13.1** Use a dummy variable to represent a qualitative explanatory variable.

LO **13.2** Use a dummy variable to capture the interaction between a qualitative explanatory variable and a quantitative explanatory variable.

LO **13.3** Estimate and interpret nonlinear regression models.

LO **13.4** Use trend regression models to make forecasts.

LO **13.5** Use trend regression models with seasonal dummy variables to make forecasts.

In this chapter, we introduce important extensions of linear regression models. So far, we analyzed only quantitative explanatory variables by measuring, for example, the contribution of an extra year of education to salary or the contribution of advertisement expenditures to sales of electronic goods. There are other important applications that use qualitative explanatory variables representing two or more categories. For instance, we may want to know if women get paid as much as men for the same education or if sales of electronic goods are higher in the 4^{th} quarter than in the other quarters. In this chapter, we analyze these issues by incorporating qualitative explanatory variables in the regression. We then estimate and interpret regression models for nonlinear relationships when the relationship between the explanatory variable and the response variable cannot be represented by a straight line. We discuss simple transformations, such as squares and natural logarithms, that allow us to capture appropriate nonlinear relationships between the variables. Finally, we turn our attention to forecasting variables such as product sales, the inflation rate, or a company's cash flows. Simple forecasting models that incorporate trend and seasonal fluctuations in the time series data are discussed.



Introductory Case

Is There Evidence of Wage Discrimination?

A few years ago, three female professors at Seton Hall University filed a lawsuit alleging that the University paid better salaries to younger instructors and male professors. Even though this particular case was eventually dismissed, other universities took notice (www.nj.com, November 23, 2010). Hannah Benson, a human resource specialist at a large liberal arts college, was asked by the college's president to test for differences in salaries due to the professor's sex or age. For 42 professors, Hannah gathered information on annual salary (Salary in $1,000s), years of experience (Experience), whether or not the professor was male or female (Male equals 1 if male, 0 otherwise), and whether or not the professor was at least 60 years of age (Age equals 1 if at least 60 years of age, 0 otherwise). A portion of the data is shown in Table 13.1.

TABLE 13.1 Salary and Other Information on 42 Professors

Salary	Experience	Male	Age
67.50	14	1	0
53.51	6	1	0
⋮	⋮	⋮	⋮
73.06	35	0	1

FILE
Professor

Hannah would like to use the sample information in Table 13.1 to

1. Determine whether salary differs by a fixed amount between male and female professors.
2. Determine whether there is evidence of age discrimination in salaries.
3. Determine whether the salary difference between male and female professors increases with experience.

A synopsis of this case is provided in Section 13.2.

13.1 DUMMY VARIABLES

Up until now, the explanatory variables used in the regression applications have been quantitative; in other words, they assume meaningful numerical values. For example, in Chapter 12, we used income and unemployment (both quantitative variables) to explain variations in consumer debt. In empirical work, however, it is common to include some variables that are qualitative. Although qualitative variables can be described by several categories, they are commonly described by only two categories. Examples include a person's sex (male or female), homeownership (own or do not own), shipment (rejected or not rejected), and admission (yes or no). In the first two sections of this chapter we focus on incorporating qualitative explanatory variables into a regression model.

QUANTITATIVE VERSUS QUALITATIVE VARIABLES IN REGRESSION

Variables employed in a regression can be quantitative or qualitative. Quantitative variables assume meaningful numerical values, whereas qualitative variables represent categories.

Given the professor salary data in the introductory case, we can estimate the model as $\hat{y} = 48.83 + 1.15x$ where y represents salary (in \$1,000s) and x is the usual quantitative variable, representing experience (in years). The sample regression equation implies that the predicted salary increases by about \$1,150 ($1.15 \times 1{,}000$) for every year of experience. Arguably, in addition to experience, variations in salary are also caused by qualitative explanatory variables such as a person's sex (male or female) and age (under or over 60 years).

A Qualitative Explanatory Variable with Two Categories

A qualitative explanatory variable with two categories can be associated with a **dummy variable**, also referred to as an **indicator variable**. A dummy variable d is defined as a variable that assumes a value of 1 for one of the categories and 0 for the other. For example, in the case of a dummy variable categorizing a person's sex, we can define 1 for male and 0 for female. Alternatively, we can define 1 for female and 0 for male, with no change in inference. Sometimes we define a dummy variable by converting a quantitative variable to a qualitative variable. In the introductory case, the qualitative variable Age (under or over 60 years) was defined from the quantitative variable Age. Similarly, in studying teen behavior, we may have access to quantitative information on age, but we can generate a dummy variable that equals 1 for ages between 13 and 19 and 0 otherwise.

A DUMMY VARIABLE

A dummy variable d is defined as a variable that takes on values of 1 or 0. It is commonly used to describe a qualitative variable with two categories.

LO 13.1

Use a dummy variable to represent a qualitative explanatory variable.

For the sake of simplicity, we will first consider a model containing one quantitative explanatory variable and one dummy variable. As we will see shortly, the model can easily be extended to include additional variables.

Consider the following model:

$$y = \beta_0 + \beta_1 x + \beta_2 d + \varepsilon,$$

where x is a quantitative variable and d is a dummy variable with values of 1 or 0. We can use sample data to estimate the model as

$$\hat{y} = b_0 + b_1 x + b_2 d.$$

For a given x and $d = 1$, we can compute the predicted value as

$$\hat{y} = b_0 + b_1 x + b_2 = (b_0 + b_2) + b_1 x.$$

Similarly, for $d = 0$,

$$\hat{y} = b_0 + b_1 x.$$

Observe that the two regression lines, $\hat{y} = (b_0 + b_2) + b_1 x$ and $\hat{y} = b_0 + b_1 x$, have the same slope b_1. Thus, the sample regression equation $\hat{y} = b_0 + b_1 x + b_2 d$ accommodates two parallel lines; that is, the dummy variable d affects the intercept but not the slope. The difference between the intercepts is b_2 when d changes from 0 to 1. Figure 13.1 shows the two regression lines when $b_2 > 0$.

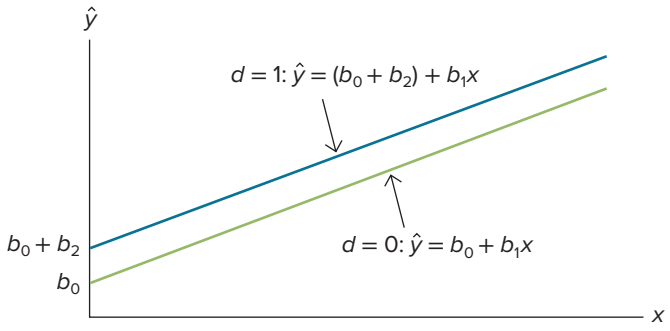


FIGURE 13.1
Using d for an intercept shift

EXAMPLE 13.1

FILE
Professor

The objective outlined in the introductory case is to determine if there are differences in salaries due to a professor's sex or age at a large liberal arts college. Use the data in Table 13.1 to answer the following questions.

a. Estimate $y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \varepsilon$, where y is the annual salary (in \$1,000s) of a professor, x is the number of years of experience, d_1 is the Male dummy variable that equals 1 if the professor is male and 0 otherwise, and d_2 is the Age dummy variable that equals 1 if the professor is at least 60 years of age and 0 otherwise.

b. Compute the predicted salary of a 50-year-old male professor with 10 years of experience. Compute the predicted salary of a 50-year-old female professor with 10 years of experience. Discuss the impact of Male on predicted salary.

c. Compute the predicted salary of a 65-year-old female professor with 10 years of experience. Discuss the impact of Age on predicted salary.

SOLUTION:

a. Table 13.2 shows a portion of the regression results.

TABLE 13.2 Regression Results for Example 13.1

	Coefficients	Standard Error	t Stat	p-Value
Intercept	40.6060	3.6919	10.999	0.000
Experience (x)	1.1279	0.1790	6.300	0.000
Male (d_1)	13.9240	2.8667	4.857	0.000
Age (d_2)	4.3428	4.6436	0.935	0.356

The estimated model is $\hat{y} = 40.6060 + 1.1279x + 13.9240d_1 + 4.3428d_2$.

b. The predicted salary of a 50-year-old male professor ($d_1 = 1$ and $d_2 = 0$) with 10 years of experience ($x = 10$) is

$$\hat{y} = 40.6060 + 1.1279(10) + 13.9240(1) + 4.3428(0) = 65.809, \text{ or } \$65{,}809.$$

The corresponding salary of a 50-year-old female professor ($d_1 = 0$ and $d_2 = 0$) is

$$\hat{y} = 40.6060 + 1.1279(10) + 13.9240(0) + 4.3428(0) = 51.885, \text{ or } \$51{,}885.$$

The predicted difference in salary between a male and a female professor with 10 years of experience is \$13,924 (65,809 − 51,885). This difference can also be inferred from the estimated coefficient 13.924 of the Male dummy variable d_1. Note that the salary difference does not change with experience. For instance, the predicted salary of a 50-year-old male with 20 years of experience is \$77,088. The corresponding salary of a 50-year-old female is \$63,164, for the same difference of \$13,924.

c. For a 65-year-old female professor ($d_1 = 0$ and $d_2 = 1$) with 10 years of experience ($x = 10$), the predicted salary is

$$\hat{y} = 40.6060 + 1.1279(10) + 13.9240(0) + 4.3428(1) = 56.228, \text{ or } \$56{,}228.$$

Prior to any statistical testing, it appears that an older female professor earns, on average, \$4,343 (56,228 − 51,885) more than a younger female professor with the same experience. Again, this difference can be inferred from the estimated coefficient of 4.343 of the Age dummy variable d_2.

Dummy variables are treated just like other explanatory variables when conducting tests of significance. In particular, we can examine whether a particular dummy variable is statistically significant by using the standard t test.

TESTING THE SIGNIFICANCE OF A DUMMY VARIABLE

In a model, $y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \varepsilon$, we can perform the t test to determine the significance of each dummy variable.

EXAMPLE 13.2

Refer to the regression results in Table 13.2.

a. Determine whether a male professor's salary differs from a female professor's salary at the 5% significance level.

b. Determine whether an older professor's salary differs from a younger professor's salary at the 5% significance level.

SOLUTION:

a. In order to test for a salary difference between male and female professors, we set up the competing hypotheses as H_0: $\beta_2 = 0$ against H_A: $\beta_2 \neq 0$. Given a value of the t_{df} test statistic of 4.857 with a p-value ≈ 0, we reject the null hypothesis at the 5% significance level and conclude that male and female professors do not make the same salary, holding other variables constant.

b. Here the competing hypotheses take the form H_0: $\beta_3 = 0$ against H_A: $\beta_3 \neq 0$. Given a value of the t_{df} test statistic of 0.935 with a p-value = 0.356, we cannot reject the null hypothesis. At the 5% significance level, we cannot conclude that an older professor's salary differs from a younger professor's salary.

We now turn our attention to selecting the preferred model for the analysis. Regression results are summarized in Table 13.3.

Model 1 uses only the quantitative variable, Experience. In addition to Experience, Model 2 includes the Male dummy variable, and Model 3 includes Experience and the two dummy variables, Male and Age. This raises an important question: which of the three models should we use for making predictions? As discussed in Chapter 12, we usually rely on adjusted R^2 to compare models with different numbers of explanatory variables. Based on the adjusted R^2 values of the models, reported in the last row of Table 13.3, we select Model 2 as the preferred model because it has the highest adjusted R^2 value of 0.7031. This is consistent with the test results that showed that the Male dummy variable is significant, but the Age dummy variable is not significant, at the 5% level.

TABLE 13.3 Summary of Model Estimates

Variable	Model 1	Model 2	Model 3
Intercept	48.8274* (0.000)	39.4333* (0.000)	40.6060* (0.000)
Experience (x)	1.1455* (0.000)	1.2396* (0.000)	1.1279* (0.000)
Male (d_1)	NA	13.8857* (0.000)	13.9240* (0.000)
Age (d_2)	NA	NA	4.3428 (0.356)
Adjusted R^2	0.5358	0.7031	0.7022

Notes: The table contains parameter estimates with p-values in parentheses; NA denotes not applicable; * represents significance at the 5% level; adjusted R^2, reported in the last row, is used for model selection.

A Qualitative Explanatory Variable with Multiple Categories

So far we have used dummy variables to describe qualitative explanatory variables with only two categories. Sometimes, a qualitative explanatory variable may be defined by more than two categories. In such cases, we use multiple dummy variables to capture all categories. For example, the mode of transportation used to commute to work may be described by three categories: Public Transportation, Driving Alone, and Car Pooling. We can then define two dummy variables d_1 and d_2, where d_1 equals 1 for Public Transportation, 0 otherwise, and d_2 equals 1 for Driving Alone, 0 otherwise. For this three-category case, we need to define only two dummy variables; Car Pooling is indicated when $d_1 = d_2 = 0$.

Consider the following regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \varepsilon,$$

where y represents commuting expenditure, x represents distance to work, and d_1 and d_2 represent the Public Transportation and Driving Alone dummy variables, respectively. We can use sample data to estimate the model as

$$\hat{y} = b_0 + b_1x + b_2d_1 + b_3d_2.$$

For $d_1 = 1$, $d_2 = 0$ (Public Transportation), $\hat{y} = b_0 + b_1x + b_2 = (b_0 + b_2) + b_1x$.
For $d_1 = 0$, $d_2 = 1$ (Driving Alone), $\hat{y} = b_0 + b_1x + b_3 = (b_0 + b_3) + b_1x$.
For $d_1 = d_2 = 0$ (Car Pooling), $\hat{y} = b_0 + b_1x$.

Here we use Car Pooling as the reference category in the estimated regression line with the intercept b_0. The intercept changes to $(b_0 + b_2)$ for Public Transportation and $(b_0 + b_3)$ for Driving Alone. Therefore, we account for all three categories with just two dummy variables.

Given the intercept term, we exclude one of the dummy variables from the regression, where the excluded variable represents the reference category against which the others are assessed. If we include as many dummy variables as there are categories, then their sum will equal one. For instance, if we add a third dummy variable d_3 that equals 1 to denote Car Pooling, then for all observations, $d_1 + d_2 + d_3 = 1$. This creates the problem of perfect multicollinearity, a topic discussed in Chapter 12; recall that such a model cannot be estimated. This situation is sometimes referred to as the **dummy variable trap**.

AVOIDING THE DUMMY VARIABLE TRAP

Assuming that the linear regression model includes an intercept, the number of dummy variables representing a qualitative variable should be *one less than the number of categories* of the variable.

EXAMPLE 13.3

In 2015, 64 groups filed a complaint with the U.S. Department of Justice claiming that Asian Americans are held to a higher standard than other students when applying for admission at elite universities (*The Los Angeles Times*, June 9, 2015). A researcher from the Center for Equal Opportunity wants to determine if SAT scores of admitted students at a large state university differed by ethnic background. She collects data on SAT scores and ethnic background for 200 admitted students. A portion of the data is shown in Table 13.4.

FILE
SAT_Ethnicity

TABLE 13.4 SAT Scores and Ethnic Background; $n = 200$

Individual	SAT	White	Black	Asian	Hispanic
1	1515	1	0	0	0
2	1530	0	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮
200	1614	1	0	0	0

a. Estimate the model $y = \beta_0 + \beta_1d_1 + \beta_2d_2 + \beta_3d_3 + \varepsilon$, where y represents a student's SAT score; d_1 equals 1 if the student is white, 0 otherwise; d_2 equals 1 if the student is black, 0 otherwise; and d_3 equals 1 if the student is Asian, 0 otherwise. Note that the reference category is Hispanic.

b. What is the predicted SAT score for an Asian student? For a Hispanic student?

c. Do SAT scores vary by ethnic background at the 5% significance level? Explain.

SOLUTION:

a. We report a portion of the regression results of this model in Table 13.5.

TABLE 13.5 Regression Results for Example 13.3

	Coefficients	Standard Error	t Stat	p-Value
Intercept	1388.8919	9.3567	148.438	0.000
White (d_1)	201.1447	12.9056	15.586	0.000
Black (d_2)	–31.4544	22.1913	–1.417	0.158
Asian (d_3)	264.8581	17.8584	14.831	0.000

b. For an Asian student, we set $d_1 = 0$, $d_2 = 0$, $d_3 = 1$ and calculate $\hat{y} = 1388.8919 + 264.8581 = 1653.75$. Thus, the predicted SAT score for an Asian student is approximately 1654. The predicted SAT score for a Hispanic student ($d_1 = d_2 = d_3 = 0$) is $\hat{y} = 1388.89$, or approximately 1389.

c. Since the p-values corresponding to d_1 and d_3 are approximately zero, we conclude at the 5% level that the SAT scores of admitted white and Asian students are different from those of Hispanic students. However, with a p-value of 0.158, we cannot conclude that the SAT scores of admitted black and Hispanic students are statistically different.

EXAMPLE 13.4

FILE
SAT_Ethnicity

Reformulate the model from Example 13.3 to determine if the SAT scores of white students are lower than the SAT scores of Asian students. Conduct the test at the 5% significance level and, as in Example 13.3, consider all ethnic categories for the analysis.

SOLUTION: We note that the regression results reported in Table 13.5 cannot be used to determine if the SAT scores of white students are lower than the SAT scores of Asian students. In order to conduct the relevant test, we must use either Asians or whites as the reference category against which the others are assessed. We estimate the model as $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \varepsilon$, where d_1 and d_2 are again dummy variables corresponding to the categories of white and black students, respectively, but now d_3 equals 1 if the student is Hispanic, 0 otherwise. Here the reference category is Asian. We report a portion of the regression results of this model in Table 13.6.

TABLE 13.6 Regression Results for Example 13.4

	Coefficient	Standard Error	t Stat	p-Value
Intercept	1653.7500	15.2111	108.720	0.000
White (d_1)	–63.7134	17.6177	–3.616	0.000
Black (d_2)	–296.3125	25.2247	–11.747	0.000
Hispanic (d_3)	–264.8581	17.8584	–14.831	0.000

For an Asian student, we set $d_1 = d_2 = d_3 = 0$ to find the predicted SAT score as 1653.75, which is the same as derived earlier. In fact, we can show that all predicted SAT scores are identical to those found in Example 13.3. This shows that the choice of the reference category does not matter for making predictions. The results in Table 13.6, however, can be used to determine if the SAT scores of white students are lower than the SAT scores of Asian students. We specify the competing

hypotheses for a left-tailed test as H_0: $\beta_1 \geq 0$ against H_A: $\beta_1 < 0$. The p-value for this one-tailed test is calculated as $0.000/2 \approx 0$. Since the p-value $< \alpha = 0.05$, we reject the null hypothesis. Therefore, we conclude that the SAT scores of admitted white students are less than the SAT scores of admitted Asian students at the 5% significance level.

EXERCISES 13.1

Mechanics

1. Consider a linear regression model where y represents the response variable, x is a quantitative explanatory variable, and d is a dummy variable. The model is estimated as $\hat{y} = 14.8 + 4.4x - 3.8d$.
   a. Interpret the dummy variable coefficient.
   b. Compute $\hat{y}$ for $x = 3$ and $d = 1$.
   c. Compute $\hat{y}$ for $x = 3$ and $d = 0$.
2. Consider a linear regression model where y represents the response variable and d_1 and d_2 are dummy variables. The model is estimated as $\hat{y} = 160 + 15d_1 + 32d_2$.
   a. Compute $\hat{y}$ for $d_1 = 1$ and $d_2 = 1$.
   b. Compute $\hat{y}$ for $d_1 = 0$ and $d_2 = 0$.
3. Using 50 observations, the following regression output is obtained from estimating $y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \varepsilon$.

	Coefficients	Standard Error	t Stat	p-Value
Intercept	–0.61	0.23	–2.75	0.007
x	3.12	1.04	3.01	0.003
d_1	–13.22	15.65	–0.85	0.401
d_2	5.35	1.25	4.27	0.000

   a. Compute $\hat{y}$ for $x = 250$, $d_1 = 1$, and $d_2 = 0$; compute $\hat{y}$ for $x = 250$, $d_1 = 0$, and $d_2 = 1$.
   b. Interpret d_1 and d_2. Are both dummy variables individually significant at the 5% level? Explain.

Applications

4. An executive researcher wants to better understand the factors that explain differences in salaries for marketing majors. He decides to estimate two models: $y = \beta_0 + \beta_1 d_1 + \varepsilon$ (Model 1) and $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \varepsilon$ (Model 2). Here y represents salary, d_1 is a dummy variable that equals 1 for male employees, and d_2 is a dummy variable that equals 1 for employees with an MBA.
   a. What is the reference group in Model 1?
   b. What is the reference group in Model 2?
   c. In the above models, would it matter if d_1 equaled 1 for female employees?
5. House price y is estimated as a function of the square footage of a house x and a dummy variable d that equals 1 if the house has ocean views. The estimated house price, measured in \$1,000s, is given by $\hat{y} = 118.90 + 0.12x + 52.60d$.
   a. Compute the predicted price of a house with ocean views and square footage of 2,000 and 3,000, respectively.
   b. Compute the predicted price of a house without ocean views and square footage of 2,000 and 3,000, respectively.
   c. Discuss the impact of ocean views on the house price.
6. FILE ***Urban.*** A sociologist is studying the relationship between consumption expenditures of families in the United States (Consumption in \$), family income (Income in \$), and whether or not the family lives in an urban or rural community (Urban = 1 if urban, 0 otherwise). She collects data on 50 families, a portion of which is shown in the accompanying table.

Consumption	Income	Urban
62336	87534	0
60076	94796	1
⋮	⋮	⋮
59055	100908	1

   a. Estimate Consumption $= \beta_0 + \beta_1$Income $+ \beta_2$Urban $+ \varepsilon$. Use the estimated model to predict the consumption expenditure of urban families with an income of \$80,000. What is the corresponding consumption expenditure of rural families?
   b. Estimate Consumption $= \beta_0 + \beta_1$Income $+ \beta_2$Rural $+ \varepsilon$ where the dummy variable Rural equals 1 if rural, 0 otherwise. Use the estimated model to predict the consumption expenditure of urban families with an income of \$80,000. What is the corresponding consumption expenditure of rural families?
   c. Interpret the results of the two models.
7. FILE ***IPO.*** One of the theories regarding initial public offering (IPO) pricing is that the initial return y (the percentage change from offer to open price) on an IPO depends on the price revision x (the percentage change from pre-offer to offer price). Another factor that may influence the initial return is a high-tech dummy variable that equals 1 for high-tech firms and 0 otherwise. The following table shows a portion of the data on 264 IPO firms from January 2001 through September 2004.

Initial Return	Price Revision	High-Tech
33.93	7.14	0
18.68	–26.39	0
⋮	⋮	⋮
0.08	–29.41	1

Source: www.ipohome.com; www.nasdaq.com.

a. Estimate $y = \beta_0 + \beta_1 x + \beta_2 d + \varepsilon$ where the dummy variable d equals 1 for firms that are high-tech. Use the estimated model to predict the initial return of a high-tech firm with a 10% price revision. Find the corresponding predicted return of a firm that is not high-tech.

b. Estimate $y = \beta_0 + \beta_1 x + \beta_2 d + \varepsilon$ where the dummy variable d equals 1 for firms that are not high-tech. Use the estimated model to predict the initial return of a high-tech firm with a 10% price revision. Find the corresponding predicted return of a firm that is not high-tech.

c. In the above two models, determine if the dummy variable is significant at the 5% level.

8. FILE ***BMI.*** According to the World Health Organization, obesity has reached epidemic proportions globally. While obesity has generally been linked with chronic disease and disability, researchers argue that it may also be linked with salaries. In other words, the body mass index (BMI) of an employee is a predictor for salary. (A person is considered overweight if his/her BMI is at least 25 and obese if BMI exceeds 30.) The accompanying table shows a portion of salary data (in $1,000s) for 30 college-educated men with their respective BMI and a dummy variable that equals 1 for a white man and 0 otherwise.

Salary	BMI	White
34	33	1
43	26	1
⋮	⋮	⋮
45	21	1

a. Estimate a model for Salary using BMI and White as the explanatory variables. Determine if BMI is significant at the 5% level.

b. What is the estimated salary of a white college-educated man with a BMI of 30? Compute the corresponding salary of a nonwhite man.

9. FILE ***Wage.*** A researcher wonders whether males get paid more, on average, than females at a large firm. She interviews 50 employees and collects data on each employee's hourly wage (Wage in $), years of higher education (EDUC), years of experience (EXPER), age (Age), and a Male dummy variable that equals 1 if male, 0 otherwise. A portion of the data is shown in the accompanying table.

Wage	EDUC	EXPER	Age	Male
37.85	11	2	40	1
21.72	4	1	39	0
⋮	⋮	⋮	⋮	⋮
24.18	8	11	64	0

a. Estimate: Wage $= \beta_0 + \beta_1$EDUC $+ \beta_2$EXPER $+ \beta_3$Age $+ \beta_4$Male $+ \varepsilon$.

b. Predict the hourly wage of a 40-year-old male employee with 10 years of higher education and 5 years experience. Predict the hourly wage of a 40-year-old female employee with the same qualifications.

c. Interpret the estimated coefficient for Male. Is the variable Male significant at the 5% level? Do the data suggest that sex discrimination exists at this firm?

10. FILE ***Nicknames.*** In the United States, baseball has always been a favorite pastime and is rife with statistics and theories. While baseball purists may disagree, to an applied statistician no topic in baseball is too small or hypothesis too unlikely. Researchers at Wayne State University showed that major league players who have nicknames live 2½ years longer than those without them (*The Wall Street Journal*, July 16, 2009). The following table shows a portion of data on the lifespan (Years) of a player and a Nickname dummy variable that equals 1 if the player had a nickname and 0 otherwise.

Years	Nickname
74	1
62	1
⋮	⋮
64	0

a. Create two subsamples, with one consisting of players with a nickname and the other one without a nickname. Calculate the average longevity for each subsample.

b. Estimate a linear regression model of Years on the Nickname dummy variable. Compute the predicted longevity of players with and without a nickname.

c. Conduct a one-tailed test at the 5% level to determine if players with a nickname live longer.

11. FILE ***SAT.*** The SAT has gone through many revisions over the years. People argue that female students generally do worse on math tests but better on writing tests. Consider the following portion of data on 20 students who took the SAT test last year. Information includes each student's score on the writing and math sections of the exam, the student's GPA, and a Female dummy variable that equals 1 if the student is female, 0 otherwise.

Writing	Math	GPA	Female
620	600	3.44	0
570	550	3.04	0
⋮	⋮	⋮	⋮
540	520	2.84	0

a. Estimate a linear regression model with Writing as the response variable and GPA and Female as the explanatory variables.
b. Compute the predicted writing score for a male student with a GPA of 3.5. Repeat the computation for a female student.
c. At the 5% significance level, determine if there is a difference in writing scores between males and females.

12. FILE ***SAT.*** Refer to the previous exercise for a description of the data set. Estimate a linear regression model with Math as the response variable and GPA and Female as the explanatory variables.
a. Compute the predicted math score for a male student with a GPA of 3.5. Repeat the computation for a female student.
b. At the 5% significance level, determine if there is a difference in math scores between males and females.

13. FILE ***Ice_Cream.*** A manager at an ice cream store is trying to determine how many customers to expect on any given day. Overall business has been relatively steady over the past several years, but the customer count seems to have ups and downs. He collects data over 30 days and records the number of customers, the high temperature (in degrees Fahrenheit), and whether the day fell on a weekend (Weekend equals 1 if weekend, 0 otherwise). A portion of the data is shown in the accompanying table.

Customers	Temperature	Weekend
376	75	0
433	78	0
⋮	⋮	⋮
401	68	0

a. Estimate: Customers $= \beta_0 + \beta_1$Temperature $+$ β_2Weekend $+ \varepsilon$.
b. How many customers should the manager expect on a Sunday with a forecasted high temperature of 80°?
c. Interpret the estimated coefficient for Weekend. Is it significant at the 5% level? How might this affect the store's staffing needs?

14. In an attempt to "time the market," a financial analyst studies the quarterly returns of a stock. He uses the model $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \varepsilon$ where y is the quarterly return of a stock, d_1 is a dummy variable that equals 1 if quarter 1 and 0 otherwise, d_2 is a dummy variable that equals 1 if quarter 2 and 0 otherwise, and d_3 is a dummy variable that equals 1 if quarter 3 and 0 otherwise. The following table shows a portion of the regression results.

	Coefficients	*Standard Error*	*t Stat*	*p-Value*
Intercept	10.62	5.81	1.83	0.08
d_1	–7.26	8.21	–0.88	0.38
d_2	–1.87	8.21	–0.23	0.82
d_3	–9.31	8.21	–1.13	0.27

a. Given that there are four quarters in a year, why doesn't the analyst include a fourth dummy variable in his model? What is the reference category?
b. At the 5% significance level, are the dummy variables individually significant? Explain.
c. Explain how you would reformulate the model to determine if the quarterly return is higher in quarter 2 than in quarter 3, still accounting for all quarters.

15. FILE ***Industry.*** The issues regarding executive compensation have received extensive media attention (*The New York Times*, February 9, 2009). Consider a regression model that links CEO compensation (in $ millions) with the total assets of the firm (in $ millions) and the firm's industry. Dummy variables are used to represent four industries: Manufacturing Technology d_1, Manufacturing Other d_2, Financial Services d_3, and Nonfinancial Services d_4. A portion of the data for the 455 highest-paid CEOs in 2006 is shown in the accompanying table.

Compensation	Assets	d_1	d_2	d_3	d_4
16.58	20917.5	1	0	0	0
26.92	32659.5	1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮
2.30	44875.0	0	0	1	0

Source: SEC website and Compustat.

a. Estimate the model: $y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \beta_4 d_3 + \varepsilon$, where y and x denote compensation and assets, respectively. Here the reference category is the nonfinancial services industry.
b. Interpret the estimated coefficients.
c. Use a 5% level of significance to determine which industries, relative to the nonfinancial services industry, have different executive compensation.
d. Reformulate the model to determine, at the 5% significance level, if compensation is higher in Manufacturing Other than in Manufacturing Technology. Your model must account for total assets and all industry types.

16. FILE ***QuickFix.*** The general manager of QuickFix, a chain of quick-service, no-appointment auto repair shops, wants to

develop a model to forecast monthly vehicles served at any particular shop based on four factors: garage bays, population within 5-mile radius (Population in 1,000s), interstate highway access (Access equals 1 if convenient, 0 otherwise), and time of year (Winter equals 1 if winter, 0 otherwise). He believes that, all else equal, shops near an interstate will service more vehicles and that more vehicles will be serviced in the winter due to battery and tire issues. A sample of 19 locations has been obtained. A portion of the data is shown in the accompanying table.

Vehicles Served	Garage Bays	Population	Access	Winter
200	3	15	0	0
351	3	22	0	1
⋮	⋮	⋮	⋮	⋮
464	6	74	1	1

a. Estimate the regression equation relating vehicles serviced to the four explanatory variables.
b. Interpret each of the slope coefficients.
c. At the 5% significance level, are the explanatory variables jointly significant? Are they individually significant? What about at the 10% significance level?
d. What proportion of the variability in vehicles served is explained by the four explanatory variables?
e. Predict vehicles serviced in a non-winter month for a particular location with 5 garage bays, a population of 40,000, and convenient interstate access.

17. FILE ***Retail.*** A government researcher is analyzing the relationship between retail sales (in $ millions) and the gross national product (GNP in $ billions). He also wonders whether there are significant differences in retail sales related to the quarters of the year. He collects 10 years of quarterly data. A portion is shown in the accompanying table.

Year	Quarter	Retail Sales	GNP
2001	1	696048	9740.5
2001	2	753211	9983.5
⋮	⋮	⋮	⋮
2009	4	985649	14442.8

Source: Retail sales obtained from www.census.gov; GNP obtained from http://research.stlouisfed.org.

a. Estimate $y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \beta_4 d_3 + \varepsilon$ where y is retail sales, x is GNP, d_1 is a dummy variable that equals 1 if quarter 1 and 0 otherwise, d_2 is a dummy variable that equals 1 if quarter 2 and 0 otherwise, and d_3 is a dummy variable that equals 1 if quarter 3 and 0 otherwise. Here the reference category is quarter 4.
b. Predict retail sales in quarters 2 and 4 if GNP equals $13,000 billion.
c. Which of the quarterly sales are significantly different from those of the 4th quarter at the 5% level?
d. Reformulate the model to determine, at the 5% significance level, if sales differ between quarter 2 and quarter 3. Your model must include GNP and account for all quarters.

13.2 INTERACTIONS WITH DUMMY VARIABLES

LO 13.2

Use a dummy variable to capture the interaction between a qualitative explanatory variable and a quantitative explanatory variable.

So far we have used a dummy variable d to allow for a shift in the intercept. In other words, d allows the predicted y to differ between the two categories of a qualitative variable by a fixed amount across the values of x. We can also use d to create an **interaction variable**, which allows the predicted y to differ between the two categories of a qualitative variable by a varying amount across the values of x. The interaction variable is a product term xd that captures the interaction between a quantitative variable x and a dummy variable d. Together, the variables d and xd allow the intercept as well as the slope of the estimated linear regression line to vary between the two categories of a qualitative variable.

Consider the following regression model:

$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 xd + \varepsilon.$$

We can use sample data to estimate the model as

$$\hat{y} = b_0 + b_1 x + b_2 d + b_3 xd.$$

For a given x and $d = 1$, we can compute the predicted value as

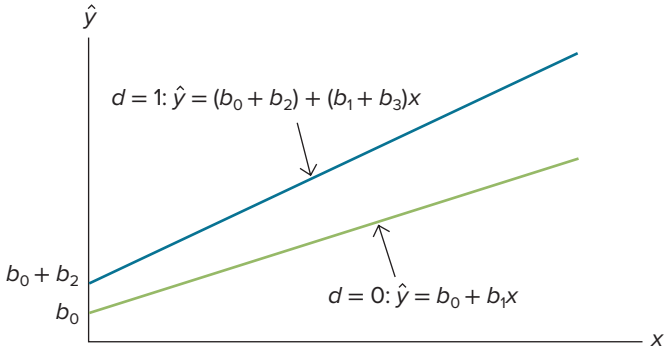
$$\hat{y} = b_0 + b_1 x + b_2 + b_3 x = (b_0 + b_2) + (b_1 + b_3)x.$$

Similarly, for $d = 0$,

$$\hat{y} = b_0 + b_1 x.$$

The use of the dummy variable d along with the interaction variable xd affects the intercept as well as the slope of the estimated regression line. Note that the estimated intercept b_0 and slope b_1 when $d = 0$ shift to $(b_0 + b_2)$ and $(b_1 + b_3)$, respectively, when $d = 1$. Figure 13.2 shows a shift in the intercept and the slope of the estimated regression line when $d = 0$ changes to $d = 1$, given $b_2 > 0$ and $b_3 > 0$.

FIGURE 13.2 Using d and xd for intercept and slope shifts



Prior to estimation, we use sample data to generate two variables, d and xd, which we use along with other explanatory variables in the regression. Tests of significance are performed as before.

FILE
Professor

EXAMPLE 13.5

In Section 13.1, we estimated a regression model to test for differences in salaries depending on a professor's sex and age. We found that the number of years of experience x and the Male dummy variable d_1 were significant in explaining salary differences; however, the Age dummy variable d_2 was insignificant. In an attempt to refine the model explaining salary, we drop d_2 and estimate three models using the data from Table 13.1, where y represents annual salary (in $1,000s).

$$\text{Model 1: } y = \beta_0 + \beta_1 x + \beta_2 d_1 + \varepsilon$$
$$\text{Model 2: } y = \beta_0 + \beta_1 x + \beta_2 x d_1 + \varepsilon$$
$$\text{Model 3: } y = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 x d_1 + \varepsilon$$

a. Estimate and interpret each of the three models.

b. Select the most appropriate model.

c. Use the selected model to predict salaries for males and females over various years of experience.

SOLUTION:

a. In order to estimate the three models, we first generate data on xd_1; Table 13.7 shows a portion of the data.

TABLE 13.7 Generating xd_1 from the Data in Table 13.1

y	x	d_1	xd_1
67.50	14	1	$14 \times 1 = 14$
53.51	6	1	$6 \times 1 = 6$
⋮	⋮	⋮	⋮
73.06	35	0	$35 \times 0 = 0$

Table 13.8 summarizes the regression results for the three models.

TABLE 13.8 Summary of Model Estimates

	Model 1	Model 2	Model 3
Intercept	39.4333* (0.000)	47.0725* (0.000)	49.4188* (0.000)
Experience (x)	1.2396* (0.000)	0.8466* (0.000)	0.7581* (0.000)
Male (d_1)	13.8857* (0.000)	NA	−4.0013 (0.422)
Experience × Male (xd_1)	NA	0.7716* (0.000)	0.9303* (0.000)
Adjusted R^2	0.7031	0.7923	0.7905

Notes: The top portion of the table contains parameter estimates with p-values in parentheses; NA denotes not applicable; * represents significance at the 5% level; Adjusted R^2, reported in the last row, is used for model selection.

Model 1 uses a Male dummy variable d_1 to allow salaries between males and females to differ by a fixed amount, irrespective of experience. It is estimated as $\hat{y} = 39.4333 + 1.2396x + 13.8857d_1$. Since d_1 is associated with a p-value ≈ 0, we conclude at the 5% level that d_1 has a statistically significant influence on salary. The estimated model implies that, on average, males earn about \$13,886 ($13.8857 \times 1{,}000$) more than females at all levels of experience.

Model 2 uses an interaction variable xd_1 to allow the difference in salaries between males and females to vary with experience. It is estimated as $\hat{y} = 47.0725 + 0.8466x + 0.7716xd_1$. Since xd_1 is associated with a p-value ≈ 0, we conclude that it is statistically significant at the 5% level. With every extra year of experience, the estimated difference in salaries between males and females increases by about \$772 ($0.7716 \times 1{,}000$).

Model 3 uses d_1 along with xd_1 to allow a fixed as well as a varying difference in salaries between males and females. The estimated regression equation is $\hat{y} = 49.4188 + 0.7581x - 4.0013d_1 + 0.9303xd_1$. Interestingly, with a p-value of 0.422, the variable d_1 is no longer statistically significant at the 5% level. However, the variable xd_1 is significant, suggesting that with every extra year of experience, the estimated difference in salaries between males and females increases by about \$930 ($0.9303 \times 1{,}000$).

b. While Model 1 shows that the Male dummy variable d_1 is significant and Model 2 shows that the interaction variable xd_1 is significant, Model 3 provides somewhat conflicting results. This raises an important question: which model should we trust? It is not uncommon to contend with such scenarios in many applications, including those pertaining to business,

engineering, and the social sciences. As discussed earlier, we usually rely on adjusted R^2 to compare models that have a different number of explanatory variables. Based on the adjusted R^2 values of the models, reported in the last row of Table 13.8, we select Model 2 as the preferred model because it has the highest value of 0.7923.

c. In order to interpret the results further, we use Model 2 to estimate salaries with varying levels of experience for both males and females. For example, with 10 years of experience, the predicted salary for males ($d_1 = 1$) is

$$\hat{y} = 47.0725 + 0.8466(10) + 0.7716(10 \times 1) = 63.25, \text{ or } \$63,250.$$

The corresponding predicted salary for females ($d_1 = 0$) is

$$\hat{y} = 47.0725 + 0.8466(10) + 0.7716(10 \times 0) = 55.54, \text{ or } \$55,540.$$

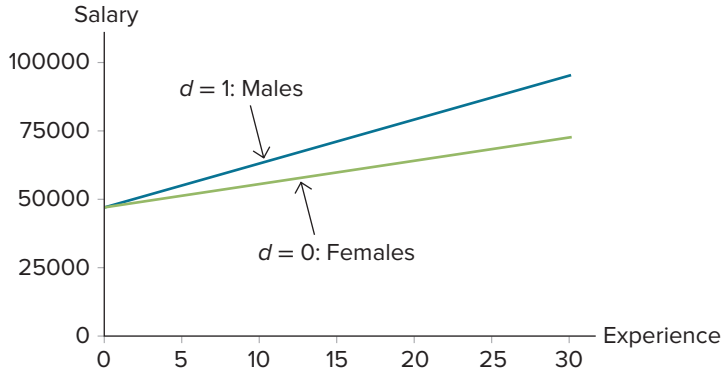
Therefore, with 10 years of experience, the salary difference between males and females is about $7,710. Predicted salaries (in $) for both males and females, and their salary difference, at other levels of experience are presented in Table 13.9.

TABLE 13.9 Predicted Salaries at Various Levels of Experience

Experience	Males	Females	Difference
1	48,690	47,920	770
2	50,310	48,770	1,540
3	51,930	49,610	2,320
4	53,550	50,460	3,090
5	55,160	51,310	3,850
10	63,250	55,540	7,710
15	71,350	59,770	11,580
20	79,440	64,000	15,440
25	87,530	68,240	19,290
30	95,620	72,470	23,150

Note that as experience increases, the salary difference between males and females becomes wider. For instance, the difference is $3,850 with 5 years of experience. However, the difference increases to $19,290 with 25 years of experience. This is consistent with the inclusion of the interaction variable in Model 2. The shift in the slope, implied by the predicted salaries in Table 13.9, is shown in Figure 13.3.

FIGURE 13.3 Predicted salaries of male and female professors



SYNOPSIS OF INTRODUCTORY CASE

A lawsuit brought against Seton Hall University by three female professors alleged that the university engaged in both age and sex discrimination with respect to salaries (www.nj.com, November 23, 2010). Despite the fact that the case was eventually dismissed, another university wonders if the same can be said about its practices. For 42 professors, information is collected on annual salary, experience, whether a professor is male or female, and whether or not the professor is at least 60 years of age. A regression of annual salary against experience, a Male dummy variable, and an Age dummy variable reveal that the professor's sex is significant in explaining variations in salary, but the professor's age is not significant.



©Hero Images/Getty Images

In an attempt to refine the model describing annual salary, various models are estimated that remove the Age dummy variable but use the Male dummy variable to allow both fixed and changing effects on salary. The sample regression line that best fits the data does not include the Male dummy variable for a fixed effect. However, the interaction variable, defined as the product of Male and experience, is significant at any reasonable level, implying that males make about $772 more than females for every year of experience. While the estimated difference in salaries between males and females is only $772 with 1 year of experience, the difference increases to $19,290 with 25 years of experience. In sum, the findings suggest that salaries do indeed differ by one's sex, and this difference increases with every extra year of experience.

EXERCISES 13.2

Mechanics

18. Consider a linear regression model where y represents the response variable and x and d are the explanatory variables; d is a dummy variable assuming values 1 or 0. A model with the dummy variable d and the interaction variable xd is estimated as $\hat{y} = 5.2 + 0.9x + 1.4d + 0.2xd$.
    a. Compute $\hat{y}$ for $x = 10$ and $d = 1$.
    b. Compute $\hat{y}$ for $x = 10$ and $d = 0$.

19. Using 20 observations, the following regression output is obtained from estimating $y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 xd + \varepsilon$.

	Coefficients	Standard Error	t Stat	p-Value
Intercept	13.56	3.31	4.09	0.001
x	4.62	0.56	8.31	0.000
d	−5.15	4.97	−1.04	0.316
xd	2.09	0.79	2.64	0.018

    a. Compute $\hat{y}$ for $x = 10$ and $d = 1$; compute $\hat{y}$ for $x = 10$ and $d = 0$.
    b. Are the dummy variable d and the interaction variable xd individually significant at the 5% level? Explain.

20. FILE **Exercise_13.20** The accompanying data file contains 20 observations on the response variable y along with the explanatory variables x and d.
    a. Estimate and interpret a regression model with the explanatory variables x and d. Compare the results with an extended model that also includes the interaction variable xd.
    b. Use the preferred model to compute the predicted value given $x = 15$, and d equal to 0 and 1.

21. FILE ***Exercise_13.21*** The accompanying data file contains 20 observations on the response variable y along with the explanatory variables x and d.
    a. Estimate and interpret a regression model with the explanatory variables x and d. Compare the results with an extended model that also includes the interaction variable xd.
    b. Use the preferred model to compute the predicted value given $x = 25$, and d equal to 0 and 1.

Applications

22. The annual salary of an employee y (in $1,000s) is estimated as a function of years of experience x; a dummy variable d that

equals 1 for college graduates and 0 for those graduating from high school but not college; and the interaction variable *xd*. The estimated salary is given by $\hat{y} = 30.3 + 1.2x + 15.5d + 2.0xd$.

a. What is the predicted salary of a college graduate who has 5 years of experience? What is the predicted salary of a college graduate who has 15 years of experience?

b. What is the predicted salary of a non-college graduate who has 5 years of experience? What is the predicted salary of a non-college graduate who has 15 years of experience?

c. Discuss the impact of a college degree on salary.

23. House price *y* is estimated as a function of the square footage of a house *x*; a dummy variable *d* that equals 1 if the house has ocean views and 0 otherwise; and the interaction variable *xd*. The estimated house price, measured in $1,000s, is given by $\hat{y} = 80 + 0.12x + 40d + 0.01xd$.

a. Compute the predicted price of a house with ocean views and square footage of 2,000 and 3,000, respectively.

b. Compute the predicted price of a house without ocean views and square footage of 2,000 and 3,000, respectively.

c. Discuss the impact of ocean views on the house price.

24. FILE ***Urban.*** A sociologist is looking at the relationship between consumption expenditures of families in the United States (Consumption in $), family income (Income in $), and whether or not the family lives in an urban or rural community (Urban = 1 if urban, 0 otherwise). She collects data on 50 families across the United States, a portion of which is shown in the accompanying table.

Consumption	Income	Urban
62336	87534	0
60076	94796	1
⋮	⋮	⋮
59055	100908	1

a. Estimate: Consumption $= \beta_0 + \beta_1$Income $+ \varepsilon$. Compute the predicted consumption expenditures of a family with income of $75,000.

b. Include a dummy variable Urban to predict consumption for a family with income of $75,000 in urban and rural communities.

c. Include a dummy variable Urban and an interaction variable (Income × Urban) to predict consumption for a family with income of $75,000 in urban and rural communities.

d. Which of the preceding models is most suitable for the data? Explain.

25. FILE ***BMI.*** According to the World Health Organization, obesity has reached epidemic proportions globally. While obesity has generally been linked with chronic disease and disability, researchers argue that it may also affect wages. In other words, the body mass index (BMI) of an employee is a predictor for salary. (A person is considered overweight if his/her BMI is at least 25 and obese if BMI exceeds 30.) The accompanying table shows a portion of salary data (in $1,000s) for 30 college-educated men with their respective BMI and a dummy variable that represents 1 for a white man and 0 otherwise.

Salary	BMI	White
34	33	1
43	26	1
⋮	⋮	⋮
45	21	1

a. Estimate a model for Salary with BMI and White as the explanatory variables.

b. Reestimate the model with BMI, White, and a product of BMI and White as the explanatory variables.

c. Which of the models is most suitable? Explain. Use this model to estimate the salary for a white college-educated man with a BMI of 30. Compute the corresponding salary for a nonwhite man.

26. FILE ***Pick_Errors.*** The distribution center for an online retailer has been experiencing quite a few "pick errors" (i.e., retrieving the wrong item). Although the warehouse manager thinks most errors are due to inexperienced workers, she believes that a training program also may help to reduce them. Before sending all employees to training, she examines data from a pilot study of 30 employees. Information is collected on the employee's annual pick errors (Errors), experience (Exper in years), and whether or not the employee attended training (Train equals 1 if the employee attended training, 0 otherwise). A portion of the data is shown in the accompanying table.

Errors	Exper	Train
13	9	0
3	27	0
⋮	⋮	⋮
4	24	1

a. Estimate two models:

$\text{Errors} = \beta_0 + \beta_1 \text{ Exper} + \beta_2 \text{ Train} + \varepsilon$, and

$\text{Errors} = \beta_0 + \beta_1 \text{ Exper} + \beta_2 \text{ Train} + \beta_3 \text{ Exper} \times \text{Train} + \varepsilon$.

b. Which model provides a better fit in terms of adjusted R^2 and the significance of the explanatory variables at the 5% level?

c. Use the chosen model to predict the number of pick errors for an employee with 10 years of experience who attended the training program, and for an employee with 20 years of experience who did not attend the training program.

d. Give a practical interpretation for the positive interaction coefficient.

27. FILE ***IPO.*** One of the theories regarding initial public offering (IPO) pricing is that the initial return (the percentage change from offer to open price) on an IPO depends on the price revision (the percentage change from pre-offer to offer price). Another factor that may influence the initial return is a dummy variable that equals 1 for high-tech firms and 0 otherwise. The following table shows a portion of data on 264 IPO firms from January 2001 through September 2004.

Initial Return	Price Revision	High Tech
33.93	7.14	0
18.68	−26.39	0
⋮	⋮	⋮
0.08	−29.41	1

Source: www.ipohome.com, www.nasdaq.com.

a. Estimate a model with the initial return as the response variable and the price revision and the high-tech dummy variable as the explanatory variables.
b. Reestimate the model with price revision along with the dummy variable and the product of the dummy variable and the price revision.
c. Which of these models is the preferred model? Explain. Use this model to estimate the initial return for a high-tech firm with a 15% price revision. Compute the corresponding initial return for a firm that is not high-tech.

28. FILE ***BP_Race.*** Important risk factors for high blood pressure reported by the *National Institute of Health* include weight and ethnicity. High blood pressure is common in adults who are overweight and are African-American. According to the American Heart Association, the systolic pressure (top number) should be below 120. In a recent study, a public policy researcher in Atlanta surveyed 150 adult men in the 55-60 age group. Data was collected on their systolic pressure, weight (in pounds), and race (Black = 1 for African-American; 0 otherwise); a portion of the data is shown in the following table.

Systolic	Weight	Black
196	254	1
151	148	0
⋮	⋮	⋮
170	228	0

a. Estimate and interpret the effect of Weight and Black on Systolic. Predict the systolic pressure of black and non-black adult men with a weight of 180 pounds.
b. Extend the model in part a to include the interaction between Weight and Black. Predict the systolic pressure of black and non-black adult men with a weight of 180 pounds.

13.3 REGRESSION MODELS FOR NONLINEAR RELATIONSHIPS

LO **13.3**
Estimate and interpret nonlinear regression models.

Regression analysis empirically validates not only whether a relationship exists between variables, but also quantifies the strength of the relationship. So far, we have considered only linear regression models. There are numerous applications where the relationship between the explanatory variable and the response variable cannot be represented by a straight line and, therefore, must be captured by an appropriate curve. In fact, the choice of a functional form is a crucial part of specifying a regression model. In this section, we discuss some common nonlinear regression models by making simple transformations of the variables. These transformations include squares, cubes, and natural logarithms, which capture interesting nonlinear relationships while still allowing easy estimation within the framework of a linear regression model.

Quadratic Regression Models

If you ever studied microeconomics, you may have learned that a firm's (or industry's) average cost curve tends to be "U-shaped." Due to economies of scale, the average cost y of a firm initially decreases as output x increases. However, as x increases beyond a certain point, its impact on y turns positive. Other applications show the influence of the explanatory variable initially positive but then turning negative, leading to an "inverted U shape." The **quadratic regression model** is appropriate when the slope, capturing the influence of x on y, changes in magnitude as well as sign.

A quadratic regression model with one explanatory variable is specified as $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$; we can easily extend it to include multiple explanatory variables. The expression $\beta_0 + \beta_1 x + \beta_2 x^2$ is the deterministic component of a quadratic regression model. In other words, conditional on x, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$.

THE QUADRATIC REGRESSION MODEL

In a quadratic regression model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$, the coefficient β_2 determines whether the relationship between x and y is U-shaped ($\beta_2 > 0$) or inverted U-shaped ($\beta_2 < 0$).

Predictions with a quadratic model are made by $\hat{y} = b_0 + b_1 x + b_2 x^2$. It is advisable to use unrounded coefficient estimates for making predictions.

FIGURE 13.4 Representative shapes of a quadratic regression model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

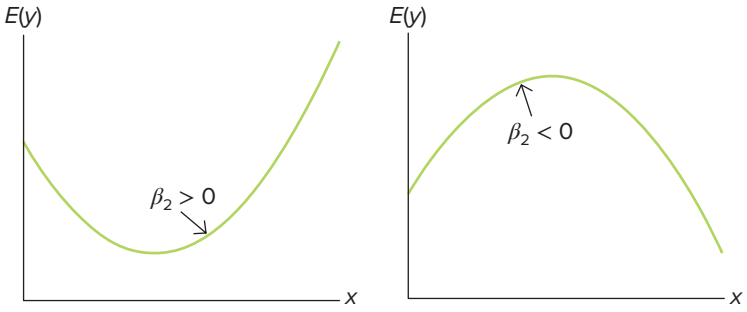


Figure 13.4 highlights some representative shapes of a quadratic regression model.

It is important to be able to determine whether a quadratic regression model provides a better fit than the linear regression model. As we learned in Chapter 12, we cannot compare these models on the basis of their respective R^2 values because the quadratic regression model uses one more parameter than the linear regression model. For comparison purposes, we use adjusted R^2, which imposes a penalty for the additional parameter. In order to estimate the quadratic regression model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$, we have to first create a variable x^2 that contains the squared values of x. The quadratic model is estimated in the usual way as $\hat{y} = b_0 + b_1 x + b_2 x^2$ where b_1 and b_2 are the estimates of β_1 and β_2, respectively.

Interpretation of coefficients in the quadratic regression model: It does not make sense to think of b_1 in the estimated quadratic regression equation as being the effect of changing x by one unit, holding the square of x constant. In nonlinear models, the sample regression equation is best interpreted by calculating, and even graphing, the predicted effect on the response variable over a range of values for the explanatory variable. We will elaborate on this point in Examples 13.6 and 13.7.

Evaluating the marginal effect of x on y in the quadratic regression model: It is important to evaluate the estimated marginal (partial) effect of the explanatory variable x on the predicted value of the response variable; that is, we want to evaluate the change in $\hat{y}$ due to a one-unit increase in x. In the estimated linear regression equation, $\hat{y} = b_0 + b_1 x$, the partial (marginal) effect is constant, estimated by the slope coefficient b_1. In a quadratic regression model, it can be shown that the partial effect of x on $\hat{y}$ can be approximated by $b_1 + 2b_2 x$. This partial effect, unlike in the case of a linear regression model, depends on the value at which x is evaluated. In addition, it can be shown that $\hat{y}$ reaches a maximum ($b_2 < 0$) or minimum ($b_2 > 0$) at $x = \frac{-b_1}{2b_2}$.

EXAMPLE 13.6

FILE
Cost_Functions

Consider the quadratic regression model: $AC = \beta_0 + \beta_1 Output + \beta_2 Output^2 + \varepsilon$, where AC is the average cost of a firm (in $) and Output is the firm's annual output (in millions of units). We estimated the model using data from 20 manufacturing firms and obtained the following regression equation: $\widehat{AC} = 10.5225 - 0.3073\,Output + 0.0210\,Output^2$.

a. What is the change in average cost going from an output level of 4 million units to 5 million units?

b. What is the change in average cost going from an output level of 8 million units to 9 million units? Compare this result to the result found in part a.

c. What is the output level that minimizes average cost?

SOLUTION:

a. The predicted average cost for a firm that produces 4 million units is:

$$\widehat{AC} = 10.5225 - 0.3073(4) + 0.0210(4^2) = \$9.63.$$

The predicted average cost for a firm that produces 5 million units is:

$$\widehat{AC} = 10.5225 - 0.3073(5) + 0.0210(5^2) = \$9.51.$$

An increase in output from 4 to 5 million units results in a $0.12 decrease in predicted average cost.

b. The predicted average cost for a firm that produces 8 million units is:

$$\widehat{AC} = 10.5225 - 0.3073(8) + 0.0210(8^2) = \$9.41.$$

The predicted average cost for a firm that produces 9 million units is:

$$\widehat{AC} = 10.5225 - 0.3073(9) + 0.0210(9^2) = \$9.46.$$

An increase in output from 8 to 9 million units results in a $0.05 increase in predicted average cost. Comparing this result to the one found in part a, we note that a one-unit change in x depends on the value at which x is evaluated.

c. Given $b_1 = -0.3073$ and $b_2 = 0.0210$, the output level that minimizes average cost is $x = \frac{-b_1}{2b_2} = \frac{-(-0.3073)}{2 \times 0.0210} = 7.32$ million units.

Let's now turn to an example with an inverted U-shaped relationship.

EXAMPLE 13.7

In the United States, age discrimination is illegal, but its occurrence is hard to prove (*Newsweek*, March 17, 2010). Even without discrimination, it is widely believed that wages of workers decline as they get older. A young worker can expect wages to rise with age only up to a certain point, beyond which wages begin to fall. Ioannes Papadopoulos works in the human resources department of a large manufacturing firm and would like to verify the quadratic effect of age on wages. He gathers data on 80 workers in his firm with information on their hourly wage (Wage, in $), years of education (Education), and age. A portion of the data is shown in Table 13.10.

FILE
Quadratic

TABLE 13.10 Data for Example 13.7 on Wage, Education, and Age; $n = 80$

Wage	Education	Age
17.54	12	76
20.93	10	61
⋮	⋮	⋮
23.66	12	49

a. Plot Wage against Age and evaluate whether a linear or quadratic regression model better captures the relationship. Verify your choice by using the appropriate goodness-of-fit measure.

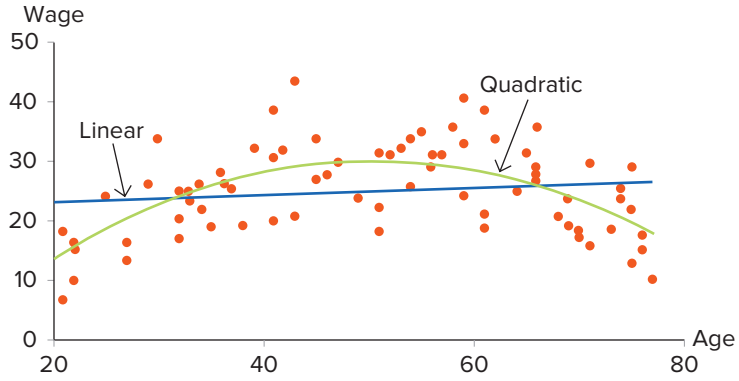
b. Use the appropriate model to predict hourly wages for someone with 16 years of education and age equal to 30, 50, or 70.

c. According to the model, at what age will someone with 16 years of education attain the highest wages?

SOLUTION:

a. Figure 13.5 shows a scatterplot of Wage against Age. We superimpose linear and quadratic trends on the scatterplot. [In order to insert the trendlines in Excel, right-click on the scatterplot and then choose Add Trendline.] It seems that the quadratic regression model provides a better fit for the data as compared to the linear regression model.

FIGURE 13.5 Scatterplot of Wage versus Age



We estimate two models.

Linear Regression Model: $\text{Wage} = \beta_0 + \beta_1\text{Education} + \beta_2\text{Age} + \varepsilon$

Quadratic Regression Model: $\text{Wage} = \beta_0 + \beta_1\text{Education} + \beta_2\text{Age} + \beta_3\text{Age}^2 + \varepsilon$

For ease of exposition, we use the same notation for the coefficients in the linear and the quadratic models even though they have a different meaning depending on the model we reference. Table 13.11 shows the relevant regression results for the linear and the quadratic regression models.

TABLE 13.11 Estimates of the Linear and the Quadratic Regression Models for Example 13.7

Variable	Linear Regression Model	Quadratic Regression Model
Intercept	2.6381 (0.268)	–22.7219* (0.000)
Education	1.4410* (0.000)	1.2540* (0.000)
Age	0.0472 (0.127)	1.3500* (0.000)
Age^2	NA	–0.0133* (0.000)
Adjusted R^2	0.6088	0.8257

Notes: Parameter estimates are in the main body of the table with the *p*-values in parentheses; NA denotes not applicable; * represents significance at the 5% level. The last row presents adjusted R^2 for model comparison.

Note that in the linear regression model, Age has an estimated coefficient of only 0.0472, which is not statistically significant (p-value = 0.127) even at the 10% significance level. However, results change dramatically when Age^2 is included along with Age. In the quadratic regression model, both of these variables, with p-values of approximately zero, are statistically significant at any reasonable level. Also, the adjusted R^2 is higher for the quadratic regression model (0.8257 > 0.6088), making it a better choice for prediction. This conclusion is consistent with our visual impression from the scatterplot in Figure 13.5, which suggested a weak linear but strong quadratic relationship between age and hourly wage.

b. The predicted hourly wage for a 30-year-old person with 16 years of education is

$$\widehat{\text{Wage}} = -22.7219 + 1.2540 \times 16 + 1.3500 \times 30 - 0.0133 \times 30^2 = \$25.87.$$

Similarly, the predicted hourly wage for a 50- and a 70-year-old person is \$31.59 and \$26.67, respectively.

c. In order to determine the optimal age at which the wage is maximized, we also solve $x = \frac{-b_2}{2b_3} = \frac{-(1.3500)}{2(-0.0133)} = 50.75$. The optimal age at which the wage is maximized is about 51 years, with a wage of about \$31.60. It is worth noting that at a different education level, predicted wages will not be the same, yet the highest wage will still be achieved at the same 51 years of age.

The quadratic regression model allows one sign change of the slope capturing the influence of x on y. It is a special case of a **polynomial regression model**. Polynomial regression models describe various numbers of sign changes. Sometimes a quadratic regression model with one sign change is referred to as a polynomial regression model of order 2. In fact, a linear regression model is a polynomial regression model of order 1, which, with a constant slope coefficient, does not allow any sign change.

The linear and the quadratic regression models are the most common polynomial regression models. Sometimes, researchers use a polynomial regression model of order 3, also called the **cubic regression model**. The cubic regression model allows for two changes in slope.

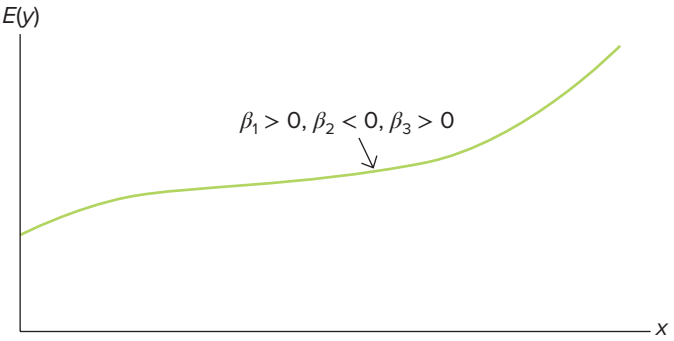
THE CUBIC REGRESSION MODEL

A cubic regression model, $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$, allows two sign changes of the slope capturing the influence of x on y.

Predictions with a cubic model are made by $\hat{y} = b_0 + b_1 x + b_2 x^2 + b_3 x^3$. It is advisable to use unrounded coefficients for making predictions.

The expression $\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ is the deterministic component of a cubic regression model; equivalently, conditional on x, $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$. The shape of a cubic relationship depends on the coefficients. Figure 13.6 highlights a representative shape of a cubic regression model when $\beta_1 > 0$, $\beta_2 < 0$, and $\beta_3 > 0$.

FIGURE 13.6
Representative shape of a cubic regression model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$



Regression Models with Logarithms

Consider an estimated linear regression of annual food expenditure y (in \$) on annual income x (in \$): $\hat{y} = 9{,}000 + 0.20x$. An estimated slope coefficient value of $b_1 = 0.20$ implies that a \$1,000 increase in annual income would lead to a \$200 increase in annual food expenditure, irrespective of whether the income increase is from \$20,000 to \$21,000 or from \$520,000 to \$521,000. This is yet another example where the linearity assumption is not justified. Since we would expect the impact to be smaller at high income levels, it may be more meaningful to analyze what happens to food expenditure as income increases by a certain percentage rather than by a certain dollar amount.

Another commonly used transformation that captures nonlinearities is based on the natural logarithm. The natural logarithm converts changes in a variable into percentage changes, which is useful since many relationships are naturally expressed in terms of percentages. For instance, it is common to log-transform variables such as incomes, house prices, and sales. On the other hand, variables such as age, experience, and scores are generally expressed in their original form. We rely both on economic intuition as well as statistical measures to find the appropriate form for the variables.

We first illustrate log models with only one explanatory variable, which we later extend to a multiple regression model.

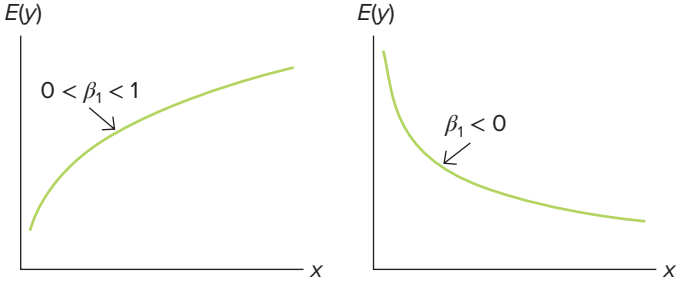
The Log-Log Model

In a **log-log model**, both the response variable and the explanatory variable are transformed into natural logs. We can write this model as

$$\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon,$$

where $\ln(y)$ is the log-transformed response variable and $\ln(x)$ is the log-transformed explanatory variable. With these transformations, the relationship between y and x is captured by a curve whose shape depends on the sign and magnitude of the slope coefficient β_1. Figure 13.7 shows a couple of representative shapes of a log-log regression model.

FIGURE 13.7
Representative shapes of a log-log model: $\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$



For $0 < \beta_1 < 1$, the log-log model implies a positive relationship between x and $E(y)$; however, as x increases, $E(y)$ increases at a slower rate. This may be appropriate in the above food

expenditure example where we expect food expenditure to react positively to changes in income, with the impact diminishing at higher income levels. If $\beta_1 < 0$, it suggests a negative relationship between x and $E(y)$; as x increases, $E(y)$ decreases at a slower rate. Finally, $\beta_1 > 1$ implies a positive and increasing relationship between x and y; this case is not shown in Figure 13.7. For any application, the estimated value of β_1 is determined by the data.

Note that while the log-log regression model is nonlinear in the variables, it is still linear in the coefficients, thus satisfying the requirement of the linear regression model. The only requirement is that we have to first transform both variables into logs before running the regression. We should also point out that in a log-log regression model, the slope coefficient β_1 measures the percentage change in y for a small percentage change in x. In other words, β_1 is a measure of elasticity. In a log-log model, if y represents the quantity demanded of a particular good and x is its unit price, then β_1 measures the price elasticity of demand, a parameter of considerable economic interest. Suppose $\beta_1 = -1.2$; then a 1% increase in the price of this good is expected to lead to about a 1.2% decrease in its quantity demanded.

Finally, even though the response variable is transformed into logs, we still make predictions in regular units. Given $\widehat{\ln(y)} = b_0 + b_1 \ln(x)$, you may be tempted to use the antilog function, to make predictions in regular units as $\hat{y} = \exp(\widehat{\ln(y)}) = \exp(b_0 + b_1 \ln(x))$, where b_0 and b_1 are the coefficient estimates. However, this transformation is known to systematically underestimate the expected value of y. One relatively simple correction is to make predictions as $\hat{y} = \exp(b_0 + b_1 \ln(x) + s_e^2/2)$, where s_e is the standard error of the estimate from the log-log model. This correction is easy to implement since virtually all statistical packages report s_e.

THE LOG-LOG REGRESSION MODEL

A log-log model is specified as $\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$, and β_1 measures the approximate percentage change in $E(y)$ when x increases by 1%.

Predictions with a log-log model are made by $\hat{y} = \exp(b_0 + b_1 \ln(x) + s_e^2/2)$, where b_0 and b_1 are the coefficient estimates and s_e is the standard error of the estimate. It is advisable to use unrounded coefficient estimates for making predictions.

The Logarithmic Model

A log-log specification transforms all variables into logs. It is also common to employ a **semi-log model**, in which not all variables are transformed into logs. We will discuss two types of semi-log models. A semi-log model that transforms only the explanatory variable is often called a **logarithmic model**, and a semi-log model that transforms only the response variable is often called an **exponential model**. We can have many variants of semi-log models when we extend the analysis to include multiple explanatory variables.

The logarithmic model is defined as

$$y = \beta_0 + \beta_1 \ln(x) + \varepsilon.$$

Like the log-log model, this model implies that an increase in x will lead to an increase ($\beta_1 > 0$) or decrease ($\beta_1 < 0$) in $E(y)$ at a decreasing rate. These models are especially attractive when only the explanatory variable is better captured in percentages. Figure 13.8 highlights some representative shapes of this model. Since the log-log and the logarithmic model can allow similar shapes, the choice between the two models can be tricky. We will compare models later in this section.

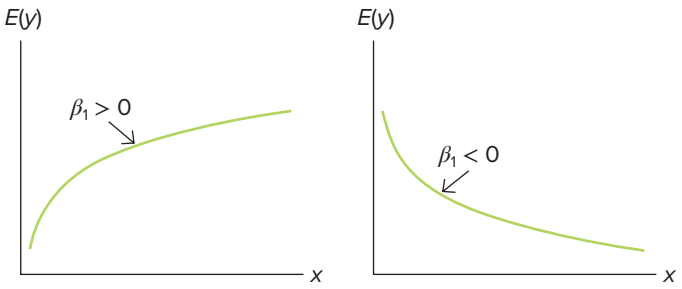


FIGURE 13.8
Representative shapes of a logarithmic model: $y = \beta_0 + \beta_1 \ln(x) + \varepsilon$

In the logarithmic model, the response variable is specified in regular units, but the explanatory variable is transformed into logs. Therefore, $\beta_1 \times 0.01$ measures the approximate unit change in $E(y)$ when x increases by 1%. For example, if $\beta_1 = 5{,}000$, then a 1% increase in x leads to a 50 unit ($= 5{,}000 \times 0.01$) increase in $E(y)$. Since the response variable is already specified in regular units, no further transformation is necessary when making predictions.

THE LOGARITHMIC MODEL

A logarithmic model is specified as $y = \beta_0 + \beta_1 \ln(x) + \varepsilon$, and $\beta_1 \times 0.01$ measures the approximate change in $E(y)$ when x increases by 1%.

Predictions with a logarithmic model are made by $\hat{y} = b_0 + b_1 \ln(x)$, where b_0 and b_1 are the coefficient estimates. It is advisable to use unrounded coefficient estimates for making predictions.

The Exponential Model

Unlike the logarithmic model, in which we were interested in finding the unit change in $E(y)$ for a 1% increase in x, the exponential model allows us to estimate the percent change in $E(y)$ when x increases by one unit. The exponential model is defined as

$$\ln(y) = \beta_0 + \beta_1 x + \varepsilon.$$

Figure 13.9 shows some representative shapes of this model.

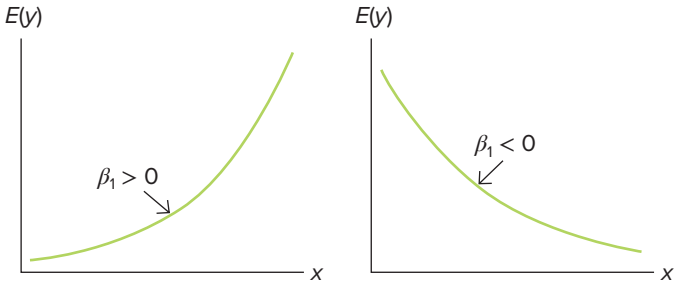


FIGURE 13.9
Representative shapes of an exponential model: $\ln(y) = \beta_0 + \beta_1 x + \varepsilon$

For an exponential model, $\beta_1 \times 100$ measures the approximate percentage change in $E(y)$ when x increases by one unit. For example, a value of $\beta_1 = 0.05$ implies that a one-unit increase in x leads to a 5% ($= 0.05 \times 100$) increase in $E(y)$. In applied work, we often use this model to describe the growth rate of certain economic variables, such as population, employment, wages, productivity, and the gross national product (GNP). As in the case of a log-log model, we make a correction for making predictions, since the response variable is measured in logs.

THE EXPONENTIAL MODEL

An exponential model is specified as $\ln(y) = \beta_0 + \beta_1 x + \varepsilon$, and $\beta_1 \times 100$ measures the approximate percentage change in $E(y)$ when x increases by one unit.

Predictions with an exponential model are made by $\hat{y} = \exp(b_0 + b_1 x + s_e^2/2)$, where b_0 and b_1 are the coefficient estimates and s_e is the standard error of the estimate. It is advisable to use unrounded coefficient estimates for making predictions.

While these log models are easily estimated within the framework of a linear regression model, care must be exercised in making predictions and interpreting the estimated slope coefficient. When interpreting the slope coefficient, keep in mind that logs essentially convert changes in variables into percentage changes. Table 13.12 summarizes the results.

TABLE 13.12 Summary of the Linear, Log-Log, Logarithmic, and Exponential Models

Model	Predicted Value	Estimated Slope Coefficient
$y = \beta_0 + \beta_1 x + \varepsilon$	$\hat{y} = b_0 + b_1 x$	b_1 measures the change in $\hat{y}$ when x increases by one unit.
$\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$	$\hat{y} = \exp(b_0 + b_1 \ln(x) + s_e^2/2)$	b_1 measures the approximate percentage change in $\hat{y}$ when x increases by 1%.
$y = \beta_0 + \beta_1 \ln(x) + \varepsilon$	$\hat{y} = b_0 + b_1 \ln(x)$	$b_1 \times 0.01$ measures the approximate change in $\hat{y}$ when x increases by 1%.
$\ln(y) = \beta_0 + \beta_1 x + \varepsilon$	$\hat{y} = \exp(b_0 + b_1 x + s_e^2/2)$	$b_1 \times 100$ measures the approximate percentage change in $\hat{y}$ when x increases by one unit.

Note: It is advisable to use unrounded coefficient estimates for making predictions with the log-transformed models.

EXAMPLE 13.8

FILE *Gender_Gap*

The salary difference between men and women has shrunk over the years, particularly among younger workers, but it still persists. According to the *Census Bureau*, women earned 83% of what men earned in 2015. Ara Lily, finishing up her MBA degree from Bentley University, would like to analyze the gender gap in salaries of project managers for a class project. From *Glassdoor*, she has read that as of November 2017, the average salary of a project manager in the Boston area is $87,277. Ara gains access to the salary of project managers in small- to middle-sized firms in the Boston area. In addition, she has data on the number of employees (Size) and whether the manager is a male (1 if male; 0 otherwise). Table 13.13 shows a portion of the data.

TABLE 13.13 Salary and Other Information on Project Managers ($n = 120$)

Salary	Size	Male
60000	123	1
213000	380	1
⋮	⋮	⋮
59000	268	1

a. Plot Salary against Size to evaluate whether the relationship is best captured by the linear or the exponential model.

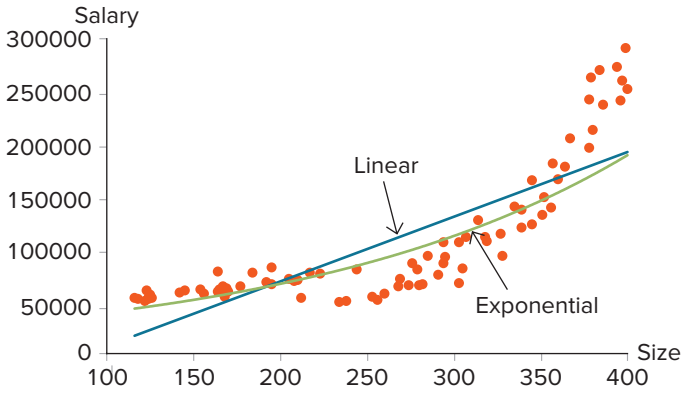
b. Estimate and interpret the linear and the exponential regression models.

c. Use the preferred model to predict the salary of a male and a female manager in a firm with 250 employees.

SOLUTION

a. In Figure 13.10, we plot Salary against Size and superimpose linear and exponential curves.

FIGURE 13.10



It appears that the relationship between Salary and Size is better captured by the exponential model; as the size increases, salary increases at an increasing rate.

b. To estimate the exponential model, we first transform Salary into the natural log of Salary; see the last column of Table 13.14.

TABLE 13.14 Salary and Other Information on Project Managers ($n = 120$)

Salary	Size	Male	ln (Salary)
60000	123	1	ln(60000) = 11.0021
213000	380	1	ln(213000) = 12.2690
⋮	⋮	⋮	
59000	268	1	ln(59000) = 10.9853

In Column 3 of Table 13.15, we present the estimates of the exponential model using ln(Salary) as the response variable; we also present the estimates of the linear model in Column 2 for reference. For both models, we use Size and Male as the explanatory variables.

TABLE 13.15 Regression Estimates for Example 13.8

Variable	Response Variable: Salary	Response Variable: ln(Salary)
Constant	−63,726.7260* (0.000)	9.9651* (0.000)
Size	605.9323* (0.000)	0.0053* (0.000)
Male	16,420.4708* (0.000)	0.1123* (0.018)
s_e	34,617.5188	0.2435

Notes: Parameter estimates are followed below in parentheses with the p-values; * represents significance at the 5% level; the standard error of the estimate s_e is provided to make predictions with the exponential model.

Given that all p-values are less than 0.05, we infer that the size and the male variables are statistically significant at the 5% level. The significance of the male dummy variable suggests gender discrimination. According to the linear model, male managers make about \$16,420 more than their female counterparts, irrespective of the firm size. The estimated exponential model, on the other hand, suggests that male managers make about 11.23% (= 0.1123 × 100) more in salaries than female managers. Regardless of the model used, there is a gender gap in salaries of project managers.

c. As discussed earlier, we incorporate the standard error of the estimate $s_e = 0.2435$ for making predictions with the estimated exponential model. We predict the salary of a male and a female manager in a firm with 250 employees as

$$\text{Male: } \widehat{\text{Salary}} = \exp(9.9651 + 0.0053 \times 250 + 0.1123 \times 1 + 0.2435^2/2) = 91{,}958$$
$$\text{Female: } \widehat{\text{Salary}} = \exp(9.9651 + 0.0053 \times 250 + 0.1123 \times 0 + 0.2435^2/2) = 82{,}189$$

Note that even though Table 13.15 reports estimates rounded to four decimal places, the above predictions are based on unrounded estimates. Also, the interpretation in part b that male managers make 11.23% (= 0.1123 × 100) more in salaries than female managers is an approximate result. Using unrounded values, the exact percentage is given by (exp(0.1123) − 1) × 100 = 11.89, which is consistent with our calculations that suggest that male managers make 11.89 % $\left(= \left(\frac{91{,}958}{82{,}189} - 1\right) \times 100\right)$ more than female managers. Similarly, the estimated exponential model implies that for every one-unit increase in size, the predicted salary increases by about 0.53% (0.0053 × 100), which is approximately the same as (exp(0.0053) − 1) × 100.

To determine which regression model is better suited for an application, we turn to economic intuition and scatterplots for direction. In Example 13.8, we conjectured that the exponential model was appropriate based on the scatterplot. We are often tempted to use formal goodness-of-fit measures like s_e, R^2, and adjusted R^2 for model comparison. It is important to note that we can use computer-generated measures only when the response variable of the competing models is the same. Therefore, we can use these measures to compare linear and logarithmic models (defined in terms of y) and exponential and log-log models (defined in terms of $\ln(y)$). We simply cannot compare, say, linear and exponential models based on computer-generated goodness-of-fit measures since the units of the two estimated models are different. For example, R^2 measures the percentage of sample variations of the response variable explained by the model; comparing the percentage of explained variations of y with that of $\ln(y)$ is like comparing apples with oranges. Similarly, starkly different s_e values of 0.2435 and 34,617.52, reported in Example 13.8, cannot be used for model comparison.

FORMAL COMPARISON OF LINEAR AND LOG-LINEAR REGRESSION MODELS

We can use computer-generated goodness-of-fit measures for model comparison only when the response variable of the competing models is the same.

For a valid comparison of the competing models, we need to first compute goodness-of-fit measures in terms of y for the estimated models that use $\ln(y)$ as the response variable, which we do not pursue in this text.

EXERCISES 13.3

Mechanics

29. Consider the estimated quadratic model $\hat{y} = 20 + 1.9x - 0.05x^2$.
    a. Predict y when x equals 10, 20, and 30.
    b. Find the value of x at which the predicted y is optimized. At this x value, is the predicted y maximized or minimized?

30. Consider the estimated quadratic model $\hat{y} = 20 - 0.72x + 0.02x^2$.
    a. Predict y when x equals 10, 20, and 30.
    b. Find the value of x at which the predicted y is optimized. At this x value, is the predicted y maximized or minimized?

31. Consider the following sample regressions for the linear, the quadratic, and the cubic models along with their respective R^2 and adjusted R^2.

	Linear	Quadratic	Cubic
Intercept	9.66	10.00	10.06
x	2.66	2.75	1.83
x^2	NA	–0.31	–0.33
x^3	NA	NA	0.26
R^2	0.810	0.836	0.896
Adjusted R^2	0.809	0.833	0.895

    a. Predict y for $x = 1$ and 2 with each of the estimated models.
    b. Select the most appropriate model. Explain.

32. Consider the following sample regressions for the linear, the quadratic, and the cubic models along with their respective R^2 and adjusted R^2.

	Linear	Quadratic	Cubic
Intercept	19.80	20.08	20.07
x	1.35	1.50	1.58
x^2	NA	–0.31	–0.27
x^3	NA	NA	–0.03
R^2	0.640	0.697	0.698
Adjusted R^2	0.636	0.691	0.689

    a. Predict y for $x = 2$ and 3 with each of the estimated models.
    b. Select the most appropriate model. Explain.

33. Consider the sample regressions for the linear, the logarithmic, the exponential, and the log-log models. For each of the estimated models, predict y when x equals 100.

	Response Variable: y		Response Variable: ln(y)	
	Model 1	Model 2	Model 3	Model 4
Intercept	240.42	–69.75	1.58	0.77
x	4.68	NA	0.05	NA
ln(x)	NA	162.51	NA	1.25
s_e	83.19	90.71	0.12	0.09

34. Consider the following sample regressions for the linear and the logarithmic models.

	Linear	Logarithmic
Intercept	6.7904	–5.6712
x	1.0607	NA
ln(x)	NA	10.5447*
s_e	2.4935	1.5231
R^2	0.8233	0.9341
Adjusted R^2	0.8013	0.9259

    a. Justify which model fits the data best.
    b. Use the selected model to predict y for $x = 10$.

35. Consider the sample regressions for the linear, the logarithmic, the exponential, and the log-log models. For each of the estimated models, predict y when x equals 50.

	Response Variable: y		Response Variable: ln(y)	
	Model 1	Model 2	Model 3	Model 4
Intercept	18.52	–6.74	1.48	1.02
x	1.68	NA	0.06	NA
ln(x)	NA	29.96	NA	0.96
s_e	23.92	19.71	0.12	0.10

Applications

36. FILE ***Crew_Size.*** The project manager at a construction company is evaluating how crew size affects the productivity of framing jobs. He has experimented with varying crew size (the number of workers) on a weekly basis over the past 27 weeks and has recorded output (jobs/week). A portion of the data is shown in the accompanying table.

Output	Crew Size
10	2
12	3
⋮	⋮
12	10

    a. Create a scatterplot of Output against Crew Size. Based on the scatterplot alone, what crew size(s) seems optimal?

b. Estimate the linear and the quadratic regression models. Evaluate the two models in terms of variable significance and adjusted R^2. Which model provides the best fit? Provide an intuitive justification for the chosen model.

c. Use the best-fitting model to predict the output produced by a crew of 5.

d. Estimate the cubic regression model. Does it improve the fit as compared to the quadratic regression model?

37. FILE ***Television.*** Numerous studies have shown that watching too much television hurts school grades. Others have argued that television is not necessarily a bad thing for children (*Mail Online*, July 18, 2009). Like books and stories, television not only entertains, it also exposes a child to new information about the world. While watching too much television is harmful, a little bit may actually help. Researcher Matt Castle gathers information on the grade point average (GPA) of 27 middle school children and the number of hours of television they watched per week. A portion of the data is shown in the accompanying table.

GPA	Hours
3.24	19
3.10	21
⋮	⋮
3.31	4

a. Estimate a quadratic regression model where the GPA of middle school children is regressed on hours and hours-squared.

b. Is the quadratic term in this model justified? Explain.

c. Find the optimal number of weekly hours of TV for middle school children.

38. FILE ***Inspection.*** A lead inspector at ElectroTech, an electronics assembly shop, wants to convince management that it takes longer, on a per-component basis, to inspect large devices with many components than it does to inspect small devices because it is difficult to keep track of which components have already been inspected. To prove her point, she has collected data on the inspection time (Time in seconds) and the number of components per device (Components) from the last 25 devices. A portion of the data is shown in the accompanying table.

Time	Components
84	32
49	13
⋮	⋮
70	23

a. Estimate the linear, quadratic, and cubic regression models. Evaluate each model in terms of variable significance and adjusted R^2. Which model provides the best fit?

b. Use the best model to predict the time required to inspect a device with 35 components.

39. FILE ***Bids.*** Consider a sample comprised of firms that were targets of tender offers during the period 1978–1985. Conduct an analysis where the response variable represents the number of bids (Bids) received prior to the takeover of the firm. The explanatory variables include the bid premium (Premium) and firm size (Size in $ billions). It is generally believed that a high initial bid premium, defined as the percentage excess of the firm's stock price, would deter subsequent bids. Moreover, while tender offers for large firms are likely to receive more media coverage and thereby attract the attention of opportunistic bidders, it also is a wealth constraint to potential bidders. A portion of the data is shown in the accompanying table.

Bids	Premium	Size
3	1.1905	0.7668
1	1.0360	0.1625
⋮	⋮	⋮
2	1.0329	3.4751

Source: Compustat and *The Wall Street Journal* Index.

a. Estimate the model, Bids $= \beta_0 + \beta_1$Premium $+ \beta_2$Size $+ \beta_2$Size$^2 + \varepsilon$.

b. Justify the inclusion of the quadratic term in the model.

c. What firm size is likely to get the highest number of bids?

40. An economist is interested in examining how an individual's cigarette consumption (C) may be influenced by the price for a pack of cigarettes (P) and the individual's annual income (I). Using data from 50 individuals, she estimates a log-log model and obtains the following regression results.

$$\begin{array}{lrrrr} \widehat{\ln(C)} = & 3.90 & - & 1.25\ln(P) & + 0.18\ln(I) \\ p\text{-values} = & (0.000) & & (0.005) & \quad (0.400) \end{array}$$

a. Interpret the value of the elasticity of demand for cigarettes with respect to price.

b. At the 5% significance level, is the price elasticity of demand statistically significant?

c. Interpret the value of the income elasticity of demand for cigarettes.

d. At the 5% significance level, is the income elasticity of demand statistically significant? Is this result surprising? Explain.

41. FILE ***Wine_Pricing.*** Professor Orley Ashenfelter of Princeton University is a pioneer in the field of wine economics. He claims that, contrary to old orthodoxy, the quality of wine can be explained mostly in terms of weather conditions. Wine romantics accuse him of undermining the whole wine-tasting culture. In an interesting co-authored paper that appeared in *Chance* magazine in 1995, he ran a multiple regression model where quality, measured by the average vintage price relative to 1961, is used as the response variable y. The explanatory

variables were the average temperature x_1 (in degrees Celsius), the amount of winter rain x_2 (in millimeters), the amount of harvest rain x_3 (in millimeters), and the years since vintage x_4. A portion of the data is shown in the accompanying table.

y	x_1	x_2	x_3	x_4
0.3684	17.1167	600	160	31
0.6348	16.7333	690	80	30
⋮	⋮	⋮	⋮	⋮
0.1359	16.0000	578	74	3

Source: www.liquidasset.com.

a. Estimate the linear model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$. What is the predicted price if $x_1 = 16$, $x_2 = 600$, $x_3 = 120$, and $x_4 = 20$?

b. Estimate the exponential model: $\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$. What is the predicted price if $x_1 = 16$, $x_2 = 600$, $x_3 = 120$, and $x_4 = 20$?

42. FILE ***Dexterity.*** A manufacturing manager uses a dexterity test on 20 current employees in order to predict watch production based on time to completion (in seconds). A portion of the data is shown in the accompanying table.

Watches	Time
23	513
19	608
⋮	⋮
20	437

a. Estimate the linear model: Watches $= \beta_0 + \beta_1$Time $+ \varepsilon$. Interpret the slope coefficient. If the time required to complete the dexterity test is 550 seconds, what is the predicted watch production?

b. Estimate the logarithmic model: Watches $= \beta_0 + \beta_1 \ln$(Time) $+ \varepsilon$. Interpret the slope coefficient. If the time required to complete the dexterity test is 550 seconds, what is the predicted watch production?

c. Which model provides a better fit? Explain.

43. FILE ***Davis_Rental.*** Chad Dobson has heard about the positive outlook for real estate investment in college towns. He is interested in investing in Davis, California, which houses one of the University of California campuses. He uses zillow.com to access data on 2011 monthly rents (in $) for 27 houses, along with three characteristics of the home: number of bedrooms (Beds), number of bathrooms (Baths), and square footage (Sqft). A portion of the data is shown in the accompanying table.

Rent	Beds	Baths	Sqft
2950	4	4	1453
2400	4	2	1476
⋮	⋮	⋮	⋮
744	2	1	930

Source: www.zillow.com.

a. Estimate a linear model that uses Rent as the response variable. Estimate an exponential model that uses log of Rent as the response variable.

b. Compute the predicted rent for a 1,500-square-foot house with three bedrooms and two bathrooms for the linear and the exponential models (ignore the significance tests).

44. FILE ***Electricity_Cost.*** The facility manager at a pharmaceutical company wants to build a regression model to forecast monthly electricity cost (Cost in $). Three main variables are thought to influence electricity cost: (1) average outdoor temperature (Temp in °F), (2) working days per month (Days), and (3) tons of product produced (Tons). A portion of the past monthly data on 80 observations is shown in the accompanying table.

Cost	Temp	Days	Tons
16747	46	22	75
7901	31	24	98
⋮	⋮	⋮	⋮
11380	56	28	84

a. Estimate the linear model: Cost $= \beta_0 + \beta_1$Temp $+ \beta_2$Days $+ \beta_3$Tons $+ \varepsilon$. What is the predicted electricity cost in a month during which the average outdoor temperature is 65°, there are 23 working days, and 76 tons are produced?

b. Estimate the exponential model: ln(Cost) $= \beta_0 + \beta_1$Temp $+ \beta_2$Days $+ \beta_3$Tons $+ \varepsilon$. What is the predicted electricity cost in a month during which the average outdoor temperature is 65°, there are 23 working days, and 76 tons are produced?

45. FILE ***Production_Function.*** Economists often examine the relationship between the inputs of a production function and the resulting output. A common way of modeling this relationship is referred to as the Cobb–Douglas production function. This function can be expressed as $\ln(Q) = \beta_0 + \beta_1 \ln(L) + \beta_2 \ln(K) + \varepsilon$, where Q stands for output, L for labor, and K for capital. The accompanying table lists a portion of data relating to the U.S. agricultural industry in the year 2004.

State	Output	Labor	Capital
AL	3.1973	2.7682	3.1315
AR	7.7006	4.9278	4.7961
⋮	⋮	⋮	⋮
WY	1.2993	1.6525	1.5206

Source: www.ers.usda.gov/Data/AgProductivity; see Tables 3, 8, 10. Values in table are indices.

Estimate $\ln(Q) = \beta_0 + \beta_1 \ln(L) + \beta_2 \ln(K) + \varepsilon$.

a. What is the predicted change in output if labor increases by 1%, holding capital constant?

b. Holding capital constant, can we conclude at the 5% level that a 1% increase in labor will increase the output by more than 0.5%?

46. FILE ***Life_Expectancy.*** Life expectancy at birth is the average number of years that a person is expected to live. There is a huge variation in life expectancies between countries, with the highest being in Japan, and the lowest in some African countries. An important factor for such variability is the availability of suitable health care. One measure of a person's access to health care is the people-to-physician ratio. We expect life expectancy to be lower for countries where this ratio is high. The accompanying table lists a portion of life expectancy of males and females in 40 countries and their corresponding people-to-physician ratio.

Country	Female Life Expectancy	People/ Physician
Argentina	74	370
Bangladesh	53	6166
⋮	⋮	⋮
Zaire	56	23193

Source: *The World Almanac and Book Facts*, 1993.

a. Construct a scatterplot of female life expectancy against the people-to-physician ratio. Superimpose a linear trend and a logarithmic trend to determine the appropriate model.
b. Estimate a simple linear regression model with life expectancy of females as the response variable and the people-to-physician ratio as the explanatory variable. What happens to life expectancy of females as the people-to-physician ratio decreases from 1,000 to 500?
c. Estimate a logarithmic regression model with the natural log of the people-to-physician ratio as the explanatory variable. What happens to the life expectancy of females as the people-to-physician ratio decreases from 1,000 to 500?
d. Use R^2 to determine which of the preceding two models is more appropriate.

13.4 TREND FORECASTING MODELS

LO 13.4
Use trend regression models to make forecasts.

In this section, we use regression models for forecasting with **time series** data. Observations of any variable recorded over time in sequential order are considered a time series. The time period can be expressed in terms of a year, a quarter, a month, a week, a day, an hour, or even a second. Examples of time series include the hourly volume of stocks traded on the New York Stock Exchange (NYSE) on four consecutive days, the number of daily visitors that frequent the Statue of Liberty over the month of June, the monthly sales for a retailer over a five-year period, and the growth rate of a country over the past 15 years.

TIME SERIES

A time series is a set of sequential observations of a variable over time.

Let $y_1, y_2, \ldots, y_T$ represent a sample of T observations of a variable y with y_t denoting the value of y at time t. With time series data, it is customary to use the notation T, instead of n, to represent the number of sample observations and to use a subscript t to identify time. For instance, if the number of daily visitors (in 1,000s) to the Statue of Liberty over five days are 100, 94, 98, 110, 102, then $y_1 = 100$, $y_2 = 94, \ldots, y_5 = 102$.

The Linear and the Exponential Trend

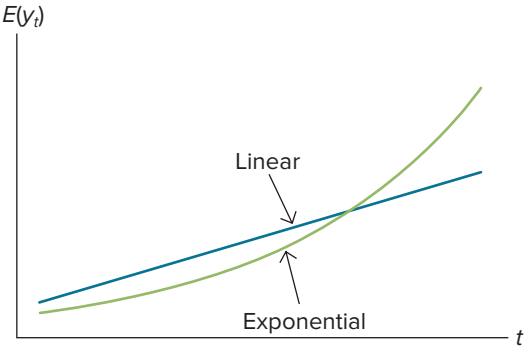
We estimate a **linear trend model** using the regression techniques described earlier. Let y_t be the value of the response variable at time t. Here we use t as the explanatory variable corresponding to consecutive time periods, such as 1, 2, 3, and so on.

THE LINEAR TREND MODEL

A linear trend model is used for a time series that is expected to grow by a fixed amount each time period. It is specified as $y_t = \beta_0 + \beta_1 t + \varepsilon_t$, where y_t is the value of the series at time t. The estimated model is used to make forecasts as $\hat{y}_t = b_0 + b_1 t$, where b_0 and b_1 are the coefficient estimates.

A linear trend model uses a straight line to capture the trend, thus implying that for each period, the value of the series is expected to change by a fixed amount, given by the estimated coefficient b_1. The **exponential trend model** is attractive when the expected increase in the series gets larger over time. Figure 13.11 compares linear and exponential trends. While both graphs have positive slopes, the exponential trend, unlike the linear trend, allows the series to grow by an increasing amount over time.

FIGURE 13.11
Linear and exponential trends



Recall from the previous section that we specify an exponential model as $\ln(y_t) = \beta_0 + \beta_1 t + \varepsilon_t$. In order to estimate this model, we first generate the series in natural logs, $\ln(y_t)$, and then run a regression of $\ln(y_t)$ on t. Since in the exponential model, the response variable is measured in logs, we make forecasts in regular units as $\hat{y}_t = \exp(b_0 + b_1 t + s_e^2/2)$, where s_e is the standard error of the estimate.

THE EXPONENTIAL TREND MODEL

An exponential trend model is commonly used for a time series that is expected to grow by an increasing amount each time period. It is specified as $\ln(y_t) = \beta_0 + \beta_1 t + \varepsilon_t$, where $\ln(y_t)$ is the natural log of y_t. The estimated model is used to make forecasts as $\hat{y}_t = \exp(b_0 + b_1 t + s_e^2/2)$, where b_0 and b_1 are the coefficient estimates and s_e is the standard error of the estimate. It is advisable to use unrounded coefficient estimates for making forecasts.

EXAMPLE 13.9

The United States continues to increase in diversity, with more than a third of its population belonging to a minority group (CNN.com, May 14, 2009). Hispanics are the fastest-growing minority segment, comprising one out of six residents in the country. Table 13.16 shows a portion of data relating to the number (Number in 1,000s) as well as the median income (Income in $) of Hispanic households from 1975 through 2007.

TABLE 13.16 Number and Median Income of Hispanics, 1975–2007

Year	Number	Income
1975	2948	8865
1976	3081	9569
⋮	⋮	⋮
2007	13339	38679

Source: United States Census Bureau.

FILE
Hispanic_Characteristics

a. Use scatterplots to determine whether the linear or the exponential trend model is better suited for making forecasts for Number as well as for Income.

b. Estimate and interpret the appropriate trend models for Number and Income.

c. Forecasts Number and Income in 2008.

SOLUTION

In order to estimate the trend models, it is advisable to first relabel the 33 years of observations from 1 to 33. In other words, make the explanatory variable t assume values 1, 2, . . . , 33 rather than 1975, 1976, . . . , 2007.

a. It is always advisable to inspect the data visually as a first step to gauge whether a linear or an exponential trend provides a better fit. Figures 13.12 and 13.13 are scatterplots, with superimposed trends, of the number and the median income of Hispanic households from period 1 (1975) through 33 (2007).

FIGURE 13.12
Number of Hispanic households with superimposed trends

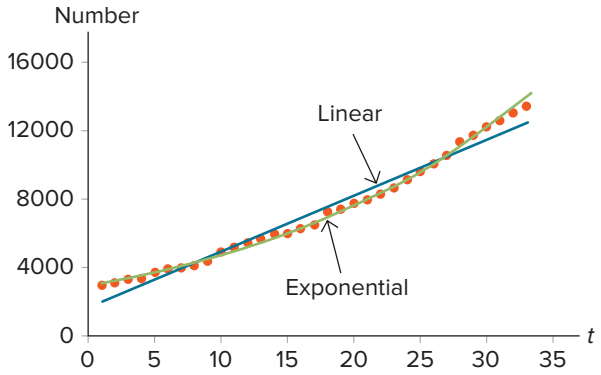
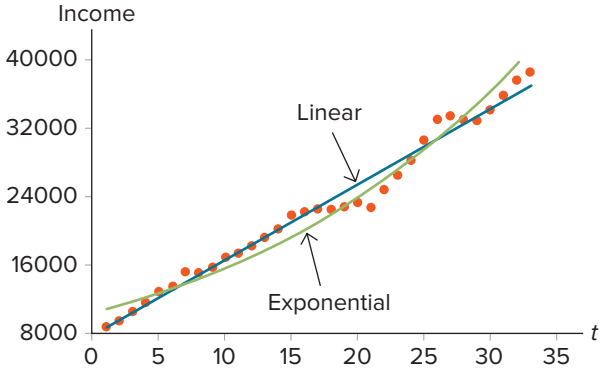


FIGURE 13.13
Median income of Hispanic households with superimposed trends



From Figures 13.12 and 13.13, we infer that the exponential trend model is appropriate for making forecasts for Number whereas the linear trend model is appropriate for making forecasts for Income.

b. In order to estimate the exponential model for Number, we transform the number series to natural logs. Table 13.17 shows a portion of the data with Number transformed into natural logs and the explanatory variable t relabeled from 1 to 33.

TABLE 13.17 Generating the Natural Log of the Series (Example 13.9)

Year	t	Number	Income	ln (Number)
1975	1	2948	8865	7.9889
1976	2	3081	9569	8.0330
⋮	⋮	⋮	⋮	⋮
2007	33	13339	38679	9.4984

Relevant regression results for the exponential trend model for Number and the linear trend model for Income are presented in Table 13.18.

TABLE 13.18 Regression Results for Example 13.9

	Response Variable: Log of Number (Regression 1)	Response Variable: Income (Regression 2)
	Coefficients	Coefficients
Intercept	7.9706* (0.000)	7,796.9186* (0.000)
t	0.0479* (0.000)	887.7249* (0.000)
s_e	0.0311	1,267.1555

Notes: Parameter estimates are followed by the p-values in parentheses; * represents significance at the 5% level. The last row shows the standard error of the estimate s_e.

The estimated trend models suggest that the number and the median income of Hispanic households are trending upward, since both regressions have positive slope coefficients. The slope coefficient from Regression 1 implies that the number of Hispanic households has grown, on average, by about 4.79% per year. Regression 2 shows that the median income for Hispanic households has grown, on average, by approximately $888 each year. In addition, both slope coefficients are significant at any level, since the p-values approximate zero in both regressions.

c. We make forecasts using unrounded coefficient estimates even though we show rounded values in the text. In order to forecast the number of Hispanic households for 2008 ($t = 34$), we compute

$$\widehat{\text{Number}}_{34} = \exp(7.9706 + 0.0479 \times 34 + 0.0311^2/2) = 14{,}750.70 \text{ (in 1,000s)}.$$

Similarly, the forecast for Hispanic median income is computed as

$$\widehat{\text{Income}}_{34} = 7{,}796.9186 + 887.7249 \times 34 = \$37{,}979.57.$$

Polynomial Trends

Sometimes a time series reverses direction, due to any number of circumstances. A common polynomial function that allows for curvature in the series is a **quadratic trend model**. A quadratic regression model was introduced in Section 13.3. For forecasting, this model is estimated as

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t.$$

The coefficient β_2 determines whether the trend is U-shaped or inverted U-shaped. In order to estimate the quadratic trend model, we generate t^2, which is simply the square of t. Then we run a multiple regression model that uses y as the response variable and both t and t^2 as the explanatory variables. The estimated model is used to make forecasts as

$$\hat{y}_t = b_0 + b_1 t + b_2 t^2.$$

Higher-order polynomial functions can be estimated similarly. For instance, consider a **cubic trend model** specified as

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \varepsilon_t.$$

In the cubic trend model, we basically generate two additional variables, t^2 and t^3, for the regression. A multiple regression model is run that uses y as the response variable and t, t^2, and t^3 as the explanatory variables. The estimated model is used to make forecasts as $\hat{y}_t = b_0 + b_1 t + b_2 t^2 + b_3 t^3$.

When comparing polynomial trend models, we simply refer to the reported adjusted R^2 since all polynomial models are based on the same response variable, y.

THE POLYNOMIAL TREND MODEL

A polynomial trend model of order q is estimated as

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \cdots + \beta_q t^q + \varepsilon_t.$$

This model specializes to a linear trend model, quadratic trend model, and cubic trend model for $q = 1$, 2, and 3, respectively. The estimated model is used to make forecasts as $\hat{y}_t = b_0 + b_1 t + b_2 t^2 + b_3 t^3 + \cdots + b_q t^q$, where $b_0, b_1, \ldots, b_q$ are the coefficient estimates. We use adjusted R^2 to compare polynomial trend models with different orders. It is advisable to use unrounded coefficient estimates for making forecasts.

EXAMPLE 13.10

An important indicator of an economy is its inflation rate, which is generally defined as the percentage change in the consumer price index over a specific period of time. Consider monthly inflation rates in the United States from January 2012 to December 2016. A portion of the data is shown in the accompanying table.

TABLE 13.19 Monthly Inflation Rates

Date	Inflation
Jan-12	2.3
Feb-12	2.2
⋮	⋮
Dec-16	2.2

Source: Bureau of Labor Statistics.

FILE
Inflation_Rates

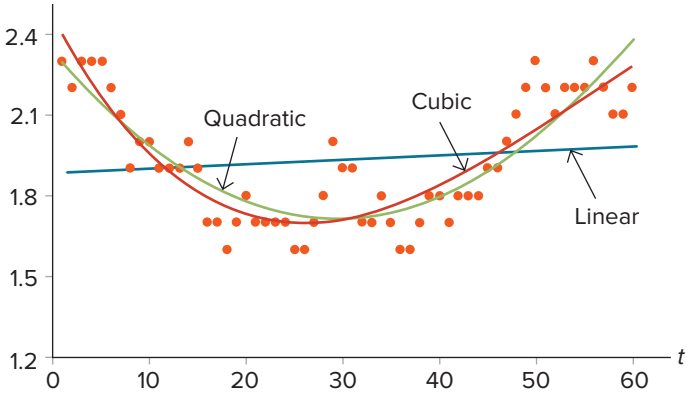
Use the sample information to

a. Evaluate the linear, quadratic, and cubic trend models for the inflation rate.
b. Use the best-fitting trend model to forecast the inflation rate for January 2017.

SOLUTION

a. A simple plot of the inflation rate from January 2012 to December 2016 is shown in Figure 13.14, where t represents the relabeled monthly observations from 1 (January 2012) through 60 (December 2016). In order to gauge whether a linear or nonlinear trend is appropriate, the linear, the quadratic, and the cubic trend models are superimposed on the inflation rate scatterplot.

FIGURE 13.14 Scatterplot of inflation (in percent) and superimposed trends



From Figure 13.14, the linear model is clearly inappropriate.

Two trend models are estimated, where y_t represents the inflation rate.

$$\text{Quadratic Model: } y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$$
$$\text{Cubic Model: } y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \varepsilon_t$$

Table 13.20 presents parameter estimates of the two models. Also included in the table is the adjusted R^2 for model comparison.

TABLE 13.20 Analysis of the Quadratic and the Cubic Trend Models

Variable	Quadratic	Cubic
Constant	2.3334* (0.000)	2.4571* (0.000)
t	−0.0421* (0.000)	−0.0654* (0.000)
t^2	0.0007* (0.000)	0.0017* (0.000)
t^3	NA	−0.00001* (0.004)
Adjusted R^2	0.7304	0.7630

Notes: The top portion of the table contains parameter estimates with p-values in parentheses. NA denotes not applicable. The asterisk designates significance at the 5% level. The last row of the table contains adjusted R^2 for model comparison.

b. The cubic trend provides the best sample fit, as it has a higher adjusted R^2 (0.7630 > 0.7304). Therefore, the estimated cubic trend

model is used with unrounded coefficient estimates to derive the forecast for January 2017 as

$$\hat{y}_{61} = 2.4571 - 0.0654 \times 61 + 0.0017 \times 61^2 - 0.00001 \times 61^3 = 2.30.$$

Inflation forecasts are widely regarded as key inputs for implementing monetary policy. Here, we employ basic trend models to project historical data on inflation.

EXERCISES 13.4

Mechanics

47. Consider the following trend models estimated using daily data for 20 days. Use them to make a forecast for the 21st day.
    a. Linear Trend: $\hat{y} = 13.54 + 1.08t$
    b. Quadratic Trend: $\hat{y} = 18.28 + 0.92t - 0.01t^2$

48. Consider the following trend models using monthly data for the last two years. Use them to make a forecast for the first month of the 3rd year.
    a. Linear Trend: $\hat{y} = 14.98 + 1.13t$
    b. Exponential Trend: $\ln(\hat{y}) = 1.81 + 0.08t$; $s_e = 0.01$

49. FILE ***Exercise_13.49.*** The accompanying data file contains 20 observations for t and y_t.
    a. Plot the series along with the superimposed linear and quadratic trends. Which trend model do you think describes the data well?
    b. Estimate the appropriate model to make a forecast for y_{21}.

50. FILE ***Exercise_13.50.*** The accompanying data file contains 20 observations for t and y_t.
    a. Plot the series along with the superimposed linear and exponential trends. Which trend model do you think describes the data well?
    b. Estimate the appropriate model to make a forecast for y_{21}.

Applications

51. FILE ***Amusement_Park.*** Despite the growth in digital entertainment, the nation's 400 amusement parks have managed to hold on to visitors. A manager collects data on the number of visitors (in millions) to amusement parks in the United States. A portion of the data is shown in the accompanying table.

Year	Visitors
2000	317
2001	319
⋮	⋮
2007	341

Source: International Association of Amusement Parks and Attractions.

    a. Estimate the linear trend model to make forecasts for 2008.
    b. Estimate the exponential trend model to make forecasts for 2008.

52. FILE ***Swine_Flu.*** The potentially deadly 2009 Swine Flu outbreak was due to a new flu strain of subtype H1N1 not previously reported in pigs. When the World Health Organization declared a pandemic, the virus continued to spread in the United States, causing illness along with regular seasonal influenza viruses. Data for week 17 to week 26 in 2009 on total Swine Flu cases in the United States are collected. A portion of the data is shown in the accompanying table.

Week	Flu Cases
17	1190
18	2012
⋮	⋮
26	1093

Source: www.cdc.gov.

    a. Plot the series. Estimate the linear and the quadratic trend models. Use their adjusted R^2 to choose the preferred model.
    b. Given the preferred model, make a forecast for the number of Swine Flu cases in the United States for week 27.

53. FILE ***Recording_Industry.*** Rapid advances in technology have had a profound impact on the United States recording industry (*The New York Times, July 28, 2008*). While cassette tapes gave vinyl records strong competition, they were subsequently eclipsed by the introduction of the compact disc (CD) in the early 1980s. Lately, the CD, too, has been in rapid decline, primarily because of Internet music stores. The following data show a portion of year-end shipment statistics on the three formats of the United States recording industry, in particular, the manufacturers' unit shipments, in millions, of vinyl, cassettes, and CDs from 1991 to 2008.

Year	Vinyl	Cassettes	CDs
1991	4.8	360.1	333.3
1992	2.3	366.4	407.5
⋮	⋮	⋮	⋮
2008	2.9	0.1	384.7

Source: www.riaa.com.

a. Plot the series for cassettes. Estimate the quadratic and the cubic trend models for this series. Make a forecast with the chosen model for 2009.

b. Plot the series for CDs. Estimate the linear and the quadratic trend models for this series. Make a forecast with the chosen model for 2009.

54. FILE ***California_Unemployment.*** Consider the following data, which lists a portion of the seasonally adjusted monthly unemployment rates (in %) in California from 2007–2010.

Date	Unemployment Rate
Jan-07	4.9
Feb-07	5.0
⋮	⋮
Dec-10	12.5

Source: Bureau of Labor Statistics.

a. Plot the series. Which polynomial trend model do you think is most appropriate?

b. Verify your answer by formally comparing the linear, the quadratic, and the cubic trend models.

c. Use the preferred model to forecast the unemployment rate in California for January 2011.

55. FILE ***TrueCar.*** Investors are always reviewing past pricing history and using it to influence their future investment decisions. On May 16, 2014, online car buying system TrueCar launched its initial public offering (IPO), raising $70 million in the stock offering. An investor, looking for a promising return, analyzes the monthly stock price data of TrueCar from June 2014 to May 2017. A portion of the data is shown in the accompanying table.

Date	Price
Jun-14	14.78
Jul-14	13.57
⋮	⋮
May-17	17.51

Source: finance.yahoo.com; data retrieved May 22, 2017.

a. Plot the stock price data with superimposed linear, quadratic, and cubic trends. Which trend do you think describes the data best?

b. Estimate the linear, the quadratic, and the cubic trend models.

c. Use the preferred model to make a forecast for June 2017.

56. FILE ***Miles_Traveled.*** The number of cars sold in the United States in 2016 reached a record high for the seventh year in a row (CNNMoney, January 4, 2017). Consider monthly total miles traveled (in billions) in the United States from January 2010 to December 2016. A portion of the data is shown in the accompanying table.

Date	Miles
Jan-10	2953.305
Feb-10	2946.689
⋮	⋮
Dec-16	3169.501

Source: Federal Reserve Bank of St. Louis.

a. Plot the data with superimposed linear, quadratic, and cubic trends.

b. Estimate the two best models based on the visual analysis in part a. Use adjusted R^2 to select the best model for making forecasts.

c. Use the chosen model to make a forecast for Miles in January 2017.

57. FILE ***Café_Sales.*** With a new chef and a creative menu, Café Venetian has witnessed a huge surge in sales. The following data show a portion of daily sales (in $) at Café Venetian in the first 100 days after the changes.

Day	Sales
1	263
2	215
⋮	⋮
100	2020

a. Plot the café's sales with superimposed linear and exponential trends. Which trend do you think describes the data better?

b. Make a forecast with the chosen model for the 101st week.

58. FILE ***Apple_Price.*** Apple Inc. has performed extremely well in the last decade. After its stock price dropped to below 90 in May 2016, it made a tremendous come back to reach about 146 by May 2017 (SeekingAlpha.com, May 1, 2017). An investor seeking to gain from the positive momentum of Apple's stock price analyzes 53 weeks of stock price data from 5/30/16 to 5/26/17. A portion of the data is shown in the accompanying table.

Date	Price
5/30/2016	97.92
6/6/2016	98.83
⋮	⋮
5/26/2017	153.57

Source: finance.yahoo.com; data retrieved May 22, 2017.

a. Plot the stock price data with superimposed linear and exponential trends. Which trend do you think describes the data better?

b. Make a forecast with the chosen model for the next week (54th week).

13.5 FORECASTING WITH TREND AND SEASONALITY

In the previous section, we employed trend forecasting models to extract long-term upward or downward movements of a time series. These models are appropriate when the time series does not exhibit seasonal variations or has been seasonally adjusted and is therefore free of seasonal variations. The **seasonal component** typically represents repetitions over a one-year period. A time series consisting of weekly, monthly, or quarterly observations often exhibits seasonal variations that repeat year after year. For instance, every year, sales of retail goods increase during the Christmas season, and the number of vacation packages goes up during the summer. In this section we will make forecasts based on the seasonal as well as the trend components of a series.

Seasonal Dummy Variables

LO 13.5

Use trend regression models with seasonal dummy variables to make forecasts.

We estimate a trend forecasting model that includes dummy variables to capture seasonal variations. In Section 13.1, we used dummy variables to describe a qualitative variable with two or more categories. Recall that a dummy variable is a binary variable that equals 1 for one of the categories and 0 for the other. Here we use dummy variables to describe seasons. For quarterly data, we need to define only three dummy variables. Let d_1, d_2, and d_3 be the dummy variables for the first three quarters, using the fourth quarter as reference. Therefore, for an observation that falls in quarter 1, we use $d_1 = 1$, $d_2 = 0$, and $d_3 = 0$. Similarly, for an observation that falls in quarter 2, $d_1 = 0$, $d_2 = 1$, and $d_3 = 0$, for an observation that falls in quarter 3, $d_1 = 0$, $d_2 = 0$, and $d_3 = 1$, and for an observation that falls in quarter 4, $d_1 = 0$, $d_2 = 0$, and $d_3 = 0$. Seasonal dummy variables for other types of seasonal data, such as monthly data, are implemented in a similar manner.

With quarterly data, the linear and the exponential trend models with seasonal dummy variables are summarized in the following captions. In the captions, we have removed the t subscript to simplify the notation.

LINEAR TREND MODEL WITH SEASONAL DUMMY VARIABLES

With quarterly data, a linear trend model with seasonal dummy variables is specified as

$$y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 t + \varepsilon.$$

Forecasts based on the estimated model are as follows:

$$\begin{aligned}
&\text{Quarter 1 } (d_1 = 1, d_2 = 0, d_3 = 0)\text{: } \hat{y}_t = (b_0 + b_1) + b_4 t \\
&\text{Quarter 2 } (d_1 = 0, d_2 = 1, d_3 = 0)\text{: } \hat{y}_t = (b_0 + b_2) + b_4 t \\
&\text{Quarter 3 } (d_1 = 0, d_2 = 0, d_3 = 1)\text{: } \hat{y}_t = (b_0 + b_3) + b_4 t \\
&\text{Quarter 4 } (d_1 = 0, d_2 = 0, d_3 = 0)\text{: } \hat{y}_t = b_0 + b_4 t
\end{aligned}$$

EXPONENTIAL TREND MODEL WITH SEASONAL DUMMY VARIABLES

With quarterly data, an exponential trend model with seasonal dummy variables is specified as

$$\ln(y) = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 t + \varepsilon.$$

Forecasts based on the estimated model are as follows:

$$\text{Quarter 1 } (d_1 = 1, d_2 = 0, d_3 = 0)\text{: } \hat{y}_t = \exp((b_0 + b_1) + b_4 t + s_e^2/2)$$
$$\text{Quarter 2 } (d_1 = 0, d_2 = 1, d_3 = 0)\text{: } \hat{y}_t = \exp((b_0 + b_2) + b_4 t + s_e^2/2)$$
$$\text{Quarter 3 } (d_1 = 0, d_2 = 0, d_3 = 1)\text{: } \hat{y}_t = \exp((b_0 + b_3) + b_4 t + s_e^2/2)$$
$$\text{Quarter 4 } (d_1 = 0, d_2 = 0, d_3 = 0)\text{: } \hat{y}_t = \exp(b_0 + b_4 t + s_e^2/2)$$

where s_e is the standard error of the estimate.

EXAMPLE 13.11

A research analyst at a small investment firm is evaluating Nike Inc.'s performance by analyzing the firm's revenues. She believes that Nike's past performance will aid in predicting its future performance and, therefore, collects quarterly data on Nike's revenue for the fiscal years 1999 through 2008. A portion of the data is shown in Table 13.21.

FILE
Nike_Revenue

TABLE 13.21 Quarterly Revenue for Nike Inc. (in $ millions)

Period	Revenue
1999:01	2505
1999:02	1913
⋮	⋮
2008:04	5088

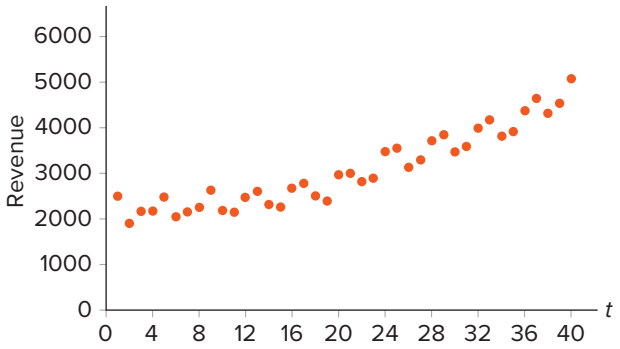
a. Plot the series to identify trend and/or seasonal variations.

b. Estimate the appropriate regression model to forecast revenue for fiscal year 2009.

SOLUTION

a. Figure 13.15 is a scatterplot of the data, where we have relabeled the 10 years of quarterly observations from 1 to 40.

FIGURE 13.15 Scatterplot of Nike's quarterly revenue (in millions $)



The graph highlights some important characteristics of Nike's revenue. First, there is a persistent upward movement in the data. Second, the trend does not seem to be linear and is better captured by an exponential model. Third, a seasonal pattern repeats itself year after year. For instance, revenue is consistently higher in the first and fourth quarters as compared to the second and third quarters. Based on these observations, we will estimate an exponential trend model with seasonal dummy variables for making forecasts.

Given quarterly data, we first construct relevant variables for the regression. Table 13.22 specifies seasonal dummy variables, along with the time variable t.

TABLE 13.22 Constructing Seasonal Dummy Variables (Example 13.11)

Period	y	ln(y)	d_1	d_2	d_3	t
1999:01	2505	7.8260	1	0	0	1
1999:02	1913	7.5564	0	1	0	2
1999:03	2177	7.6857	0	0	1	3
1999:04	2182	7.6880	0	0	0	4
2000:01	2501	7.8245	1	0	0	5
2000:02	2060	7.6305	0	1	0	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2008:03	4544	8.4216	0	0	1	39
2008:04	5088	8.5346	0	0	0	40

As mentioned earlier, we use the exponential model to capture trend: $\ln(y) = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \beta_4 t + \varepsilon$. Relevant estimates of the regression model are presented in Table 13.23.

TABLE 13.23 Regression Results for Example 13.11

	Coefficients	Standard Error	t Stat	p-Value
Intercept	7.5929	0.0261	290.568	0.000
d_1	0.0501	0.0268	1.870	0.070
d_2	–0.1036	0.0267	–3.874	0.000
d_3	–0.0985	0.0267	–3.687	0.000
t	0.0218	0.0008	26.533	0.000
The standard error of the estimate s_e equals 0.0597.				

The coefficients for the seasonal dummy variables indicate that relative to the fourth quarter, the revenue is about 5% higher in the first quarter and about 10% lower in the second and third quarters. The estimated coefficient of 0.0218 for the trend variable t suggests that the predicted quarterly increase in revenue is about 2.18%.

The estimated equation for the exponential trend model with seasonal dummy variables is $\hat{y}_t = \exp(7.5929 + 0.0501d_1 - 0.1036d_2 - 0.0985d_3 + 0.0218t + 0.0597^2/2)$. Using unrounded values for the estimated coefficients, we derive the forecast for the first quarter ($d_1 = 1$, $d_2 = 0$, $d_3 = 0$) of Nike's revenue in 2009 ($t = 41$) as

$$\begin{aligned}\hat{y}_{2009:01} &= \hat{y}_{41} = \exp(7.5929 + 0.0501 + 0.0218 \times 41 + 0.0597^2/2)\\ &= \$5{,}107.16 \text{ million.}\end{aligned}$$

Similarly, the forecasts of Nike's revenue for the remaining quarters of 2009 are computed as

$$\begin{aligned}\hat{y}_{2009:02} &= \hat{y}_{42} = \exp(7.5929 - 0.1036 + 0.0218 \times 42 + 0.0597^2/2)\\ &= \$4{,}475.87 \text{ million,}\\ \hat{y}_{2009:03} &= \hat{y}_{43} = \exp(7.5929 - 0.0985 + 0.0218 \times 43 + 0.0597^2/2)\\ &= \$4{,}598.06 \text{ million, and}\\ \hat{y}_{2009:04} &= \hat{y}_{44} = \exp(7.5929 + 0.0218 \times 44 + 0.0597^2/2) = \$5{,}185.62 \text{ million.}\end{aligned}$$

The total revenue forecasted for fiscal year 2009 is \$19.37 billion. Interestingly, this forecast, based on a simple time series analysis, is quite close to the actual revenue of \$19.20 billion reported by Nike.

Note: We can easily incorporate a polynomial trend, discussed in Section 13.4, with seasonal data. For example, a quadratic trend model with seasonal dummy variables, estimated as $\hat{y}_t = b_0 + b_1 d_1 + b_2 d_2 + b_3 d_3 + b_4 t + b_5 t^2$, can be used to make forecasts with quarterly data.

EXERCISES 13.5

Mechanics

59. Consider a linear trend model with seasonal dummy variables where t represents time and d_i represents the i^{th} quarter. Using five years of quarterly data, the model is estimated as $\hat{y} = 1.45 + 1.01t - 3.87d_1 - 1.55d_2 + 5.66d_3$. Make a forecast for all four quarters of the sixth year.

60. Consider an exponential trend model with seasonal dummy variables where t represents time and d_i represents the i^{th} quarter. Using five years of quarterly data, the model is estimated as $\widehat{\ln(y)} = 0.28 + 0.15t + 0.18d_1 + 0.12d_2 - 0.08d_3$; $s_e = 0.28$. Make a forecast for all four quarters of the sixth year.

61. FILE ***Exercise_13.61.*** The accompanying file contains ten years of quarterly data for the time series y along with the time variable, t, and dummy variables for each quarter, d_i. Estimate a linear trend model with seasonal dummy variables to forecast y for all four quarters of the 11^{th} year.

62. FILE ***Exercise_13.62.*** The accompanying file contains ten years of quarterly data for the time series y along with the time variable, t, and dummy variables for each quarter, d_i. Estimate an exponential trend model with seasonal dummy variables to forecast y for all four quarters of the 11^{th} year.

63. FILE ***Exercise_13.63.*** The accompanying file contains five years of monthly data for the time series y. Create the time variable and the monthly dummy variables to estimate a linear trend model with seasonal dummy variables to forecast y for January, February and December of the 6^{th} year.

64. FILE **Exercise_13.64.** The accompanying file contains five years of monthly data for the time series y. Create the time variable and the monthly dummy variables to estimate an exponential trend model with seasonal dummy variables to forecast y for January, February and December of the 6^{th} year.

Applications

65. FILE ***Treasury_Bonds.*** Consider a portion of monthly return data (in %) on 20-year Treasury Bonds from 2006–2010.

Date	Return
Jan-06	4.65
Feb-06	4.73
⋮	⋮
Dec-10	4.16

Source: Federal Reserve Bank of Dallas.

Estimate a linear trend model with seasonal dummy variables to make forecasts for the first three months of 2011.

66. FILE ***Expenses.*** The controller of a small construction company is attempting to forecast expenses for the next year. He collects quarterly data on expenses (in $1,000s) over the past five years, a portion of which is shown in the accompanying table.

Year	Quarter	Expenses
2008	1	96.50
2008	2	54.00
⋮		⋮
2017	4	22335.30

a. Estimate a linear trend model with seasonal dummy variables and forecast expenses for all four quarters of 2018.
b. Estimate an exponential trend model with seasonal dummy variables and and forecast expenses for all four quarters of 2018.
c. Plot expenses over time to determine if the forecasts made with the linear or the exponential model are more suitable.

67. FILE ***Blockbuster.*** Blockbuster Inc. faced challenges by the growing online market (CNNMoney.com, March 3, 2009). Its revenue from rental stores sagged as customers increasingly got their movies through the mail or high-speed Internet connections. The following table contains a portion of the quarterly revenue from rentals of all formats of movies at Blockbuster Inc. (in $ millions) from 2001 through 2008.

Year	Quarter	Revenue
2001	1	1.403683
2001	2	1.287625
⋮	⋮	⋮
2008	4	1.097712

a. Estimate a linear trend model with seasonal dummy variables.
b. Estimate a quadratic trend model with seasonal dummy variables.
c. Use the appropriate model to make quarterly forecasts for 2009.

68. FILE ***Weekly_Earnings.*** Data on weekly earnings are collected as part of the Current Population Survey, a nationwide sample survey of households in which respondents are asked how much each wage and salary worker usually earns. The following table contains a portion of quarterly data on weekly earnings (Earnings, adjusted for inflation) in the U.S. from 2010–2017.

Year	Quarter	Earnings
2010	1	347
2010	2	340
⋮		⋮
2017	4	347

Source: U.S. Bureau of Labor Statistics

a. Plot the series to determine if the linear or the quadratic trend model is more suitable.

b. Create the time variable and the quarterly dummy variables to estimate the preferred seasonal trend model. Forecast expenses for the first quarter of 2018.

69. FILE ***Consumer_Sentiment.*** The following table lists a portion of the University of Michigan's Consumer Sentiment Index. This index is normalized to have a value of 100 in 1965 and is used to record changes in consumer morale.

Date	Consumer Sentiment
Jan-05	95.5
Feb-05	94.1
⋮	⋮
Oct-10	67.7

Source: Federal Reserve Bank of St. Louis.

Estimate a quadratic trend model with monthly dummy variables. Forecast consumer sentiment for November and December of 2010.

70. FILE ***Housing_Starts.*** Housing starts are the number of new residential construction projects that have begun during any given month. It is considered to be a leading indicator of economic strength. The following table contains a portion of monthly data on housing starts (in 1,000s) in the U.S. from 2011–2017.

Period	Housing Starts
Jan-11	40.2
Feb-11	35.4
⋮	⋮
Dec-17	81.3

Source: U.S. Bureau of the Census

Create the time and the monthly dummy variables to estimate an exponential seasonal trend model. Forecast housing starts for January and February of 2018.

WRITING WITH STATISTICS



©Photodisc/Getty Images

Numerous attempts have been made to relate happiness to various factors. Since there is no unique way to quantify happiness, researchers generally rely on surveys to capture a subjective assessment of well-being. One study relates happiness with age and finds that holding everything else constant, people seem to be least happy when they are in their mid- to upper-40s (*The Economist*, December 16, 2010). Perhaps with greater age comes maturity that contributes to a better sense of overall well-being. With regard to the influence of money, a study from Princeton University's Woodrow Wilson School suggests that money does buy happiness, but its effect diminishes as incomes rise above $75,000 a year (*Time Magazine*, September 6, 2010). Perhaps people do not need more than $75,000 to do what matters most to their emotional well-being, such as spending time with friends and family and meeting their basic food, health, and leisure needs. Nick Fisher is a young business school graduate who is fascinated by these reports. He decides to collect his own data to better comprehend and also verify the results of these studies. He surveys working adults in his hometown and inputs information on the respondent's self-assessed happiness on a scale of 0 to 100, along with age and family income. A portion of the data is shown in Table 13.24.

TABLE 13.24 Happiness, Age, and Income Data, $n = 100$

Happiness	Age	Family Income
69	49	52000
83	47	123000
⋮	⋮	⋮
79	31	105000

FILE ***Happiness***

Nick would like to use the above sample information to

1. Find the appropriate functional form to capture the influence of age and family income on happiness.
2. Predict happiness associated with varying levels of age for a family with income of $80,000.
3. Predict happiness associated with varying levels of family income for a 60-year-old working adult.

Sample Report—Understanding Happiness

In a survey of 100 working adults, respondents were asked to report their age and family income, as well as rate their happiness on a scale of 0 to 100. This report summarizes the analysis of several regression models that examine the influence of age and income on the perceived happiness of respondents. The models used various transformations to capture interesting nonlinearities suggested by recent research reports. For example, one such report shows that people get happier as they get older, despite the fact that old age is associated with a loss of hearing, vision, and muscle tone (*The New York Times*, May 31, 2010). In addition, while people start out feeling pretty good about themselves in their 20s, their self-assessed happiness deteriorates until around age 50 and then improves steadily thereafter. In order to quantify this possible quadratic effect, both age and age-squared variables are used for the regression. Also, the natural log of income is considered in order to capture the possible diminishing effect on happiness of incomes above $75,000 (*Time Magazine*, September 6, 2010). The results of the various regression models are summarized in Table 13.A.

TABLE 13.A Regression Results

	Model 1	Model 2	Model 3	Model 4
Intercept	49.1938* (0.00)	118.5285* (0.00)	−81.0939* (0.00)	−13.3021 (0.39)
Age	0.2212* (0.00)	−2.4859* (0.00)	0.2309* (0.00)	−2.4296* (0.00)
Age-squared	NA	0.0245* (0.00)	NA	0.0241* (0.00)
Income	0.0001* (0.00)	0.0001* (0.00)	NA	NA
ln(Income)	NA	NA	12.6761* (0.00)	12.7210* (0.00)
Adjusted R^2	0.4863	0.6638	0.5191	0.6907

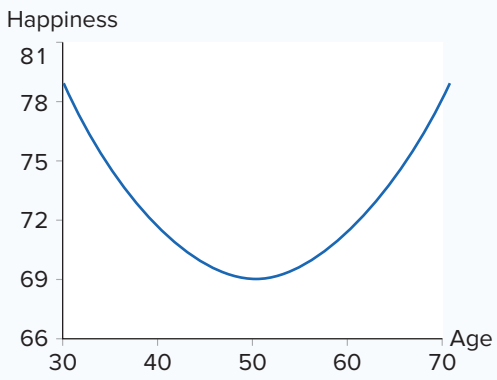
Notes: Parameter estimates are in the top portion of the table with the *p*-values in parentheses; NA denotes not applicable;
* represents significance at the 5% level. The last row presents the adjusted R^2 values for model comparison.

Model 4 was selected as the most appropriate model because it has the highest adjusted R^2 value of 0.6907. The estimated parameters of this model were used to make predictions. For instance, with family income equal to $80,000, the predicted happiness for a 30-, 50-, and 70-year-old is 79.09, 69.00, and 78.17, respectively. Note that these results are consistent with those suggesting that happiness first decreases and then increases with age. Specifically, using the estimated coefficients for Age, a person is least happy at 50.4 years of age. These results are shown graphically in Panel a of Figure 13.A, where Happiness is plotted against Age, holding Income fixed at $80,000.

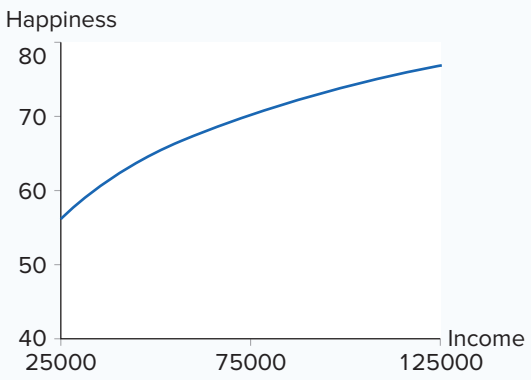
The regression results were also used to analyze the income effect. For instance, for a 60-year-old, the predicted happiness with family income of $50,000, $75,000, and $100,000 is 65.20, 70.36, and 74.02, respectively. Note that there is a greater increase in Happiness when income increases from $50,000 to $75,000 than when it increases from $75,000 to $100,000. These results are shown in Panel b of Figure 13.A, where predicted Happiness is plotted against Income, holding Age fixed at 60 years. Overall, the results support research findings.

FIGURE 13.A Predicted Happiness using Model 4 regression results

a. Happiness against Age with $80,000 in Income



b. Happiness against Income at 60 years of Age



CONCEPTUAL REVIEW

LO 13.1 Use a dummy variable to represent a qualitative explanatory variable.

A **dummy variable *d*** is defined as a variable that takes on values of 1 or 0. Dummy variables are used to represent categories of a qualitative variable. The number of dummy variables needed should be one less than the number of categories of the variable.

A regression model with a quantitative variable x and a dummy variable d is specified as $y = \beta_0 + \beta_1 x + \beta_2 d + \varepsilon$. We estimate this model to make predictions as $\hat{y} = (b_0 + b_2) + b_1 x$ for $d = 1$ and as $\hat{y} = b_0 + b_1 x$ for $d = 0$. We can perform a standard t test to determine whether differences exist between the two categories of the qualitative variable d.

LO 13.2 Use a dummy variable to capture the interaction between a qualitative explanatory variable and a quantitative explanatory variable.

A regression model with a dummy variable d, a quantitative variable x, and an interaction variable xd is specified by $y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 xd + \varepsilon$. We estimate this model to make predictions as $\hat{y} = (b_0 + b_2) + (b_1 + b_3)x$ for $d = 1$, and as $\hat{y} = b_0 + b_1 x$ for $d = 0$. The interaction variable xd allows the predicted y to differ between the two categories of a qualitative variable by a varying amount across the values of x.

LO 13.3 Estimate and interpret nonlinear regression models.

In a **quadratic regression model**, $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$, the sign of the coefficient β_2 determines whether the relationship between x and $E(y)$ is U-shaped ($\beta_2 > 0$) or inverted U-shaped ($\beta_2 < 0$). Predictions are made by $\hat{y} = b_0 + b_1 x + b_2 x^2$ where b_0, b_1, and b_2 are the coefficient estimates. In a quadratic regression model, the partial effect of x on $\hat{y}$ is approximated by $b_1 + 2b_2 x$; the maximum (if $b_2 < 0$) or minimum (if $b_2 > 0$) is reached at $x = \frac{-b_1}{2b_2}$.

In a **log-log model**, $\ln(y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$, β_1 measures the approximate percentage change in $E(y)$ when x increases by 1%. Predictions are made by $\hat{y} = \exp(b_0 + b_1 \ln(x) + s_e^2/2)$, where b_0 and b_1 are the coefficient estimates and s_e is the standard error of the estimate.

In a **logarithmic model**, $y = \beta_0 + \beta_1 \ln(x) + \varepsilon$, $\beta_1 \times 0.01$ measures the approximate change in $E(y)$ when x increases by 1%. Predictions are made by $\hat{y} = b_0 + b_1 \ln(x)$, where b_0, b_1 and b_2 are the coefficient estimates.

In an **exponential model**, $\ln(y) = \beta_0 + \beta_1 x + \varepsilon$, $\beta_1 \times 100$ measures the approximate percentage change in $E(y)$ when x increases by one unit. Predictions are made by $\hat{y} = \exp(b_0 + b_1 x + s_e^2/2)$, where b_0 and b_1 are the coefficient estimates and s_e is the standard error of the estimate.

LO 13.4 Use trend regression models to make forecasts.

Observations of any variable recorded over time in sequential order are considered a **time series**. The purpose of any forecasting model is to forecast the time series at time t, or $\hat{y}_t$.

A **linear trend model**, $y_t = \beta_0 + \beta_1 t + \varepsilon_t$, is appropriate when y_t is expected to grow by a fixed amount each time period. We estimate this model to make forecasts as $\hat{y}_t = b_0 + b_1 t$ where b_0 and b_1 are the coefficient estimates.

An **exponential trend model**, $\ln(y_t) = \beta_0 + \beta_1 t + \varepsilon_t$, is appropriate when y_t is expected to grow by an increasing amount each time period. We estimate this model to make forecasts as $\hat{y}_t = \exp(b_0 + b_1 t + s_e^2/2)$ where b_0 and b_1 are the coefficient estimates and s_e is the standard error of the estimate.

A **polynomial trend model** of order q, $y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_q t^q + \varepsilon_t$, specializes to a linear trend model for $q = 1$, to a quadratic trend model for $q = 2$, and to a cubic trend model for $q = 3$. We estimate this model to make forecasts as $\hat{y}_t = b_0 + b_1 t + b_2 t^2 + \cdots + b_q t^q$ where $b_0, b_1, \ldots b_q$ are the coefficient estimates.

LO 13.5 Use trend regression models with seasonal dummy variables to make forecasts.

We incorporate **seasonal dummy variables** in a multiple regression model to capture trend along with seasonal variations.

A linear trend model with quarterly data is estimated as $\hat{y}_t = b_0 + b_1 d_1 + b_2 d_2 + b_3 d_3 + b_4 t$. Forecasts are made as: $\hat{y}_t = b_0 + b_1 + b_4 t$ for Quarter 1, $\hat{y}_t = b_0 + b_2 + b_4 t$ for Quarter 2, $\hat{y}_t = b_0 + b_3 + b_4 t$ for Quarter 3, and $\hat{y}_t = b_0 + b_4 t$ for Quarter 4.

An exponential trend model with quarterly data is estimated as $\widehat{\ln(y_t)} = b_0 + b_1 d_1 + b_2 d_2 + b_3 d_3 + b_4 t$. Forecasts are made as: $\hat{y}_t = \exp(b_0 + b_1 + b_4 t + s_e^2/2)$ for Quarter 1, $\hat{y}_t = \exp(b_0 + b_2 + b_4 t + s_e^2/2)$ for Quarter 2, $\hat{y}_t = \exp(b_0 + b_3 + b_4 t + s_e^2/2)$ for Quarter 3, and $\hat{y}_t = \exp(b_0 + b_4 t + s_e^2/2)$ for Quarter 4.

Forecasts with quadratic and cubic models and/or with monthly data are made similarly.

71. FILE ***Magellan.*** A financial analyst collects 10 years of quarterly data on the return of Fidelity's Magellan mutual fund. A portion of the data is shown in the accompanying table.

Year	Quarter	Return	d_1	d_2	d_3
2000	1	4.85	1	0	0
2000	2	−3.96	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮
2009	4	4.06	0	0	0

Source: http://finance.yahoo.com.

a. Estimate $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \varepsilon$, where y is Magellan's quarterly return, d_1 is a dummy variable that equals 1 if quarter 1 and 0 otherwise, d_2 is a dummy variable that equals 1 if quarter 2 and 0 otherwise, and d_3 is a dummy variable that equals 1 if quarter 3 and 0 otherwise.

b. Interpret the slope coefficients of the dummy variables.

c. Predict Magellan's stock return in quarters 2 and 4.

72. FILE ***Hiring.*** In a seminal study, researchers documented race-based hiring in the Boston and Chicago labor markets (*American Economic Review*, September 2004). They sent out identical resumes to employers, half with traditionally African American names and the other half with traditionally Caucasian names. Interestingly, there was a 53% difference in call-back rates between the two groups of people. A research fellow at an institute in Santa Barbara decides to repeat the same experiment with names along with age in the Los Angeles labor market. She repeatedly sends out resumes for sales positions in the city that are identical except for the difference in the names and ages of the applicants. She also records the call-back rate for each candidate. The accompanying table shows a portion of data on call-back rate (in %), age, and a Caucasian dummy that equals 1 for a Caucasian-sounding name, 0 otherwise.

Call-back	Age	Caucasian
12	60	1
9	56	0
⋮	⋮	⋮
15	38	0

a. Estimate a linear regression model with call-back as the response variable, and age and the Caucasian dummy variable as the explanatory variables.

b. Compute the call-back rate for a 30-year-old applicant with a Caucasian-sounding name. What is the corresponding call-back rate for a non-Caucasian?

c. Conduct a test for race discrimination at the 5% significance level.

73. An analyst studies quarterly data on the relationship between retail sales (y, in $ millions), gross national product (x, in $ billions), and a quarterly dummy d that equals 1 if the sales are for the 4th quarter, 0 otherwise. He estimates the model $y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 xd + \varepsilon$. Relevant regression results are shown in the accompanying table.

	Coefficients	*Standard Error*	*t Stat*	*p-Value*
Intercept	186553.3	56421.1	3.31	0.002
x	55.0	4.6	12.08	0.000
d	112605.8	117053.0	0.96	0.342
xd	−4.7	9.3	−0.50	0.618

a. Interpret the dummy variable, d. Is it significant at the 5% level?

b. Interpret the interaction variable. Is it significant at the 5% level?

74. FILE ***Study.*** A researcher in the education department wants to determine if the number of hours that business students study per week at a state university varies by term. He conducts a survey where business students are asked how much they study per week in each of the three terms. He defines a dummy variable Fall that equals 1 if the survey was conducted in the fall term and 0 otherwise. The dummy variables Winter and Spring are defined similarly. The accompanying table shows a portion of the data for 120 students.

Study Hours	Fall	Winter	Spring
15	0	0	1
16	0	1	0
⋮	⋮	⋮	⋮
14	0	0	1

a. Estimate the appropriate model to determine, at the 5% significance level, if students study the least in the spring term.

b. Find the predicted number of hours that students study per week in the fall, winter, and spring terms.

75. FILE ***Longevity.*** According to the Center for Disease Control and Prevention, life expectancy at age 65 in America is about 18.7 years. Medical researchers have argued that while excessive drinking is detrimental to health, drinking a little alcohol every day, especially wine, may be associated with an increase in life expectancy. Others have also linked longevity with income and a person's sex. The accompanying table shows a portion of data relating to the length of life after 65, average income (in $1,000s) at a retirement age of 65, a Female dummy variable, that equals 1 if the individual is female, 0 otherwise, and the average number of alcoholic drinks consumed per day.

Life	Income	Female	Drinks
19.00	64	0	1
19.30	43	1	3
⋮	⋮	⋮	⋮
20.24	36	1	0

a. Use the data to model life expectancy at 65 on the basis of Income, Female, and Drinks.
b. Conduct a one-tailed test at $\alpha = 0.01$ to determine if females live longer than males.
c. Predict the life expectancy at 65 of a male with an income of $40,000 and an alcoholic consumption of two drinks per day; repeat the prediction for a female.

76. FILE ***Shifts.*** The manager of a diner wants to reevaluate his staffing needs depending on variations in customer traffic during the day. He collects data on the number of customers served, along with four dummy variables representing the morning, afternoon, evening, and night shifts. The dummy variable Morning equals 1 if the number of customers served was from the morning shift and 0 otherwise; other dummy variables are defined similarly. The accompanying table shows a portion of the data.

Customers	Morning	Afternoon	Evening	Night
99	0	0	0	1
148	0	1	0	0
⋮	⋮	⋮	⋮	⋮
111	0	1	0	0

a. Estimate a regression model using the number of customers as the response variable and the shift dummy variables as the explanatory variables; use Night as the reference category.
b. What is the predicted number of customers served during the morning, afternoon, evening, and night shifts?

77. FILE ***Overweight.*** According to the U.S. Department of Health and Human Services, African American women have the highest rates of being overweight compared to other groups in the United States. Individuals are considered overweight if their body mass index (BMI) is 25 or greater. Data are collected from 120 individuals. The following table shows a portion of data on each individual's BMI, a Female dummy variable that equals 1 if the individual is female, 0 otherwise, and a Black dummy variable that equals 1 if the individual is African American, 0 otherwise.

BMI	Female	Black
28.70	0	1
28.31	0	0
⋮	⋮	⋮
24.90	0	1

a. Estimate the model, $\text{BMI} = \beta_0 + \beta_1\text{Female} + \beta_2\text{Black} + \beta_3(\text{Female} \times \text{Black}) + \varepsilon$, to predict the BMI for white males, white females, black males, and black females.
b. Is the difference between white females and white males statistically significant at the 5% level?
c. Is the difference between white males and black males statistically significant at the 5% level?

78. FILE ***Compensation.*** To encourage performance, loyalty, and continuing education, the human resources department at a large company wants to develop a regression-based compensation model (Comp in $ per year) for mid-level managers based on three variables: (1) business unit-profitability (Profit in $1,000's per year), (2) years with the company (Years), and (3) whether or not the manager has a graduate degree (Grad equals 1 if graduate degree, 0 otherwise). The accompanying table shows a portion of data collected for 36 managers.

Comp	Profit	Years	Grad
118100	4500	37	1
90800	5400	5	1
⋮	⋮	⋮	⋮
85000	4200	29	0

a. Estimate the following model for compensation: $\text{Comp} = \beta_0 + \beta_1\text{Profit} + \beta_2\text{Years} + \beta_3\text{Grad} + \beta_4\text{Profit} \times \text{Grad} + \beta_5\text{Years} \times \text{Grad} + \varepsilon$.
b. At the 5% significance level, is the overall regression model significant?
c. Which predictor variables and interaction terms are significant at $\alpha = 0.05$?
d. Use the (full) model to determine compensation for a manager having 15 years with the company, a graduate degree, and a business-unit profit of $4,800(000) last year.

79. FILE ***Fertilizer2.*** A horticulturist is studying the relationship between tomato plant height and fertilizer amount. Thirty tomato plants grown in similar conditions were subjected to various amounts of fertilizer (in ounces) over a four-month period, and then their heights (in inches) were measured. A portion of the results is shown in the accompanying table.

Height	Fertilizer
20.4	1.9
29.1	5.0
⋮	⋮
36.4	3.1

a. Estimate the linear regression model: $Height = \beta_0 + \beta_1 \text{Fertilizer} + \varepsilon$.

b. Estimate the quadratic regression model: $Height = \beta_0 + \beta_1 \text{Fertilizer} + \beta_2 \text{Fertilizer}^2 + \varepsilon$. Find the fertilizer amount at which the height reaches a minimum or maximum.

c. Use the best-fitting model to predict, after a four-month period, the height of a tomato plant that received 3.0 ounces of fertilizer.

80. FILE ***Circuit_Boards.*** The operators manager at an electronics company believes that the time required for workers to build a circuit board is not necessarily proportional to the number of parts on the board. He wants to develop a regression model to predict time (in minutes) based on part quantity. He has collected data for the last 25 boards. A portion of this data is shown in the accompanying table.

Time	Parts
30.8	62
9.8	32
⋮	⋮
29.8	60

a. Estimate the linear regression model to predict time as a function of the number of parts (Parts). Then estimate the quadratic regression model to predict time as a function of Parts and Parts squared.

b. Evaluate the two models in terms of variable significance ($\alpha = 0.05$) and adjusted R^2.

c. Use the best-fitting model to predict how long it would take to build a circuit board consisting of 48 parts.

81. FILE ***Inventory_Cost.*** The inventory manager at a warehouse distributor wants to predict inventory cost (Cost in $) based on order quantity (Quantity in units). She thinks it may be a nonlinear relationship since its two primary components move in opposite directions: (1) order processing cost (costs of procurement personnel, shipping, transportation), which *decreases* as order quantity increases (due to fewer orders needed), and (2) holding cost (costs of capital, facility, warehouse personnel, equipment), which *increases* as order quantity increases (due to more inventory held). She has collected monthly inventory costs and order quantities for the past 36 months. A portion of the data is shown in the accompanying table.

Cost	Quantity
54.4	844
52.1	503
⋮	⋮
55.5	870

a. Create a scatterplot of inventory cost as a function of quantity. Superimpose a linear trendline and quadratic trendline.

b. Estimate the linear regression model to predict inventory cost as a function of order quantity. Then estimate the quadratic regression model to predict inventory cost as a function of order quantity and order quantity squared.

c. Evaluate the two models in terms of significance tests ($\alpha = 0.05$) and adjusted R^2.

d. Use the best-fitting model to predict monthly inventory cost for an order quantity of 800 units.

82. FILE ***Learning_Curve.*** Learning curves are used in production operations to estimate the time required to complete a repetitive task as an operator gains experience. Suppose a production manager has compiled 30 time values (in minutes) for a particular operator as she progressed down the learning curve during the first 100 units. A portion of this data is shown in the accompanying table.

Time per Unit	Unit Number
18.30	3
17.50	5
⋮	⋮
5.60	100

a. Create a scatterplot of time per unit against units built. Superimpose a linear trendline and a logarithmic trendline to determine visually the best-fitting model.

b. Estimate the linear regression model and the logarithmic regression model for predicting time per unit using unit number as the explanatory variable.

c. Based on R^2, use the best-fitting model to predict the time that was required for the operator to build Unit 50.

83. FILE ***Smoking.*** A nutritionist wants to understand the influence of income and healthy food on the incidence of smoking. He collects 2009 data on the percentage of smokers in each state in the U.S., the percentage of the state's population that regularly eats fruits and vegetables and the state's median income (in $). A portion of the data is shown in the accompanying table.

State	Smoke	Fruits/ Vegetables	Income
AK	14.6	23.3	61604
AL	16.4	20.3	39980
⋮	⋮	⋮	⋮
WY	15.2	23.3	52470

Source: Centers for Disease Control and Prevention and U.S. Census Bureau.

a. Estimate: Smoke $= \beta_0 + \beta_1$Fruits/Vegetables $+$ β_2Income $+ \varepsilon$.
b. Compare this model with a model that log-transforms the income variable.

84. FILE ***GasPrice_Forecast.*** Consider the following portion of monthly data on the price per gallon of regular unleaded gasoline in the United States from January 2009 to December 2010.

Date	Price Per Gallon
Jan-09	1.79
Feb-09	1.92
⋮	⋮
Dec-10	2.99

Source: U.S. Energy Information Administration.

a. Plot the series to identify an appropriate polynomial trend model. You may ignore seasonality.
b. Compare the adjusted R^2 of the linear, the quadratic, and the cubic trend models.
c. Use the appropriate model to make a forecast for the price of regular unleaded gasoline for January and February of 2011.

85. FILE ***Country_Rap.*** The following table lists a portion of the percentage (share) of total shipment of music that falls in the category of country and rap/hip-hop rock music from 1990–2008.

Year	Country	Rap/Hip-hop
1990	9.6	8.5
1991	12.8	10.0
⋮	⋮	⋮
2008	11.9	10.7

Source: www.riaa.com.

a. Plot each of the series and comment on the respective trend.
b. Estimate a linear, a quadratic, and a cubic trend model for the share of country music in the United States. Use adjusted R^2 to choose the preferred model, and, with this model, make a forecast for 2009.
c. Estimate a linear, a quadratic, and a cubic trend model for the share of rap/hip-hop music in the United States. Use adjusted R^2 to choose the preferred model, and, with this model, make a forecast for 2009.

86. FILE ***Loans.*** Consider the following portion of data on real estate loans granted by FDIC-insured Commercial Banks in the United States (in $ billions, base = 2007) from 1972 to 2007.

Year	Loans
1972	489.27
1973	567.26
⋮	⋮
2007	3604.03

Source: www2.fdic.gov.

a. Plot the series and comment on the growth of real estate loans.
b. Estimate the exponential trend model to forecast loans in 2008.

87. FILE ***Inventory_Sales.*** While U.S. inventory levels remain low, there is a slight indication of an increase in the U.S. business inventory-to-sales ratio, due to higher sales (*The Wall Street Journal*, December 15, 2010). The accompanying table shows a portion of seasonally adjusted inventory-to-sales ratios from January 2008 to October 2010.

Date	Inventory-to-Sales
Jan-08	1.28
Feb-08	1.30
⋮	⋮
Oct-10	1.27

Source: U.S. Department of Commerce.

a. Plot the series. Which polynomial trend model do you think is most appropriate?
b. Verify your answer by formally comparing the linear, the quadratic, and the cubic trend models.
c. Make a forecast for the inventory-to-sales ratio for November and December of 2010.

88. FILE ***Lowe's_Sales.*** The following data represent a portion of quarterly net sales (in $ millions) of Lowe's Companies, Inc., over five years.

Year	Quarter	Net Sales
2004	1	8861
2004	2	10169
⋮	⋮	⋮
2008	1	9984

Source: All data retrieved from Annual Reports for Lowe's Companies, Inc.

Estimate an exponential trend model with quarterly dummy variables to forecast net sales for fiscal year 2009.

89. FILE ***Revenue_Miles.*** Revenue passenger-miles are calculated by multiplying the number of paying passengers by the distance flown in thousands. The accompanying table shows a portion of monthly data on revenue passenger-miles (in millions) from January 2006 through September 2010.

Date	Revenue
Jan-06	43.1652
Feb-06	44.0447
⋮	⋮
Sep-10	43.7704

Source: Bureau of Transportation Statistics.

Estimate a linear trend model with monthly dummy variables to forecast revenue for the last three months of 2010.

CASE STUDIES

CASE STUDY 13.1 Jack Sprague is the relocation specialist for a real estate firm in the town of Arlington, Massachusetts. He has been working with a client who wishes to purchase a single-family home in Arlington. After seeing the information that Jack provided, the client is perplexed by the variability of home prices in Arlington. She is especially puzzled by the premium that a colonial house commands. (A colonial house is a style dating back to the time of the American colonies, with a simple rectangular structure and a peaked roof.) Despite Jack's eloquent explanations, it seems that the client will not be satisfied until she understands the quantitative relationship between house prices and house characteristics. Jack decides to use a multiple regression model to provide the client with the necessary information. He collects data on the prices (in $) for 36 single-family homes in Arlington sold in the first quarter of 2009. Also included in the data is the information on square footage, the number of bedrooms, the number of bathrooms, and whether or not the house is a colonial (1 for colonial, 0 otherwise). A portion of the data is shown in the accompanying table.

Data for Case Study 13.1 Sales Information of Single-Family Homes in Arlington, MA

FILE *Arlington*

Price	Square feet	Bedrooms	Baths	Colonial
840000	2768	4	3.5	1
822000	2500	4	2.5	1
⋮	⋮	⋮	⋮	⋮
307500	850	1	1	0

Source: NewEnglandMoves.com.

In a report, use the sample information to

1. Estimate and interpret three models, where d is the colonial dummy variable.

   Model 1: $\text{Price} = \beta_0 + \beta_1\text{Sqft} + \beta_2\text{Beds} + \beta_3\text{Baths} + \beta_4 d + \varepsilon$.

   Model 2: $\text{Price} = \beta_0 + \beta_1\text{Sqft} + \beta_2\text{Beds} + \beta_3\text{Baths} + \beta_4(\text{Sqft} \times d) + \varepsilon$.

   Model 3: $\text{Price} = \beta_0 + \beta_1\text{Sqft} + \beta_2\text{Beds} + \beta_3\text{Baths} + \beta_4 d + \beta_5(\text{Sqft} \times d) + \varepsilon$.

2. Choose which model is most reliable in predicting the price of a house. Provide at least one reason for your choice. Are price differences between colonial homes versus other styles fixed and/or changing at the 5% significance level?
3. Using the preferred model, make predictions for a colonial home versus other styles, given the average values of the explanatory variables.

CASE STUDY 13.2 Brendan Connolly, a statistician for a Major League Baseball (MLB) team, wants to elaborate on the salary of baseball players. Excluding pitchers from his analysis, he believes that a baseball player's batting average (BA), runs batted in (RBI), and years of experience playing professional baseball (Experience) are the most important factors that influence a player's salary. Further, he believes that salaries rise with experience only up to a point, beyond which they begin to fall; in other words, experience has a quadratic effect on salaries. Brendan collects data on salary (in $1,000s), BA, RBI, and experience for 138 outfielders in 2008. A portion of the data is shown in the accompanying table.

FILE
MLB_Salary

Data for Case Study 13.2 Major League Baseball Outfielder Data, $n = 138$

Player	Salary	BA	RBI	Experience
Nick Markakis	455	299	87	3
Adam Jones	390	261	23	3
⋮	⋮	⋮	⋮	⋮
Randy Winn	8875	288	53	11

Notes: All data collected from usatoday.com or espn.com; BA and RBI are averages over the player's professional life through 2008. For exposition, BA has been multiplied by 1,000.

In a report, use the sample information to

1. Estimate a quadratic regression model using Salary as the response variable and BA, RBI, Experience, and Experience^2 as the explanatory variables.
2. Determine the optimal level of experience at which the salary is maximized.

CASE STUDY 13.3 Madelyn Davis is a research analyst for a large investment firm. She has been assigned the task of forecasting sales for Walmart Stores, Inc., for fiscal year 2011. She collects quarterly sales for Walmart Stores, Inc. (in $ millions) for the 10-year period 2001 through 2010, a portion of which is shown in the accompanying table.

FILE
Walmart_Sales

Data for Case Study 13.3 Walmart Quarterly Sales (in millions $)

Period	Sales
2001:01	42985
2001:02	46112
⋮	⋮
2010:04	112826

Source: Annual Reports for Walmart Stores Inc.

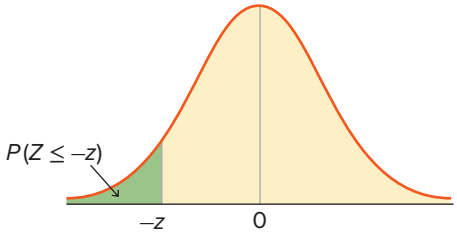
In a report, use the sample information to

1. Construct a scatterplot and determine which model best depicts trend for Walmart's sales.
2. Estimate the appropriate trend model with seasonal dummy variables to provide forecast values for the four quarters of 2011 as well as total projected sales for fiscal year 2011.

Tables

TABLE 1 Standard Normal Curve Areas

Entries in this table provide cumulative probabilities, that is, the area under the curve to the left of $-z$. For example, $P(Z \leq -1.52) = 0.0643$.

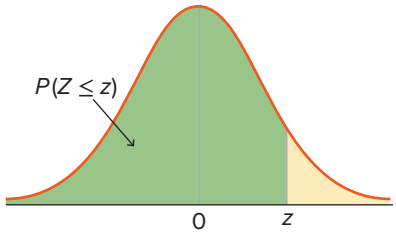


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
−3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
−3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Source: Probabilities calculated with Excel.

TABLE 1 (*Continued*)

Entries in this table provide cumulative probabilities, that is, the area under the curve to the left of z. For example, $P(Z \leq 1.52) = 0.9357$.

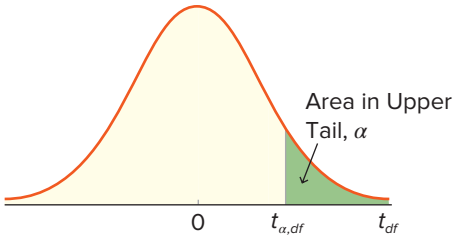


z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Source: Probabilities calculated with Excel.

TABLE 2 Student's *t* Distribution

Entries in this table provide the values of $t_{\alpha,df}$ that correspond to a given upper-tail area α and a specified number of degrees of freedom *df*. For example, for $\alpha = 0.05$ and $df = 10$, $P(T_{10} \geq 1.812) = 0.05$.



	α					
df	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	1.376	3.078	6.314	12.706	31.821	63.657
2	1.061	1.886	2.920	4.303	6.965	9.925
3	0.978	1.638	2.353	3.182	4.541	5.841
4	0.941	1.533	2.132	2.776	3.747	4.604
5	0.920	1.476	2.015	2.571	3.365	4.032
6	0.906	1.440	1.943	2.447	3.143	3.707
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1.397	1.860	2.306	2.896	3.355
9	0.883	1.383	1.833	2.262	2.821	3.250
10	0.879	1.372	1.812	2.228	2.764	3.169
11	0.876	1.363	1.796	2.201	2.718	3.106
12	0.873	1.356	1.782	2.179	2.681	3.055
13	0.870	1.350	1.771	2.160	2.650	3.012
14	0.868	1.345	1.761	2.145	2.624	2.977
15	0.866	1.341	1.753	2.131	2.602	2.947
16	0.865	1.337	1.746	2.120	2.583	2.921
17	0.863	1.333	1.740	2.110	2.567	2.898
18	0.862	1.330	1.734	2.101	2.552	2.878
19	0.861	1.328	1.729	2.093	2.539	2.861
20	0.860	1.325	1.725	2.086	2.528	2.845
21	0.859	1.323	1.721	2.080	2.518	2.831
22	0.858	1.321	1.717	2.074	2.508	2.819
23	0.858	1.319	1.714	2.069	2.500	2.807
24	0.857	1.318	1.711	2.064	2.492	2.797
25	0.856	1.316	1.708	2.060	2.485	2.787
26	0.856	1.315	1.706	2.056	2.479	2.779
27	0.855	1.314	1.703	2.052	2.473	2.771
28	0.855	1.313	1.701	2.048	2.467	2.763
29	0.854	1.311	1.699	2.045	2.462	2.756
30	0.854	1.310	1.697	2.042	2.457	2.750

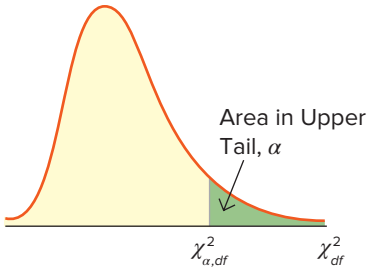
TABLE 2 (*Continued*)

	α					
df	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
31	0.853	1.309	1.696	2.040	2.453	2.744
32	0.853	1.309	1.694	2.037	2.449	2.738
33	0.853	1.308	1.692	2.035	2.445	2.733
34	0.852	1.307	1.691	2.032	2.441	2.728
35	0.852	1.306	1.690	2.030	2.438	2.724
36	0.852	1.306	1.688	2.028	2.434	2.719
37	0.851	1.305	1.687	2.026	2.431	2.715
38	0.851	1.304	1.686	2.024	2.429	2.712
39	0.851	1.304	1.685	2.023	2.426	2.708
40	0.851	1.303	1.684	2.021	2.423	2.704
41	0.850	1.303	1.683	2.020	2.421	2.701
42	0.850	1.302	1.682	2.018	2.418	2.698
43	0.850	1.302	1.681	2.017	2.416	2.695
44	0.850	1.301	1.680	2.015	2.414	2.692
45	0.850	1.301	1.679	2.014	2.412	2.690
46	0.850	1.300	1.679	2.013	2.410	2.687
47	0.849	1.300	1.678	2.012	2.408	2.685
48	0.849	1.299	1.677	2.011	2.407	2.682
49	0.849	1.299	1.677	2.010	2.405	2.680
50	0.849	1.299	1.676	2.009	2.403	2.678
51	0.849	1.298	1.675	2.008	2.402	2.676
52	0.849	1.298	1.675	2.007	2.400	2.674
53	0.848	1.298	1.674	2.006	2.399	2.672
54	0.848	1.297	1.674	2.005	2.397	2.670
55	0.848	1.297	1.673	2.004	2.396	2.668
56	0.848	1.297	1.673	2.003	2.395	2.667
57	0.848	1.297	1.672	2.002	2.394	2.665
58	0.848	1.296	1.672	2.002	2.392	2.663
59	0.848	1.296	1.671	2.001	2.391	2.662
60	0.848	1.296	1.671	2.000	2.390	2.660
80	0.846	1.292	1.664	1.990	2.374	2.639
100	0.845	1.290	1.660	1.984	2.364	2.626
150	0.844	1.287	1.655	1.976	2.351	2.609
200	0.843	1.286	1.653	1.972	2.345	2.601
500	0.842	1.283	1.648	1.965	2.334	2.586
1000	0.842	1.282	1.646	1.962	2.330	2.581
∞	0.842	1.282	1.645	1.960	2.326	2.576

Source: *t* values calculated with Excel.

TABLE 3 χ^2 (Chi-Square) Distribution

Entries in this table provide the values of $\chi^2_{\alpha,df}$ that correspond to a given upper-tail area α and a specified number of degrees of freedom *df*. For example, for $\alpha = 0.05$ and $df = 10$, $P(\chi^2_{10} \geq 18.307) = 0.05$.



	α									
df	**0.995**	**0.990**	**0.975**	**0.950**	**0.900**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672

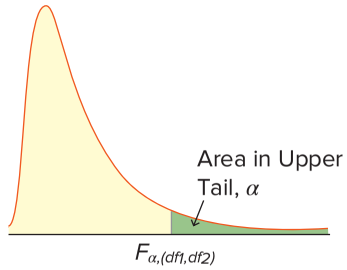
TABLE 3 (*Continued*)

	α									
df	**0.995**	**0.990**	**0.975**	**0.950**	**0.900**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	23.110	43.745	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.586	19.960	22.106	24.075	26.492	48.363	52.192	55.668	59.893	62.883
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.996	21.426	23.654	25.695	28.196	50.660	54.572	58.120	62.428	65.476
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
41	21.421	22.906	25.215	27.326	29.907	52.949	56.942	60.561	64.950	68.053
42	22.138	23.650	25.999	28.144	30.765	54.090	58.124	61.777	66.206	69.336
43	22.859	24.398	26.785	28.965	31.625	55.230	59.304	62.990	67.459	70.616
44	23.584	25.148	27.575	29.787	32.487	56.369	60.481	64.201	68.710	71.893
45	24.311	25.901	28.366	30.612	33.350	57.505	61.656	65.410	69.957	73.166
46	25.041	26.657	29.160	31.439	34.215	58.641	62.830	66.617	71.201	74.437
47	25.775	27.416	29.956	32.268	35.081	59.774	64.001	67.821	72.443	75.704
48	26.511	28.177	30.755	33.098	35.949	60.907	65.171	69.023	73.683	76.969
49	27.249	28.941	31.555	33.930	36.818	62.038	66.339	70.222	74.919	78.231
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
55	31.735	33.570	36.398	38.958	42.060	68.796	73.311	77.380	82.292	85.749
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
65	39.383	41.444	44.603	47.450	50.883	79.973	84.821	89.177	94.422	98.105
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
75	47.206	49.475	52.942	56.054	59.795	91.061	96.217	100.839	106.393	110.286
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
85	55.170	57.634	61.389	64.749	68.777	102.079	107.522	112.393	118.236	122.325
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
95	63.250	65.898	69.925	73.520	77.818	113.038	118.752	123.858	129.973	134.247
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Source: χ^2 values calculated with Excel.

TABLE 4 *F* Distribution

Entries in this table provide the values of $F_{\alpha,(df_1,df_2)}$ that correspond to a given upper-tail area α and a specified number of degrees of freedom in the numerator df_1 and degrees of freedom in the denominator df_2. For example, for $\alpha = 0.05$, $df_1 = 8$, and $df_2 = 6$, $P(F_{(8,6)} \geq 4.15) = 0.05$.



		df_1														
df_2	α	1	2	3	4	5	6	7	8	9	10	15	25	50	100	500
1	0.10	39.86	49.50	53.59	55.83	57.24	58.2	58.91	59.44	59.86	60.19	61.22	62.05	62.69	63.01	63.26
	0.05	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95	249.26	251.77	253.04	254.06
	0.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	984.87	998.08	1008.12	1013.17	1017.24
	0.01	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6157.28	6239.83	6302.52	6334.11	6359.50
2	0.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42	9.45	9.47	9.48	9.49
	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.46	19.48	19.49	19.49
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43	39.46	39.48	39.49	39.50
	0.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.43	99.46	99.48	99.49	99.50
3	0.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20	5.17	5.15	5.14	5.14
	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.63	8.58	8.55	8.53
	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25	14.12	14.01	13.96	13.91
	0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.87	26.58	26.35	26.24	26.15
4	0.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87	3.83	3.80	3.78	3.76
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.77	5.70	5.66	5.64
	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.50	8.38	8.32	8.27
	0.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20	13.91	13.69	13.58	13.49
5	0.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.24	3.19	3.15	3.13	3.11
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.52	4.44	4.41	4.37
	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.27	6.14	6.08	6.03
	0.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72	9.45	9.24	9.13	9.04
6	0.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87	2.81	2.77	2.75	2.73
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.83	3.75	3.71	3.68
	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.11	4.98	4.92	4.86
	0.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.30	7.09	6.99	6.90
7	0.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63	2.57	2.52	2.50	2.48
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.40	3.32	3.27	3.24
	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.40	4.28	4.21	4.16
	0.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.06	5.86	5.75	5.67

TABLE 4 (*Continued*)

		df_1														
df_2	α	1	2	3	4	5	6	7	8	9	10	15	25	50	100	500
8	0.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46	2.40	2.35	2.32	2.30
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.11	3.02	2.97	2.94
	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	3.94	3.81	3.74	3.68
	0.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.26	5.07	4.96	4.88
9	0.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34	2.27	2.22	2.19	2.17
	0.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.89	2.80	2.76	2.72
	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.60	3.47	3.40	3.35
	0.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.71	4.52	4.41	4.33
10	0.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24	2.17	2.12	2.09	2.06
	0.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.73	2.64	2.59	2.55
	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.35	3.22	3.15	3.09
	0.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.31	4.12	4.01	3.93
11	0.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.17	2.10	2.04	2.01	1.98
	0.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.60	2.51	2.46	2.42
	0.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33	3.16	3.03	2.96	2.90
	0.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.01	3.81	3.71	3.62
12	0.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.10	2.03	1.97	1.94	1.91
	0.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.50	2.40	2.35	2.31
	0.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18	3.01	2.87	2.80	2.74
	0.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.76	3.57	3.47	3.38
13	0.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.05	1.98	1.92	1.88	1.85
	0.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.41	2.31	2.26	2.22
	0.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05	2.88	2.74	2.67	2.61
	0.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.57	3.38	3.27	3.19
14	0.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.01	1.93	1.87	1.83	1.80
	0.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.34	2.24	2.19	2.14
	0.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95	2.78	2.64	2.56	2.50
	0.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.41	3.22	3.11	3.03
15	0.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97	1.89	1.83	1.79	1.76
	0.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.28	2.18	2.12	2.08
	0.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.69	2.55	2.47	2.41
	0.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.28	3.08	2.98	2.89
16	0.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.94	1.86	1.79	1.76	1.73
	0.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.23	2.12	2.07	2.02
	0.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.79	2.61	2.47	2.40	2.33
	0.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.16	2.97	2.86	2.78

TABLE 4 (*Continued*)

		df_1														
df_2	α	1	2	3	4	5	6	7	8	9	10	15	25	50	100	500
17	0.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.91	1.83	1.76	1.73	1.69
	0.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.18	2.08	2.02	1.97
	0.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.72	2.55	2.41	2.33	2.26
	0.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31	3.07	2.87	2.76	2.68
18	0.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.89	1.80	1.74	1.70	1.67
	0.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.14	2.04	1.98	1.93
	0.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.67	2.49	2.35	2.27	2.20
	0.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	2.98	2.78	2.68	2.59
19	0.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.86	1.78	1.71	1.67	1.64
	0.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.11	2.00	1.94	1.89
	0.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.62	2.44	2.30	2.22	2.15
	0.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15	2.91	2.71	2.60	2.51
20	0.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.84	1.76	1.69	1.65	1.62
	0.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.07	1.97	1.91	1.86
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.40	2.25	2.17	2.10
	0.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.84	2.64	2.54	2.44
21	0.10	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.83	1.74	1.67	1.63	1.60
	0.05	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.18	2.05	1.94	1.88	1.83
	0.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.53	2.36	2.21	2.13	2.06
	0.01	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.03	2.79	2.58	2.48	2.38
22	0.10	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.81	1.73	1.65	1.61	1.58
	0.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.02	1.91	1.85	1.80
	0.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.50	2.32	2.17	2.09	2.02
	0.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98	2.73	2.53	2.42	2.33
23	0.10	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.80	1.71	1.64	1.59	1.56
	0.05	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.13	2.00	1.88	1.82	1.77
	0.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.47	2.29	2.14	2.06	1.99
	0.01	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	2.93	2.69	2.48	2.37	2.28
24	0.10	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.78	1.70	1.62	1.58	1.54
	0.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	1.97	1.86	1.80	1.75
	0.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.44	2.26	2.11	2.02	1.95
	0.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	2.89	2.64	2.44	2.33	2.24

TABLE 4 (*Continued*)

		df_1														
df_2	α	1	2	3	4	5	6	7	8	9	10	15	25	50	100	500
25	0.10	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.77	1.68	1.61	1.56	1.53
	0.05	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.09	1.96	1.84	1.78	1.73
	0.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.41	2.23	2.08	2.00	1.92
	0.01	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.85	2.60	2.40	2.29	2.19
26	0.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.76	1.67	1.59	1.55	1.51
	0.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.94	1.82	1.76	1.71
	0.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.39	2.21	2.05	1.97	1.90
	0.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.81	2.57	2.36	2.25	2.16
27	0.10	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.75	1.66	1.58	1.54	1.50
	0.05	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.06	1.92	1.81	1.74	1.69
	0.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.36	2.18	2.03	1.94	1.87
	0.01	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.78	2.54	2.33	2.22	2.12
28	0.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.74	1.65	1.57	1.53	1.49
	0.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.91	1.79	1.73	1.67
	0.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.34	2.16	2.01	1.92	1.85
	0.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.75	2.51	2.30	2.19	2.09
29	0.10	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.73	1.64	1.56	1.52	1.48
	0.05	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.03	1.89	1.77	1.71	1.65
	0.025	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.32	2.14	1.99	1.90	1.83
	0.01	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.73	2.48	2.27	2.16	2.06
30	0.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.72	1.63	1.55	1.51	1.47
	0.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.88	1.76	1.70	1.64
	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.12	1.97	1.88	1.81
	0.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.45	2.25	2.13	2.03
50	0.10	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73	1.63	1.53	1.44	1.39	1.34
	0.05	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.73	1.60	1.52	1.46
	0.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32	2.11	1.92	1.75	1.66	1.57
	0.01	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.42	2.17	1.95	1.82	1.71
100	0.10	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66	1.56	1.45	1.35	1.29	1.23
	0.05	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77	1.62	1.48	1.39	1.31
	0.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18	1.97	1.77	1.59	1.48	1.38
	0.01	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22	1.97	1.74	1.60	1.47
500	0.10	2.72	2.31	2.09	1.96	1.86	1.79	1.73	1.68	1.64	1.61	1.50	1.39	1.28	1.21	1.12
	0.05	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.69	1.53	1.38	1.28	1.16
	0.025	5.05	3.72	3.14	2.81	2.59	2.43	2.31	2.22	2.14	2.07	1.86	1.65	1.46	1.34	1.19
	0.01	6.69	4.65	3.82	3.36	3.05	2.84	2.68	2.55	2.44	2.36	2.07	1.81	1.57	1.41	1.23

Source: *F*-values calculated with Excel.

Answers to Selected Even-Numbered Exercises

Chapter 1

1.2 35 is likely the estimated average age. It would be rather impossible to reach all video game players.

1.4 a. The population consists of all marketing managers.

b. No, the average salary was likely computed from a sample in order to save time and money.

1.6 Answers will vary depending on when data are retrieved. The numbers represent time series data.

1.8 Answers will vary depending on when data are retrieved. The numbers represent cross-sectional data.

1.10 Answers will vary depending on when data are retrieved. The numbers represent cross-sectional data.

1.12 Structured

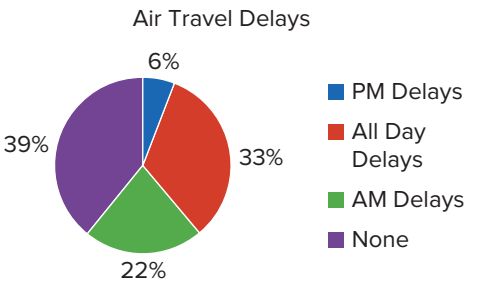
1.14 Structured; time series

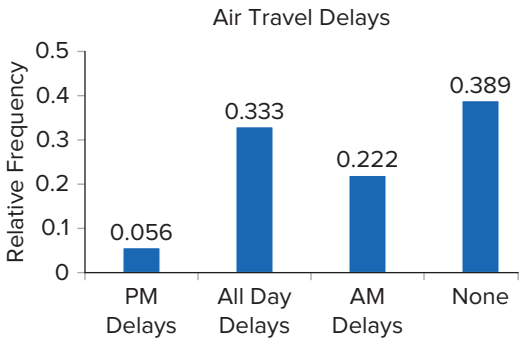
Chapter 2

2.4 a.

Delays	Frequency	Relative Frequency
PM Delays	1	0.056
All Day Delays	6	0.333
AM Delays	4	0.222
None	7	0.389

b.



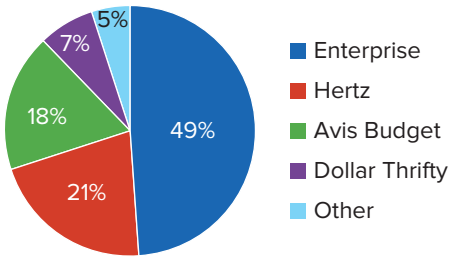


2.8 a.

Company	Market Share
Enterprise	0.489
Hertz	0.215
Avis Budget	0.183
Dollar Thrifty	0.068
Other	0.046

b. Hertz accounted for 21.5% of sales.

c.



2.12 It does not. With a relatively high upper limit on the vertical axis ($500), the rise in stock price appears dampened.

2.14 It does not. The vertical axis has been stretched and the increase in sales appears more pronounced than warranted.

2.22 a.

Assets (in billions)	Frequency
40 up to 70	9
70 up to 100	8
100 up to 130	2
130 up to 160	0
160 up to 190	1

b.

Assets (in billions)	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
40 up to 70	0.45	9	0.45
70 up to 100	0.40	17	0.85
100 up to 130	0.10	19	0.95
130 up to 160	0.00	19	0.95
160 up to 190	0.05	20	1.00

c. Two funds had assets of at least 100 but less than 130 (in $ billions); 19 funds had assets less than $160 billion.

d. 40% of the funds had assets of at least $70 but less than $100 (in billions); 95% of the funds had assets less than $130 billion.

e. The distribution is positively skewed.

2.24 a.

Temperature	Frequency
60 up to 70	2
70 up to 80	7
80 up to 90	14
90 up to 100	10

b.

Temperature	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
60 up to 70	0.061	2	0.061
70 up to 80	0.212	9	0.273
80 up to 90	0.424	23	0.697
90 up to 100	0.303	33	1.000

c. 9 cities had temperatures less than 80°.

d. 42.4% of the cities recorded temperatures of at least 80° but less than 90°; 69.7% of the cities had temperatures less than 90°.

e. The distribution is slightly negatively skewed.

2.26 a.

Vacancy Rate (%)	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0 up to 3	0.10	5	0.10
3 up to 6	0.20	15	0.30
6 up to 9	0.40	35	0.70
9 up to 12	0.20	45	0.90
12 up to 15	0.10	50	1.00

b. 45 cities had a vacancy rate of less than 12%; 40% of the cities had a vacancy rate of at least 6% but less than 9%; 70% of the cities had a vacancy rate of less than 9%.

c. The distribution is symmetric.

2.30 a. No, it is positively skewed

b. Minimum is at least 50; maximum is at most 450.

c. 50–150 class

2.32 a. 70%

b. 85%

2.38

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

No, it is positively skewed. Most of the numbers are in the lower stems of –8 and –7.

2.40

Stem	Leaf
7	3 4 6 7 8 8
8	0 1 2 3 4 4 4 4 7 8
9	0 0 0 1 1 2 2 2 3 3 4 4 4 4 4 5 6 6 6 8 8 9
10	6 7

Temperatures ranged between 73 and 107. The distribution is negatively skewed, with most temperatures in the 90s.

2.42 Spain

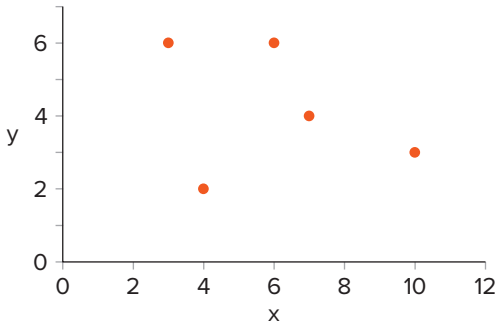
Stem	Leaf
2	1 1 1 2 3 3 4 4 5 5 5 6 7 8 9 9 9
3	0 0 2

Netherlands

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

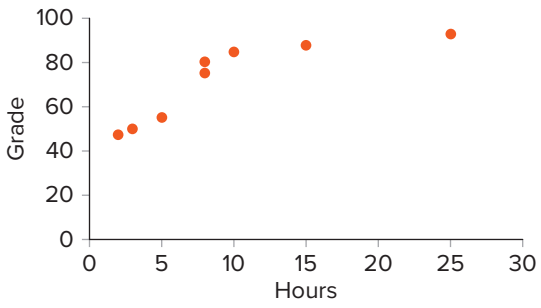
Spain has a relatively younger team compared to the Netherlands. The majority of players on both teams are in their 20s.

2.44



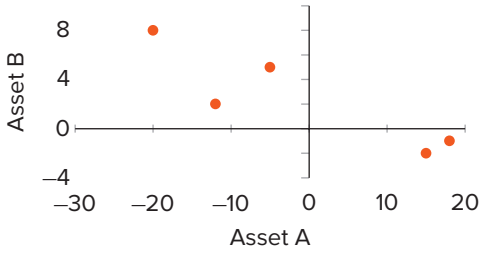
There is no evident relationship between *x* and *y*.

2.46



There is a positive relationship between number of hours spent studying and grades.

2.48



There is a negative relationship between the two asset returns. For diversification, the investor should include both assets in her portfolio.

2.50 a.

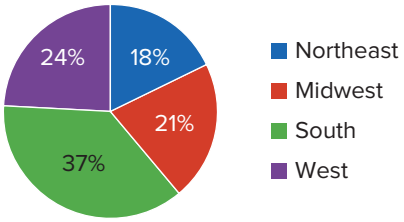
Responses	Utah	Kentucky
Yes	0.10	0.45
No	0.90	0.55

Relative to Utah, Kentucky is more lenient in allowing smoking at home.

b.

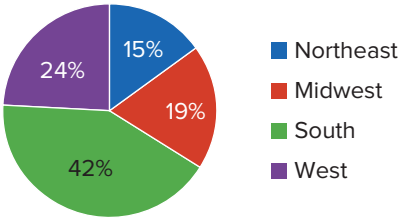


2.58 a. Percentage of People in Each Region



The highest percentage of people live in the South and the lowest percentage live in the Northeast.

b. Percentage of People Below Poverty Level



The percentage of people living in poverty is highest in the South and lowest in the Northeast. Furthermore, relative to the population, there are more (less) people living in poverty in the South (Northeast).

2.62 a. 16%

b. 76%

c.

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

The distribution is negatively skewed. The majority of ages range between 60s and 70s.

2.64 a.

Type of House	Frequency
Colonial	6
Contemporary	4
Ranch	6
Other	4

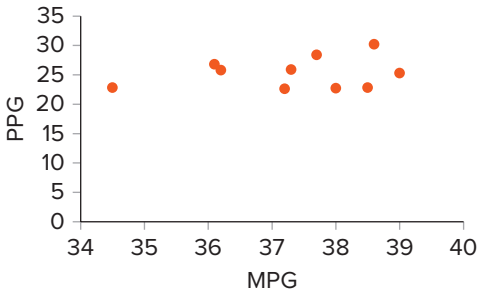
The majority (60%) of houses were either Ranch or Colonial.

b.

Classes	Frequency
300 up to 350	4
350 up to 400	6
400 up to 450	4
450 up to 500	2
500 up to 550	3
550 up to 600	1

The most frequent house price is in the $350,000 up to $400,000 range. The distribution is positively skewed.

2.66



The scatterplot reveals a mild positive relationship between PPG and MPG.

Chapter 3

3.2 Mean = −2.67; Median = −3.5; Mode = −4

3.4 Mean = 18.33; Median = 20; Two Modes: 15 and 20

3.6 Mean: "=AVERAGE(A1:A20)" = −0.25;
Median: "=MEDIAN(A1:A20)" = −1.50

3.8 a. Mean = 3.6; Median = 3.5; Mode = 3

b. Mode

3.12 a. 101.27

b. 107.42

3.14 a. Market capitalization: Mean = 164.10; Median = 167.50.

b. Total return: Mean = 40.71%; Median = 6.05%.

c. Either of the two for market capitalization; median for total return.

3.16 Mean = 3.90; Median = 3.87; Mode = 3.89

3.22 a. 25% (75%) of the observations are less than 54 (78).

b. IQR = 24; no outliers

c. Relatively symmetric

3.24 a. 25th Percentile = −0.19; 50th Percentile = 13.67; 75th Percentile = 17.46

b. IQR = 26.48; Lower limit = −26.67; Upper limit = 43.94; −35.97 is an outlier.

c. No, it is negatively skewed.

3.26 a. Approximately 25% of the employees earn less than $46,702 and 75% percent of the employees earn less than $116,288.

b. The value $221,086 is an outlier.

c. The distribution positively skewed.

3.28 a. No outliers; the distribution is positively skewed.

b. Two outliers; the distribution is positively skewed.

3.30 a. 18
b. 4.8
c. 36.80
d. 6.07

3.32 a. 22
b. 7.33
c. $s^2 = 81.2$; $s = 9.01$

3.34 MAD: "=AVEDEV(A1:A20)" = 5.625; Sample Variance: "=VAR.S(A1:A20)" = 41.1447 Sample Standard Deviation: "=STDEV.S(A1:A20)" = 6.4114

3.36 a. For Starbucks, $s^2 = 4.57$; $s = 2.14$.
For Panera Bread Co, $s^2 = 77.87$; $s = 8.82$.
b. Panera (8.82 against 2.14)
c. Panera (0.042 against 0.037)

3.38 a. $CV_{MKT} = \frac{74.01}{164.10} = 0.45$.

b. $CV_{Return} = \frac{105.07}{40.71} = 2.58$.

c. Total return

3.40 a. Investment B provides a higher return. Investment A provides a lower risk.

b. $\text{Sharpe}_A = \frac{10 - 1.4}{5} = 1.72$;

$\text{Sharpe}_B = \frac{15 - 1.4}{10} = 1.36$.

Investment A provides a higher reward per unit of risk.

3.42 a. Investment 2
b. Investment 1
c. $\text{Sharpe}_1 = \frac{3 - 1.2}{5.29} = 0.34$; $\text{Sharpe}_2 = \frac{5 - 1.2}{9} = 0.42$
Investment 2 performs better because it offers more reward per risk.

3.44 a. Latin America Fund
b. Latin America Fund
c. Sharpe (Latin America) = 0.26;
Sharpe (Canada) = 0.34;
Canada Fund has a higher reward per unit of risk.

3.46 a. At least 75%
b. At least 89%

3.48 a. 450 to 550
b. 425 to 575

3.50 a. 16%
b. 80

3.56 a. At least 75%
b. At least 89%

3.58 a. 68%
b. 2.5%
c. 16%

3.60 a. At least 75%
b. About 95%

3.62 a. No outliers
b. No outliers

3.64 a. $\bar{x} = 65.86$
b. $s^2 = 88.95$; $s = 9.43$

3.66 $\bar{x} = 3.36$; $s^2 = 3.87$; $s = 1.97$.

3.68 a. $\bar{x} = 7.69$
b. $s^2 = 3.35$; $s = 1.83$

3.72 a. −12.3.
b. −0.96; strong negative linear relationship

3.74 Sample Covariance: "=COVARIANCE.S(A2:A21, B2:B21)"= −6.6737; Correlation Coefficient: "=CORREL(A2:A21,B2:B21)"= −0.8420

3.76 a. 631.39; positive linear relationship
b. 0.45; moderate, positive linear relationship

3.78 a. 35; positive linear relationship
b. 0.95; strong positive linear relationship

3.82 Mean = 809.14; Median = 366; Mode = *not available*. The median best reflects the typical sales as the value 3,300 is clearly an outlier that pulls the mean up.

3.84 a. 1,817
b. $s^2 = 113{,}065$; $s = 336.25$

3.88 a. 136.62; positive linear relationship
b. 0.81; strong positive linear relationship

Chapter 4

4.2 a. 0.30
b. 0.55
c. 0

4.6 Empirical probability. It is considered accurate since it is based on a very large sample of 65,000 subscribers.

4.10 a. You may not get an offer from either firm.
b. You may get an offer from both firms.

4.18 a. No, because $P(A|B) \neq (A)$
b. No, because $P(A \cap B) \neq 0$.
c. $P((A \cup B)^C) = 0.185$

4.20 a. $P(B) = 0.60$
b. $P(A|B) = 0.133$
c. $P(B|A) = 0.32$

4.22 a. $P(A^C|B^C) = 0.48$
b. $P(A^C \cup B^C) = 0.86$
c. $P(A^C \cap B^C) = 0.76$

4.24 Let event O correspond to "students who ever go to their professor during office hours" and events MI and MA to "minor clarification" and "major clarification," respectively. We have $P(O) = 0.20$, $P(MI|O) = 0.3$, $P(MA|O) = 0.7$
a. $P(MI \cap O) = 0.06$
b. $P(MA \cap O) = 0.14$

4.26 Let event R correspond to "Reduction in unemployment in the US" and event E to "Recession in Europe." We have $P(R) = 0.18$, $P(R|E) = 0.06$
a. $P(R^C) = 0.82$
b. $P(R^C \cap E) = 0.0752$

4.28 Let F correspond to "Foreign student" and S to "Smoke." We have $P(F \cap S) = 0.05$ and $P(S|F) = 0.50$.
$P(F) = 0.10$, or 10%

4.30 a. $P(A) = 0.70$, $P(A^C) = 0.30$, $P(B) = 0.50$
b. No, because only 6 shirts are neither white nor blue.
c. $A \cap B$

4.34 For $i = 1,2$, let event A_i be "the i-th selected member is in favor of the bonus."

a. $P(A_1 \cap A_2) = \frac{10}{15} \times \frac{9}{14} = 0.4286$

b. $P(A_1^c \cap A_2^c) = \frac{5}{15} \times \frac{4}{14} = 0.0952$

4.38 Let event H correspond to "Woman faces sexual harassment" and event T to "Woman uses public transportation." We have $P(H) = 0.6667$, $P(H|T) = 0.82$, and $P(T) = 0.28$.

a. $P(H \cap T) = 0.2296$

b. $P(T|H) = 0.3444$

4.46 a. Joint probability table:

Global Warming	Political Affiliation		
	Democrat (*D*)	Republican (*R*)	Total
Yes (*Y*)	280	120	400
No (*N*)	120	280	400
Total	400	400	800

b. $P(Y|R) = 0.30$

c. $P(N) = 0.50$

d. $P(D|Y) = 0.70$

4.50 a. $P(B^C) = 0.40$

b. $P(A \cap B) = 0.48$
$P(A \cap B^C) = 0.04$

c. $P(A) = 0.52$

d. $P(B|A) = 0.9231$

4.56 Let event D correspond to "Experience a decline" and event N to "Ratio is negative." We have $P(D) = 0.20$, $P(N|D) = 0.70$, and $P(N|D^C) = 0.15$.
$P(D|N) = \frac{P(N \cap D)}{P(N)} = 0.54$.

4.58 Let event O correspond to "Teen owns a cell phone" and event T to "Older teens." We have $P(O|T) = 0.90$, $P(O|T^c) = 0.60$, and $P(T) = 0.70$.

a. $P(O) = 0.81$

b. $P(T|O) = 0.7778$

c. $P(T^c|O) = 0.2222$

4.60 Let F = "Player is fully fit to play," S = "Player is somewhat fit to play," N = "Player is not able to play," and W = "The Lakers win the game."

a. $P(W) = P(W \cap F) + P(W \cap S) + P(W \cap N) = 0.62$

b. $P(F|W) = 0.52$

4.62 a. Empirical Probabilities: $P(A) = 0.10$, $P(B) = 0.44$, and $P(B|A) = 0.60$

b. Not mutually exclusive since $P(A \cap B) > 0$
Not exhaustive since $P(A \cup B) < 1$

c. Not independent since $P(B|A) \neq P(B)$

d. $P(A|B) = 0.136$

4.64 Let event A correspond to "own a mobile phone" and B to "own a smartphone." We have $P(A) = 0.883$ and $P(B|A) = 0.84$. $P(B) = P(B \cap A) = 0.742$, or 74.2%.

4.68 Let event S correspond to "Biggest smilers," F to "Biggest frowners," and D to "Divorced." We have $P(D|S) = 0.11$ and $P(D|F) = 0.31$.

a. $P(S) = 0.1818$

b. $P(F \cap D) = 0.0775$

4.70 a. $\frac{17}{20} = 0.85$

b. $\frac{17}{20} \times \frac{16}{19} = 0.7158$

c. $\frac{3}{20} \times \frac{2}{19} = 0.0158$

4.72 Let event O correspond to "Optimism about the global economy," U to "Respondents from the U.S.," and A to "Respondents from Asia." We have $P(O) = 0.18$, $P(O|U) = 0.22$, and $P(O|A) = 0.09$.

a. $P(O^c|A) = 0.91$.

b. $P(O \cap U) = 0.0616$

c. $P(A|O) = 0.11$

4.76

	Survived for Discharge (*S*)	Did not Survive for Discharge (S^c)	Total
Day or Evening (*D*)	0.1338	0.5417	0.6755
Graveyard Shift (*G*)	0.0477	0.2768	0.3245
Total	0.1815	0.8185	1.00

a. $P(G) = 0.3245$

b. $P(S) = 0.1815$

c. $P(S|G) = 0.1470$

d. $P(G|S) = 0.2628$

e. No since $P(S|G) \neq P(S)$

4.78 Let event A correspond to "US economy performs well" and B to "Asian countries perform well." We have $P(A) = 0.40$, $P(B|A) = 0.80$, and $P(B|A^C) = 0.30$.

a. $P(A \cap B) = 0.32$

b. $P(B) = P(B \cap A) + P(B \cap A^C) = 0.50$

c. $P(A|B) = 0.64$

4.80 Let event M correspond to "Men," W to "Women," and H to "Healthy weight." We have $P(H|W) = 0.365$, $P(H|M) = 0.266$, and $P(W) = 0.5052$.

a. $P(H) = P(H \cap W) + P(H \cap M) = 0.3160$

b. $P(W|H) = 0.5835$

c. $P(M|H) = 0.4165$

4.82 Let event R correspond to "Republican," D to "Democrat," I to "Independent," and S to "Support marijuana legalization." We have $P(R) = 0.27$, $P(D) = 0.30$, $P(I) = 0.43$, $P(S|R) = 0.41$, $P(S|D) = 0.66$, $P(S|I) = 0.63$.

a. $P(S \cap R) = 0.1107$

b. $P(S \cap D) = 0.1980$

c. $P(S \cap I) = 0.2709$

d. $P(S) = P(S \cap R) + P(S \cap D) + P(S \cap I) = 0.5796$

e. $P(R|S) = 0.1910$

Chapter 5

5.4 a. $P(X \leq 0) = 0.5$

b. $P(X = 50) = 0.25$

c. Yes

5.8 Let X represent performance.

a. The analyst has a somewhat pessimistic view. There is a 57% chance that the performance will be poor or very poor.

b.

x	$P(X \leq x)$
1 (Very poor)	0.14
2 (Poor)	0.57
3 (Neutral)	0.79
4 (Good)	0.95
5 (Very good)	1

c. $P(X \geq 4) = 0.21$.

5.10 Let X represent confidence score.
a. $P(X = 2) = 0.20$.
b. $P(2 \le X \le 3) = 0.25$.
5.14 $\mu = 10.75$
$\sigma^2 = 28.19$ $\sigma = 5.31$
5.16 a. $\mu = 0.95$
b. $\sigma = 0.80$
5.20 2.2
5.22 $3,600
5.28 a. $P(X = 0) = 0.1160$
b. $P(X = 1) = 0.3124$
c. $P(X \le 1) = 0.4284$
5.30 a. $P(3 < X < 5) = 0.1569$
b. $P(3 < X \le 5) = 0.2160$
c. $P(3 \le X \le 5) = 0.4828$
5.32 a. "= BINOMDIST(50,150,0.36,1)" = 0.2776
b. "= BINOMDIST(40,150,0.36,0)" = 0.0038
c. "= 1 – BINOMDIST(60,150,0.36,1)" = 0.1348
d. "= 1 – BINOMDIST(54,150,0.36,1)" = 0.4630
5.36 a. $P(X < 2) = 0.7213$
b. $P(X < 2) = 0.4580$
5.38 a. $E(X) = 2{,}850$; $SD(X) = 35.01$
b. $E(X) = 2{,}150$; $SD(X) = 35.01$
c. $P(X = 6) = 0.1780$
5.40 a. $P(X > 2) = 0.3125$
b. $P(X > 2) = 0.5276$
c. $P(X > 2) = 0.1362$
5.42 a. $P(X \ge 1) = 0.4375$; her statement is not correct
b. $P(X \ge 1) = 0.5781$; her statement is correct
5.44 a. $P(X = 10) = 0.1171$
b. $P(X \le 10) = 0.8725$
c. $P(X \ge 15) = 0.0016$
5.46 a. $P(X = 1) = 0.3347$
b. $P(X = 2) = 0.2510$
c. $P(X \ge 2) = 0.4422$
5.50 a. "= POISSON.DIST(13,20,1)" = 0.0661
b. "= 1 – POISSON.DIST(19,20,1)" = 0.5297
c. "= POISSON.DIST(25,20,0)" = 0.0446
d. "= POISSON.DIST(23,20,1) – POISSON.DIST(17,20,1)" = 0.4905
5.52 a. Poisson
b. Not Poisson
c. Poisson
d. Not Poisson
5.54 a. $\mu = 6$; $P(X = 2) = 0.0446$
b. $\mu = 6$; $P(X \ge 2) = 0.9826$
c. $\mu = 60$; $P(X = 40) = 0.001$
5.58 a. $P(X \le 425) = 0.8980$
b. $P(X \ge 375) = 0.8998$
5.60 a. $\mu = 304$; $P(X > 320) = 0.1717$
b. $\mu = 2{,}128$; $P(X > 2{,}200) = 0.0586$
5.62 a. $P(X = 0) = 0.5783$
b. $P(X = 1) = 0.3652$
c. $P(X \le 1) = 0.9435$
5.66 $P(X \ge 8) = 0.0777$
$E(X) = 5$; $SD(X) = 1.7408$
5.70 a. $P(X = 3) = 0.2696$
b. $P(X \ge 2) = 0.7549$
5.72 $P(X = 2) = 0.0316$
5.74 a. $P(X = 2) = 0.0495$
b. $P(X = 5) = 0.0000002$
c. $P(X = 1) = 0.0256$
d. $0.0000002 \times 0.0256 = 0.00000000512$
5.76 a. $E(X) = -\$700{,}000$
5.78 a. $E(X) = 2.79$; $SD(X) = 1.3137$
b. $120 \times 41.85 = 5{,}022$
5.82 a. $P(X = 2) = 0.3747$
b. $P(X = 4) = 0.0677$
c. $E(X) = 51$; $SD(X) = 4.999$
5.84 a. $P(X = 10) = 0.0272$
b. $P(10 \le X \le 20) = 0.0451$
c. $P(X \le 8) = 0.8996$
5.88 a. $P(X = 3) = 0.0129$
b. $P(X \le 2) = 0.9871$
c. $2P(X = 3) = 0.0258$
5.92 a. $P(X = 6) = 0.0115$
b. $P(X \ge 5) = 0.0647$
c. $P(X \le 2) = 0.5206$
d. $E(X) = 2.5$

Chapter 6

6.2 a. 0.30
b. 0.16
c. 0.70
6.4 a. $f(x) = 0.0333$
b. $\mu = 20$; $\sigma = 8.66$
c. $P(X > 10) = 0.8325$
6.6 a. $\mu = 20$; $\sigma = 5.77$
b. $P(X > 22) = 0.4$
c. $P(15 \le X \le 23) = 0.4$
6.8 a. $\mu = 16$
b. $P(X < 15.5) = 0.4375$
c. $P(X > 14) = 0.75$
6.12 a. $P(X \ge 25) = 0.4167$
b. $P(X \le 20) = 0.1667$
6.14 a. $P(Z > 1.32) = 0.0934$
b. $P(Z \le -1.32) = 0.0934$
c. $P(1.32 \le Z \le 2.37) = 0.0845$
d. $P(-1.32 \le Z \le 2.37) = 0.8977$
6.20 a. $P(X \le 0) = P(Z \le -2.5) = 0.0062$
b. $P(X > 2) = P(Z > -2) = 0.9772$
c. $P(4 \le X \le 10) = P(-1.5 \le Z \le 0) = 0.4332$
d. $P(6 \le X \le 14) = P(-1 \le Z \le 1) = 0.6826$
6.26 a. "= 1 – NORM.DIST(–12,–15,9,1)" = 0.3694
b. "= NORM.DIST(5,–15,9,1) – NORM.DIST(0,–15,9,1)" = 0.0347
c. "= NORM.INV(0.25,–15,9)" = –21.0704
d. "= NORM.INV(0.75,–15,9)" = –8.9296

6.28 a. $P(84 < X < 116) = P(-1 < Z < 1) = 0.6826$
b. $P(X < 68) = P(Z < -2) = 0.0228$
c. Given $P(X > x) = 0.01$, $x = 137.22$

6.30 a. $P(60 < X < 100) = P(-2 < Z < 2) = 0.9544$
b. $P(X > 100) = P(Z > 2) = 0.0228$; 1.87 games

6.34 a. $P(X \geq 40) = P(Z \geq 1.77) = 0.0384$
b. $P(30 \leq X \leq 35) = P(-1.09 \leq Z \leq 0.34) = 0.4952$
c. Given $P(X \leq x) = 0.99$, $x = 41.94$

6.36 No, since $P(X > 16) = P(Z > 0.67) = 0.2514 \neq 0.15$

6.38 a. $P(X > 19) = P(Z > -1.5) = 0.9332$
b. $P(X > 19) = P(Z > 1.5) = 0.0668$
c. $P(23 \leq X \leq 25) = P(0.5 \leq Z \leq 1.5) = 0.2417$
d. $P(23 \leq X \leq 25) = P(3.5 \leq Z \leq 4.5) = 0.0002$

6.44 a. $P(50 \leq X \leq 80) = P(-0.5 \leq Z \leq 1) = 0.5328$
b. $P(20 \leq X \leq 40) = P(-2 \leq Z \leq -1) = 0.1359$
c. Given $P(X \geq x) = 0.15$, $x = 80.72$
d. Given $P(X < x) = 0.10$, $x = 34.36$

6.48 a. Riskier Fund: $P(X < 0) = P(Z < -0.57) = 0.2843$
Less Risky Fund: $P(X < 0) = P(Z < -0.8) = 0.2119$
Pick the less risky fund.
b. Riskier Fund: $P(X > 8) = P(Z > 0) = 0.5$
Less Risky Fund: $P(X > 8) = P(Z > 0.8) = 0.2119$
Pick the riskier fund.

6.50 a. $P(X < 200) = P(Z < -0.58) = 0.2810$
b. $P(X \geq 266.5) = P(Z \geq 2.74) = 0.0031$
c. Given $P(X < x) = 0.10$, $x = 186.06$
d. Given $P(X < x) = 0.99$, $x = 258.22$

6.54 a. $\mu = 2.5$
b. $\lambda = 0.4$
c. $P(1 \leq X \leq 2) = 0.2210$

6.56 $E(X) = SD(X) = 0.20$

6.58 a. $P(X < 2.3) = 0.8412$
b. $P(1.5 \leq X \leq 5.5) = 0.2889$
c. $P(X > 7) = 0.0037$

6.60 a. $\mu = 6$
b. $P(X > 15) = 0.0820$
c. $P(15 \leq X \leq 20) = 0.0463$

6.62 a. $\mu = 0.0028$ (in hours)
b. $P\left(X < \frac{10}{3600}\right) = 0.6321$

6.64 a. $P(X \leq 24) = 0.0469$
b. $(P(X \leq 24))^2 = 0.0022$

6.68 a. $E(X) = 3$; $Var(X) = 1.3333$
b. $P(X > 4) = 0.25$
c. $P(X < 2.5) = 0.375$

6.70 a. $P(80 \leq X \leq 90) = P(0.1 \leq Z \leq 1.1) = 0.3245$
b. $P(120 \leq X \leq 139) = P(-0.29 \leq Z \leq 0.82) = 0.4080$

6.74 Given $P(X > x) = 0.03$, $x = 100.22$

6.76 $Q_1 = 5.99$; $Q_2 = 6.00$; $Q_3 = 6.01$

6.78 a. $P(X \leq 10) = P(Z \leq -1.2) = 0.1151$
b. $1{,}000 \times 271.225 = \$271{,}225$

6.82 a. $\mu = 7.5$
b. $P(X \leq 5) = 0.4865$

6.84 a. $\mu = 0.365$
b. $\lambda = 2.7397$
c. $P(X \leq 1) = 0.9354$

6.86 a. $\mu = 0.8333$
b. $P(X \leq 1) = 0.6988$
c. $P(1 \leq X \leq 2) = 0.2105$

Chapter 7

7.2 Nonresponse bias if some people are less likely to stop at the booth. Selection bias since the booth is only open on the weekend.

7.4 a. Nonresponse bias if the people who respond are systematically different from those who do not respond.
b. Selection bias since those who frequent the store in the morning are likely to prefer an earlier opening time.
c. Selection bias since not everyone reads a newspaper. Nonresponse bias if the people who respond are systematically different from those who do not respond.

7.8 a. $E(\overline{X}) = 80$; $SE(\overline{X}) = 1.4$
b. $P(77 \leq \overline{X} \leq 85) = P(-2.14 \leq Z \leq 3.57) = 0.9836$
c. $P(\overline{X} > 84) = P(Z > 2.86) = 0.0021$

7.12 a. $P(\overline{X} \geq 18) = P(Z \geq 1.85) = 0.0322$
b. $P(\overline{X} \geq 17.5) = P(Z \geq 2.03) = 0.0212$
c. Janice; her findings are more likely if a representative sample is used.

7.14 a. The sample mean has a normal distribution because the population is normally distributed.
b. $P(\overline{X} > 25) = P(Z > 2.4) = 0.0082$
c. $P(18 \leq \overline{X} \leq 24) = P(-3.20 \leq Z \leq 1.60) = 0.9445$

7.18 a. $P(X > 1{,}000{,}000) = P(Z > 0.80) = 0.2119$
b. $P(\overline{X} > 1{,}000{,}000) = P(Z > 1.60) = 0.0548$

7.20 a. $P(X < 90) = P(Z < -0.63) = 0.2643$
b. $P(\overline{X} < 90) = P(Z < -1.25) = 0.1056$
c. $(0.2643)^4 = 0.0049$

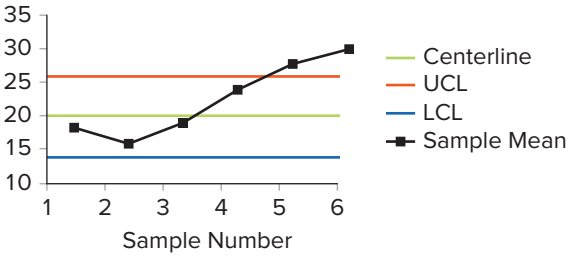
7.24 a. $P(\overline{P} < 0.30) = P(Z < 1.29) = 0.9015$
b. $p = 0.74$; $P(\overline{P} > 0.75) = P(Z > 0.32) = 0.3745$

7.26 a. $E(\overline{P}) = 0.17$ and $se(\overline{P}) = 0.0266$; the normal approximation criteria are met because $np = 34 > 5$ and $n(1 - p) = 166 > 5$.
b. $P(\overline{P} > 0.20) = P(Z > 1.13) = 0.1292$

7.34 $n = 120$, $N = 1{,}000$; apply the finite population correction; $E(\overline{P}) = 0.6667$ and $se(\overline{P}) = 0.0404$; $P(\overline{P} > 0.625) = P(Z > -1.03) = 0.8485$

7.36 a. No, since n is less than 5% of N.
b. No because we do not know if the population has a normal distribution and $n < 30$.
c. $E(\overline{X}) = 10.32$; $se(\overline{X}) = 2.8232$
d. The normal approximation is not justified (see part b).

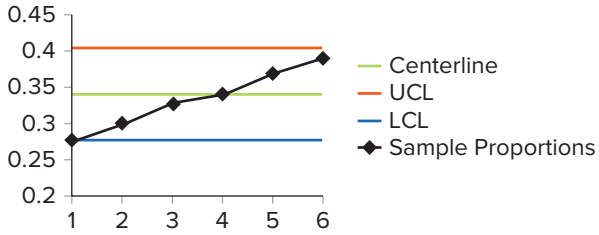
7.40 a. Centerline: $\mu = 20$
UCL = 26
LCL = 14
b.



c. The last two points are outside the upper control limit. There is also an upward trend, suggesting the process is becoming increasingly out of control. The process should be adjusted.

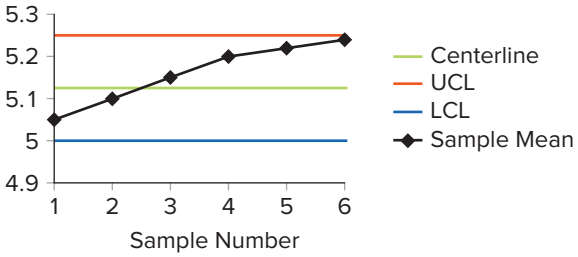
7.42 a. Centerline: $p = 0.34$
UCL = 0.404
LCL = 0.276

b.



c. Although there are no points outside the control limits, the positive trend suggests that the process may become out of control if the upward trend continues.

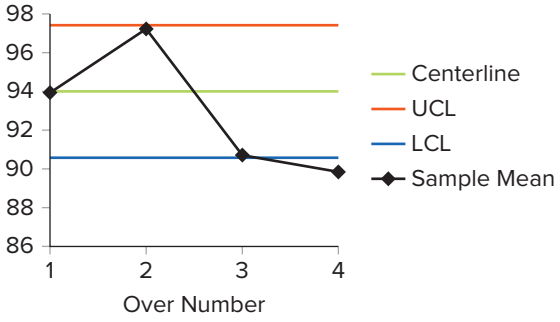
7.44 a.



Centerline: $\mu = 5.125$
UCL = 5.25
LCL = 5

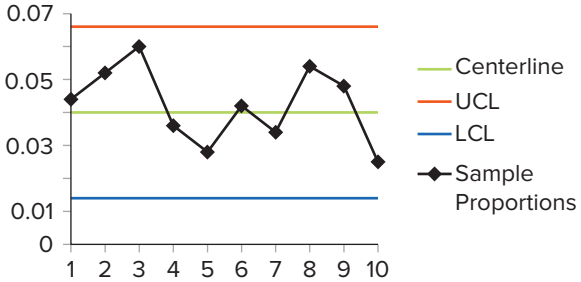
b. There are no points outside the control limits. It appears that the process is under control, but the positive trend suggests the process may become out of control if the trend continues.

7.46 a.



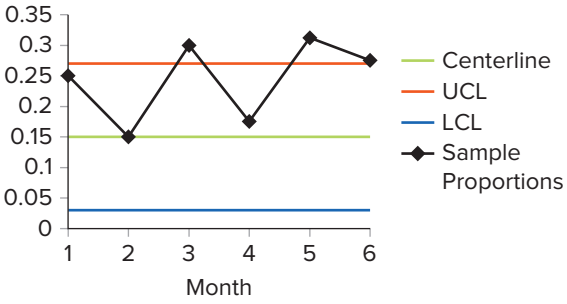
b. Kalwant's average speed is out of control; the coach's concern is justified.

7.48 a.



b. Yes since all sample proportions are within the control limits and there is no apparent trend.

7.50 a.



b. 3 out of 6 months were out of the control limits, which is a good justification for why Dell chose to direct customers away from India call centers.

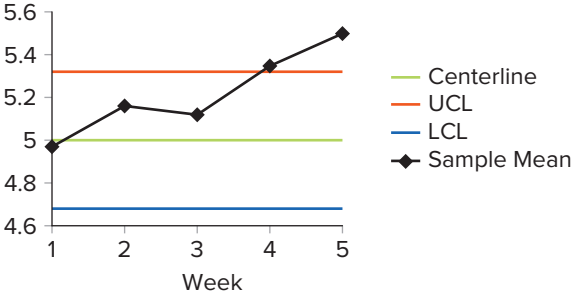
7.54 a. $P(\overline{P} > 0.50) = P(Z > 0.71) = 0.2389$
b. $P(\overline{P} > 0.50) = P(Z > 0.28) = 0.3897$

7.56 a. $P(X < 79) = P(Z < -0.5) = 0.3085$
b. $P(\overline{X} < 79) = P(Z < -1.58) = 0.0571$
c. $P(\overline{X} < 79) = P(Z < -2.74) = 0.0031$

7.58 a. $P(X > 500) = P(Z > 0.59) = 0.2776$
b. $P(\overline{X} > 500) = P(Z > 1.17) = 0.1210$
c. $(0.2776)^4 = 0.0059$

7.60 a. $P(\overline{P} > 0.80) = P(Z > 0.87) = 0.1922$
b. $P(\overline{P} < 0.70) = P(Z < -2.04) = 0.0207$

7.62 a.



b. The last two points are outside the upper control limit, and there is a positive trend, suggesting that the process is out of control.

Chapter 8

8.2 a. For 89%, $z_{\alpha/2} = 1.598$
b. For 92%, $z_{\alpha/2} = 1.751$
c. For 96%, $z_{\alpha/2} = 2.054$

8.6 Lower limit: "=AVERAGE(A1:A20) − NORM.S.INV(0.975)*12/SQRT(20)" = 117.6909; Upper limit: "=AVERAGE(A1:A20)+NORM.S.INV(0.975)*12/SQRT(20)" = 128.2091

8.8 a. $\overline{x} = 78.1$
b. $1.645\frac{4.5}{\sqrt{50}} = 1.05$
c. 78.1 ± 1.05 or [77.05, 79.15]

8.12 a. $2.576\frac{500}{\sqrt{100}} = 128.80$
b. 7,790 ± 128.80 or [7,661.20, 7,918.80]

8.16 90%: [18.81, 21.61]
99%: [18.02, 22.40]
The 99% confidence interval is wider.

8.24 Lower limit: "=AVERAGE(A1:A20) −T.INV(0.995,19) *STDEV.S(A1:A20)/SQRT(20)" = −4.3535;
Upper limit: "=AVERAGE(A1:A20) +T.INV(0.995,19) *STDEV.S(A1:A20)/SQRT(20)" = 3.8535

8.26 a. $2.11\frac{9.2}{\sqrt{18}} = 4.58$
b. 12.5 ± 4.58 or [7.92, 17.08]

8.28 a. $2.724\frac{10}{\sqrt{36}} = 4.54$
b. 100 ± 4.54 or [95.46, 104.54]

8.30 a. $17.25 \pm 3.499\ \frac{5.95}{\sqrt{8}}$ or [9.89, 24.61].
b. The population is normally distributed.

8.36 a. Electronic: $18 \pm 4.604\frac{20.70}{\sqrt{5}}$ or [−24.62, 60.62]
Utilities: $14.8 \pm 4.604\frac{6.50}{\sqrt{5}}$ or [1.42, 28.18]
b. Annual return of each fund has a normal distribution

8.38 $1{,}080 \pm 2.032\frac{260}{\sqrt{35}}$ or [990.70, 1,169.30]
The manager is wrong since the value 1,200 is not within the 95% confidence interval.

8.40 a. Microeconomics: [68.74, 74.91]
Macroeconomics: [66.16, 74.64]
b. The widths are different because of the differences in the sample standard deviations for microeconomics and macroeconomics.

8.44 a. $0.6 \pm 1.960\sqrt{\frac{0.6(1-0.6)}{50}}$ or [0.464, 0.736]
b. $0.6 \pm 1.960\sqrt{\frac{0.6(1-0.6)}{200}}$ or [0.532, 0.668].
With larger n, the interval is narrower.

8.46 a. $\bar{p} = 0.40$.
b. 90%: $0.40 \pm 1.645\sqrt{\frac{0.40(1-0.40)}{100}}$ or [0.319, 0.481]
99%: $0.40 \pm 2.576\sqrt{\frac{0.40(1-0.40)}{100}}$ or [0.274, 0.526]
c. Yes, since the value 0.5 does not fall within the interval.
d. No, since the value 0.5 falls within the interval.

8.48 $0.51 \pm 1.960\sqrt{\frac{0.51(1-0.51)}{1{,}079}}$ or [0.480, 0.540].

8.50 a. $0.37 \pm 1.645\sqrt{\frac{0.37(1-0.37)}{5{,}324}}$ or [0.359, 0.381]
b. $0.37 \pm 2.576\sqrt{\frac{0.37(1-0.37)}{5{,}324}}$ or [0.353, 0.387]
c. The margin of error in part b is greater because it uses a higher confidence level.

8.54 a. $0.275 \pm 1.645\sqrt{\frac{0.275(1-0.275)}{400}}$ or [0.238, 0.312]
b. No, because the value 0.30 falls in the interval.

8.56 a. $0.30 \pm 1.645\sqrt{\frac{0.30(1-0.30)}{1{,}026}}$ or [0.276, 0.324]
b. $0.10 \pm 1.645\sqrt{\frac{0.10(1-0.10)}{1{,}026}}$ or [0.085, 0.115]

8.60 $n = 61.47$, rounded up to 62

8.62 With $E = 1.2$, $n = 23.02$, rounded up to 24.
With $E = 0.7$, $n = 67.65$, rounded up to 68.

8.64 With $E = 0.08$, $n = 138.3$, rounded up to 139.
With $E = 0.12$, $n = 61.47$, rounded up to 62.

8.68 a. $n = 101.89$, rounded up to 102.
b. $n = 39.34$, rounded up to 40.
c. With higher standard deviation, Fund A requires a larger sample size to achieve the same margin of error.

8.72 $n = 1{,}680.44$, rounded up to 1,681.

8.76 $10 \pm 2.262\frac{15}{\sqrt{10}}$ or [−0.73, 20.73]

8.78 a. $16 \pm 1.971\frac{12}{\sqrt{225}}$ or [14.42, 17.58]
b. Yes, because the interval does not include the value 14.

8.86 a. [77.04, 79.81]
b. It differs since the value 81.84 is not contained in the interval.

8.88 a. 7.89 (Monday), 4.41 (Tuesday); The margin of error for Monday is greater because of its higher sample standard deviation.
b. [214.48, 230.25] (Monday), [185.65, 194.48] (Tuesday)
c. For both Monday and Tuesday, the population mean differs from 200 because the value 200 does not belong to either of the two confidence intervals.

8.90 a. 0.018
b. 0.121 ± 0.018 or [0.103, 0.139].

8.96 $n = 259.35$, rounded up to 260

Chapter 9

9.4 a. Incorrect; we never accept the null hypothesis.
b. Correct
c. Incorrect; we establish a claim only if the null hypothesis is rejected.
d. Correct

9.10 a. Type I error; the new software is purchased even though it does not reduce assembly costs.
b. Type II error; the new software is not purchased even though it reduces assembly costs.

9.12 a. Type I error; the restaurant is incorrectly implicated for using higher fat content.
b. Type II error; the restaurant escapes being implicated for using higher fat content.

9.14 a. $z = -2$
b. p-value = 0.0456
c. Reject H_0; at the 10% significance level, we conclude that the population mean differs from 100.

9.16 a. H_0: $\mu \leq 45$; H_A: $\mu > 45$
b. $z = 1.50$
c. p-value = 0.0668
d. Do not reject H_0; at the 5% significance level, we cannot conclude that the population mean is greater than 45.

9.18 $z = 1.41$; p-value = 0.0793. Do not reject H_0; at the 5% significance level, we cannot conclude that the population mean is greater than −5.

9.20 $z = -3.57$; p-value = 0.0004. Reject H_0; at the 1% significance level, we conclude that the population mean differs from −100.

9.22 H_0: $\mu \geq 125$; H_A: $\mu < 125$; z: "=(AVERAGE(A1:A20) −125)/(12/SQRT(20))" = −0.7640; p-value: "=NORM.S.DIST(−0.7640,1)" = 0.2224; do not reject H_0; at the 5% significance level, we cannot conclude that the population mean is less than 125.

9.26 a. H_0: $\mu \leq 90$; H_A: $\mu > 90$
b. $z = 1.58$; p-value = 0.0571
c. Do not reject H_0; the manager's claim is not supported at the 1% significance level.

9.30 a. H_0: $\mu = 30$; H_A: $\mu \neq 30$
b. $z = 2.40$; p-value = 0.0164
c. Reject H_0; at the 5% significance level, we conclude that the average weekly price differs from $30.

9.38 H_0: $\mu = 16$; H_A: $\mu \neq 16$; $t_{31} = -7.54$; p-value = 0 (approximately). Reject H_0; at the 1% significance level, we conclude that the population mean differs from 16.

9.44 H_0: $\mu = 1$; H_A: $\mu \neq 1$; t_{19}: "=(AVERAGE(A1:A20) −1)/((STDEV.S(A1:A20)/SQRT(20))" = −0.8715; p-value: "=2*(T.DIST.RT(0.8715,19))" = 0.3944; do not reject H_0; at the 5% significance level, we cannot conclude that the population mean differs from 1.

9.46 a. H_0: $\mu \leq 5$; H_A: $\mu > 5$

b. $t_6 = 0.643$; p-value = 0.272; population normally distributed

c. Do not reject H_0; at the 10% significance level, we cannot conclude that the average waiting time is more than 5 minutes.

9.48 a. H_0: $\mu = 12$; H_A: $\mu \neq 12$

b. No, since $n > 30$

c. $t_{47} = -1.732$; p-value = 0.09

d. Do not reject H_0; at the 5% significance level, we cannot conclude that the bottling process has fallen out of adjustment.

9.50 H_0: $\mu \leq 7$; H_A: $\mu > 7$; $t_{33} = 2.915$; p-value = 0.003; reject H_0. At the 1% significance level, we conclude that the mean drop of home prices in San Diego is greater than the 7% drop in Los Angeles.

9.54 a. H_0: $\mu = 95$; H_A: $\mu \neq 95$

b. $t_{24} = 0.71$; p-value = 0.484

c. Do not reject H_0; at the 5% significance level, we cannot conclude that the average MPG differs from 95.

9.60 a. $z = -0.30$; p-value = 0.7642

b. $z = 2.05$; p-value = 0.0404

c. $z = 1.08$; p-value = 0.2802

d. $z = 1.73$; p-value = 0.0836

9.68 a. H_0: $p \leq 0.20$; H_A: $p > 0.20$

b. $z = 2.18$; p-value = 0.0146

c. Reject H_0; at the 5% significance level, the economist's concern is supported.

9.70 H_0: $p \leq 0.75$; H_A: $p > 0.75$; $z = 1.01$; p-value = 0.1562. Do not reject H_0; at the 5% significance level, we cannot conclude that more than 75% of financial institutions are prone to fraud.

9.82. H_0: $p \geq 0.35$; H_A: $p < 0.35$

Case 1: $z = -1.33$; p-value = 0.0918. Do not reject H_0; at the 5% significance level, we cannot conclude that the percentage of Americans who feel that the country is headed in the right direction is below 35%.

Case 2: $z = -1.88$; p-value = 0.0301. Reject H_0; at the 5% significance level, we conclude that the percentage of Americans who feel that the country is headed in the right direction is below 35%.

9.86 a. H_0: $p = 0.17$; H_A: $p \neq 0.17$

b. $z = 2.07$; p-value = 0.0384

c. Reject H_0; at the 5% significance level, we conclude that the proportion of households in the rural South is not representative of the national proportion.

9.88 a. H_0: $\mu = 13{,}500$; H_A: $\mu \neq 13{,}500$

b. $t_{49} = 2.593$; p-value = 0.012

c. Reject H_0; at the 10% significance level, we conclude that the average number of miles driven by Midwesterners differs from the U.S. average.

Chapter 10

10.4 a. $t_{20} = 1.719$; p-value = 0.051. Do not reject H_0; at the 5% significance level, we cannot conclude that μ_1 is greater than μ_2.

b. At the 10% significance level, we conclude that μ_1 is greater than μ_2.

10.8 a. H_0: $\mu_1 - \mu_2 = 0$; H_A: $\mu_1 - \mu_2 \neq 0$

b. $t_8 = \dfrac{98.3333 - 111.6667}{\sqrt{\frac{16.2686^2}{6} + \frac{10.9118^2}{6}}} = -1.667$;

p-value = 0.134

c. Do not reject H_0; at the 10% significance level, we cannot conclude that the population means differ.

10.10 a. H_0: $\mu_1 - \mu_2 \geq 0$; H_A: $\mu_1 - \mu_2 < 0$

b. $z = -5.81$; p-value = 0 (approximately)

c. Reject H_0; at the 5% significance level, we conclude that there is a "community college penalty" at Lucille's university.

10.12 a. H_0: $\mu_1 - \mu_2 = 0$; H_A: $\mu_1 - \mu_2 \neq 0$

b. $z = 1.53$; p-value = 0.126

c. Do not reject H_0; at the 5% or the 10% significance levels, we cannot conclude the mean profitability differs between condominiums and apartment buildings.

10.14 a. H_0: $\mu_1 - \mu_2 \leq 0$; H_A: $\mu_1 - \mu_2 > 0$ (Population 1 = New Process and Population 2 = Old Process)

b. $t_{16} = \dfrac{(2613.63 - 2485.10) - 0}{\sqrt{15{,}963.10\left(\frac{1}{8} + \frac{1}{10}\right)}} = 2.145$;

p-value = 0.024

c. Reject H_0; at the 5% significance level, we conclude that the mean output rate of the new process exceeds that of the old process.

d. Do not reject H_0; at the 1% significance level, we cannot conclude that the mean output rate of the new process exceeds that of the old process.

10.18 a. H_0: $\mu_1 - \mu_2 = 0$; H_A: $\mu_1 - \mu_2 \neq 0$ (Population 1 = Day Searches and Population 2 = Evening Searches)

b. $t_{16} = \frac{(98{,}817.30 - 110{,}204.17) - 0}{1{,}570.7702} = -7.249$;

p-value = 0 (approximately)

c. Reject H_0; at the 5% significance level, we conclude that the mean number of website searches differs between the day and evening advertisements.

10.20 a. H_0: $\mu_1 - \mu_2 \geq 0$; H_A: $\mu_1 - \mu_2 < 0$ (Population 1 = New Method and Population 2 = Old Method)

b. $t_{66} = \frac{(32.0938 - 32.7500) - 0}{0.4571} = -1.436$;

p-value = 0.078

c. We conclude that the mean assembly time using the new method is less than the old method only at the 10% significance level.

10.28 a. H_0: $\mu_D \leq 0$; H_A: $\mu_D > 0$

b. $t_{34} = 1.868$; p-value = 0.035

c. Reject H_0; at the 5% significance level, we conclude that there is a positive mean difference.

10.30 a. H_0: $\mu_D = 0$; H_A: $\mu_D \neq 0$ (Mean difference between Method A and Method B)

b. $t_6 = \frac{-1.8571 - 0}{2.3401/\sqrt{7}} = -2.10$

c. p-value = 0.08

d. Reject H_0; at the 10% significance level, the manager's assertion is supported by the data.

10.34 a. H_0: $\mu_D \geq 0$; H_A: $\mu_D < 0$ (Mean time difference between New Processor and Existing Processor)

b. $t_6 = \frac{-0.2771 - 0}{0.1991/\sqrt{7}} = -3.682$; p-value = 0.005

c. Reject H_0; at the 5% significance level, we conclude that the mean difference between the new and the existing processing time is less than zero. Yes, there is evidence the new processor is faster than the old processor.

10.36 a. H_0: $\mu_D \leq 100$; H_A: $\mu_D > 100$ (Mean difference between the competitor's and Insure-Me premiums)

b. $t_{49} = \frac{197.06 - 100}{443.9387/\sqrt{50}} = 1.546$; p-value = 0.064

c. We conclude, only at the 10% significance level, that the mean premium difference between the competitor and Insure-Me is more than $100.

10.46 a. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

H_A: Not all population means are equal.

c. $F_{(5,54)} = 1.371$; p-value = 0.250; do not reject H_0. At the 5% significance level, we cannot conclude that not all population means are equal.

10.48 a. H_0: $\mu_1 = \mu_2 = \mu_3$

H_A: Not all population means are equal.

b. $F_{(2,21)} = 2.804$; p-value = 0.083; do not reject H_0. At the 5% significance level, we cannot conclude that there are differences in the effectiveness of the three detergents.

10.54 a. H_0: $\mu_{Low} = \mu_{Medium} = \mu_{High}$

H_A: Not all population means are equal.

b. $F_{(2,117)} = 10.591$; p-value = 0 (approximately). At the 1% and 5% significance levels, we conclude that the mean fill volumes are not equal.

10.58 H_0: $\mu_1 = \mu_2 = \mu_3$

H_A: Not all population means are equal.

$F_{(2,87)} = 11.479$; p-value = 0 (approximately); reject H_0. At the 10% significance level, we conclude that the average job satisfaction differs by field.

10.60 a. H_0: $\mu_1 - \mu_2 \leq 0$; H_A: $\mu_1 - \mu_2 > 0$; (Population 1 = Men and Population 2 = Women)

b. $z = 3.53$

c. p-value = 0.0002

d. Reject H_0; at the 1% significance level, we conclude that, on average, men spend more money than women on St. Patrick's Day.

10.62 H_0: $\mu_1 - \mu_2 = 0$; H_A: $\mu_1 - \mu_2 \neq 0$; (Population 1 = Men and Population 2 = Women)

$t_{28} = \frac{(194.48 - 188.88) - 0}{1,7045} = 3.285$; p-value = 0.002

At the 1% significance level, we conclude that the mean cholesterol levels for men and women are different.

10.64 a. H_0: $\mu_D \leq 30$; H_A: $\mu_D > 30$ (Mean difference between after and before pregnancy weight); $t_{39} = \frac{36 - 30}{9.6503/\sqrt{40}} = 3.932$; p-value = 0 (approximately); reject H_0. At the 5% significance level, we conclude that the mean weight gain of women due to pregnancy is more than 30 pounds.

b. H_0: $\mu_D \leq 35$; H_A: $\mu_D > 35$ (Mean difference between after and before pregnancy weight) $t_{39} = \frac{36 - 35}{9.6503/\sqrt{40}} = 0.655$; p-value = 0.258; do not reject H_0. At the 5% significance level, we cannot conclude that the mean weight gain of women due to pregnancy is more than 35 pounds.

10.68 a. H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_A: Not all population means are equal.

b. p-value = 0 (approximately); reject H_0. At the 5% significance level, we conclude that the average salaries of the four different transportation operators differ.

10.72 H_0: $\mu_1 = \mu_2 = \mu_3$

H_A: Not all population means are equal.

p-value = 0.020; reject H_0. At the 5% significance level, we conclude that the average strength of the plywood boards differs by the type of glue used.

10.74 H_0: $\mu_A = \mu_B = \mu_C$

H_A: Not all population means are equal.

p-value = 0.003; reject H_0. At the 5% significance level, we conclude that the mean P/E ratios of these three industries differ.

10.76 H_0: $\mu_{\text{Internet}} = \mu_{\text{Phone}} = \mu_{\text{Mail-in}}$

H_A: Not all population means are equal.

p-value = 0.651; do not reject H_0. At the 5% significance level, we cannot conclude that the mean purchase amounts differ across the three purchase sources.

10.78 H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$

H_A: Not all population means are equal.

p-value = 0 (approximately); reject H_0. At the 5% significance level, we conclude that the mean visits to the website differ by day of the week.

Chapter 11

11.2 $(0.25 - 0.28) \pm 1.96\sqrt{\frac{0.25(1-0.25)}{200} + \frac{0.28(1-0.28)}{250}}$, or [−0.112, 0.052]

At the 5% significance level, we cannot conclude that the population proportions differ.

11.4 a. $z = \frac{0.40 - 0.43}{\sqrt{0.4185(1 - 0.4185)\left(\frac{1}{250} + \frac{1}{400}\right)}} = -0.75$

b. p-value = 0.453

c. Do not reject H_0; at the 5% significance level, we cannot conclude that the population proportions differ.

11.10 a. H_0: $p_1 - p_2 \leq 0$; H_A: $p_1 - p_2 > 0$ (Population 1 = Boys and Population 2 = Girls)

b. $z = \frac{0.27 - 0.14}{\sqrt{0.205(1 - 0.205)\left(\frac{1}{500} + \frac{1}{500}\right)}} = 5.09$;

p-value = 0 (approximately)

Reject H_0; at the 5% significance level, we conclude that the proportion of boys growing out of asthma is more than that of girls.

c. H_0: $p_1 - p_2 \leq 0.10$; H_A $p_1 - p_2 > 0.10$;

$z = \frac{(0.27 - 0.14) - 0.10}{\sqrt{\frac{0.27(1-0.27)}{500} + \frac{0.14(1-0.14)}{500}}} = 1.19$;

p-value = 0.117

Do not reject H_0; at the 5% significance level, we cannot conclude that the proportion of boys who grow out of asthma exceeds that of girls by more than 0.10.

11.14 a. H_0: $p_1 - p_2 \geq 0$; H_A: $p_1 - p_2 < 0$ (Population 1 = African American Men and Population 2 = Caucasian Men)

$z = \frac{0.2769 - 0.3444}{\sqrt{0.3161(1 - 0.3161)\left(\frac{1}{130} + \frac{1}{180}\right)}} = -1.26$;

p-value = 0.1038; do not reject H_0. At the 5% significance level, we cannot conclude that the proportion of obese African American men is less than their Caucasian counterparts.

b. $H_0: p_1 - p_2 \leq 0$; $H_A: p_1 - p_2 > 0$ (Population 1 = African American Women and Population 2 = Caucasian Women)

$$z = \frac{0.3889 - 0.2583}{\sqrt{0.3143(1 - 0.3143)\left(\frac{1}{90} + \frac{1}{120}\right)}} = 2.02;$$

p-value = 0.0217; reject H_0.

At the 5% significance level, we conclude that the proportion of obese African American women is greater than their Caucasian counterparts.

c. $H_0: p_1 - p_2 = 0$; $H_A: p_1 - p_2 \neq 0$ (Population 1 = African American Adults and Population 2 = Caucasian Adults)

$$z = \frac{0.3227 - 0.3100}{\sqrt{0.3154(1 - 0.3154)\left(\frac{1}{220} + \frac{1}{300}\right)}} = 0.31;$$

p-value = 0.7566; do not reject H_0.

At the 5% significance level, we cannot conclude that the proportion of obese African American adults differs from their Caucasian counterparts.

11.18 $H_0: p_1 - p_2 \leq 0.10$; $H_A: p_1 - p_2 > 0.10$ (Population 1 = Male Students and Population 2 = Female Students)

$$z = \frac{(0.57 - 0.32) - 0.10}{\sqrt{\frac{0.57(1 - 0.57)}{100} + \frac{0.32(1 - 0.32)}{100}}} = 2.21;$$

p-value = 0.0136 ; reject H_0.

At the 5% significance level, we conclude that there is a greater than 10 percentage point difference between the proportion of male and female students who think it is not feasible for men and women to be just friends.

11.26 a. $H_0: p_1 = p_2 = p_3 = 1/3$; H_A: Not all population proportions are equal to 1/3.

b. $\chi^2_2 = 6.968$; p-value = 0.031; reject H_0. At the 5% significance level, we conclude that Zimbabwe visitors are not equally represented by Europe, North America, and the rest of the world.

11.28 $H_0: p_1 = 0.38, p_2 = 0.33, p_3 = 0.29$;
H_A: At least one of the p_i (i = 1,2,3) differs from its hypothesized value
$\chi^2_2 = 2.179$; p-value = 0.336; do not reject H_0.
At the 5% significance level, the researcher cannot conclude that car preferences have changed since the Associated Press-GfK Poll.

11.30 a. $H_0: p_1 = p_2 = p_3 = p_4 = 0.25$;
H_A: At least one p_i (i = 1,2,3,4) differs from 0.25.

b. $\chi^2_3 = 7.720$; p-value = 0.052

c. At the 10% significance level, we conclude that the proportion of bags filled by at least one chute differs from 0.25. Yes, the conclusion changes at the 5% significance level.

11.32 H_0: The two categories are independent.
H_A: The two categories are dependent.
$\chi^2_6 = 1.249$; p-value = 0.974; do not reject H_0. At the 1% significance level, we cannot conclude that Category 1 and Category 2 are dependent.

11.36 a. H_0: Vehicle brand and union membership are independent.
H_A: Vehicle brand and union membership are dependent.

b. $\chi^2_1 = 14.915$; p-value = 0 (approximately); reject H_0. At the 10% significance level, we conclude that vehicle brand and union membership are dependent.

c. The conclusion in part b is not sensitive to the choice of significance level.

11.40 H_0: Breakup reasons and one's sex are independent.
H_A: Breakup reasons and one's sex are dependent.
$\chi^2_4 = 19.463$; p-value = 0 (approximately); reject H_0. At the 1% significance level, we conclude that breakup reason and one's sex are dependent.

11.42 a. $H_0: p_1 - p_2 \leq 0.05$; $H_A: p_1 - p_2 > 0.05$ (Population 1 = JFK and Population 2 = O'Hare)

b. $$z = \frac{(0.70 - 0.63) - 0.05}{\sqrt{\frac{0.70(1 - 0.70)}{200} + \frac{0.63(1 - 0.63)}{200}}} = 0.42;$$
p-value = 0.3372

c. Do not reject H_0; at the 5% significance level, we cannot conclude that the proportion of on-time flights at JFK is more than 5 percentage points higher than that of O'Hare.

11.44 a. $H_0: p_1 = 0.40, p_2 = 0.30, p_3 = 0.20, p_4 = 0.10$
H_A: At least one of the p_i (i = 1,2,3,4) differs from its hypothesized value.

b. $\chi^2_3 = 8.182$

c. p-value = 0.042; do not reject H_0. At the 1% significance level, we cannot conclude that the market shares have changed.

11.46 $H_0: p_1 = 0.47, p_2 = 0.30, p_3 = 0.04, p_4 = 0.05, p_5 = 0.14$
H_A: At least one of the p_i (i = 1,2,3,4,5) differs from its hypothesized value.
$\chi^2_4 = 9.961$; p-value = 0.041; reject H_0. At the 5% significance level, we conclude that the researcher's results are inconsistent with the survey results conducted by Facebook.

11.48 a. H_0: Surviving a cardiac arrest is independent of the time of the cardiac arrest.
H_A: Surviving a cardiac arrest is dependent on the time of the cardiac arrest.

b. $\chi^2_1 = 333.462$; p-value = 0 (approximately)

c. Reject H_0; at the 1% significance level, we conclude that a patient surviving a cardiac arrest is dependent on the time that it happens. Hospitals need to ensure that patients have equal chances of surviving a cardiac arrest, regardless of when it happens.

11.52 H_0: A household's delinquency in payment is independent of the type of heating.
H_A: A household's delinquency in payment is dependent on the type of heating.
$\chi^2_3 = 23.82$; p-value = 0 (approximately); reject H_0. At the 5% significance level, we conclude that a household's delinquency in payments and the type of heating it uses are dependent.

11.54 a. $H_0: p_1 = 0.80, p_2 = 0.09, p_3 = 0.09, p_4 = 0.02$
H_A: At least one of the p_i (i = 1,2,3,4) differs from its hypothesized value.

b. $\chi^2_3 = 8.436$; p-value = 0.038

c. At the 5% significance level, we conclude that the management's goals are not being met. Yes, the conclusion changes at the 1% significance level.

Chapter 12

12.2 a. $\hat{y} = 40$

b. $\hat{y}$ increases by 25

12.8 a. $\widehat{\text{GPA}} = 0.4256 + 0.0041\text{GRE}$

b. $\widehat{\text{GPA}} = 3.34$

12.10 a. $\widehat{\text{Consumption}} = 8550.675 + 0.686\text{Disposable Income}$

b. Consumers spend 68.6% of an increase in income on consumption.

c. $\widehat{\text{Consumption}} = 47,652.68$

12.14 a. $\widehat{\text{Final}} = 27.5818 + 0.6774\text{Midterm}$

b. $\widehat{\text{Final}} = 81.77$

12.18 a. $\hat{y} = 568$

b. If x_2 increases by 1 unit, $\hat{y}$ decreases by 47.2 units, holding x_1 constant.

12.20 a. $b_1 = 30$; if x_1 increases by 1 unit, $\hat{y}$ increases by 30 units, holding x_2 constant.

b. $\hat{y} = 21.97 + 30x_1 - 1.88x_2$

c. $\hat{y} = 884.37$

12.24 a. The positive sign for the Poverty coefficient is as expected; the slope coefficient for Income is not as expected.

b. As Poverty increases by 1%, Crime is predicted to rise by 53.16, holding Income constant.

c. $\widehat{\text{Crime}} = 1009.08$

12.32 a. $\widehat{\text{Rent}} = 300.4116 + 225.81\text{Bed} + 89.2661\text{Bath} + 0.2096\text{Sqft}$

b. For every additional bathroom, the predicted rent increases by \$89.27, holding number of bedrooms and square feet constant.

c. $\widehat{\text{Rent}} = 1155.70$

12.34 a. $s_e = 0.8629$

b. $R^2 = 0.6111$

12.36 Model 2, since it has a smaller s_e and a higher R^2. We need not use adjusted R^2 since both models have the same number of explanatory variables.

12.38 Model 2, since it has a smaller s_e and a higher adjusted R^2.

12.40 a. 82.99% of the sample variability in sales is explained.

b. 17.01% of the sample variability in sales is not explained.

12.44 Model 2, since it has a smaller s_e (1475 versus 1509) and a higher adjusted R^2 (0.8283 versus 0.8204).

12.46 a. $s_e = 12.7697$

b. $R^2 = 0.1726$

c. Adjusted $R^2 = 0.1113$

12.56 a. H_0: $\beta_1 \leq 0$; H_A: $\beta_1 > 0$

b. $t_{18} = 8.40$; p-value = 0 (approximately)

c. Reject H_0; at the 5% significance level, we conclude that advertising expenditures and sales have a positive linear relationship.

12.58 a. $\widehat{\text{Time}} = 13.353 - 0.0477\text{Height}$

b. H_0: $\beta_1 = 0$; H_A: $\beta_1 \neq 0$

c. $t_5 = -2.926$

d. p-value = 0.033; reject H_0. At the 5% significance level, Height is significant.

12.62 a. $\widehat{\text{Return}} = -12.0243 + 0.1459\text{P/E} + 5.4417\text{P/S}$

b. H_0: $\beta_1 = \beta_2 = 0$; H_A: At least one $\beta_1 \neq 0$

$F_{(2,27)} = 2.817$; p-value = 0.077; reject H_0. At the 10% significance level, the two explanatory variables are jointly significant.

c. At the 10% significance level, P/E is not significant but P/S is significant.

12.64 $\widehat{\text{Taxes}} = 6,499.4126 + 6.8063\text{Size}$

H_0: $\beta_1 = 0$; and H_A $\beta_1 \neq 0$

p-value = 0 (approximately); reject H_0. At the 5% significance level, home size is significant.

12.68 a. $\widehat{\text{Cost}} = 14{,}039.1873 + 92.7827\text{Temp} + 446.1406\text{Days} - 27.0033\text{ Tons}$

b. H_0: $\beta_1 = \beta_2 = \beta_3 = 0$; H_A: At least one $\beta_j \neq 0$

p-value = 0.026; reject H_0. At the 10% significance level, the explanatory variables are jointly significant.

c. At the 10% significance level, the average temperature is significant, the number of work days is not significant, and the tons produced is not significant.

12.70 a. $\widehat{\text{Price}} = 153,348.2664 + 95.8559\text{Sqft} + 556.8907\text{Beds} + 92{,}022.9126\text{Baths}$

b. At the 5% significance level, the explanatory variables are jointly significant.

c. At the 5% significance level, square footage is significant, number of beds is not significant, and number of baths is significant.

12.74 b. Since the residuals fan out when plotted against x, it suggests a problem of changing variability (heteroskedasticity). This means that the estimators are not efficient and the significance tests are not valid. A common solution is to use robust standard errors for conducting significance tests.

12.76 The scatterplot shows that a simple linear regression model is not appropriate as GPA is positively related to Hours at lower levels but negatively related at higher levels of Hours.

12.78 a. Perfect multicollinearity, since Study + Sleep + Leisure = 24; the proposed model cannot be estimated.

b. Drop the Sleep variable.

12.80 a. Experienced (older) employees are likely to have more variability in salaries because not all employees reach the same level of success over time.

b. The residuals fan out when plotted against experience, confirming the changing variability (heteroskedasticity) problem.

12.82 There does not appear to be an issue with correlated observations, as the residuals do not show any pattern around the horizontal axis.

12.86 a. $\widehat{\text{Happiness}} = 56.1772 + 0.2845\text{Age}$.

b. $\widehat{\text{Happiness}} = 68.98$.

c. 32.67% of the sample variation in Happiness is explained by the estimated model.

d. $H_0: \beta_1 = 0$
$H_A: \beta_1 \neq 0$
The p-value = 0.0035; reject H_0. Age has a significant influence on Happiness at the 1% significance level.

12.88 a. $\widehat{\text{Ownership}} = 78.9791 - 0.0002\text{Income}$. For a \$1,000 increase in income, the predicted homeownership rate decreases by 2%. This negative relationship is surprising.
b. $s_e = 5.77$
c. 6.18% of the sample variation in homeownership is explained by the sample regression equation.

12.90 a. $\widehat{\text{Return}} = -33.3966 + 3.9674(\text{P/E}) - 3.3681(\text{P/S})$. The signs are as expected. Holding the other variable constant, as P/E increases, the predicted return increases, and as P/S increases, the predicted return decreases.
b. As the P/S ratio increases by 1 unit, the predicted return of the firm decreases by 3.37%, holding P/E constant.
c. $\widehat{\text{Return}} = -0.46\%$
d. $s_e = 13.64$
e. 40.28% of the sample variation in return is explained by the sample regression equation.
f. $H_0: \beta_1 = \beta_2 = 0$
H_A: At least one $\beta_j \neq 0$
The p-value ≈ 0; reject H_0. At the 5% significance level, the explanatory variables are jointly significant.
g. The hypotheses for each test of individual significance would be
$H_0: \beta_j = 0$
$H_A: \beta_j \neq 0$
At the 5% significance level, the P/E ratio is significant, but the P/S ratio is not significant.

12.92 a. $\widehat{\text{Startups}} = 0.4190 + 0.0087\text{Research} + 0.0517\text{Patents} - 0.0194\text{Duration}$
b. $\widehat{\text{Startups}} = 1.49\text{startups}$
c. A \$1 million increase in research expenditure results in a predicted increase in the number of startups by 0.0087, holding everything else constant. Thus, approximately \$114.94 million $\left(\frac{1}{0.0087} = 114.94\right)$ in additional research expenditures would be needed to have 1 additional predicted startup, everything else being the same.

12.94 a. $\widehat{\text{Rent}} = 300.41 + 225.81\ \text{Bed} + 89.27\ \text{Bath} + 0.21\ \text{Sqft}$.
b. 80.91% of the sample variation in Rent is explained by the sample regression equation; 19.09% of the sample variation in Rent is unexplained by the sample regression equation.
c. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$
H_A: At least one $\beta_j \neq 0$
The p-value ≈ 0; reject H_0. At the 5% level, the explanatory variables are jointly significant.
d. The hypotheses for each test of individual significance would be
$H_0: \beta_j = 0$
$H_A: \beta_j \neq 0$
At the 5% significance level, the number of bedrooms and the square footage are significant; the number of bathrooms is not significant.

12.96 a. $\widehat{\text{Return}} = 7.2359 + 0.1074\text{Turnover} - 3.1338\text{Expense}$
At the 5% significance level, the explanatory variables are not jointly or individually significant. George's theory is not valid.
b. Multicollinearity is not likely a problem since the sample correlation coefficient between Turnover and Expense is only −0.247. There does not seem to be a problem of changing variability since the residuals appear randomly dispersed around zero when plotted against Turnover and Expense.

Chapter 13

13.2 a. $\hat{y} = 160 + 15 \times 1 + 32 \times 1 = 207$
b. $\hat{y} = 160 + 15 \times 0 + 32 \times 0 = 160$

13.4 a. Female employees
b. Female employees without an MBA
c. No; the inferences would not change.

13.6 a. $\widehat{\text{Consumption}} = 13,007.2568 + 0.4444\text{Income} + 6,544.4264\text{Urban}$
Urban: = \$55,103.68.
Rural: = \$48,559.26.
b. $\widehat{\text{Consumption}} = 19,551.6832 + 0.4444\text{Income} - 6,544.4264\text{Rural}$
Urban: = \$55,103.68.
Rural: = \$48,559.26.

13.8 a. $\widehat{\text{Salary}} = 62.3383 - 0.9605\text{BMI} + 4.4855\text{White}$
At the 5% significance level, we conclude that BMI influences Salary.
b. White man: $\widehat{\text{Salary}} = 38.01$
c. Non-white man: $\widehat{\text{Salary}} = 33.52$

13.16 a. $\widehat{\text{Vehicles}} = 135.3913 + 23.5056\text{Garage Bays} + 0.5955\text{Population} + 84.5998\text{Access} + 77.4646\text{Winter}$
c. At the 5% (and 10%) significance level, we conclude the explanatory variables are jointly significant.
At the 5% significance level, only the access and the winter variables are significant.
At the 10% significance level, the garage bays, the access, and the winter variables are significant.
d. 87.39%
e. $\widehat{\text{Vehicles}} = 361.34$

13.20 a. The model with the interaction variable xd is preferred as it has a highest adjusted R^2 (0.5578 > 0.4184).
b. $\hat{y} = -0.5194 + 0.3791x + 13.1749d - 0.2758xd$
For $x = 15$ and $d = 1$, $\hat{y} = 14.205$
For $x = 15$ and $d = 0$, $\hat{y} = 5.167$

13.22 a. 5 years: $\widehat{\text{Salary}} = 61.8$, or \$61,800
15 years: $\widehat{\text{Salary}} = 93.8$, or \$93,800
b. 5 years: $\widehat{\text{Salary}} = 36.3$, or \$36,300
15 years: $\widehat{\text{Salary}} = 48.3$, or \$48,300
c. Higher salary with a college degree, where the salary gap increases with experience.

13.26 a. Model 1: $\widehat{\text{Errors}} = 37.9305 - 1.2814\text{Exper} - 7.4241\text{Train}$
Model 2: $\widehat{\text{Errors}} = 42.7764 - 1.6991\text{Exper} - 23.1111\text{Train} + 0.9785(\text{Exper} \times \text{Train})$

b. Model 2; higher adjusted R^2 (0.6035 > 0.5621) and the explanatory variables are significant.

c. 10 years, training: $\widehat{\text{Errors}} = 12.46$; 20 years, no training: $\widehat{\text{Errors}} = 8.79$

d. Less experienced employees benefit more from the training program than more experienced employees.

13.30 a. At $x = 10$, $\hat{y} = 20 - 0.72 \times 10 + 0.02 \times 10^2 = 14.80$
At $x = 20$, $\hat{y} = 20 - 0.72 \times 20 + 0.02 \times 20^2 = 13.60$
At $x = 30$, $\hat{y} = 20 - 0.72 \times 30 + 0.02 \times 30^2 = 16.40$

b. Given $b_2 = 0.02 > 0$, $\hat{y}$ is minimized at $x = \frac{0.72}{2(0.02)} = 18$.

13.32 a. Linear Model: when $x = 2$, $\hat{y} = 19.80 + 1.35 \times 2 = 22.50$; when $x = 3$, $\hat{y} = 19.80 + 1.35 \times 3 = 23.85$
Quadratic Model: when $x = 2$, $\hat{y} = 20.08 + 1.50 \times 2 - 0.31 \times 2^2 = 21.84$; when $x = 3$, $\hat{y} = 20.08 + 1.50 \times 3 - 0.31 \times 3^2 = 21.79$
Cubic Model: when x = 2, $\hat{y} = 20.07 + 1.58 \times 2 - 0.27 \times 2^2 - 0.03 \times 2^3 = 21.91$; when $x = 3$, $\hat{y} = 20.07 + 1.58 \times 3 - 0.27 \times 3^2 - 0.03 \times 3^3 = 21.57$

b. Since the quadratic model has the highest adjusted R^2 (0.691 > 0.689 > 0.636), it is the most appropriate model.

13.34 a. The logarithmic model, with a higher R^2 (0.9341 > 0.8233), provides a better fit than the linear model.

b. $\hat{y} = 18.61$

13.36 a. From the scatterplot, crew sizes between 6 and 7 seem optimal.

b. Linear Model: $\widehat{\text{Jobs}} = 13.0741 + 0.1111(\text{Crew Size})$
Quadratic Model: $\widehat{\text{Jobs}} = 2.1111 + 4.5960(\text{Crew Size}) - 0.3737(\text{Crew Size})^2$
The quadratic model provides a better fit than the linear model. It has a higher adjusted R^2 (0.6307 > −0.0284) and also has statistically significant explanatory variables.

c. $\widehat{\text{Jobs}} = 15.75$ jobs/week

d. $\widehat{\text{Jobs}} = 0.6852 + 5.5407(\text{Crew Size}) - 0.5505(\text{Crew Size})^2 + 0.0098(\text{Crew Size})^3$
The quadratic model, with a higher adjusted R^2 (0.6307 > 0.6170), provides a better fit than the cubic model.

13.44 a. $\widehat{\text{Cost}} = 16{,}776.0365 + 147.7790 \times 65 + 475.7513 \times 23 - 271.9524 \times 76 = 16{,}655.57$

b. $\widehat{\text{Cost}} = \exp(9.7378 + 0.0093 \times 65 + 0.0301 \times 23 - 0.0181 \times 76 + 0.2029^2/2) = 15{,}912.73$

13.48 a. $\hat{y} = 14.98 + 1.13 \times 25 = 43.23$

b. $\hat{y} = \exp\left(1.81 + 0.08 \times 25 + \frac{0.01^2}{2}\right) = 45.15$

13.54 a. The scatterplot suggests the cubic trend model.

b. Again, the cubic model is preferred as it has a highest adjusted R^2 of 98.79%.

c. $\hat{y} = 5.5027 - 0.2123 \times 49 + 0.0235 \times 49^2 - 0.0003 \times 49^3 = 11.48\%$.

13.58 a. The linear and the exponential trends are almost indiscernible; we will use the exponential trend model for making forecasts.

b. $\hat{y} = \exp(4.5347 + 0.0091 \times 54 + 0.0405^2/2) = 152.72$

13.60 Year 6, Quarter 1:
$\hat{y} = \exp\left(0.28 + 0.15 \times 21 + 0.18 + \frac{0.28^2}{2}\right) = 38.44$
Year 6, Quarter 2:
$\hat{y} = \exp\left(0.28 + 0.15 \times 22 + 0.12 + \frac{0.28^2}{2}\right) = 42.06$
Year 6, Quarter 3:
$\hat{y} = \exp\left(0.28 + 0.15 \times 23 - 0.08 + \frac{0.28^2}{2}\right) = 40.01$
Year 6, Quarter 4:
$\hat{y} = \exp\left(0.28 + 0.15 \times 24 + \frac{0.28^2}{2}\right) = 50.36$

13.62 Year 11, Quarter 1:
$\hat{y} = \exp\left(2.7099 + 0.5642 + 0.1431 \times 41 + \frac{0.4940^2}{2}\right) = 10{,}537.25$
Year 11, Quarter 2:
$\hat{y} = \exp\left(2.7099 + 0.3775 + 0.1431 \times 42 + \frac{0.4940^2}{2}\right) = 10{,}087.55$
Year 11, Quarter 3:
$\hat{y} = \exp\left(2.7099 - 0.7295 + 0.1431 \times 43 + \frac{0.4940^2}{2}\right) = 3{,}847.47$
Year 11, Quarter 4:
$\hat{y} = \exp\left(2.7099 + 0.1431 \times 44 + \frac{0.4940^2}{2}\right) = 9{,}206.98$

13.64 Year 6, Month 1:
$\hat{y} = \exp\left(6.3047 - 0.4901 + 0.0502 \times 61 + \frac{0.0428^2}{2}\right) = 7{,}186.15$
Year 6, Month 2:
$\hat{y} = \exp\left(6.3047 - 0.6134 + 0.0502 \times 62 + \frac{0.0428^2}{2}\right) = 6{,}679.55$
Year 6, Month 12:
$\hat{y} = \exp\left(6.3047 + 0.0502 \times 72 + \frac{0.0428^2}{2}\right) = 20{,}385.28$

13.68 a. The scatterplot suggests a quadratic trend model with seasonal dummy variables.

b. $\hat{y} = 346.2045 + 2.1853 - 1.5461 \times 33 + 0.0588 \times 33^3 = 361.43$

13.70 January 2018:
$\hat{y}_t = \exp\left(3.7961 - 0.0273 + 0.0091 \times 85 + \frac{0.1002^2}{2}\right) = 94.29$
February 2018:
$\hat{y}_t = \exp\left(3.7961 - 0.0168 + 0.0091 \times 86 + \frac{0.1002^2}{2}\right) = 96.15$

13.74 a. At the 5% significance level, we conclude that the students study more in the fall quarter and the winter quarter than in the spring quarter.

b. $\widehat{\text{Study Hours}} = 13.7647 + 4.5144\text{Fall} + 1.6074\text{Winter}$
Fall quarter: $\widehat{\text{Study Hours}} = 18.28$
Winter quarter: $\widehat{\text{Study Hours}} = 15.37$
Spring quarter: $\widehat{\text{Study Hours}} = 13.76$

13.78 a. $\widehat{\text{Compensation}} = 2{,}677.1892 + 10.3154\text{Profit} + 1{,}227.6866\text{Years} + 36{,}655.1363\text{Grad} - 0.5227(\text{Profit} \times \text{Grad}) - 193.1612(\text{Years} \times \text{Grad})$

b. At the 5% significance level, we conclude that the explanatory variables are jointly significant.

c. At the 5% significance level, we conclude that the explanatory variables are individually significant, except for the two interaction terms, Profit × Grad and Years × Grad.

d. $\widehat{\text{Compensation}} = \$101{,}855$

13.80 a. Linear Model: $\widehat{\text{Time}} = -14.4886 + 0.7502\text{Parts}$

Quadratic Model: $\widehat{\text{Time}} = -6.7165 + 0.4476\text{Parts} + 0.0025\text{Parts}^2$

b. All the explanatory variables are significant in both models. The quadratic model, with a higher adjusted R^2 (0.9966 > 0.9916), provides a better fit than the linear model.

c. $\widehat{\text{Time}} = 20.53$ minutes

13.86 a. Real estate loans have been growing exponentially.

b. $\hat{y}_t = \exp(6.0545 + 0.0543 \times 37 + 0.1082^2/2) = \$3{,}189.63$ billion

13.88 Quarter 1: $\hat{y}_t = \exp(9.0170 + 0.1340 + 0.0161 \times 21 + 0.0730^2/2) = \$13{,}237.42$ million

Quarter 2: $\hat{y}_t = \exp(9.0170 + 0.2739 + 0.0161 \times 22 + 0.0730^2/2) = \$15{,}471.72$ million

Quarter 3: $\hat{y}_t = \exp(9.0170 + 0.0924 + 0.0161 \times 23 + 0.0730^2/2) = \$13{,}112.69$ million

Quarter 4: $\hat{y}_t = \exp(9.0170 + 0.0161 \times 24 + 0.0730^2/2) = \$12{,}148.44$ million

The forecast for the fiscal year 2009 is $53,970.26 million.

A

Acceptance sampling A statistical quality control technique in which a portion of the completed products is inspected.

Addition rule The probability that A or B occurs, or that at least one of these events occurs, is $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Adjusted R^2 A modification of the coefficient of determination that imposes a penalty for using additional explanatory variables in the linear regression model.

Alpha In the capital asset pricing model (CAPM), it measures whether abnormal returns exist.

Alternative hypothesis (H_A) In a hypothesis test, the alternative hypothesis contradicts the default state or status quo specified in the null hypothesis.

Analysis of variance (ANOVA) A statistical technique used to determine if differences exist between three or more population means.

Arithmetic mean The average value of a data set; the most commonly used measure of central location, also referred to as the mean or the average.

Assignable variation In a production process, the variation that is caused by specific events or factors that can usually be identified and eliminated.

Average See *Arithmetic mean*.

B

Bar chart A graph that depicts the frequency or relative frequency of each category of qualitative data as a series of horizontal or vertical bars, the lengths of which are proportional to the values that are to be depicted.

Bayes' theorem The rule for updating probabilities is $P(B|A) = \dfrac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$, where $P(B)$ is the prior probability and $P(B|A)$ is the posterior probability.

Bell curve See *Normal curve*.

Bell-shaped distribution See *Normal distribution*.

Bernoulli process A series of n independent and identical trials of an experiment such that each trial has only two possible outcomes, and each time the trial is repeated, the probabilities of success and failure remain the same.

Beta In the capital asset pricing model (CAPM), it measures the sensitivity of the stock's return to changes in the level of the overall market.

Between-treatments variance In ANOVA, a measure of the variability between sample means.

Bias The tendency of a sample statistic to systematically overestimate or underestimate a population parameter.

Big data A massive volume of both structured and unstructured data that are often difficult to manage, process, and analyze using traditional data processing tools.

Binomial distribution A description of the probabilities associated with the possible values of a binomial random variable.

Binomial random variable The number of successes achieved in the n trials of a Bernoulli process.

Box plot A graphical display of the minimum value, quartiles, and the maximum value of a data set.

C

***c* chart** A control chart that monitors the count of defects per item in statistical quality control.

Capital asset pricing model (CAPM) A regression model used in finance to examine an investment return.

Centerline In a control chart, the centerline represents a variable's expected value when the production process is in control.

Central limit theorem (CLT) The CLT states that the sum or mean of a large number of independent observations from the same underlying distribution has an approximate normal distribution.

Chance variation In a production process, the variation that is caused by a number of randomly occurring events that are part of the production process.

Changing variability In regression analysis, a violation of the assumption that the variance of the error term is the same for all observations. It is also referred to as heteroskedasticity.

Chebyshev's theorem For any data set, the proportion of observations that lie within k standard deviations from the mean will be at least $1 - 1/k^2$, where k is any number greater than 1.

Chi-square test of a contingency table See *Test for independence*.

Chi-square (χ^2) distribution A family of distributions where each distribution depends on its particular degrees of freedom *df*. It is positively skewed, with values ranging from zero to infinity, but becomes increasingly symmetric as *df* increase.

Classes Intervals for a frequency distribution of quantitative data.

Classical probability A probability often used in games of chance. It is based on the assumption that all outcomes are equally likely.

Cluster sampling A population is first divided up into mutually exclusive and collectively exhaustive groups of observations, called clusters. A cluster sample includes observations from randomly selected clusters.

Coefficient of determination (R^2) The proportion of the sample variation in the response variable that is explained by the sample regression equation.

Coefficient of variation (CV) The ratio of the standard deviation of a data set to its mean; a relative measure of dispersion.

Complement The complement of event A, denoted A^c, is the event consisting of all outcomes in the sample space that are not in A.

Complement rule The probability of the complement of an event is $P(A^c) = 1 - P(A)$.

Conditional probability The probability of an event given that another event has already occurred.

Confidence coefficient The probability that the estimation procedure will generate an interval that contains the population parameter of interest.

Confidence interval A range of values that, with a certain level of confidence, contains the population parameter of interest.

Consistency An estimator is consistent if it approaches the unknown population parameter being estimated as the sample size grows larger.

Contingency table A table that shows frequencies for two qualitative (categorical) variables, x and y, where each cell represents a mutually exclusive combination of the pair of x and y values.

Continuous (random) variable A variable that assumes uncountable values in an interval.

Continuous uniform distribution A distribution describing a continuous random variable that has an equally likely chance of assuming a value within a specified range.

Control chart A plot of statistics of a production process over time.

Correlated observations In regression analysis, a violation of the assumption that the observations are uncorrelated. It is also referred to as serial correlation.

Correlation coefficient A measure that describes the direction and strength of the linear relationship between two variables.

Covariance A measure that describes the direction of the linear relationship between two variables.

Critical value In a hypothesis test, the critical value is a point that separates the rejection region from the nonrejection region.

Cross-sectional data Values of a characteristic of many subjects at the same point in time or approximately the same point in time.

Cubic regression model In regression analysis, a model that allows two sign changes of the slope capturing the influence of the explanatory variable on the response variable.

Cubic trend model In time series analysis, a model that allows for two changes in the direction of the series.

Cumulative distribution function A probability that the value of a random variable X is less than or equal to a particular value x, $P(X \leq x)$.

Cumulative frequency distribution A distribution of quantitative data recording the number of observations that falls below the upper limit of each class.

Cumulative relative frequency distribution A distribution of quantitative data recording the fraction (proportion) of observations that falls below the upper limit of each class.

D

Degrees of freedom The number of independent pieces of information that go into the calculation of a given statistic. Many probability distributions are identified by the degrees of freedom.

Dependent events The occurrence of one event is related to the probability of the occurrence of the other event.

Descriptive statistics The summary of a data set in the form of tables, graphs, or numerical measures.

Detection approach A statistical quality control technique that determines at which point the production process does not conform to specifications.

Deterministic relationship A relationship in which the value of the response variable is uniquely determined by the values of the explanatory variables.

Discrete uniform distribution A symmetric distribution where the random variable assumes a finite number of values and each value is equally likely.

Discrete (random) variable A variable that assumes a countable number of values.

Dummy variable A variable that takes on values of 0 or 1.

Dummy variable trap A regression model where the number of dummy variables equals the number of categories of a qualitative variable; the resulting model cannot be estimated.

E

Efficiency An unbiased estimator is efficient if its standard error is lower than that of other unbiased estimators.

Empirical probability A probability value based on observing the relative frequency with which an event occurs.

Empirical rule Given a sample mean $\bar{x}$, a sample standard deviation s, and a relatively symmetric and bell-shaped distribution, approximately 68% of all observations fall in the interval $\bar{x} \pm s$; approximately 95% of all observations fall in the interval $\bar{x} \pm 2s$; and almost all observations fall in the interval $\bar{x} \pm 3s$.

Endogeneity See *Excluded variables*.

Error sum of squares (SSE) In ANOVA, a measure of the degree of variability that exists even if all population means are the same. In regression analysis, it measures the unexplained variation in the response variable.

Estimate A particular value of an estimator.

Estimator A statistic used to estimate a population parameter.

Event A subset of a sample space.

Excluded variables In regression analysis, a situation where important explanatory variables are excluded from the regression. It often leads to the violation of the assumption that the error term is uncorrelated with the (included) explanatory variables.

Exhaustive events When all possible outcomes of an experiment are included in the events.

Expected value A weighted average of all possible values of a random variable.

Experiment A process that leads to one of several possible outcomes.

Explanatory variables In regression analysis, the variables that influence the response variable. They are also called the independent variables, predictor variables, control variables, or regressors.

Exponential distribution A continuous, nonsymmetric probability distribution used to describe the time that has elapsed *between* occurrences of an event.

Exponential regression model A regression model in which only the response variable is transformed into natural logs.

Exponential trend model A regression model used for a time series that is expected to grow by an increasing amount each period.

F

***F* distribution** A family of distributions where each distribution depends on two degrees of freedom: the numerator degrees of freedom df_1 and the denominator degrees of freedom df_2. It is positively skewed, with values ranging from zero to infinity, but becomes increasingly symmetric as df_1 and df_2 increase.

Finite population correction factor A correction factor that accounts for the added precision gained by sampling a larger percentage of the population. It is implemented when the sample constitutes at least 5% of the population.

Frequency distribution A table that groups qualitative data into categories, or quantitative data into intervals called classes, where the number of observations that fall into each category or class is recorded.

G

Goodness-of-fit test A chi-square test used to determine if the sample proportions resulting from a multinomial experiment differ from the hypothesized population proportions specified in the null hypothesis.

Grand mean In ANOVA, the sum of all observations in a data set divided by the total number of observations.

H

Heteroskedasticity See *Changing variability*.

Histogram A graphical depiction of a frequency or relative frequency distribution; it is a series of rectangles where the width and height of each rectangle represent the class width and frequency (or relative frequency) of the respective class.

Hypergeometric distribution A description of the probabilities associated with the possible values of a hypergeometric random variable.

Hypergeometric random variable The number of successes achieved in the n trials of a two-outcome experiment, where the trials are not assumed to be independent.

Hypothesis test A statistical procedure to resolve conflicts between two competing claims (hypotheses) on a particular population parameter of interest.

I

Independent events The occurrence of one event does not affect the probability of the occurrence of the other event.

Independent random samples Two (or more) random samples are considered independent if the process that generates one sample is completely separate from the process that generates the other sample.

Indicator variable See *dummy variable*.

Inferential statistics The practice of extracting useful information from a sample to draw conclusions about a population.

Interaction variable In a regression model, a product of two explanatory variables. For example, xd captures the interaction between a quantitative variable x and a dummy variable d.

Interquartile range (IQR) The difference between the third and first quartiles.

Intersection The intersection of two events A and B, denoted $A \cap B$, is the event consisting of all outcomes in A and B.

Interval (scale) data Values of a quantitative variable that can be categorized and ranked, and in which differences between values are meaningful.

Interval estimate See *Confidence interval*.

Inverse transformation A standard normal variable Z can be transformed to the normally distributed random variable X with mean μ and standard deviation σ as $X = \mu + Z\sigma$.

J

Joint probabilities The values in the interior of a joint probability table, representing the probabilities of the intersection of two events.

Joint probability table A contingency table whose frequencies have been converted to relative frequencies.

K

Kurtosis coefficient A measure of whether data is more or less peaked than a normal distribution.

L

Law of large numbers In probability theory, if an experiment is repeated a large number of times, its empirical probability approaches its classical probability.

Left-tailed test In hypothesis testing, when the null hypothesis is rejected on the left side of the hypothesized value of the population parameter.

Linear trend model A regression model used for a time series that is expected to grow by a fixed amount each time period.

Logarithmic regression model A regression model in which only the explanatory variable is transformed into natural logs.

Log-log regression model A regression model in which both the response variable and the explanatory variable(s) are transformed into natural logs.

Lower control limit In a control chart, the lower control limit indicates excessive deviation below the expected value of the variable of interest.

M

Margin of error A value that accounts for the standard error of the estimator and the desired confidence level of the interval.

Marginal probabilities The values in the margins of a joint probability table that represent unconditional probabilities.

Matched-pairs sample When a sample is matched or paired in some way.

Mean See *Arithmetic mean*.

Mean absolute deviation (MAD) The average of the absolute differences between the observations and the mean.

Mean square error (MSE) The average of the error (residual) sum of squares, where the residual is the difference between the observed and the predicted value of a variable.

Mean square regression The average of the sum of squares due to regression.

Mean-variance analysis The idea that the performance of an asset is measured by its rate of return, and this rate of return is evaluated in terms of its reward (mean) and risk (variance).

Median The middle value of a data set.

Method of least squares See *Ordinary least squares* (*OLS*).

Mode The most frequently occurring value in a data set.

Multicollinearity In regression analysis, a situation where two or more explanatory variables are correlated.

Multinomial experiment A series of n independent and identical trials, such that on each trial there are k possible outcomes, called categories; the probability p_i associated with the ith category remains the same; and the sum of the probabilities is one.

Multiple linear regression model In regression analysis, more than one explanatory variable is used to explain the variability in the response variable.

Multiplication rule The probability that A and B both occur is $P(A \cap B) = P(A|B)P(B)$.

Mutually exclusive events Events that do not share any common outcome of an experiment.

N

Negatively skewed (left-skewed) distribution A distribution in which extreme values are concentrated in the left tail of the distribution.

Nominal (scale) data Values of a qualitative variable that differ merely by name or label.

Nonresponse bias A systematic difference in preferences between respondents and nonrespondents of a survey or a poll.

Normal curve A graph depicting the normal probability density function; also referred to as the bell curve.

Normal (probability) distribution The most extensively used probability distribution in statistical work and the cornerstone of statistical inference. It is symmetric and bell-shaped and is completely described by the mean and the variance.

Null hypothesis (H_0) In a hypothesis test, the null hypothesis corresponds to a presumed default state of nature or status quo.

O

Ogive A graph of the cumulative frequency or cumulative relative frequency distribution in which lines connect a series of neighboring points, where each point represents the upper limit of each class and its corresponding cumulative frequency or cumulative relative frequency.

One-tailed hypothesis test A test in which the null hypothesis is rejected only on one side of the hypothesized value of the population parameter.

One-way ANOVA A statistical technique that analyzes the effect of one categorical variable (factor) on the mean.

Ordinal (scale) data Values of a qualitative variable that can be categorized and ranked.

Ordinary least squares (OLS) A regression technique for fitting a straight line whereby the error (residual) sum of squares is minimized.

Outliers Extreme small or large data values.

P

$\bar{p}$ chart A control chart that monitors the proportion of defectives (or some other characteristic) of a production process.

***p*-value** In a hypothesis test, the likelihood of observing a sample mean that is at least as extreme as the one derived from the given sample, under the assumption that the null hypothesis is true.

Parameter See *Population parameter*.

Percentile The pth percentile divides a data set into two parts: approximately p percent of the observations have values less than the pth percentile and approximately $(100 - p)$ percent of the observations have values greater than the pth percentile.

Pie chart A segmented circle portraying the categories and relative sizes of some qualitative variable.

Point estimate The value of the point estimator derived from a given sample.

Point estimator A function of the random sample used to make inferences about the value of an unknown population parameter.

Poisson distribution A description of the probabilities associated with the possible values of a Poisson random variable.

Poisson process An experiment in which the number of successes within a specified time or space interval equals any integer between zero and infinity; the numbers of successes counted in nonoverlapping intervals are independent from one another; and the probability that success occurs in any interval is the same for all intervals of equal size and is proportional to the size of the interval.

Poisson random variable The number of successes over a given interval of time or space in a Poisson process.

Polygon A graph of a frequency or relative frequency distribution in which lines connect a series of neighboring points, where each point represents the midpoint of a particular class and its associated frequency or relative frequency.

Polynomial regression model In regression analysis, a model that allow sign changes of the slope capturing the influence of an explanatory variable on the response variable.

Population The complete collection of items of interest in a statistical problem.

Population parameter A characteristic of a population.

Positively skewed (right-skewed) distribution A distribution in which extreme values are concentrated in the right tail of the distribution.

Posterior probability The updated probability, conditional on the arrival of new information.

Prior probability The unconditional probability before the arrival of new information.

Probability A numerical value between 0 and 1 that measures the likelihood that an event occurs.

Probability density function The probability density function provides the probability that a continuous random variable falls within a particular range of values.

Probability distribution Every random variable is associated with a probability distribution that describes the variable completely. It is used to compute probabilities associated with the variable.

Probability mass function The probability mass function provides the probability that a discrete random variable takes on a particular value.

Probability tree A graphical representation of the various possible sequences of an experiment.

Q

Quadratic regression model In regression analysis, a model that allows one sign change of the slope capturing the influence of the explanatory variable on the response variable.

Quadratic trend model In time series analysis, a model that captures either a U-shaped trend or an inverted U-shaped trend.

Qualitative variable A variable that uses labels or names to identify the distinguishing characteristics of observations.

Quantitative variable A variable that assumes meaningful numerical values for observations.

Quartiles Any of the three values that divide the ordered data into four equal parts, where the first, second, and third quartiles refer to the 25th, 50th, and 75th percentiles, respectively.

R

***R* chart** A control chart that monitors the variability of a production process.

Random error In regression analysis, random error is due to the omission of factors that influence the response variable.

Random variable A function that assigns numerical values to the outcomes of an experiment.

Range The difference between the maximum and the minimum values in a data set.

Ratio (scale) data Values of a quantitative variable that can be categorized and ranked, and in which differences between values are meaningful; in addition, a true zero point (origin) exists.

Regression analysis A statistical method for analyzing the relationship between variables.

Rejection region In a hypothesis test, a range of values such that if the value of the test statistic falls into this range, then the decision is to reject the null hypothesis.

Relative frequency distribution A frequency distribution that shows the fraction (proportion) of observations in each category of qualitative data or class of quantitative data.

Residual (*e*) In regression analysis, the difference between the observed value and the predicted value of the response variable, that is, $e = y - \hat{y}$.

Residual plots In regression analysis, the residuals are plotted sequentially or against an explanatory variable to identify model inadequacies. The model is adequate if the residuals are randomly dispersed around the zero value.

Response variable In regression analysis, the variable that is influenced by the explanatory variable(s). It is also called the dependent variable, the explained variable, the predicted variable, or the regressand.

Right-tailed test In hypothesis testing, when the null hypothesis is rejected on the right side of the hypothesized value of the population parameter.

Risk-averse consumer Someone who takes risk only if it entails a suitable compensation and may decline a risky prospect even if it offers a positive expected gain.

Risk-loving consumer Someone who may accept a risky prospect even if the expected gain is negative.

Risk-neutral consumer Someone who is indifferent to risk and makes his/her decisions solely on the basis of the expected gain.

S

***s* chart** A control chart that monitors the variability of a production process.

Sample A subset of a population of interest.

Sample correlation coefficient A sample measure that describes both the direction and strength of the linear relationship between two variables.

Sample covariance A sample measure that describes the direction of the linear relationship between two variables.

Sample space A record of all possible outcomes of an experiment.

Sample statistic A random variable used to estimate the unknown population parameter of interest.

Sampling distribution The probability distribution of an estimator.

Scatterplot A graphical tool that helps in determining whether or not two variables are related in some systematic way. Each point in the diagram represents a pair of known or observed values of the two variables.

Seasonal dummy variables Dummy variables used to capture the seasonal component from a time series.

Selection bias A systematic underrepresentation of certain groups from consideration for a sample.

Semi-log model A regression model in which not all variables are transformed into logs.

Serial correlation See *Correlated observations*.

Sharpe ratio A ratio calculated by dividing the difference of the mean return from the risk-free rate by the asset's standard deviation.

Significance level The allowed probability of making a Type I error.

Simple linear regression model In regression analysis, one explanatory variable is used to explain the variability in the response variable.

Simple random sample A sample of n observations that has the same probability of being selected from the population as any other sample of n observations.

Skewness coefficient A measure that determines if the data are symmetric about the mean. Symmetric data have a skewness coefficient of zero.

Social-desirability bias A systematic difference between a group's "socially acceptable" responses to a survey or poll and this group's ultimate choice.

Standard deviation The positive square root of the variance; a common measure of dispersion.

Standard error The standard deviation of an estimator.

Standard error of the estimate The standard deviation of the residual; used as a goodness-of-fit measure for regression analysis.

Standard normal distribution A special case of the normal distribution with a mean equal to zero and a standard deviation (or variance) equal to one.

Standard normal table See *z table*.

Standard transformation A normally distributed random variable X with mean μ and standard deviation σ can be transformed into the standard normal random variable Z as $Z = (X - \mu)/\sigma$.

Standardize A technique used to convert a value into its corresponding z-score.

Statistic See *Sample statistic*.

Statistical quality control Statistical techniques used to develop and maintain a firm's ability to produce high-quality goods and services.

Stem-and-leaf diagram A visual method of displaying quantitative data where each value of a data set is separated into two parts: a stem, which consists of the leftmost digits, and a leaf, which consists of the last digit.

Stochastic relationship A relationship in which the value of the response variable is not uniquely determined by the values of the explanatory variables.

Stratified random sampling A population is first divided up into mutually exclusive and collectively exhaustive groups, called strata. A stratified sample includes randomly selected observations from each stratum. The number of observations per stratum is proportional to the stratum's size in the population. The data for each stratum are eventually pooled.

Structured data Data that conform to a predefined row-column format.

Student's *t* distribution See *t distribution*.

Subjective probability A probability value based on personal and subjective judgment.

Sum of squares due to regression (SSR) In regression analysis, it measures the explained variation in the response variable.

Sum of squares due to treatments (SSTR) In ANOVA, a weighted sum of squared differences between the sample means and the overall mean of the data.

Symmetry When one side of a distribution is a mirror image of the other side.

T

***t* distribution** A family of distributions that are similar to the z distribution except that they have broader tails. They are identified by their degrees of freedom *df*.

Test for independence A goodness-of-fit test analyzing the relationship between two qualitative variables. Also called a chi-square test of a contingency table.

Test of individual significance In regression analysis, a test that determines whether an explanatory variable has an individual statistical influence on the response variable.

Test of joint significance In regression analysis, a test to determine whether the explanatory variables have a joint statistical influence on the response variable.

Test statistic A sample-based measure used in hypothesis testing.

Time series A set of sequential observations of a variable over time.

Total probability rule A rule that expresses the unconditional probability of an event, $P(A)$, in terms of probabilities conditional on various mutually exclusive and exhaustive events. The total probability rule conditional on two events B and B^c is $P(A) = P(A \cap B) + P(A \cap B^c) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c)$.

Total sum of squares (SST) In regression analysis, it measures the total variation in the response variable.

Trend A long-term upward or downward movement of a time series.

Two-tailed hypothesis test A test in which the null hypothesis can be rejected on either side of the hypothesized value of the population parameter.

Type I error In a hypothesis test, this error occurs when the decision is to reject the null hypothesis when the null hypothesis is actually true.

Type II error In a hypothesis test, this error occurs when the decision is to not reject the null hypothesis when the null hypothesis is actually false.

U

Unbiased An estimator is unbiased if its expected value equals the unknown population parameter being estimated.

Unconditional probability The probability of an event without any restriction.

Union The union of two events A and B, denoted $A \cup B$, is the event consisting of all outcomes in A or B.

Unstructured data Data that do not conform to a predefined row-column format.

Upper control limit In a control chart, the upper control limit indicates excessive deviation above the expected value of the variable of interest.

V

Variable A general characteristic being observed on a set of people, objects, or events, where each observation varies in kind or degree.

Variance The average of the squared differences from the mean; a common measure of dispersion.

W

Weighted mean When some observations contribute more than others in the calculation of an average.

Within-treatments variance In ANOVA, a measure of the variability within each sample.

X

$\bar{x}$ chart A control chart that monitors the central tendency of a production process.

Z

***z*-score** The relative position of a value within a data set; it is also used to detect outliers.

***z* table** A table providing cumulative probabilities for positive or negative values of the standard normal random variable *Z*.

n refers to information in notes at the bottom of pages.

A

AARP, 282
Acceptance sampling, 240
AccuWeather.com, 25
Addition rule of probability, 114–116
Adidas, Inc., 105, 124, 126, 371, 386, 389
Adjusted coefficient of determination (R^2), 419–420
Affordable Care Act, 163
Albrecht, Karl, 43
Allstate Insurance Co., 236, 286, 316
Alpha *(α),* of stock, 425–426
Alstead, Troy, 4
Alternative hypothesis, 294–297, 354. *See also* Hypothesis testing
American Heart Association, 473
Amtrak, 112
Analysis of variance (ANOVA)
  between treatments variance estimate in, 352–353
  coefficient of determination (R^2) derived from, 418
  F distribution as basis for, 349–350
  one-way ANOVA table for, 355–356
  one-way test in, 350–352
  for regression, 428
  within treatments variance estimate in, 353–355
Annual Urban Mobility Report (Texas Transportation Institute), 359–360
Apple, Inc., 70, 122, 177, 224, 452, 494
Arithmetic mean, 62–64
Ashenfelter, Orley, 485–486
Assignable variation, in quality control, 240
Assigning probability, 109–113
Associated Press, 101, 317
Associated Press/GfK Poll, 385
Association, measures of, 92–94
Asymptotic, normal distribution as, 189
AT&T, 212, 398
Auburn University, 173
Average, 62–64. *See also* Means

B

Bar charts, 21, 23. *See also* Tabular and graphical methods for qualitative data
Bayes' theorem, 131–134
Bayes, Thomas, 131
Bell-shaped curve, smoothed histogram as, 33
Bell-shaped distribution, 188–189
Bell Telephone Laboratories, 241
Bentley University, 481
Bernoulli, James, 156
Bernoulli process, 156–157, 169, 379
Beta *(β),* of stock, 425–426
Between treatments variance. *See* Analysis of variance (ANOVA)
Bias, in sampling, 220–222
Big data, 8
Binomial distribution
  Excel for binomial probabilities, 161–162
  expected value, variance, and standard deviation of, 159–161
  normal approximation of, 199
  overview, 156–159
Binomial random variable, 156–157, 159
Blockbuster, Inc., 498
Bloomberg Businessweek, 236
BLS (Bureau of Labor Statistics), 6, 8, 202, 250, 253
Boston Globe, The, 4
Boston Security Analysts Society, Inc. (BSAS), 91
Boxplots, 73–75
Brady, Tom, 389
Brin, Sergey, 43
Brown, Scott, 4
Bureau of Economic Analysis, 8, 252
Bureau of Labor Statistics (BLS), 6, 8, 202, 250, 253
Bureau of Transportation Statistics, 288
Bush, George W., 289
Business Week, 9

C

Capital asset pricing model (CAPM), 425–426
CareerBuilder.com, 390
Case studies
  compression-gear market, 105, 126
  continuous probability distributions, 183, 199
  discrete probability distributions, 145, 167
  house prices in Southern California, 19, 32
  hypothesis testing, 293, 310
  interval estimation, 259, 281
  investment decisions, 61, 83
  means, 329, 345
  proportions, 371, 389
  regression analysis, 403, 429
  regression analysis extensions, 457, 470
  sampling, 219, 236
  tween survey, 3, 15

Caterpillar, Inc., 27, 432–433
Causation, fallacy of correlation-to-, 5
CBS News, 26, 50, 278
Census's Population Survey, 122
Center for Equal Opportunity, 462
Centerline, in control charts, 241–245, 248
Centers for Disease Control and Prevention (CDC), 50, 214, 307, 377, 504
Central limit theorem (CLT)
  sample mean, 228–230
  sample proportion, 233–236
Central location, measures of
  Excel to calculate, 67–69
  mean, 62–64
  median, 64–65
  mode, 65–66
  weighted mean, 66
CGMA Economic Index, 140
Chance magazine, 485
Chance variation, in quality control, 240
Chartered Financial Analyst (CFA) designation, 273, 338
Charts and graphs. *See* Tabular and graphical methods
Chebyshev, Pavroty, 85
Chebyshev's theorem, 85–86
Chicago data portal (data.cityofchicago.org), 8–9
Children's Online Privacy Protection Act of 1998, 143
Chi-square (χ^2) distribution
  overview, 378–379
  table of, 514–515
  values and probabilities in, 379–383
Chi-square test for independence, 385–390
Classes, in frequency distributions, 27–29
Classical probabilities, 109, 111–112
Clinton, Hillary, 221–223
CLT (central limit theorem). *See* Central limit theorem (CLT)
Cluster sampling, 223
CNN, 307, 339
CNNMoney.com, 277
Coakley, Martha, 4
Coalition for Airline Passengers Rights, Health, and Safety, 169
Cobb-Douglas production function, 486
Coca-Cola, Inc., 142
Coefficients
  confidence (1-α), 262–263
  correlation, 92–94
  of determination (R^2), 417–419
  of determination (R^2), adjusted, 419–420
  nonzero slope, 425–426
  in quadratic regression model, 474
  regression (b_1 and b_0), 406–407
  sample correlation (Multiple R), 419
  of variation (CV), as measure of dispersion, 79–80
Cold chain distribution system, 214
Complement rule of probability, 113–114
Complements of events, 108
Conditional probability, 116–118
Confidence coefficient (1-α), 262–263
Confidence intervals. *See also* Interval estimation
  for difference between proportions, 372–373
  for difference between two means, 330–332
  in matched-pairs sampling, 341
  in two-tailed hypothesis testing, 304–305
Consumer confidence index, 151
Consumer Sentiment Index (University of Michigan), 499
Consumption function (Keynes), 410, 444
Contingency tables, 123–126, 385
Continuous probability distributions, 182–217. *See also* Discrete probability distributions; Probability
  continuous random variables and the uniform distribution, 184–187
  Excel for normal distribution, 199–200
  exponential distribution, 204–207
  introductory case study, 183, 199
  normal approximation of binomial distribution, 199
  normal distribution characteristics, 188–190
  normal random variables, transformation of, 195–199
  overview, 182
  probability, z value for, 193–195
  software packages on, 215–217
  standard normal distribution, 190–191
  writing with statistics on, 209–210
  z value, probability of, 191–193
Continuous random variables, 146, 184–187
Continuous uniform distribution, 185–187
Continuous variables, 10
Control charts, for quality control, 241–245
Correlated observations, as violation of linear regression model, 440–441
Correlation coefficient, 92–94, 419
Correlation-to-causation fallacy, 5
Costco, Inc., 307
Covariance, as measure of association, 92–94
Critical value approach to hypothesis testing, 324–326
Cross-sectional data, 6
Cubic regression model, 477–478
Cubic trend model, 491–492
Cumulative distribution function, 147, 185
Cumulative frequency distribution, 30
Cumulative relative frequency distribution, 31
Current Population Survey, 498

D

Daily Mail, 208
Data. *See also* Tabular and graphical methods for qualitative data; Tabular and graphical methods for quantitative data
  big data, 8
  cross-sectional, 6
  structured and unstructured, 7–8

time series, 6–7, 487
on Web, 8–9
Degrees of freedom *(df)*
in chi-square (χ^2) distribution, 378–379
in *t* distribution, 268–269
Dell Computer, Inc., 246
De Moivre, Abraham, 188*n*
Dependent events, 118–119
Descriptive measures, numerical. *See* Numerical descriptive measures
Descriptive statistics, 5
Detection approach to quality control, 240
Deterministic relationship between variables, 404
Discrete probability distributions, 144–181. *See also* Continuous probability distributions; Probability
binomial distribution, 156–161
Excel for binomial probabilities, 161–162
Excel for hypergeometric probabilities, 171–172
Excel for Poisson probabilities, 167–168
expected value in, 152
hypergeometric distribution, 169–171
introductory case study, 145, 167
overview, 144
Poisson distribution, 164–167
random variables and, 146–150
risk neutrality and risk aversion, 153–154
variance and standard deviation in, 152–153
writing with statistics on, 173–175
Discrete random variables, 146, 152
Discrete uniform distributions, 149
Discrete variables, 10
Dow Jones Industrial Average (DJIA), 5, 11, 42, 53–54, 76, 321
Dummy variables
interactions with, 467–470
qualitative explanatory variable with multiple categories, 461–464
qualitative explanatory variable with two categories, 458–461
seasonal, 495–497
Dunkin' Donuts, 219

E

Economist, The, 9
EEOC (Equal Employment Opportunity Commission), 169
Effects, marginal, 474
eMarketer.com, 139
Empirical probabilities, 109–112
Empirical rule, 85–86
Environmental Protection Agency (EPA), 281
Equal Employment Opportunity Commission (EEOC), 169
Error sum of squares (SSE), 353, 406
Estimates, 225, 254–255, 416–417. *See also* Interval estimation
Events
exhaustive, 107
independent and dependent, 118–119
mutually exclusive, 107, 115–116
total probability rule conditional on, 129
Venn diagrams to illustrate, 107–109
Excel
for binomial probabilities, 161–162
for chart construction of qualitative data, 24–25
for chi-square distribution, 382–383
for confidence intervals of population mean when variance is known, 265–266
for confidence intervals of population mean when variance is unknown, 271–272
for exponential distribution, 207
to generate simple random sample, 224
for histogram construction, 36–38
for hypergeometric probabilities, 171–172
for hypothesis testing in matched-pairs sampling, 344–345
for hypothesis testing of population mean when population variance is known, 305–306
for hypothesis testing of population mean when population variance is unknown, 309–310
for hypothesis test of difference between two means, 334–336
for measures of central location, 67–69
for measures of dispersion, 80
for multiple linear regression model, 412–413
for normal distribution, 199–200
for ogive construction, 38
for one-way ANOVA table construction, 355–356
for Poisson probabilities, 167–168
for polygon construction, 38
for regression coefficients, 406–407
for residual plot construction, 442
for scatterplot construction, 46, 408
for simple linear regression model, 408
Excluded variables, as violation of linear regression model, 441
Exhaustive events, 107
Expected frequencies, 386–390
Expected value
of binomial random variable, 159
in discrete probability distributions, 152
of hypergeometric random variable, 170
of Poisson random variable, 165
of sample mean, 226–227
of sample proportion, 232–236
Experiments
goodness-of-fit test for multinomial, 378–383
overview, 106–107

Explanatory variables, 404
Exponential distribution, 204–207
Exponential nonlinear regression models, 479–483
Exponential trend forecasting models, 487–490
Exponential trend model with seasonal dummy variables, 495–496

F

Facebook.com, 7–8, 162, 275, 391, 395, 449
Fahrenheit scale for temperature, 13–14
FCC (Federal Communications Commission), 237
F distribution
  in analysis of variance (ANOVA), 349–350
  table of, 516–519
Federal Bureau of Investigation (FBI), 213
Federal Reserve, 116, 448
FICO scores, 281
Fidelity Funds, 84, 98
Fidelity Gold Fund, 289
Fidelity Select Automotive Fund, 289
Fidelity's Electronics and Utilities funds, 274
Fidelity's Magellan mutual fund, 503
Fidelity's Select Electronic and Select Utilities mutual funds, 348
Fidelity's Strategic Income fund, 214
Field Poll, 143
FINA Congress, 366
Finite population correction factor, 237–239
Fisher, Ronald, 349*n*
Forbes magazine, 9
Fortune 500, 70–71, 81
Fortune magazine, 9
Frequency distributions
  cumulative, 30
  cumulative relative, 31
  of qualitative data, 20–21
  of quantitative data, 27–30
  relative, 21, 31

G

Gallup, George, 221
Gallup-Healthways Well-Being Index, 323
Gallup Organization, 5, 221
Gauss, Carl Friedrich, 188*n*
Gaussian distribution, 188
General Electric Corp., 6
Glassdoor.com, 481
Goodness-of-fit measures
  for multinomial experiments, 378–383
  in regression analysis, 416–420
Google.com, 8–9, 275, 449
GoogleFinance, 9
Gossett, William S., 268*n*
Graduate Record Examination (GRE), 94, 409
Grand mean of data set, 352
Graphs and charts. *See* Tabular and graphical methods
Great Depression of 1930s, 220
Great Recession of 2007–2008, 7, 312–313, 318, 320
Grouped data, 29, 89–90
Guinness Brewery, 268*n*
Gulf Oil, 169, 208

H

Haas School of Business, University of California at Berkeley, 224
Happiness index data, 9
Harris Interactive, 121, 127, 390
Harvard Medical School, 377
Harvard University, 322
Health of Boston, 337, 394
Helu, Carlos Slim, 43
Histograms, 32–33, 36–38
Home Depot, Inc., 282–283, 308
Hoover, Herbert, 220
Hypergeometric distribution, 169–172
Hypergeometric random variable, 170
Hypothesis testing, 292–327
  critical value approach to, 324–326
  of difference between proportions, 373–375
  of difference between two means, 332–336
  introduction to, 294–298
  introductory case study, 293, 310
  in matched-pairs sampling, 342–345
  overview, 292
  for population mean when population variance is known, 300–306
  for population mean when population variance is unknown, 308–310
  for population proportion, 313–315
  software on, 326–327
  for within treatments variance, 354
  writing with statistics on, 317–318

I

IDC, Inc., 287
Independence
  chi-square test for, 385–390
  of events, 118–120
  multiplication rule for, 387
Independent random samples, 330

Indiana University, 152
Individual significance,
tests of, 422–426
Inferential statistics, 5
Interaction variable, 467–470
Internal Revenue Service (IRS), 178
Interquartile range (IQR), 73–74
Intersection of two events, 108, 117
Interval estimation, 258–291
introductory case study, 259, 281
overview, 258
for population mean when population variance is known, 260–266
for population mean when population variance is unknown, 268–272
for population proportion, 275–277
sample size requirements, 278–280
software on, 290–291
writing with statistics on, 282–283
Interval scale of measurement, 13–14. *See also* Tabular and graphical methods for quantitative data
Inverse transformation, 197–199
IQR (interquartile range), 73–74
IRS (Internal Revenue Service), 178

J

Janus Capital Group, 101
JMP software
on binomial distribution, 181
for chart construction, 57–58
on comparison of means, 368
on control charts, 256
on exponential distribution, 217
on hypergeometric distribution, 181
on hypothesis testing, 327
on multicollinearity, 454
on multiple linear regression, 454
on normal distribution, 216–217
for numerical descriptive measures, 103
on Poisson distribution, 181
on population mean estimation, 290–291
on random samples, 256
on residual plots, 454
on simple linear regression, 454
on tests of independence, 400
Johnson & Johnson (J&J), 24, 425–426, 444
Joint probabilities, 125, 129, 131
Joint significance, tests of, 427–428
Journal of the American Medical Association, 113, 140, 395, 397

K

Kennedy, Ted, 4
Keynes, John Maynard, 410
Kiplinger's, 39

L

Landon, Alf, 220
Law of large numbers, 112
LCL (lower control limit), in control charts, 241–245, 247–248
Linear regression. *See* Regression analysis
Linear trend forecasting models, 487–490
Linear trend model with seasonal dummy variables, 495
LinkedIn.com, 7, 377
Literary Digest case of 1936 ("bad" sample), 220–221
Logarithmic regression model, 479–480
Logarithms in nonlinear regression models, 478–480
Log-linear regression models, linear *versus,* 483
Log-log regression model, 478–479
Los Angeles Lakers basketball team, 135
Lower control limit (LCL), in control charts, 241–245, 247–248
Lowe's Companies, Inc., 282–283, 504–505

M

MAD (mean absolute dispersion), 77–78
Major League Baseball (MLB), 101, 246, 410, 444, 508
Marginal effects, 474
Marginal probabilities, 125
Margin of error, 261
Massachusetts Community & Banking Council, 391
Matched-pairs sampling
confidence interval for mean difference in, 341
hypothesis testing for mean difference in, 342–345
overview, 340
recognizing, 341
McDonalds, Inc., 165, 204
Mean absolute dispersion (MAD), 77–78
Means, 328–369. *See also* Population mean *(μ);* Sample mean
difference between two, 330–336
differences among many, 349–356
of exponential distribution, 205
introductory case study, 329, 345
in matched-pairs sampling, 340–345
as measure of central location, 62–64
overview, 328
software on comparing, 367–369
writing with statistics on, 359–360

Mean square error (MSE), 353–354, 427
Mean square regression (MSR), 427
Mean squares for treatments (MSTR), 352–354
Mean-variance analysis, 81–83
Measurement scales
  interval, 13–14
  nominal, 11
  ordinal, 12–13
  ratio, 14–15
Measures, numerical descriptive. *See* Numerical descriptive measures
Median, as measure of central location, 64–65
Merck & Co., 127–128, 215
Merrill Lynch, 139
Method of least squares, 406
Michigan State University, 127, 338
Microsoft Corporation, 70, 176, 275, 449, 452
Minitab software
  on binomial distribution, 179
  for chart construction, 55–56
  on comparison of means, 367
  on control charts, 256
  on exponential distribution, 215
  on goodness-of-fit test, 399
  on hypergeometric distribution, 180
  on hypothesis testing, 326–327
  on multicollinearity, 453
  on multiple linear regression, 453
  on normal distribution, 215
  for numerical descriptive measures, 102
  on Poisson distribution, 180
  on population mean estimation, 290
  on population proportion estimation, 290
  on proportion difference testing, 399
  on random samples, 256
  on residual plots, 453
  on simple linear regression, 453
  on tests of independence, 399
  on uniform distribution, 215
MLB (Major League Baseball), 101, 246, 410, 444, 508
Mode, as measure of central location, 65–66
Money magazine, 51
Monster.com, 287
Morningstar ratings, for companies, 16
Mortgage Bankers Association, 202
MSE (mean square error), 353–354, 427
MSR (mean square regression), 427
MSTR (mean squares for treatments), 352–354
Multicollinearity, as violation of linear regression model, 436–438
Multinomial experiments, 378–383
Multiple linear regression analysis, 411–413
Multiple *R* (sample correlation coefficient), 419
Multiplication rule of probability, 119–120
Mutually exclusive events, 107, 115–116

N

National Association of Business Economists (NABE), 312
National Association of Colleges and Employers' Summer 2010 Salary Survey, 251
National Association of Securities Dealers Automated Quotations (NASDAQ), 11, 396
National Basketball Association (NBA), 41, 53
National Climatic Data Center (NCDC), 8
National Geographic Kids, 26
National Geographic News, 267
National Health and Nutrition Examination Survey, 135, 141
National High Blood Pressure Education Program, 212
National Hockey League (NHL), 253
National Institutes of Health (NIH), 473
National Science Foundation, 173, 338
National Sporting Goods Association (NSGA), 91, 99
NBA (National Basketball Association), 41, 53
NBC News/*Wall Street Journal* poll, 250, 277, 289
NBC-TV, 26
NCDC (National Climatic Data Center), 8
Negative linear relationship between variables, 405
Negatively skewed distribution, 33
New England Journal of Medicine, 339
New York City Youth Risk Behavior Survey, 178
New York Stock Exchange (NYSE), 11, 487
New York Times, 9, 178, 305
NHL (National Hockey League), 253
Nielsen, Inc., 169
NIH (National Institutes of Health), 473
Nike, Inc., 105, 124–126, 359, 371, 386, 389, 496–497
95% confidence intervals, 262
No linear relationship between variables, 405
Nominal scale of measurement, 11. *See also* Tabular and graphical methods for qualitative data
Nonlinear patterns, as violation of linear regression model, 435–436
Nonlinear regression models
  exponential, 480–483
  logarithms in, 478–480
  quadratic, 473–478
Nonresponse bias, in sampling, 221
Nonzero slope coefficient, 425–426
Normal distribution, 33, 188–191
Normal population, sampling from, 227–228
Normal random variables, transformation of, 195–199
NSGA (National Sporting Goods Association), 91, 99
Null hypothesis, 294–297, 354. *See also* Hypothesis testing
Numerical descriptive measures, 60–103
  association, measures of, 92–94
  boxplots, 73–75
  coefficient of variation, as measure of dispersion, 79–80
  Excel to calculate measures of central location, 67–69

Excel to calculate measures of dispersion, 80
grouped data, 89–90
introductory case study, 61, 83
mean, as measure of central location, 62–64
mean absolute dispersion, 77–78
mean-variance analysis and Sharpe ratio, 81–83
median, as measure of central location, 64–65
mode, as measure of central location, 65–66
overview, 60
percentiles, 71–73
range, as measure of dispersion, 76–77
relative location, analysis of, 84–87
symmetry, 69
variance and standard deviation, as measures of dispersion, 78–79
weighted mean, as measure of central location, 66
writing with statistics on, 95–96
NYSE (New York Stock Exchange), 11, 487

O

Obama, Barack, 27, 223, 250, 289
Obama, Michelle, 443
O'Connor dexterity test, 411
OECD (Organization for Economic Cooperation and Development), 285
Ogives, 35–36, 38
Ohio State University, The, 5
OLS (ordinary least squares) method, 406
One-tailed hypothesis test, 295–297, 302–305, 325
One-way ANOVA. *See* Analysis of variance (ANOVA)
Ordinal scale of measurement, 12–13. *See also* Tabular and graphical methods for qualitative data
Ordinary least squares (OLS) method, 406
Organization for Economic Cooperation and Development (OECD), 285
Outliers, 63, 73–74

P

Panera Bread Co., 81, 274
Parameter, population mean as, 63
$\bar{p}$ chart, to monitor proportion of defects, 241
PEG (price-to-earnings growth) ratio, 53
Percent frequency, 21
Percentiles, 71–73
Pew Forum on Religion & Public Life, 398
Pew Research Center, 135–136, 141, 162, 251, 288, 376, 391
Pie charts, 21–22. *See also* Tabular and graphical methods for qualitative data
Point estimators, 225, 254–255. *See also* Interval estimation
Poisson, Simeon, 164
Poisson distribution, 164–168, 204–205
Poisson random variable, 165
Polling, statistics in, 4
Polygons, 34, 38
Polynomial regression model, 477
Polynomial trend forecasting models, 490–493
Population
finite population correction factor, 237–239
inferential statistics to draw conclusions on, 5–6
parameters to describe, 5–6
sample *versus,* 220
sampling from normal, 227–228
Population mean *(μ). See also* Interval estimation
calculating, 63
hypothesis testing when population variance is known, 300–306
hypothesis testing when population variance is unknown, 308–310
interval estimation when population variance is known, 260–266
interval estimation when population variance is unknown, 268–272
normal distribution described by, 189
sample size required to estimate, 279–280
Population proportion
confidence interval estimation for, 275–277
hypothesis testing for, 313–315
sample size required to estimate, 280
software to estimate, 290
Population standard deviation, 263–264
Population variance (δ^2). *See also* Interval estimation
hypothesis testing for population mean with known, 300–306
hypothesis testing for population mean with unknown, 308–310
normal distribution described by, 189
Positive linear relationship between variables, 405
Positively skewed distribution, 33
Posterior probability, 131
Powerball jackpot game, 173
Precision of confidence interval width, 263–264
Price-to-earnings growth (PEG) ratio, 53
Princeton University, 485, 499
Prior probability, 131
Probability, 104–143. *See also* Continuous probability distributions; Discrete probability distributions; Interval estimation
addition rule of, 114–116
assigning, 109–113
Bayes' theorem, 131–134
chi-square (χ^2) values and, 379–383
complement rule of, 113–114
conditional, 116–118
contingency tables and, 123–126

Probability—*Cont.*
definition and terminology, 106
of events, 107–109, 118–119
F distribution values and, 349–350
introductory case study, 105, 126
multiplication rule of, 119–120
overview, 104
total probability rule, 128–131
writing with statistics on, 135–137
Probability density function, 147, 184, 189–190
Probability mass function, 147
Probability trees, 129–131, 157–158
Proportions, 370–401. *See also* Population proportion; Sample proportion
chi-square test for independence, 385–390
difference between two, 372–375
goodness-of-fit test for multinomial experiments, 378–383
introductory case study, 371, 389
overview, 370
software guidelines for, 399–400
writing with statistics on, 392–393
*p*th percentile, 72–73
Putnam's mutual funds, 99
p-value
computer-generated test statistic and, 425
hypothesis testing with, 300–304, 354

Q

Quadratic nonlinear regression models, 473–478
Quadratic trend model, 490–492, 497
Qualitative explanatory variable
with multiple categories, 461–464
with two categories, 458–461
Qualitative variables. *See also* Tabular and graphical methods for qualitative data
definition of, 10
nominal and ordinal scales for, 11–13
population proportion described by, 313
in regression, 458
Quantitative variables. *See also* Tabular and graphical methods for quantitative data
definition of, 10
interval and ratio scales for, 13–14
population mean and standard deviation described by, 313
in regression, 458

R

Random samples
independent, 330
simple, 222, 224
stratified, 222–223
Random variables *(X)*
binomial, 156–157
continuous, 184–187
discrete probability distributions and, 146–150
hypergeometric, 170
overview, 146–150
Poisson, 165
transformation of normal, 195–199
Range, as measure of dispersion, 76–77
Rate parameter (λ) of exponential distribution, 205
Ratio scale of measurement, 14–15. *See also* Tabular and graphical methods for quantitative data
R chart, to monitor variability, 241
Regression analysis, 402–455
Excel for residual plot construction, 442
goodness-of-fit measures, 416–420
introductory case study, 403, 429
linear *versus* log-linear models, 483
model assumptions, 433–435
model violations, 435–442
multiple linear, 411–413
overview, 402
reporting results of, 429
simple linear, 404–408
software guidelines on, 453–455
tests of individual significance, 422–426
tests of joint significance, 427–428
writing with statistics on, 444–445
Regression analysis, extensions of, 456–509
dummy variables, interactions with, 467–470
exponential model of nonlinear relationships, 480–483
introductory case study, 457, 470
logarithms in nonlinear regression models, 478–480
overview, 456
quadratic regression models of nonlinear relationships, 473–478
qualitative explanatory variable with multiple categories, 461–464
qualitative explanatory variable with two categories, 458–461
trend forecasting models, 487–492
trend forecasting models with seasonality, 495–497
writing with statistics on, 499–501
Regression sum of squares (SSR), 418–419
Relative frequency distributions, 21, 31
Relative location, analysis of, 84–87
Residual plots, 434–435, 442
Response variables, 404
Risk neutrality and risk aversion, 153–154
Romney, Mitt, 27
Roosevelt, Franklin D., 220
R software
on binomial distribution, 181
on comparison of means, 369
on control charts, 257

on exponential distribution, 217
on goodness-of-fit test, 400
on hypergeometric distribution, 181
on hypothesis testing, 327
on multicollinearity, 455
on multiple linear regression, 455
on normal distribution, 217
for numerical descriptive measures, 103
on Poisson distribution, 181
on population mean estimation, 291
on random samples, 257
on residual plots, 455
on simple linear regression, 455
on tests of independence, 400
on uniform distribution, 217

S

Sample correlation coefficient (Multiple *R*), 419
Sample mean
central limit theorem for, 229–230
derivation of, 253–254
description of, 63
finite population correction factor for, 237–238
sampling distribution of, 225–230
Sample proportion
central limit theorem for, 233–236
derivation of, 254
finite population correction factor for, 238–239
sampling distribution of, 232–236
Sample regression equation, 405–407, 411
Sample space, 106
Sample statistics, 5
Sampling, 218–257
"bad" sample (*Literary Digest* case of 1936), 220–221
Excel to generate simple random sample, 224
finite population correction factor, 237–239
inferential statistics to draw conclusions based on, 5–6
introductory case study, 219, 236
methods of, 222–223
overview, 218
point estimator properties, 254–255
random independent, 330
sample mean derivation, 253–254
sample proportion derivation, 254
sample size requirements, 278–280
sampling distribution of the sample mean, 225–230
sampling distribution of the sample proportion, 232–236
software packages for, 255–257
in statistical quality control, 240–245
Trump victory in 2016 and, 221–222
writing with statistics on, 246–247
Sarkozy, Nicolas, 236
SAT scores, percentiles of, 71–72
Scales of measurement, 11–15
Scatterplots, 44–46
s chart, to monitor variability, 241
Search engines, for Web data, 8–9
Sears, Inc., 151, 176–177
Seasonal component, in trend forecasting models, 495–497
Second Skins, Inc., 105
Securities and Exchange Commission (SEC), 239
Selection bias, in sampling, 221
Semi-log regression model, 479
Seton Hall University, 457, 470
Sharpe, William, 82
Sharpe ratio, 81–83
Shewhart, Walter A., 241
Significance
of dummy variable, 460–461
individual, 422–426
joint, 427–428
level of, 301, 354
Simple linear regression analysis, 404–408
Simple random sample, 222, 224
Skewed distributions, 33, 69
Social-desirability bias, 222
Social media, unstructured data on, 7
Sperling Manufacturing, 338
Spine Patient Outcomes Research Trial (SPORT), 122
Sporting Goods Manufacturers Association, 358
SPSS software
on binomial distribution, 179
for chart construction, 56–57
on comparison of means, 368
on control charts, 256
on exponential distribution, 216
on goodness-of-fit test, 399
on hypergeometric distribution, 180
on hypothesis testing, 327
on multicollinearity, 454
on multiple linear regression, 453
on normal distribution, 216
for numerical descriptive measures, 103
on Poisson distribution, 180
on population mean estimation, 290
on residual plots, 453–454
on simple linear regression, 453
on tests of independence, 400
on uniform distribution, 216
Spurious correlation, 5
SSE (error sum of squares), 353, 406
SSR (regression sum of squares), 418–419
SST (total sum of squares), 355
SSTR (sum of squares due to treatments), 352

Standard deviation
of binomial random variable, 159
confidence interval width and population, 263–264
in discrete probability distributions, 152–153
of hypergeometric random variable, 170
as measure of dispersion, 78–79
of Poisson random variable, 165
Standard error
central limit theorem for, 233–236
of the estimate, 416–417
of the estimator, 261
of sample mean, 226–227
of sample proportion, 232–236
Standardizing data, with *z*-scores, 87
Standard normal curve, table of, 190, 510–511
Standard normal probability density function, 191
Standard transformation, 195–197
Starbucks, Inc., 4, 81, 142, 146, 166–167, 219, 230, 236–237, 329, 345
Statistic, sample mean as, 63
Statistical Abstract of the United States, 2010, 51
Statistical quality control, 240–245
Statistics, introduction to, 2–17
definition of, 5–9
introductory case, 3, 15
measurement scales in, 11–15
overview, 2
relevance of, 4–5
variables in, 10–11
Stem-and-leaf diagrams, 42–44
Stochastic relationship between variables, 404
Stock's alpha *(α),* 425–426
Stratified random sample, 222–223
Structured data, 7–8
Student's *t* distribution, 268–272, 512–513
Subjective probabilities, 109–110, 112
Sum of squares due to treatments (SSTR), 352
Symmetric distribution, 33, 189
Symmetry, 69
Syracuse University, 173

T

Tables
chi-square (χ^2) distribution, 514–515
chi-square (χ^2) test of contingency, 385–386
contingency, 123–126
F distribution, 516–519
one-way ANOVA, 355–356
standard normal curve, 190, 510–511
Student's *t* distribution, 512–513
z values, 190–192, 194
Tabular and graphical methods for qualitative data
Excel for chart construction, 24–25
frequency distributions, 20–21
interpreting charts and graphs, 24
introductory case study, 19, 32
overview, 18
pie charts and bar charts, 21–23
Tabular and graphical methods for quantitative data
frequency distributions, 27–30
histograms, 32–33, 36–38
introductory case study, 19, 32
ogives, 35–36, 38
overview, 18
polygons, 34, 38
relative frequency distributions, 31
in reports, 47–49
scatterplots, 44–46
stem-and-leaf diagrams, 42–44
Target Stores, Inc., 282
t distribution, 268–272
Temperature, Fahrenheit scale of, 13–14
Test statistic
computer-generated, *p*-value and, 425
for difference between proportions, 373–375
for difference between two means, 333–334
for goodness-of-fit test for multinomial experiments, 381–382
for matched-pairs sampling, 342–345
for one-way ANOVA, 354
for population mean when population variance is known, 300–301
for population mean when population variance is unknown, 308–309
for population proportion, 314
for test for independence, 388
for test of individual significance, 423–425
Texas Transportation Institute, 359
Texas Workforce Commission, 288
Time series data, 6–7, 487. *See also* Trend forecasting models
Total probability rule, 128–131, 133
Total sum of squares (SST), 355
Trader Joe's, Inc., 251
tradingeconomics.com, 274
Transamerica Center for Health Studies, 163
Trend forecasting models
linear and exponential, 487–490
overview, 487
polynomial, 490–493
with seasonality, 495–497
TrueCar online car buying system, 494
True zero point, in ratio scales, 14
Trump, Donald, 151, 221–222
Tukey, John, 42

Twitter.com, 7
Two-tailed hypothesis test, 295–297, 302–303, 325
Type I (α) and Type II (β) errors, in hypothesis testing, 297–298, 301

U

UCL (upper control limit), in control charts, 241–245, 247–248
Unconditional probability, 117
Under Armour, Inc., 105, 123–126, 371, 386, 389
"Underwater" mortgages, 177
Uniform distribution, 184–187
Union of two events, 107, 114–115
University of California, Berkeley, 224
University of California, Davis, 486
University of Illinois, 250
University of Michigan, 8, 287, 499
University of New Hampshire, 286
University of Notre Dame Mendoza College of Business, 164
University of Pennsylvania Medical Center, 4
University of Utah, 398
University of Wisconsin, 66
Unstructured data, 7–8
Upper control limit (UCL), in control charts, 241–245, 247–248
USA Today, 9, 392
USA Today/Gallup Poll, 27
U.S. Census Bureau, 6, 8, 64, 164, 338, 481
U.S. Census Current Population Survey, 142
U.S. Department of Health and Human Services (HHS), 504
U.S. Department of Transportation, 81, 177, 394

V

Vanguard Balanced Index Fund, 200
Vanguard's Balanced Index and European Stock Index mutual funds, 362
Vanguard's Growth and Value Index mutual funds, 61, 83–84
Vanguard's Precious Metals and Mining Fund, 214, 321
Variability changes, as violation of linear regression model, 438–439
Variables. *See also* Dummy variables
  binomial random, 156–157
  continuous, 10
  continuous random, 184–187
  description of, 10–11
  excluded, 441
  explanatory, 404
  hypergeometric random, 170
  interaction, 467–470
  Poisson random, 165
  qualitative and quantitative, 313
  random, 146–150
  response, 404
  transformation of normal random, 195–199
Variance. *See also* Analysis of variance (ANOVA); Population variance (δ^2)
  of binomial random variable, 159
  in discrete probability distributions, 152–153
  of hypergeometric random variable, 170
  hypothesis testing for population mean with known population, 300–306
  hypothesis testing for population mean with unknown population, 308–310
  interval estimation for population mean with known population, 260–266
  interval estimation for population mean with unknown population, 268–272
  as measure of dispersion, 78–79
  of Poisson random variable, 165
Variation, coefficient of, 79–80
Variation, in quality control, 240
Venn, John, 107
Venn diagrams
  for addition rule of probability, 114–116
  for conditional probability, 117
  to illustrate events, 107–109
  for total probability rule, 129
Vodafone, Ltd., 316
Vons Supermarkets, 224

W

Wall Street Journal, 9, 52, 101
Walmart Stores, Inc., 508
Walt Disney, Inc., 443
Washington Post–Kaiser Family Foundation, 142
Watson Wyatt consulting, 177
Wayne State University, 465
Web, data on, 8–9
Weighted mean, as measure of central location, 66
Wharton School of Business, 339
WHO (World Health Organization), 465, 472, 493
Within treatments variance. *See* Analysis of variance (ANOVA)
Woodrow Wilson School, Princeton University, 499
World Cup Soccer, 44
World Health Organization (WHO), 465, 472, 493
World Wealth Report, The, 384
Writing with statistics
  continuous probability distributions, 209–210
  discrete probability distributions, 173–175

Writing with statistics—*Cont.*
hypothesis testing, 317–318
interval estimation, 282–283
means, 359–360
numerical descriptive measures in, 95–96
probability, 135–137
proportions, 392–393
regression analysis, 444–445
regression analysis extensions, 499–501
sampling, 246–247
tabular and graphical methods for, 47–49

X

$\bar{x}$ chart, to monitor central tendency, 241

Y

YahooFinance, 9
YouTube.com, 7

Z

Zillow.com, 9
z values
in confidence interval estimation for population mean, 263
for given probability, 193–195
probability of, 191–193
z-Scores, 86–87
z table of, 190–192, 194

Mc Graw Hill Education
5011333