



Michael Sullivan III

FUNDAMENTALS OF

STATISTICS

Informed Decisions
Using Data

SIXTH EDITION



Sullivan's Pathway to Making an Informed Decision

Begin your journey . . .

- **Making an Informed Decision** projects at the start of each chapter allow you to work with data in order to make informed decisions that impact your life.
- **Putting It Together** overviews show how material you are about to cover relates to prior material.

Preparation is key . . .

- **Preparing for This Section** lists all of the skills needed to be successful.
- **Preparing for This Section Quizzes** are available as a digital MyLab assignment or as a print quiz to help you check your mastery.
- **Each Objective** is listed at the beginning of the section and then repeated in the text for easy reference.

Look at the model then practice, practice, practice . . .

- **Step-by-Step Annotated Examples** illustrate new concepts and methods in 3 steps:
 1. Problem
 2. Approach
 3. Solution
- **Examples** point to **Now Work Exercises** so you can solve similar exercises on your own.

Exercise Sets . . .

- **Putting It Together** exercises use skills you've acquired in various chapters. (*See facing page*)
- **You Explain It!** exercises ask you to provide an interpretation of statistical results.
- **Threaded Tornado Problems** allow you to analyze a single data set throughout the entire semester. (*See facing page*)
- **Retain Your Knowledge** exercises help you to maintain the skills you have acquired earlier in the course.

Check where you've been and test your mastery . . .

- **Putting It Together Sections** require you to decide which technique to use. (*See facing page*)
- **End-of-Chapter Objectives** are listed with page references for easy review.
- **Chapter Tests** provide an opportunity to test your knowledge.

Apply yourself . . .

- **In-Class Activities** in the Student Activity Workbook allow you to experience statistics in a fun and exciting way by experiencing the process firsthand.
- **Making an Informed Decision** projects require you to use data and statistical techniques learned in the chapter to make important life decisions.
- **End-of-Chapter Case Studies** tie statistical concepts together within an interesting application.

Sullivan's Guide to Putting It Together

Putting It Together Sections	Objective	Page(s)
5.7 Putting It Together: Which Method Do I Use?	① Determine the appropriate probability rule to use ② Determine the appropriate counting technique to use	287–289 289–290
9.3 Putting It Together: Which Method Do I Use?	① Determine the appropriate confidence interval to construct	425–426
10.4 Putting It Together: Which Method Do I Use?	① Determine the appropriate hypothesis test to perform (one sample)	470
11.4 Putting It Together: Which Method Do I Use?	① Determine the appropriate hypothesis test to perform (two samples)	513
Putting It Together Exercises	Skills Utilized	Section(s) Covered
1.2.26 Passive Smoke	Variables, observational studies, designed experiments	1.1, 1.2
1.4.37 Comparing Sampling Methods	Simple random sampling and other sampling techniques	1.3, 1.4
1.4.38 Thinking about Randomness	Random sampling	1.3, 1.4
2.1.29 Online Homework	Variables, designed experiments, bar graphs	1.1, 1.2, 1.6, 2.1
2.2.42 Time Viewing a Webpage	Graphing data	2.2
2.2.43 Red Light Cameras	Variables, population vs. sample, histograms, dot plots	1.1, 2.2
2.2.44 Which Graphical Summary?	Choosing the best graphical summary	2.1, 2.2
2.2.45 Shark!	Graphing data	2.3
3.1.42 Shape, Mean, and Median	Discrete vs. continuous data, histograms, shape of a distribution, mean, median, mode, bias	1.1, 1.4, 2.2, 3.1
3.5.18 Paternal Smoking	Observational studies, designed experiments, lurking variables, mean, median, standard deviation, quartiles, boxplots	1.2, 1.6, 3.1, 3.2, 3.4, 3.5
3.5.19 Taxi Ride	Bar graphs, histograms, boxplots, range, standard deviation	2.1, 2.2, 3.2, 3.5
4.2.29 Housing Prices	Scatter diagrams, correlation, linear regression	4.1, 4.2
4.2.30 Smoking and Birth Weight	Observational study vs. designed experiment, prospective studies, scatter diagrams, linear regression, correlation vs. causation, lurking variables	1.2, 4.1, 4.2
4.3.16 Exam Scores	Building a linear model	4.1, 4.2, 4.3
4.3.17 Cigarette Smuggling	Scatter diagrams, correlation, least-squares regression	4.1, 4.2, 4.3
4.4.15 Sullivan Survey II	Relative frequency distributions, bar graphs, pie charts, contingency tables, conditional distributions	2.1, 4.4
5.1.52 Drug Side Effects	Variables, graphical summaries of data, experiments, probability	1.1, 1.6, 2.1, 5.1
5.2.44 Speeding Tickets	Contingency tables, marginal distributions, empirical probabilities	4.4, 5.1
5.2.45 Red Light Cameras	Variables, relative frequency distributions, bar graphs, mean, standard deviation, probability, Simpson's Paradox	1.1, 2.1, 3.1, 3.2, 4.4, 5.1, 5.2
6.1.37 Sullivan Statistics Survey I	Mean, standard deviation, probability, probability distributions	3.1, 3.2, 5.1, 6.1
6.2.55 A Drug Study	Types of variables, experimental design; binomial probabilities	1.1, 1.2, 1.6, 6.2
6.2.56 Beating the Stock Market	Expected value, binomial probabilities	6.1, 6.2
7.2.52 Birth Weights	Relative frequency distribution, histograms, mean and standard deviation from grouped data, normal probabilities	2.1, 2.2, 3.3, 7.2
7.3.13 Disney's Dinosaur Ride	Histograms, distribution shape, normal probability plots	2.2, 7.3
8.1.34 Bike Sharing	Histograms, mean, standard deviation, distribution shape, sampling distribution of the mean	2.2, 3.1, 3.2, 8.1
8.1.35 Playing Roulette	Probability distributions, mean and standard deviation of a random variable, sampling distributions	6.1, 8.1
9.1.47 Hand Washing	Observational studies, bias, confidence intervals	1.2, 1.5, 9.1
9.2.47 Smoking Cessation Study	Experimental design, confidence intervals	1.6, 9.1, 9.2
10.2.40 Lupus	Observational studies, retrospective vs. prospective studies, bar graphs, confidence intervals, hypothesis testing	1.2, 2.1, 9.1, 10.2
10.2.41 Naughty or Nice?	Experimental design, determining null and alternative hypotheses, binomial probabilities, interpreting <i>P</i> -values	1.6, 6.2, 10.1, 10.2
11.1.36 Salk Vaccine	Completely randomized design, hypothesis testing	1.6, 11.1

(continued)

Putting It Together Exercises	Skills Utilized	Section(s) Covered	Page(s)
11.2.19 Glide Testing	Matched pairs design, hypothesis testing	1.6, 11.2	500–501
11.3.23 Online Homework	Completely randomized design, confounding, hypothesis testing	1.6, 11.3	512
12.1.27 The V-2 Rocket in London	Mean of discrete data, expected value, Poisson probability distribution, goodness-of-fit	6.1, 6.3, 12.1	537
12.1.28 Weldon's Dice	Addition Rule for Disjoint Events, classical probability, goodness-of-fit	5.1, 5.2, 12.1	537
12.2.22 Women, Aspirin, and Heart Attacks	Population, sample, variables, observational study vs. designed experiment, experimental design, compare two proportions, chi-square test of homogeneity	1.1, 1.2, 1.6, 11.1, 12.2	552–553
12.2.23 Corequisite College Algebra	Comparing two independent means, comparing two independent proportions, chi-square test for independence	11.1, 11.3, 12.2	553
12.4.19 Predicting Intelligence	Scatter diagrams, linear correlation coefficient, least-squares regression, normal probability plots, inference on least-squares regression, confidence and prediction intervals	4.1, 4.2, 4.3, 7.3, 12.3, 12.4	574
B.3.27 Psychological Profiles	Standard deviation, sampling methods, two-sample <i>t</i> -test, Central Limit Theorem, one-way Analysis of Variance	1.4, 3.2, 8.1, 11.2, B.3	B-28

Threaded Tornado Problems

Throughout the text a single, large data set that measures various variables on all tornadoes that struck the United States in 2017 is utilized. The problems are marked with a  icon. The table below shows the sections, problems, topics covered, and page for the Threaded Tornado Problems.

Section	Problem(s)	Topics	Page(s)
1.1	47, 48	Types of variables; types of data	13
2.1	25	Frequency & relative frequency distributions; bar charts; pie charts	74–75
2.2	41	Frequency & relative frequency distributions; histogram; dot plots	91
3.1	41	Mean, median, distribution shape	120
3.2	51	Range, standard deviation	138
3.4	29	Quartiles, interquartile range, outliers	155
3.5	20	Boxplots	163
4.3	15	Scatter diagrams, correlation, least-squares regression, coefficient of determination, residual analysis	205
5.1	49	Probability models; unusual events	240
8.1	33	Describe the distribution of the sample mean from a non-normal population	382–383
9.1	33	Confidence interval for a population proportion	408
9.2	37	Confidence interval for a population mean	422–423
10.2	33	Hypothesis test for a population proportion	456
10.2B	25	Hypothesis test for a population proportion	10.2AB.24
10.3	35	Hypothesis test for a population mean	468
11.1	29	Compare two population proportions (independent samples)	489
11.3	17	Compare two population means (independent samples)	511
12.3	17	Inference on least-squares regression; prediction intervals	573–574
B.3	29	One-way Analysis of Variance (ANOVA)	B-28

FUNDAMENTALS OF STATISTICS

INFORMED DECISIONS USING DATA 6E

Michael Sullivan, III
Joliet Junior College



Content Management: Suzanna Bainbridge
Content Production: Robert Carroll, Jean Choe, Peggy McMahon
Product Management: Karen Montgomery
Product Marketing: Alicia Wilson
Rights and Permissions: Pavithra Gunasekaran, Integra

Please contact <https://support.pearson.com/getsupport/s/contactsupport> with any queries on this content.

MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS MAKE NO REPRESENTATIONS ABOUT THE SUITABILITY OF THE INFORMATION CONTAINED IN THE DOCUMENTS AND RELATED GRAPHICS PUBLISHED AS PART OF THE SERVICES FOR ANY PURPOSE. ALL SUCH DOCUMENTS AND RELATED GRAPHICS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND. MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS HEREBY DISCLAIM ALL WARRANTIES AND CONDITIONS WITH REGARD TO THIS INFORMATION, INCLUDING ALL WARRANTIES AND CONDITIONS OF MERCHANTABILITY, WHETHER EXPRESS, IMPLIED OR STATUTORY, FITNESS FOR A PARTICULAR PURPOSE, TITLE AND NON-INFRINGEMENT. IN NO EVENT SHALL MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF INFORMATION AVAILABLE FROM THE SERVICES.

THE DOCUMENTS AND RELATED GRAPHICS CONTAINED HEREIN COULD INCLUDE TECHNICAL INACCURACIES OR TYPOGRAPHICAL ERRORS. CHANGES ARE PERIODICALLY ADDED TO THE INFORMATION HEREIN. MICROSOFT AND/OR ITS RESPECTIVE SUPPLIERS MAY MAKE IMPROVEMENTS AND/OR CHANGES IN THE PRODUCT(S) AND/OR THE PROGRAM(S) DESCRIBED HEREIN AT ANY TIME. PARTIAL SCREEN SHOTS MAY BE VIEWED IN FULL WITHIN THE SOFTWARE VERSION SPECIFIED.

MICROSOFT® WINDOWS®, AND MICROSOFT OFFICE® ARE REGISTERED TRADEMARKS OF THE MICROSOFT CORPORATION IN THE U.S.A. AND OTHER COUNTRIES. THIS BOOK IS NOT SPONSORED OR ENDORSED BY OR AFFILIATED WITH THE MICROSOFT CORPORATION.

Copyright © 2022, 2018, 2014 by Pearson Education, Inc. or its affiliates, 221 River Street, Hoboken, NJ 07030. All Rights Reserved. Manufactured in the United States of America. This publication is protected by copyright, and permission should be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise. For information regarding permissions, request forms and the appropriate contacts within the Pearson Education Global Rights & Permissions Department, please visit www.pearsoned.com/permissions/. Acknowledgements of third party content appear on page C-1, which constitutes an extension of this copyright page.

PEARSON, ALWAYS LEARNING, MYLAB™ are exclusive trademarks owned by Pearson Education, Inc. or its affiliates in the United States and/or other countries.

Unless otherwise indicated herein, any third-party trademarks that may appear in this work are the property of their respective owners and any references to third-party trademarks, logos or other trade dress are for demonstrative or descriptive purposes only. Such references are not intended to imply any sponsorship, endorsement, authorization, or promotion of Pearson's products by the owners of such marks, or any relationship between the owner and Pearson Education, Inc. or its affiliates, authors, licensees or distributors.

Library of Congress Cataloging-in-Publication Data

Names: Sullivan, Michael, III, 1967- author.

Title: Fundamentals of statistics : informed decisions using data / Michael Sullivan, III, Joliet Junior College.

Description: 6th edition. | Hoboken : Pearson, 2022. | Includes index. |

Summary: "Popular college introductory text to the subject of Statistics that follows GAISE guidelines"-- Provided by publisher.

Identifiers: LCCN 2020035728 | ISBN 9780136807346 (hardback)

Subjects: LCSH: Statistics--Textbooks.

Classification: LCC QA276.12 .S846 2022 | DDC 519.5--dc23

LC record available at <https://lccn.loc.gov/2020035728>

ScoutAutomatedPrintCode



ISBN 10: 0-13-680734-8
ISBN 13: 978-0-13-680734-6

To My Wife Yolanda
and My Children
Michael, Kevin, and Marissa

This page intentionally left blank

Contents

Preface to the Instructor ix

Resources for Success xv

Applications Index xvii

PART 1 Getting the Information You Need 1



Data Collection 2

- 1.1 Introduction to the Practice of Statistics 3
 - 1.2 Observational Studies versus Designed Experiments 14
 - 1.3 Simple Random Sampling 23
 - 1.4 Other Effective Sampling Methods 30
 - 1.5 Bias in Sampling 38
 - 1.6 The Design of Experiments 44
- Chapter 1 Review* 55
Chapter Test 59
Making an Informed Decision: What College Should I Attend? 60

PART 2 Descriptive Statistics 61



Summarizing Data in Tables and Graphs 62

- 2.1 Organizing Qualitative Data 63
 - 2.2 Organizing Quantitative Data 76
 - 2.3 Graphical Misrepresentations of Data 92
- Chapter 2 Review* 100
Chapter Test 104
Making an Informed Decision: Tables or Graphs? 106



Numerically Summarizing Data 107

- 3.1 Measures of Central Tendency 108
 - 3.2 Measures of Dispersion 121
 - 3.3 Measures of Central Tendency and Dispersion from Grouped Data 139
 - 3.4 Measures of Position and Outliers 145
 - 3.5 The Five-Number Summary and Boxplots 155
- Chapter 3 Review* 164
Chapter Test 167
Making an Informed Decision: What Car Should I Buy? 169



Describing the Relation between Two Variables 170

- 4.1 Scatter Diagrams and Correlation 171
- 4.2 Least-Squares Regression 187
- 4.3 The Coefficient of Determination 201
- 4.4 Contingency Tables and Association 206

Chapter 4 Review 217

Chapter Test 222

Making an Informed Decision: Relationships among Variables on a World Scale 224

PART 3 Probability and Probability Distributions 225



Probability 226

- 5.1 Probability Rules 227
- 5.2 The Addition Rule and Complements 242
- 5.3 Independence and the Multiplication Rule 253
- 5.4 Conditional Probability and the General Multiplication Rule 260
- 5.5 Counting Techniques 269
- 5.6 Simulating Probability Experiments 282
- 5.7 Putting It Together: Which Method Do I Use? 287

Chapter 5 Review 293

Chapter Test 297

Making an Informed Decision: The Effects of Drinking and Driving 298



Discrete Probability Distributions 299

- 6.1 Discrete Random Variables 300
- 6.2 The Binomial Probability Distribution 312

Chapter 6 Review 327

Chapter Test 330

Making an Informed Decision: Should We Convict? 331



The Normal Probability Distribution 332

- 7.1 Properties of the Normal Distribution 333
- 7.2 Applications of the Normal Distribution 343
- 7.3 Assessing Normality 355
- 7.4 The Normal Approximation to the Binomial Probability Distribution 360

Chapter 7 Review 365

Chapter Test 367

Making an Informed Decision: Stock Picking 368

PART 4 Inference: From Samples to Population 369**Sampling Distributions 370**

- 8.1** Distribution of the Sample Mean 371
 - 8.2** Distribution of the Sample Proportion 384
- Chapter 8 Review 392*
Chapter Test 393
Making an Informed Decision: How Much Time Do You Spend in a Day . . . ? 394

**Estimating the Value of a Parameter 395**

- 9.1** Estimating a Population Proportion 396
 - 9.2** Estimating a Population Mean 410
 - 9.3** Putting It Together: Which Method Do I Use? 425
 - 9.4** Estimating a Population Standard Deviation (eText)
 - 9.5** Estimating with Bootstrapping (eText)
- Chapter 9 Review 429*
Chapter Test 432
Making an Informed Decision: How Much Should I Spend for this House? 433

**Hypothesis Tests Regarding a Parameter 434**

- 10.1** The Language of Hypothesis Testing 435
 - 10.2** Hypothesis Tests for a Population Proportion 443
 - 10.2A** Using Simulation to Perform Hypothesis Tests on a Population Proportion (eText)
 - 10.2B** Hypothesis Tests for a Population Proportion Using the Normal Model (eText)
 - 10.3** Hypothesis Tests for a Population Mean 458
 - 10.3A** Using Simulation and the Bootstrap to Perform Hypothesis Tests on a Population Mean (eText)
 - 10.4** Putting It Together: Which Method Do I Use? 470
 - 10.5** Hypothesis Tests for a Population Standard Deviation (eText)
- Chapter 10 Review 472*
Chapter Test 475
Making an Informed Decision: Selecting a Mutual Fund 476



Inference on Two Population Parameters 477

- 11.1** Inference about Two Population Proportions 478
- 11.1A** Using Randomization Techniques to Compare Two Proportions (eText)
- 11.2** Inference about Two Means: Dependent Samples 490
- 11.2A** Using Bootstrapping to Conduct Inference on Two Dependent Means (eText)
- 11.3** Inference about Two Means: Independent Samples 501
- 11.3A** Using Randomization Techniques to Compare Two Independent Means (eText)
- 11.4** Putting It Together: Which Method Do I Use? 513
- 11.5** Inference about Two Population Standard Deviations (eText)
 - Chapter 11 Review* 517
 - Chapter Test* 520
- Making an Informed Decision: Which Car Should I Buy?* 522



Additional Inferential Methods 523

- 12.1** Goodness-of-Fit Test 524
- 12.2** Tests for Independence and the Homogeneity of Proportions 538
- 12.3** Testing the Significance of the Least-Squares Regression Model 554
- 12.3A** Using Randomization Techniques on the Slope of the Least-Squares Regression Line (eText)
- 12.4** Confidence and Prediction Intervals 569
 - Chapter 12 Review* 575
 - Chapter Test* 579
- Making an Informed Decision: Benefits of College* 581

Credits C-1

Appendix A Tables A-1

Appendix B (eText) B-1

B.1 Lines B-1

B.2 Inference about Two Population Proportions: Dependent Samples B-6

B.3 Comparing Three or More Means (One-Way Analysis of Variance) B-12

Answers ANS-1

Subject Index I-1

Tables and Formulas for Sullivan, *Fundamentals of Statistics*, 6e

Preface to the Instructor

Capturing a Powerful and Exciting Discipline in a Textbook

Statistics is a powerful subject, and it is one of my passions. Bringing my passion for the subject together with my desire to create a text that would work for me, my students, and my school led me to write the first edition of this textbook. It continues to motivate me as I reflect on changes in students, in the statistics community, and in the world around us.

When I started writing, I used the manuscript of this text in class. My students provided valuable, insightful feedback, and I made adjustments based on their comments. In many respects, this text was written by students and for students. I also received constructive feedback from a wide range of statistics faculty, which has refined ideas in the book and in my teaching. I continue to receive valuable feedback from both faculty and students, and this text continues to evolve with the goal of providing clear, concise, and readable explanations, while challenging students to think statistically.

In writing this edition, I continue to make a special effort to abide by the Guidelines for Assessment and Instruction in Statistics Education (GAISE) for the college introductory course endorsed by the American Statistical Association (ASA). The GAISE Report gives six recommendations for the course:

1. Emphasize statistical literacy and develop statistical thinking
2. Use real data in teaching statistics
3. Stress conceptual understanding
4. Foster active learning
5. Use technology for developing conceptual understanding
6. Use assessments to improve and evaluate student learning

Changes to this edition and the hallmark features of the text reflect a strong adherence to these important GAISE guidelines.

New to This Edition

- **Over 350 New and Updated Exercises** The sixth edition makes a concerted effort to require students to write a few sentences that explain the results of their statistical analysis. To reflect this effort, the answers in the back of the text provide recommended explanations of the statistical results. Not all the exercises are computational or require statistical analysis. Many of the exercises have been written to require students to explain statistical concepts or understand pitfalls in faulty statistical analysis.
- **Over 100 New and Updated Examples** The examples continue to engage and provide clear, concise explanations for the students while following the *Problem, Approach, Solution* presentation. Problem lays out the scenario of the example, Approach provides insight into the thought process behind the methodology used to solve the problem, and Solution goes through the solution utilizing the methodology suggested in the approach.
- **Threaded Tornado Problems** Throughout the text a single, large data set that measures various variables on all tornadoes that struck the United States in 2017 is utilized. The problems are marked with a  icon. The table on the front inside cover shows the sections, problems, topics covered and pages for the Threaded Tornado Problems. In addition, the author wrote corresponding MyLab problems around this data set. The problems may serve as a semester-long project for your students.
- **Updated MyLab Problems** New MyLab problems written by Michael Sullivan utilize real data that is randomly generated from a larger data set. He also wrote new applet exercises that allow students to explore statistical concepts.
- **Optional Simulation & Randomization Sections** Simulation and randomization methods are a new approach to hypothesis testing. New to this edition are optional sections on using simulation to test hypotheses for a population proportion (Section 10.2A) and population mean (Section 10.3A), and randomization methods for testing hypotheses on two independent proportions (Section 11.1A), two independent means (Section 11.3A), and the slope of the least-squares regression model (Section 12.3A).
- **Classroom Notes** Written by Heidi Lyne and Michael Sullivan, new to this edition are classroom notes, which may be used by the instructor to deliver lectures to students. Students may print these notes out and bring them to the classroom, which facilitates good note-taking and allows them to focus on the concepts. The examples and activities in the classroom notes are different from those in the text and Instructor's Resource Guide.
- **Videos** New lightboard videos featuring the author, Michael Sullivan, develop statistical concepts for students. New animated videos explain concepts or tie material learned earlier in the course with the upcoming chapter or section. And finally, new Excel video solutions for any example in which Excel may be used to obtain statistical results are available.
- **R Technology Guide** Written by Patrick Murphy (nephew of the author) and Michael Sullivan, the R Technology Guide provides a chapter-by-chapter discussion of R commands needed for each topic. The R Technology Guide may be found under Learning Tools in MyLab or at www.sullystats.com/statistics-6e/r-guidebook.
- **Learning Catalytics** Learning Catalytics allows students to use their own mobile devices in the classroom for real-time engagement. Search "SullivanStats" in Learning Catalytics to add pre-made questions written by Michael Sullivan for Sullivan's *Statistics* series.

Hallmark Features

- **Putting It Together** When students are learning statistics, they often struggle with seeing the big picture of how it all fits together. One of my goals is to help students learn not just the important concepts and methods of statistics but also how to put them together and see how the methods work together. On the inside front cover, you'll see a pathway that provides a guide for students as they navigate through the process of learning statistics. The features and chapter organization in the sixth edition reinforce this important process. There are two categories of "Putting It Together."
 - **Putting It Together Sections** appear in Chapters 5, 9, 10, and 11. The problems in these sections are meant to help students identify the correct approach to solving a problem. Many exercises in these sections mix in inferential techniques from earlier sections. Plus, there are problems that require students to identify the inferential technique that may be used to answer the research objective (but no analysis is required). For example, see Problems 20 to 25 in Section 10.4.
 - **Putting It Together Problems** appear throughout the text. The purpose of these problems is to tie concepts together and see the entire statistical process. For example, problems on hypothesis testing may require students to first identify the data collection method (such as observational study or designed experiment, the explanatory and response variables, the role of randomization, the role of control) prior to completing the data analysis.
- **Student Activity Workbook** The student activity workbook now contains an outline for a semester-long project and suggestions for how to use the StatCrunch survey tool to develop a survey that could result in a semester-long project. Plus, there are ten new activities included in the activity workbook along with suggested answers in the corresponding instructor's guide.
- **Retain Your Knowledge** These problems occur periodically at the end of section exercises and are meant to assist students in retaining skills learned earlier in the course. This way, the material is fresh for the final exam.
- **MyLab Technology Help** Online homework problems that may be analyzed using statistical packages now have an updated technology help feature. Marked with a  icon, this feature provides step-by-step instructions on how to obtain results using StatCrunch, TI-84 Plus/TI-84 Plus CE, and Excel.
- **Instructor's Resource Guide** Written by Michael Sullivan, the Instructor's Resource Guide provides an overview of the chapter. It also details points to emphasize within each section and suggestions for presenting the material. In addition, the guide provides examples that may be used in the classroom. Many new examples have been added to this edition.
- Because the use of **Real Data** piques student interest and helps show the relevance of statistics, great efforts have been made to extensively incorporate real data in the exercises and examples.
- **Step-by-Step Annotated Examples** guide a student from problem to solution in three easy-to-follow steps.
- "Now Work" problems follow most examples so students can practice the concepts shown.
- Multiple types of **Exercises** are used at the end of sections and chapters to test varying skills with progressive levels of difficulty. These exercises include **Vocabulary and Skill Building**, **Applying the Concepts**, and **Explaining the Concepts**.
- **Chapter Review** sections include:
 - **Chapter Summary**.
 - A list of key chapter **Vocabulary**.
 - A list of **Formulas** used in the chapter.
 - **Chapter Objectives** listed with corresponding review exercises.
 - **Review Exercises** with all answers available in the back of the book.
 - **Chapter Test** with all answers available in the back of the book. In addition, the Chapter Test problems have **video solutions** available.
- Each chapter has **Case Studies** available at www.pearsonhighered.com/sullivanstats that help students apply their knowledge and promote active learning.

Integration of Technology

This book can be used with or without technology. Should you choose to integrate technology in the course, the following resources are available for your students:

- Technology Step-by-Step guides are included in applicable sections that show how to use Minitab®, Excel®, the TI-83/84, and StatCrunch to complete statistics processes. The Technology Step-by-Step for StatCrunch was written by Michael Sullivan.
- Any problem that has 12 or more observations in the data set has a  icon indicating that data set is included on the companion website (<http://www.pearsonhighered.com/sullivanstats>) in various formats.
- Where applicable, exercises and examples incorporate output screens from various software including Minitab, the TI-83/84 Plus CE, Excel, and StatCrunch.
- Applets are included on the companion website and connected with certain activities from the Student Activity Workbook, allowing students to manipulate data and interact with animations.
- A technology manual is available that contains detailed tutorial instructions and worked out examples and exercises for the TI-83/84. There is also a new R Technology Manual should you choose to incorporate R into your class.

Companion Website Contents

The companion website is
<http://www.pearsonhighered.com/sullivanstats>.

- Data Sets
- Applets
- Formula Cards and Tables in PDF format
- Additional Topics Folder including:
 - Sections 9.4, 9.5, 10.2A, 10.2B, 10.3A, 10.5, 11.1A, 11.2A, 11.3A, 11.5, 12.3A
 - Appendix A
 - Appendix B
 - Lines
 - Inference about Two Proportions (Dependent Samples)
 - Comparing Three or More Means (One-Way Analysis of Variance)
- Case Studies for each chapter in the text.
- A copy of the questions asked on the Sullivan Statistics Survey I and Survey II
- Consumer Reports projects that were formerly in the text
- The author has also created a website at <https://www.sullystats.com>. This site has chapter-by-chapter suggestions for teaching the material, links to interesting data sets, and much more.

Key Chapter Content Changes

Chapter 1 Data Collection

Section 1.2 now includes a discussion of obtaining data through web scraping and how to obtain data from the Internet. Section 1.6 expands on the discussion of the placebo effect.

Chapter 5

Section 5.1 now distinguishes the Law of Large Numbers from the nonexistent Law of Averages. There is a new Section 5.6 on simulating probability experiments. This material is very helpful in allowing students to see the role of randomness in probability experiments. It also foreshadows topics such as sampling distributions and inference.

Chapter 9

There is an expanded discussion on the normality condition for constructing confidence intervals for the population mean using Student's t -distribution in Section 9.2.

Chapter 10

Chapter 10 now contains optional sections on simulation methods for conducting inference. The organization of Chapter 10 allows for presenting simulation along with traditional inference, or simply presenting traditional inference. Should you decide to present only the traditional approach to inference, simply cover Section 10.2 from the text.

If you decide to present hypothesis testing using simulation, skip Section 10.2 in the text and cover Sections 10.2A and 10.2B (available in MyLab or the companion website as pdfs). Section 10.3A (MyLab) presents hypothesis testing on a mean using simulation and bootstrapping. This section is optional and may be skipped without loss of continuity.

Chapter 11

Chapter 11 has new optional sections on randomization methods. Section 11.1A (available in MyLab or the companion website as a pdf) presents randomization tests for two independent proportions. If you choose to present randomization methods, we recommend presenting Section 11.1A prior to Section 11.1. Section 11.2A (MyLab) presents hypothesis tests on dependent means using bootstrapping. This section is optional and may be skipped without loss of continuity. Section 11.3A (MyLab) presents randomization tests for two independent means. We recommend covering this section prior to Section 11.3, if you choose to discuss this approach to hypothesis testing.

Chapter 12

Chapter 12 has a new optional section on randomization. Section 12.3A (available in MyLab or the companion website as a pdf) presents randomization tests for the slope of the least-squares regression model. If you choose to cover this section, do so prior to Section 12.3.

Flexible to Work with Your Syllabus

To meet the varied needs of diverse syllabi, this book has been organized to be flexible.

You will notice the “Preparing for This Section” material at the beginning of each section, which will tip you off to dependencies within the course. The two most common variations within an introductory statistics course are the treatment of regression analysis and the treatment of probability.

- **Coverage of Correlation and Regression** The text was written with the descriptive portion of bivariate data (Chapter 4) presented after the descriptive portion of univariate data (Chapter 3). Instructors who prefer to postpone the discussion of bivariate data can skip Chapter 4 and return to it before covering Sections 12.3 and 12.4.
- **Coverage of Probability** The text allows for light to extensive coverage of probability. Instructors wishing to minimize probability may cover Section 5.1 and skip the remaining sections. A mid-level treatment of probability can be accomplished by covering Sections 5.1 through 5.3. Instructors who will cover the chi-square test for independence will want to cover Sections 5.1 through 5.3. In addition, an instructor who will cover binomial probabilities will want to cover independence in Section 5.3 and combinations in Section 5.5.

Acknowledgments

Textbooks evolve into their final form through the efforts and contributions of many people. First and foremost, I would like to thank my family, whose dedication to this project was just as much as mine: my wife, Yolanda, whose words of encouragement



and support were unabashed, and my children, Michael, Kevin, and Marissa, who have been supportive throughout their childhood and now into adulthood (my how time flies). I owe each of them my sincerest gratitude. I would also like to thank the entire Mathematics Department at Joliet Junior College and my colleagues who provided support, ideas, and encouragement to help me complete this project. From Pearson Education: I thank Suzanna Bainbridge, whose editorial expertise has been an invaluable asset; Peggy McMahon, who provided

organizational skills that made this project go smoothly; Alicia Wilson and Demetrius Hall, for their marketing savvy and dedication to getting the word out; Vicki Dreyfus and Jean Choe, for their dedication in organizing all the media; Rose Kernan for her ability to control the production process; Dana Bettez for her editorial skill with the Instructor's Resource Guide and Student Activity Workbook; and the Pearson sales team, for their confidence and support of this book.

I also want to thank Ryan Cromar, Susan Herring, Craig Johnson, Kathleen McLaughlin, Patrick Murphy, Heidi Lyne, and Dorothy Wakefield for their help in creating supplements. A big thank-you goes to Cindy Trimble and Associates, who assisted in verifying answers for the back of the text and helped in proofreading. I would also like to acknowledge Kathleen Almy and Heather Foes for their help and expertise in developing the Student Activity Workbook. Finally, I would like to thank George Woodbury for helping me with the incredible suite of videos that accompanies the text. Many thanks to all the reviewers, whose insights and ideas form the backbone of this text. I apologize for any omissions.

CALIFORNIA Charles Biles, Humboldt State University • Carol Curtis, Fresno City College • Jacqueline Faris, Modesto Junior College • Freida Ganter, California State University–Fresno • Jessica Kramer, Santiago Canyon College • Sherry Lohse, Napa Valley College • Craig Nance, Santiago Canyon College • Diane Van Deusen, Napa Valley College **COLORADO** Roxanne Byrne, University of Colorado–Denver • Monica Geist, Front Range Community College **CONNECTICUT** Kathleen McLaughlin, Manchester Community College • Dorothy Wakefield, University of Connecticut • Cathleen M. Zucco Teveloff, Trinity College **DISTRICT OF COLUMBIA** Monica Jackson, American University • Jill McGowan, Howard University **FLORIDA** Randall Allbritton, Daytona Beach Community College • Greg Bloxom, Pensacola State College • Anthony DePass, St. Petersburg College Clearwater • Kelcey Ellis, University of Central Florida • Franco Fedele, University of West Florida • Laura Heath, Palm Beach Community College • Perrian Herring, Okaloosa Walton College • Marilyn Hixson, Brevard Community College • Daniel Inghram, University of Central Florida • Philip Pina, Florida Atlantic University • Mike Rosenthal, Florida International University • James Smart, Tallahassee Community College **GEORGIA** Virginia Parks, Georgia Perimeter College • Chandler Pike, University of Georgia • Jill Smith, University of Georgia • John Weber, Georgia Perimeter College **HAWAII** Eric Matsuoka at Leeward Community College • Leslie Rush, University of Hawaii **IDAHO** K. Shane Goodwin, Brigham Young University • Craig Johnson, Brigham Young University • Brent Timothy, Brigham Young University • Kirk Trigsted, University of Idaho **ILLINOIS** Grant Alexander, Joliet Junior College • Kathleen Almy, Rock Valley College • John Bialas, Joliet Junior College • Linda Blanco, Joliet Junior College • Kevin Bodden, Lewis & Clark Community College • Rebecca Goad, Joliet Junior College • Joanne Brunner, Joliet Junior College • Robert Capetta, University of Illinois at Chicago • Elena Catoiu, Joliet Junior College • Faye Dang, Joliet Junior College • Laura Egner, Joliet Junior College • Jason Eltrevoog, Joliet Junior College • Erica Egizio, Lewis University • Heather Foes, Rock Valley College • Randy Gallaher, Lewis & Clark Community College • Melissa Gaddini, Robert Morris University • Iraj Kalantari, Western Illinois University • Donna Katula, Joliet Junior College • Diane Koenig, Rock Valley College • Diane Long, College of DuPage • Heidi Lyne, Joliet Junior College • Jean McArthur, Joliet Junior College • Patricia McCarthy, Robert Morris University • David McGuire, Joliet Junior College • Angela McNulty, Joliet Junior College • James Morgan, Joliet Junior College • Andrew Neath, Southern Illinois University–Edwardsville • Linda Padilla, Joliet Junior College • David Ruffato, Joliet Junior College • Patrick Stevens, Joliet Junior College • Robert Tuskey, Joliet Junior College • Stephen Zuro, Joliet Junior College **INDIANA** Susitha Karunaratne, Purdue University North Central • Jason Parcon, Indiana University–Purdue University Ft. Wayne • Henry Wakhungu, Indiana University **KANSAS** Donna Gorton, Butler Community College • Ingrid Peterson, University of Kansas **LOUISIANA** Melissa Myers, University of Louisiana at Lafayette **MARYLAND** Nancy Chell, Anne Arundel Community College • John Climent, Cecil Community College • Rita Kolb, The Community College of Baltimore County • Jignasa Rami, Community College of Baltimore County • Mary Lou Townsend, Wor-Wic Community College **MASSACHUSETTS** Susan McCourt, Bristol Community College • Daniel Weiner, Boston University • Pradipta Seal, Boston University of Public Health **MICHIGAN** Margaret M. Balachowski, Michigan Technological University • Diane Krasnewich, Muskegon Community College • Susan Lenker, Central Michigan University • Timothy D. Stebbins, Kalamazoo Valley Community College • Sharon Stokero, Michigan Technological University • Alana Tuckey, Jackson Community College **MINNESOTA** Mezbhur Rahman, Minnesota State University **MISSOURI** Carroll Tim

Wright, University of Missouri–Columbia **NEBRASKA** Jane Keller, Metropolitan Community College **NEW YORK** Jacob Amidon, Finger Lakes Community College • Stella Aminova, Hunter College • Jennifer Bergamo, Onondaga Community College • Kathleen Cantone, Onondaga Community College • Pinyuen Chen, Syracuse University • Sandra Clarkson, Hunter College of CUNY • Rebecca Daggar, Rochester Institute of Technology • Bryan Ingham, Finger Lakes Community College • Anne M. Jowsey, Niagara County Community College • Maryann E. Justinger, Erie Community College–South Campus • Bernadette Lanciaux, Rochester Institute of Technology • Kathleen Miranda, SUNY at Old Westbury • Robert Sackett, Erie Community College–North Campus • Sean Simpson, Westchester Community College • Bill Williams, Hunter College of CUNY **NORTH CAROLINA** Fusan Akman, Coastal Carolina Community College • Mohammad Kazemi, University of North Carolina–Charlotte • Janet Mays, Elon University • Marilyn McCollum, North Carolina State University • Claudia McKenzie, Central Piedmont Community College • Said E. Said, East Carolina University • Karen Spike, University of North Carolina–Wilmington • Jeanette Szwee, Cape Fear Community College **NORTH DAKOTA** Myron Berg, Dickinson State University • Ronald Degges, North Dakota State University **OHIO** Richard Einsporn, The University of Akron • Michael McCraith, Cuyahoga Community College **OREGON** Daniel Kim, Southern Oregon University • Jong Sung Kin, Portland State University **PENNSYLVANIA** Mary Brown, Harrisburg Area Community College • LeAnne Conaway, Harrisburg Area Community College **SOUTH CAROLINA** Diana Asmus, Greenville Technical College • Dr. William P. Fox, Francis Marion University • Cheryl Hawkins, Greenville Technical College • Rose Jenkins, Midlands Technical College • Lindsay Packer, College of Charleston • Laura Shick, Clemson University • Erwin Walker, Clemson University **TENNESSEE** Tim Britt, Jackson State Community College • Nancy Pevey, Pellissippi State Technical Community College • David Ray, University of Tennessee–Martin **TEXAS** Edith Aguirre, El Paso Community College • Ivette Chuca, El Paso Community College • Aaron Gutknecht, Tarrant County College • Jada Hill, Richland College • David Lane, Rice University • Alma F. Lopez, South Plains College • Shanna Moody, University of Texas at Arlington • Jasdeep Pannu, Lamar University **UTAH** Joe Gallegos, Salt Lake City Community College • Alia Maw, Salt Lake City Community College **VIRGINIA** Kim Jones, Virginia Commonwealth University • Vasanth Solomon, Old Dominion University **WEST VIRGINIA** Mike Mays, West Virginia University **WISCONSIN** William Applebaugh, University of Wisconsin–Eau Claire • Carolyn Chapel, Western Wisconsin Technical College • Beverly Dretzke, University of Wisconsin–Eau Claire • Jolene Hartwick, Western Wisconsin Technical College • Thomas Pomykalski, Madison Area Technical College • Walter Reid, University of Wisconsin–Eau Claire

Michael Sullivan, III
Joliet Junior College



Pearson
MyLab

MyLab Statistics is available to accompany Pearson's market-leading text options, including *Fundamentals of Statistics*, 6e by Michael Sullivan (access code required).

MyLab™ is the teaching and learning platform that empowers you to reach every student. MyLab Statistics combines trusted author content—including full eText and assessment with immediate feedback—with digital tools and a flexible platform to personalize the learning experience and improve results for each student. Integrated with StatCrunch®, a web-based statistical software program, students learn the skills they need to interact with data in the real world.

Integrated Review for Corequisite Courses

This MyLab™ includes an additional eText written by the author on prerequisite skills and concepts. There are also prebuilt (and editable) MyLab quizzes that populate personalized homework assignments for gaps in skills for that chapter. These resources may be used in a corequisite course model, or simply to help underprepared students master prerequisite skills and concepts.

The screenshot shows Chapter 2 of the eText titled "Integrated Review: Getting Ready for Numerically Summarizing Data". The chapter outline includes sections on Exponents and the Order of Operations, Square Roots, Simplifying Algebraic Expressions, and Summation Notation. A specific section, 2.IR1 Exponents and the Order of Operations, is highlighted. It contains objectives for evaluating exponential expressions and applying the rules of order of operations. A callout box provides an example of using exponential notation for repeated multiplication, comparing it to writing out the expression 3 multiplied by itself eight times.

The screenshot shows Examples 1 and 2 from the eText. Example 1 discusses a human study where the explanatory variable of interest was cell phone usage. Example 2 discusses a rat study where the explanatory variable was radiofrequency radiation, with three possible levels: no RFR, GSM, or CDMA. The examples are presented with cartoon illustrations of a woman and a man.

The screenshot shows a video player interface. A man in a checkered shirt is speaking while pointing at a scatter plot on a screen. The scatter plot has four quadrants labeled I, II, III, and IV. A regression line is drawn through the points. Below the plot, the formula for the correlation coefficient $r = \sqrt{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$ is displayed. The video player controls at the bottom include a play button, a progress bar showing 02:08 / 04:06, and other standard video controls.

NEW! Videos

In addition to existing Author in the Classroom, StatTalk, and Example Videos, the following **video types** were added to this edition.

- **Innovative lightboard videos** featuring Mike Sullivan guide students towards deeper conceptual understanding of certain key topics.
- **Excel video solutions** were added to the existing suite of StatCrunch, TI-83/84 Plus, and by-hand videos for examples in the text.
- **Animation videos** remind students of where they have been and where they are going in the upcoming objectives.



Resources for Success

Student Resources

Each student learns at a different pace. Personalized learning pinpoints the precise areas where each student needs practice, giving all students the support they need — when and where they need it — to be successful.

StatCrunch

StatCrunch® is powerful web-based statistical software that allows users to collect, crunch, and communicate with data. The vibrant online community offers tens of thousands of shared data sets for students and instructors to analyze, in addition to all of the data sets in the text or online homework. StatCrunch is integrated directly into MyLab Statistics or it can be purchased separately. Learn more at www.statcrunch.com.

Data Sets

All data sets from the textbook are available in MyLab Statistics. They can be analyzed in StatCrunch or downloaded for use in other statistical software programs.

Statistical Software Support

Instructors and students can copy data sets from the text and MyLab Statistics exercises directly into software such as StatCrunch or Excel®. Students can also access instructional support tools including tutorial videos, Study Cards, and manuals for a variety of statistical software programs including StatCrunch, Excel, Minitab®, JMP®, R, SPSS, and TI 83/84 calculators.

Student Solutions Manual

This manual provides detailed, worked-out solutions to all odd-numbered text exercises, as well as all solutions for the Chapter Reviews and Chapter Tests. It is available in print and can be downloaded from MyLab Statistics. (ISBN-13: 9780136969761)

Instructor Resources

Your course is unique. So whether you'd like to build your own assignments, teach multiple sections, or set prerequisites, MyLab gives you the flexibility to easily create your course to fit your needs.

Annotated Instructor's Edition

Includes answers to all text exercises, as well as teaching tips and common student errors. (ISBN-13: 9780136807315; ISBN-10: 0136807313)

Instructor Solutions Manual

Contains worked-out solutions to all text exercises and case study answers. It can be downloaded from MyLab Statistics or from www.pearson.com.

PowerPoint Lecture Slides

PowerPoint Lecture Slides include key graphics and follow the sequence and philosophy of the text. They can be downloaded from MyLab Statistics or from www.pearson.com.

TestGen

TestGen® (www.pearson.com/testgen) enables instructors to build, edit, print, and administer tests using a computerized bank of questions developed to cover all the objectives of the text. TestGen is algorithmically based, allowing instructors to create multiple but equivalent versions of the same question or test with the click of a button. Instructors can also modify test bank questions or add new questions. The software and test bank are available for download from Pearson's online catalog, www.pearson.com. The questions are also assignable in MyLab Statistics.

Data Analytics / Early Alerts

Instructors have a comprehensive gradebook with enhanced reporting functionality that makes it easier to understand which students are struggling, and which topics they struggle with most. **New Early Alerts** use predictive analytics to identify struggling students as early as possible—even if their assignment scores are not a cause for concern.

Resources for Success



Pearson
MyLab

Instructor Resources (continued)

Question Libraries

MyLab Statistics includes a number of question libraries that instructors can incorporate into their regular assignments. StatCrunch Projects consist of questions about large data sets in StatCrunch. The Conceptual Question Library offers 1,000 questions to help students learn concepts and how to think critically. Finally, the StatTalk Video Library includes questions associated to the video series by statistician Andrew Vickers.

Minitab and Minitab Express™

Bundling Minitab software with educational materials ensures students have access to the software they need in the classroom, around campus, and at home. And having 12-month access to Minitab and Minitab Express ensures students can use the software for the duration of their course. (ISBN-13: 9780134456409; ISBN-10: 0134456408) (access card only; not sold as stand alone.)

JMP Student Edition

An easy-to-use, streamlined version of JMP desktop statistical discovery software from SAS Institute, Inc. is available for bundling with the text. (ISBN-13: 9780134679792; ISBN-10: 0134679792)

XLSTAT™

An Excel add-in that enhances the analytical capabilities of Excel. XLSTAT is used by leading businesses and universities around the world. It is available to bundle with this text. For more information go to www.pearsonhighered.com/xlstatupdate. (ISBN-13: 9780321759320; ISBN-10: 032175932X)

Accessibility

Pearson works continuously to ensure our products are as accessible as possible to all students. We are working toward achieving WCAG 2.0 Level AA and Section 508 standards, as expressed in the Pearson Guidelines for Accessible Educational Web Media, www.pearson.com/mylab/statistics/accessibility.

Applications Index

Accounting

client satisfaction, 25–27

Aeronautics

moonwalkers, 11
O-ring failures on Columbia, 114
Spacelab, 503

Agriculture

corn production, B24
optimal level of fertilizer, 48–49
orchard damage, 59
yield
 of orchard, 37
 soybean, 134, B24

Animals/Nature

American black bears, weight and length of, 182–183, 185, 205, 567, 573
shark attacks, 91–92, 98, 219

Anthropometrics

upper arm length, 341
upper leg length, 341, 380

Astronomy

life on Mars, 298

Banking

ATM withdrawals, 382, 465
credit-card debt, 409, 474
credit cards, 390, 471

Biology

alcohol effects, 53, 58
alopecia, 497
blood types, 237
cholesterol level, 38, 520
DNA sequences, 280, 298
growth plates, 352
HDL cholesterol, 423, 568
hemoglobin
 in cats, 154
LDL cholesterol, B27
reaction time, 53, 58, 497–498, 510
testosterone levels, 472, 509

Business. *See also Work*

acceptance sampling, 267, 281, 390
advertising
 campaign, 37
 effective commercial, 101
 humor in, 58
airline customer opinion, 36
bolts production, 57, 152–153
buying new cars, 537
car rentals, 499
car sales, 86
CEO performance, 183, 198, 567, 573
coffee sales, 296
customer satisfaction, 32–33
customer service, 359
Disney World statistics conference, 240
employee morale, 37
entrepreneurship, 393
name vs. store brand, 488
new store opening decision, 42
oil change time, 381

online groceries, 74
packaging error, 267, 281, 296
quality control, 36, 37, 59, 258, 393–394, 471
salaries, 120, 136
shopping habits of customers, 42
shopping online, 73, 118–119, 154
Speedy Lube, 354
stocks on the NASDAQ, 280
stocks on the NYSE, 280
Target demographic information gathering, 38
traveling salesperson, 280
unemployment and inflation, 87
union membership, 98
waiting in line, 309, 498
waiting time for restaurant seating, 89
worker injury, 99
worker morale, 30

Chemistry

calcium in rainwater, 467
diversity and pH, B28
pH in rain, 421–422, 521, B27
pH in water, 117, 134
reaction time, 340, 497–498, 510

Combinatorics

arranging flags, 296
clothing option, 280
combination locks, 280
committee, 267
committee formation, 280
committee selection, 281
license plate numbers, 280, 296
passwords, 298
seating arrangements, 291
starting lineups, 291

Communication(s)

cell phone, 58
 in bathroom, 408
 body mass index and, 22–23
 brain tumors and, 14–15
 conversations, 488
 crime rate and, 186
 servicing, 441
 screen time, 421
do-not-call registry, 43
e-mail, 430
high-speed Internet service, 57, 408
length of phone calls, 340
newspaper article analysis, 13–14, 22, 221–222
social media, 252, 267
teen, 268
text messaging
 number of texts, 73
 while driving, 471
voice-recognition systems, B10–B11

Computer(s). *See also Internet*

defective chips, 269
download time, 37
DSL Internet connection speed, 37
e-mail, 430
FBI ID numbering, 281–282
fingerprint identification, 259
passwords, 282
toner cartridges, 168
user names, 280

Construction

concrete, 205
concrete mix, 116, 133
new road, 104

Consumers

Coke or Pepsi preferences, 54
taste test, 21

Crime(s)

burglaries, 92–93
fingerprints, 259
fraud identity, 71–72
larceny theft, 238–239
murder conviction by race, 216
rate of cell phones, 186
robberies, 98
speeding, 38
weapons used in murder/homicide, 101

Criminology

FBI ID numbering, 281–282
fraud detection, 154, 155

Demographics

births
 live, 101–102, 310
 per capita income and rates of, 105
 per woman, 91
 proportion born each day of week, B25
childless women, 390
family size, 100
handedness and mortality, 23
households speaking foreign language as primary language, 42
left-handedness, 410
life expectancy, 9, 91, 185, 258
living alone, 455, 537
marital status and happiness, 216
number of live births, 50- to 54-year-old mothers, 310
population
 age of, 145
 of selected countries, 9
 shifts in, 536
southpaws, 257

Drugs. *See also Pharmaceuticals*

AndroGel, 408
Aspirin, 552–553
Celebrex, 551
experimental, 324
marijuana use, 291
Nexium, 454
thalidomide on TEN, 327
Viagra, 241
Zoloft, 521

Economy

abolishing the penny, 408
health care expenditures, 99
poverty, 71
unemployment and inflation, 87
unemployment rates, 220

Education. *See also Test(s)*

advanced degrees, 393
bachelor's degree, elapsed time to earn, 509

birthrate and, 181
 board work, 267
 bullying in schools, 11
 calculus exam score, 11
 class average, 144
 college
 application, 168
 campus safety, 31
 complete rate, 441, 490
 drug use among students, 11
 exam skills, 512
 literature selection, 29
 survey, 72–73, 238
 textbook packages required, 42
 corequisite remediation and study skills, 552, 553
 course redesign, 471
 course selection, 29
 day care, 3-year-old, 237
 delivery methods, B24–B25
 developmental math, 51
 dropping course, 552
 exam grades/scores, 58, 117, 119, 133, 205–206, 253
 study time, 196
 exam time, 116, 133
 faculty evaluation, 186
 faculty opinion poll, 29
 gender bias in grades, 516–517
 GPA, 103, 144, 167
 vs. seating choice, 577–578
 grade distribution, 536
 grade inflation, 103
 graduation rates, 186, 199, 432, 512
 health and, 550
 illicit drug use among students, 11, 42
 income and, 196, 215
 invest in, 567, 573
 journal costs, 118
 learning community curriculum, 455–456
 level of (educational attainment), 96–97, 215
 marriage and, 292
 mathematics
 studying college, 428, 475
 teaching, 455
 TIMMS exam, 182
 TIMS report and Kumon, 514
 music's impact on learning, 49–50
 online homework, 75, 512
 premature birth and, 577
 quality of, 455
 reading and math ability of fourth-graders, 407
 school
 admissions, 152
 dropouts, 266
 e-cigs usage, 442, 551
 enrollment, 88
 illegal drug use in, 11
 multitasking, 515–516
 National Honor Society, 292
 seat selection in classroom, 296
 student loans, 359, 471
 seating arrangements, 534–535, 577–578
 self-injurious behaviors, 330
 self-study format, 475
 sleep disorders and academic performance, 521
 student age, 469
 student loans, 90
 student opinion poll/survey, 29, 37, 38

student retention, 475
 student services fees, 38
 study time, 468
 teacher evaluations, 515
 teaching reading, 51
 time spent on homework, 105
 typical student, 106
 visual vs. textual learners, 511

Electricity

Christmas lights, 257
 light bulbs, 166, 281

Electronics

televisions in the household, 89

Energy

carbon dioxide emissions and energy production, 196–197
 consumption, 466
 gas price hike, 100
 oil reserves, 99
 during pregnancy, 392–393

Engineering

batteries and temperature, 54
 battery charge life, 393
 bearing failure rate, 168
 bolts production, 57
 concrete strength, 566, 573, B27
 driving under the influence (DUI) simulator, 499
 glide testing, 500–501
 hardness testing, 498–499
 linear rotary bearing, 474
 O-ring thickness, 359
 Prolong engine treatment, 442
 ramp metering, 510
 tire design, 54
 triple modular redundancy, 258
 valve pressure, 441

Entertainment. *See also* Leisure and recreation

Academy Award winners, 382
 Disney's Dinosaur Ride, 359
 movie popcorn, 510
 movie ratings, 56
 neighborhood party, 267
 People Meter measurement, 35
 raffle, 13
 social drinking, 59
 student survey, 154
 television
 in bedroom, obesity and, 21
 hours of watching, 359, 423
 luxury or necessity, 407
 number of, 393
 watching, 382
 theme park spending, 427
 tickets to concert, 24–25

Environment

Flint water crisis, 162
 pH in rain, 421–422, B27
 Secchi disk, 432, 498

Exercise

gender differences in, 490
 routines, 292

Family

gender income inequality, 456
 ideal number of children, 90, 102, 310, 431, 472
 imprisoned members of, 432
 infidelity among married men, 455
 smarter kids, 471
 spanking, 326
 structure, 250, 550
 values, 407

Farming. *See also* Agriculture

incubation times for hen eggs, 340–341, 352, 353

Fashion

women's preference for shoes, 103

Finance. *See also* Investment(s)

ATM withdrawals, 382, 465–466
 cash/credit, 514–515
 cigarette tax rates, 90, 144, 206
 cost
 of kids, 99
 credit-card debt, 409, 474
 credit cards, 311, 390, 471
 credit scores, 205, 552, 565, 573
 deficit reduction, 408, 489
 derivatives, 258
 estate tax returns, 427
 federal debt, 90
 FICO credit score, 182, 197, 466, 565
 Gini index, 89
 health care expenditures, 99
 income
 adjusted gross income, 103
 age vs., 186
 average, 89
 children vs. parents, 258
 distribution, 103, 169
 educational attainment and, 215
 gender inequality in, 456
 household, 33–34, 42
 median, 89, 98, 181
 per capita, 105
 by region, 266
 student survey, 154
 taxes, 93–94, B10
 IRS audits, 259
 loan application, 552
 net worth, 120, 394
 property tax, 428
 retirement savings, 442–443
 stock analysis, 454
 stock market, 183–184, 198
 student loans, 90
 taxing, 422
 tax rates, 93–94, 423, 511
 tax revenue, 99

Firearms

gun laws, B11
 muzzle velocity, 166, 427, 497

Food. *See also* Nutrition

accuracy of drive thru orders, 455
 allergies, 394
 candy weight, 118, 134, 162
 cauliflower, 409
 cheeseburgers, fat and calories, 218–219
 chewing number and consumption, 50–51, 428
 chocolates, 144

consumption of popcorn, 442

cookies

- Chips Ahoy, 359
- chocolate chip, 162, 352, 353–354
- diameter of, 102

dining out, 73

fast-food restaurants, 354

insect fragments, 381

McDonald's drive-through traffic, 104

meatloaf, 168

M&M, 533–534

number of drinks, 424–425

nut mix, 144

para-nonylphenol in processing and packaging of, 466

peanuts, 534

pizza, 441

popcorn, 510

priming for healthy food, 53–54

quality control, 442

soda preferences, 457

takeout, 239

tea tasting, 259

time spent eating and drinking, 421

Tootsie Pop, 421

Yelp ratings for restaurant, 104

Gambling. See also Game(s)

betting on sports, 280

craps, 286

lotteries, 280

- double jackpot, 257

- instant winner, 292

- Pick 3, 296

- Pick 4, 296

- Pick 5, 296

- Powerball, 311

- state, 296

roulette, 238, 251, 295, 311, 383, 576

video poker, 311

Game(s). See also Gambling

Blackjack, 310, 353

card drawing, 250, 266, 281, 296, 324

coin toss, 237, 257

Dictator Game, 54

die/dice, 537

- loaded, 240

- rolling, 86, 237, 240, 257, 287

five-card stud, 296

Lingo, 292

Little Lotto, 281

Mega Millions, 281

poker

- flush, 268–269

- royal flush, 269

- seven-card stud, 237

- three-card, 329

- winning, 381

The Price Is Right, 286

Solitaire, 296

Text Twist, 292

Gardening

planting tulips, 268

Gender

lupus and, 457

wage gap, 517

weight gain and, 259

Genetics

Huntington's disease, 238

sickle-cell anemia, 238

Geography

highest elevation for continents, 75

Geology

density of Earth, 366

earthquakes, 88

Old Faithful geyser (Calistoga, California), 118, 135, 162, 205

Old Faithful geyser (Yellowstone Park), 168, 381

Government

federal debt, 90

IRS audits, 259

New Deal policies, 407

Social Security numbers, 280

Social Security reform, 390

state, 11

trust in, 326

type of, 9

waste, 12

Health. See also Exercise; Medicine

alcohol abstention, 488

alcohol dependence treatment, 51

alcohol effects on brain, 466

allergy sufferers, 325, 326

blood alcohol concentration, 119

blood types, 237

body mass index, 488

bone mineral density and cola consumption, 60, 198

brain tumors and cell phones, 14–15

burning calories, 98

calories vs. sugar, 578

cancer, 21

- breast, 268

- cell phones and brain tumors, 14–15

- cholesterol, 38

- death in, 266

- lung, 17, 23

- passive smoke and lung cancer, 23

- power lines and, 22

- skin, coffee consumption and, 21

- survival rates, 119

cardiac arrest, 342–343

cholesterol levels and green tea, 53

commuting time and, 136, 166, 182, 197, 205

doctor visits, 251

drug side effects, 241

education and, 550

effect of Lipitor on cardiovascular

- disease, 46, 47

emergency room visit, 329, 474

exercises, 87

fitness club member satisfaction, 37

flu shots for seniors, 16

ginkgo and memory, 52

handwashing behavior, 409–410

happiness and, 21, 216, 550

hazardous activities, B10

headache, 161

health care expenditures, 99

hearing/vision problems, 251

heart attacks, 552–553

heart disease and baldness, 21

HIV test false positives, 257

hospital-acquired conditions, 250

hospital admissions, 119, 516

hygiene habits, 11

hypertension, 12, 54, 431

insomnia, 52

LDL cholesterol, B27

life expectancy, 185

Lipitor, 454

live births, 101–102, 144

lung cancer and, 17, 23

Lyme disease vs. drownings, 185

marriage/cohabitation and weight gain, 21

migraine, 442

obesity, 185

- social well-being and, 550–551

- television in the bedroom and, 21

osteoporosis treatment, 520

overweight, 42, 100, 441

pulse rates, 117, 134

self-injurious behaviors, 330

shrinking stomach and diet, 52

skinfold thickness procedure, 167

sleep disorders, 521

sleeping habits, 42, 471

smoking, 12, 268

- birth weight, 199–200

- cessation program, 424, 521

- cigar, 251

- e-cig study, B29

- educational attainment and, 522

- lung cancer and, 17, 23

- paternal, 162–163

- during pregnancy, 456

- profile of, 551

- survival rates, 216

- tar and nicotine levels in cigarettes, 566, 573

- weight gain and, 522

sneezing habits, 325, 364, 474

television stations and life expectancy, 185

testosterone levels, 472

tooth whitener, 51, 57

vitamins, 162

weight loss, role of diet and drugs in, 54

weight of college students, 42

women, aspirin, and heart attacks, 552–553

Height(s)

arm span vs., 519–520

father and son, 498

females

- five-year-old, 342

- 20 years of age, 466

- vs. males, 152

head circumference vs., 182, 197, 205,

- 565–566, 573

10-year-old males, 341

Houses and housing

apartments, 219, 291, 578

females living at home, 364

garage door code, 280

home ownership, 102, 408

household winter temperature, 144

increase in assessments, 391

males living at home, 364

number of people living in, 87

pricing, 199

rents, 310, 428

rooms, 250

single-family home price, 441
square footage, 143
Zestimate, 568

Insurance

collision claims, 514
credit scores and, 552
life, 310
Medicare fines, 250

Intelligence

brain size and, 184
IQ scores, 86, 119, 135, 136–137, 153, 184, 222, 430, 469, 472
predictions, 574

Internet

download time, 118
frequency of use of, 73
high speed access, 408
linear transformations, 120, 136
online dating, 287
online homework, 75, 512
online search, 58
time viewing a Web page, 91
Web page design, 22, 517

Investment(s)

attitudes toward, 13
bear markets, 567–568, 573
bull markets, 384
comparing stock sectors, 516
dispersion in the market, 490
diversification, 137, 185
in education, 567, 573
hot stock tips, 310
mutual funds, 135
rate of return on, 135, 138, 259, 327, 381, 511, B25
return on, 90
savings, 143
stock price, 74, 239, 408, 471–472
Super Bowl effect, 456
trade volume, 422
volume of stock
Altria Group, 90
PepsiCo, 422
Starbucks, 468

Landscaping

golf course, 280

Language

foreign, 390
spoken at home, 251

Law(s)

chief justices, 166
death penalty, 409, 488
driver's license, 12
fair packaging and labeling, 441
gun control, 42
jury selection, 281, 325

Law enforcement

age of death-row inmates, 466
racial profiling, 535

Leisure and recreation. See also Entertainment

Boy Scouts merit badge requirement, 29
dining out, 73, 74

kids and, 511
Six Flags over Mid-America, 240

Manufacturing

ball bearings, 353
bolts production, 152–153
copper tubing, 393
products made in America, 72, 215, 266–267
Prolong engine treatment, 442
steel rods, 353
tire production, 366

Marriage

age and, 88, 196, 309
age difference, married couples, 514
couples at work, 251
divorce, 71
education and, 292
happiness and, 216
longevity, 266
infidelity/extramarital affairs, 42, 455, 515
testosterone's influence on, 509
unemployment rates, 220

Math

Benford's Law of frequency of digits, 534

Media

death penalty, 488

Medicine. See also Drugs; Health; Pharmaceuticals

abortion, 215–216
alcohol dependence treatment, 51
alcohol effects on brain, 466
allergy sufferers, 325
Alzheimer's disease treatment, 56
AndroGel, 408
baby delivery methods, 21
bacteria in hospital, 510
blood alcohol concentration, 119
blood types, 74, 237
Cancer Prevention Study II, 57
cardiac arrest, 342–343
carpal tunnel syndrome, 21
cholesterol level, 38, 105
cortical blindness, 455
Covid-19 vaccine, 54
drug side effects, 241
Ebola vaccine, 54
effect of Lipitor on cardiovascular disease, 46, 47
flu season, 71
folate and hypertension, 12
gum disease, 427
hair growth and platelet-rich plasma, 52–53
HDL cholesterol, 423, 568
heart attacks, 552–553
LDL cholesterol, B27
Lipitor, 454
live births, 144
lupus and, 457
Medicare fines, 250
metastatic melanoma, 489
migraine, 442
placebo effect, 54, 251–252
poison ivy ointments, B11
Salk vaccine, 490
side effects, 488–489
sleep apnea, 427
wart treatment, 12

Military

atomic bomb, protection from, 471
Iraq War, 489
night vision goggles, 57
peacekeeping missions, 37
satellite defense system, 259
V-2 rocket hits in London, 537

Miscellaneous

aluminum bottle, 512
birthdays, 238, 250, 268, 286
diameter of Douglas fir trees, 431
filling bottles, 467–468
fingerprints, 259
journal article results, B28
random-number generator, 340
relationship deal-breakers, 489–490
reproducibility of primary studies, 455
selling yourself, 70–71
sleeping, 384, 421
tattoos, 488
toilet flushing, 95, 325–326, 364
wet suits, 515, 517

Money. See also Finance; Investment(s)

abolishing the penny, 408
cash/credit, 514–515
credit-card debt, 474
FICO credit score, 182, 197, 466
income taxes, B10
retirement and, 442–443

Morality

state of, in U.S., 287, 325, 364, 428
unwed women having children, 515

Mortality

bicycle deaths, 535
Gallup Organization, 325
pedestrian death, 536
Titanic disaster, 577

Motor vehicle(s). See also Transportation

accident
fatal traffic, 454–455
red-light camera programs, 252–253
autonomous vehicles, 427
blood alcohol concentration (BAC) for drivers involved in, 421, 454–455
BMW's, 12
braking distance, 499
buying car, 135, 137, 537
discrimination in, B26
car accidents, 98
car color, 74, 325
carpoolers, 161
car prices, 166
car rentals, 499
collision coverage claims, 514
collision data, 521
crash data, B26–B27
crash test results, 422
driving under influence, 258–259
fatalities
alcohol-related, 87
driver, 252, 267
traffic, 295
flight time, 116, 133
gas mileage/fuel economy, 471

gas prices, 100, 136
 male vs. female drivers, 184, 199
 miles per gallon, 116, 133, 342, 416–417
 minimum driving age, 329–330
 new cars, 292
 new vs. used car, 219
 octane in fuel, 499–500
 oil change, 381
 seat belts, 427
 SMART car, 154
 speeding tickets, 252, 267
 SUV vs. car, 498
 wearing helmets, 534

Music

arranging songs, 280
 effect on learning, 49–50
 playing songs, 267, 280, 281

Nutrition. See also Food

bone mineral density and cola consumption, 60, 198
 caffeinated sports drinks, 431
 calories
 burning of, 98
 cheeseburgers, 218–219
 vs. sugar, 578
 dissolving rates of vitamins, 162
 eating together, 455
 fat in
 cheeseburgers, 218–219
 green tea and cholesterol levels, 53, 428
 overweight, 100, 441
 salt and hypertension, 54

Obstetrics. See also Pediatrics

birth(s)
 by day of week, B25
 gestation period, 152, 341–342, 352–353, 380
 multiple, 239
 premature, 577
 weight, 341
 prenatal care, 550
 sleeping patterns during pregnancy, 475

Pediatrics. See also Obstetrics

age of mother at childbirth, 145, 161
 birth weight, 144, 152, 199–200, 341, 354
 gestation period vs., 295
 maternal age and, 537
 36 weeks of, 340
 crawling babies, 422
 energy during pregnancy, 392–393
 head circumference vs. heights, 182, 197, 205

Pets

talking to, 455

Pharmaceuticals. See also Drugs; Medicine

alcohol dependence treatment, 51
 Aspirin, 552–553
 Celebrex, 551
 cholesterol research, 38
 cold medication, 51
 drug effectiveness, 53
 Lipitor, 46, 454
 memory drug, 52
 Nexium, 454
 Prevnar, 488
 skin ointment, 60

Physics

muzzle velocity, 166, 427, 497

Politics

affiliation, 102, 268, 552
 age and, B27
 capitalism, 552
 decisions, 471
 elections
 county, 37
 predictions, 390, 515
 Senate, 312, 409
 simulation, 286–287, 326
 estate taxes, 36
 exit polls, 43
 Future Government Club, 30, 37
 health care and health insurance, 38
 House of Representative gender
 composition, 410
 mayor and small business owners, 57
 poll, 38
 presidents
 age at inauguration, 100, 161
 inaugural addresses, 167
 random sample of, 29
 public knowledge about, 11
 public policy survey, 59
 pundit predictions, 454
 questionnaire wording, 489
 Republican voters, 456–457
 Roosevelt vs. Landon, 577
 socialism, 551
 views, 119
 village poll, 30
 voter polls, 37, 38

Polls and surveys

abortion, 215–216
 about gun-control laws, 42
 annoying behavior, 427
 blood donation, 407
 boys are preferred, 364
 children and childcare, B11
 of city residents, 37
 college, 72–73, 238
 Current Population Survey, 43
 on desirability attributes, 72, 215
 dream job, 73
 dropping course, 552
 election, 44, 390
 e-mail survey, 42
 exit, 43
 faculty opinion, 29
 on family values, 407
 on frequency of having sex, 42
 gender of children in family, 280
 gun control, 409
 happiness and health, 216
 on hazardous activities, B10
 on high-speed Internet service, 57
 informed opinion, 44
 liars, 364
 on life satisfaction, 390
 on long life, 430, 469
 on marriage being obsolete, 390
 number of drinks, 424–425
 order of the questions, 43–44
 police department, 42
 political, 38

population, 43

random digit dialing, 43
 reading number of books, 423
 registered voters, 324
 response rate, 42, 43
 retirement planning, 427
 robocalling, 43
 rotating choices, 43
 sample independence, 269
 seat belts, 427
 speaker evaluation, 37
 student opinion, 29, 30, 37
 student sample for, 29
 tattoos, 488
 on televisions in the household, 89, 90, 309
 TVaholics, 466
 village, 30
 wording of questions, 43
 working hours, 420–421

Psychiatry

attention deficit-hyperactivity disorder, 390

Psychology

ESP and, 454
 ideomotor action, 51–52
 insomnia relief, 52
 profiles and, B28
 rationalized lies, 535–536
 reaction time, B25–B26
 risk handling, B11
 stressful commute, 407, 565, 573

Psychometrics

IQ scores, 86, 119, 136, 153, 184, 222,
 430, 469, 472

Reading

America reads, 366
 at bedtime, 469
 number of books read, 408, 410, 421, 423
 rates, 352, 353, 380–381, 467
 SAT scores, 442

Religion

in Congress, 534
 teen prayer, 475
 trust in, 295–296

Research

level of measurement in, 13

Sex and sexuality

family structure and, 550
 sexual intercourse frequency, 42

Society

abortion issue, 215–216
 affirmative action, 409
 civic duty, 577
 death penalty, 409, 488
 divorce
 opinion regarding, 71
 dog ownership, 259
 marijuana use, 291
 online dating, 287
 path to success, 269
 poverty, 71, 94–95, 144
 racial profiling, 326
 reincarnation belief, 390–391
 social well-being and obesity, 550–551

superstition, 407
unwed women having children, 515
Valentine's Day, 407
volunteers and, 237

Sports

athletics participation, 326
baseball
 batting averages, 11, 136, 152, 241
 cold streaks, 258
 ERA, 152
 factory production, 366
 fastball, 427, 467, 515
 homeruns, 89, 183, 198, 205, 237, 296, 467, 553, 574
 Ichiro's Hit Parade, 309
 injuries, 249–250
 jersey numbers, 13
 most valuable player, 71, 100
 no-hitter, 167
 safety, 324
 sprint speed of players, 88
 starting lineup, 281
 variability, 393
 winning percentage, 185
World Series, 309, 576

basketball
 free throws, 88, 237, 324
 point spread, 218, 353
 salaries, 121

betting on, 259

bowling, 258

caffeinated sports drinks, 431

car racing, INDY 500, 280

football
 completion rate for passes, 56
 extra point, 311
 fans, 409
 fumbles, 391
National Football League combine, 240, B27
spread accuracy, 456
Super Bowl effect on investing, 456

golf
 balls, 268, 474
 pitching wedge, 342

hockey
 National Hockey League, 535
 Stanley Cup, 329

human growth hormone (HGH) use among high school athletes, 36

organized play, 237

soccer, 96, 291

swimming, 152
team captains, 29
television commentator, 409
tennis, Wimbledon tournament, 432–433
triathlon, 152

Statistics

classifying probability, 240
coefficient of skewness, 137
coefficient of variation, 138
Fish Story, 134
geometric probability distribution, 326
in media, 456
midrange, 120
negative binomial probability distribution, 326–327
number of tickets issued, 166
practical significance, 468
probability, 237
shape, mean and median, 120
simulation, 286–287, 296, 311, 326, 409, 424, 457, 469
trimmed mean, 120

Temperature

household winter, 144
human, 474

Test(s)

ACT scores, 466
crash results, 422
essay, 292
FICO score, 120, 466, 565
IQ scores, 86, 119, 136, 153, 184, 222, 430, 472
multiple-choice, 58, 298
SAT scores, 135–136, 153, 184, 292, 311, 442, 466, 468
Wechsler Intelligence Scale, 366

Time

cab ride average, 163
drive-through service, 381, 382, 421, 466–467, 520
eating and drinking, 421
eruptions vs. length of eruption, 205
exam, 116, 133
flight, 116, 133, 324–325, 364
oil change, 381
online, 102–103
on phone, 421
reaction, 53, 58, 497–498, 510, B25–B26
study, 468
travel, 117–118, 134, 155
waiting, 90, 118, 154, 340, 422, 511, 520

Transportation. *See also Motor vehicle(s)*

alcohol-related traffic fatalities, 87
bike sharing, 383
fear of flying, 390
flight overbooking, 326
flight time, 324–325, 364
moving violations, 91
on-time flights, 324–325, 364
parking and camera violation fines, 79
red light cameras, 499
roundabout vs. four-way stop, 512
time spent in drive-through, 381
traffic lights, 239–240

Travel

airline reservations, 391
creative thinking during, 367
on-time flights, 324–325
taxes, 422
text while driving, 471
Titanic survivors, 577
walking in airport, 509–510

Weather

forecast, 292
hurricanes, 91, 183, 198, 205, 566, 573
Memphis snowfall, 359
temperatures, 135, 474
tornadoes, 13, 74–75, 91, 120, 138, 155, 163, 205, 240, 382, 408, 422–423, 456, 468, 489, 511, 573–574, B28

Weight(s)

American Black Bears, 185, 197–198, 205, 567, 573
birth, 340, 354
 smoking and, 199–200
body mass index, 488
car vs. miles per gallon, 184–185, 198, 205
coins, 168
gaining, 259, 428
gestation period vs., 295
male vs. female, 167

Work. *See also Business*

commuting time, 136, 166, 182, 197, 205, 239
employee morale, 30
getting to, 251
married couples, 251
multiple jobs, 268
rate of unemployment, 220
unemployment, 87
walk to, 409

PART



Getting the Information You Need

Statistics is a process—a series of steps that leads to a goal. This text is divided into four parts to help see the process of statistics.

Part 1 is focused on the first step in the process, which is to determine the research objective or question to be answered. Then information is obtained to answer the questions stated in the research objective.

CHAPTER 1 Data Collection



Data Collection

Outline

- 1.1** Introduction to the Practice of Statistics
- 1.2** Observational Studies versus Designed Experiments
- 1.3** Simple Random Sampling
- 1.4** Other Effective Sampling Methods
- 1.5** Bias in Sampling
- 1.6** The Design of Experiments

Making an Informed Decision



It is your senior year of high school. You will have a lot of exciting experiences in the upcoming year, plus a major decision to make—which college should I attend? The choice you make may affect many aspects of your life—your career, where you live, your significant other, and so on, so you don't want to simply choose the college that everyone else picks. You need to design a questionnaire to help you make an informed decision about college. In addition, you want to know how well the college you are considering educates its students. See Making an Informed Decision on page 60.

Putting It Together

Statistics plays a major role in many aspects of our lives. It is used in sports, for example, to help a general manager decide which player might be the best fit for a team. It is used in politics to help candidates understand how the public feels about various policies. And statistics is used in medicine to help determine the effectiveness of new drugs.

Used appropriately, statistics can enhance our understanding of the world around us. Used inappropriately, it can lend support to inaccurate beliefs. Understanding statistical methods will provide you with the ability to analyze and critique studies and the opportunity to become an informed consumer of information. Understanding statistical methods will also enable you to distinguish solid analysis from bogus “facts.”

To help you understand the features of this text and for hints to help you study, read the **Pathway to Success** on the front inside cover of the text.

1.1 Introduction to the Practice of Statistics



Objectives

- ① Define statistics and statistical thinking
- ② Explain the process of statistics
- ③ Distinguish between qualitative and quantitative variables
- ④ Distinguish between discrete and continuous variables
- ⑤ Determine the level of measurement of a variable

① Define Statistics and Statistical Thinking

What is statistics? Many people say that statistics is numbers. After all, we are bombarded by numbers that supposedly represent how we feel and who we are. For example, we hear on the radio that 50% of first marriages, 67% of second marriages, and 74% of third marriages end in divorce (Forest Institute of Professional Psychology, Springfield, MO).

Another interesting consideration about the “facts” we hear or read is that two different sources can report two different results. For example, a January 12, 2019 poll by Rasmussen Reports indicated that 45% of Americans believed the country was on the right track. However, a January 16, 2019 Monmouth poll indicated that 37% of Americans believed the country was on the right track. Is it possible that the percent of Americans who believe the country is on the right track could decrease by 8% in four days, or is something else going on? Statistics helps to provide the answer.

Certainly, statistics has a lot to do with numbers, but this definition is only partially correct. Statistics is also about where the numbers come from (that is, how they were obtained) and how closely the numbers reflect reality.

Definition

Statistics is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

Let’s break this definition into four parts. The first part states that statistics involves the collection of information. The second refers to the organization and summarization of information. The third states that the information is analyzed to draw conclusions or answer specific questions. The fourth part states that results should be reported using some measure that represents how convinced we are that our conclusions reflect reality.

What is the information referred to in the definition? The information is **data**, which the *American Heritage Dictionary* defines as “a fact or proposition used to draw a conclusion or make a decision.” Data can be numerical, as in height, or nonnumerical, as in gender. In either case, data describe characteristics of an individual.

Analysis of data can lead to powerful results. Data can be used to offset anecdotal claims, such as the suggestion that cellular telephones cause brain cancer. After carefully collecting, summarizing, and analyzing data regarding this phenomenon, it was determined that there is no link between cell phone usage and brain cancer. See Examples 1 and 2 in Section 1.2.

Because data are powerful, they can be dangerous when misused. The misuse of data usually occurs when data are incorrectly obtained or analyzed. For example, radio or television talk shows regularly ask poll questions for which respondents must call in or use the Internet to supply their vote. Most likely, the individuals who are going to call in are those who have a strong opinion about the topic. This group is not likely to be representative of people in general, so the results of the poll are not meaningful. Whenever we look at data, we should be mindful of where the data come from.

IN OTHER WORDS

Anecdotal means that the information being conveyed is based on casual observation, not scientific research.

Even when data tell us that a relation exists, we need to investigate. For example, a study showed that breast-fed children have higher IQs than those who were not breast-fed. Does this study mean that a mother who breast-feeds her child will increase the child's IQ? Not necessarily. It may be that some factor other than breast-feeding contributes to the IQ of the children. In this case, it turns out that mothers who breast-feed generally have higher IQs than those who do not. Therefore, it may be genetics that leads to the higher IQ, not breast-feeding.* This illustrates an idea in statistics known as the *lurking variable*. A good statistical study will have a way of dealing with lurking variables.

A key aspect of data is that they vary. Consider the students in your classroom. Is everyone the same height? No. Does everyone have the same color hair? No. So, within groups there is variation. Now consider yourself. Do you eat the same amount of food each day? No. Do you sleep the same number of hours each day? No. So even considering an individual there is variation. Data vary. One goal of statistics is to describe and understand the sources of variation. Variability in data may help to explain the different results obtained by the Rasmussen Reports and Monmouth polls described at the beginning of this section.

Because of this variability, the results that we obtain using data can vary. In a mathematics class, if Bob and Jane are asked to solve $3x + 5 = 11$, they will both obtain $x = 2$ as the solution when they use the correct procedures. In a statistics class, if Bob and Jane are asked to estimate the average commute time for workers in Dallas, Texas, they will likely get different answers, even though both use the correct procedure. The different answers occur because they likely surveyed different individuals, and these individuals have different commute times. Bob and Jane would get the same result if they both asked *all* commuters or the same commuters about their commutes, but how likely is this?

So, in mathematics when a problem is solved correctly, the results can be reported with 100% certainty. In statistics, when a problem is solved, the results do not have 100% certainty. In statistics, we might say that we are 95% confident that the average commute time in Dallas, Texas, is between 20 and 23 minutes. Uncertain results may seem disturbing now but will feel more comfortable as we proceed through the course.

Without certainty, how can statistics be useful? Statistics can provide an understanding of the world around us because recognizing where variability in data comes from can help us to control it. Understanding the techniques presented in this text will provide you with powerful tools that will give you the ability to analyze and critique media reports, make investment decisions, or conduct research on major purchases. This will help to make you an informed citizen, consumer of information, and critical and statistical thinker.

② Explain the Process of Statistics

Consider the following scenario.

NOTE

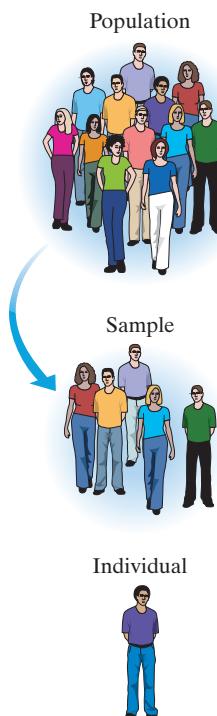
Obtaining a truthful response to a question such as this is challenging. In Section 1.5, we present some techniques for obtaining truthful responses to sensitive questions.

You are walking down the street and notice that a person walking in front of you drops \$100. Nobody seems to notice the \$100 except you. Since you could keep the money without anyone knowing, would you keep the money or return it to the owner?

Suppose you wanted to use this scenario as a gauge of the morality of students at your school by determining the percent of students who would return the money. How might you do this? You could attempt to present the scenario to every student at the school, but this would be difficult or impossible if the student body is large. A second possibility is to present the scenario to 50 students and use the results to make a statement about all the students at the school.

*In fact, a study found that a gene called FADS2 is responsible for higher IQ scores in breast-fed babies. Source: Duke University, "Breastfeeding Boosts IQ in Infants with 'Helpful' Genetic Variant," *Science Daily* 6 November 2007.

Figure 1

**Definitions**

The entire group to be studied is called the **population**. An **individual** is a person or object that is a member of the population being studied. A **sample** is a subset of the population that is being studied. See Figure 1.

In the \$100 study presented, the population is all the students at the school. Each student is an individual. The sample is the 50 students selected to participate in the study.

Suppose 39 of the 50 students stated that they would return the money to the owner. We could present this result by saying that the percent of students in the survey who would return the money to the owner is 78%. This is an example of a *descriptive statistic* because it describes the results of the sample without making any general conclusions about the population.

Definitions

A **statistic** is a numerical summary of a sample. **Descriptive statistics** consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs.

So 78% is a statistic because it is a numerical summary based on a sample. Descriptive statistics make it easier to get an overview of what the data are telling us.

If we extend the results of our sample to the population, we are performing *inferential statistics*.

Definition

Inferential statistics uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.

The generalization contains uncertainty because a sample cannot tell us everything about a population. Therefore, inferential statistics includes a level of confidence in the results. So rather than saying that 78% of all students would return the money, we might say that we are 95% confident that between 74% and 82% of all students would return the money. Notice how this inferential statement includes a *level of confidence* (measure of reliability) in our results. It also includes a range of values to account for the variability in our results.

One goal of inferential statistics is to use statistics to estimate *parameters*.

Definition

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

EXAMPLE 1**Parameter versus Statistic**

- (a) Suppose 48.2% of all students on your campus own a car. This value represents a parameter because it is a numerical summary of a population. Suppose a sample of 100 students is obtained, and from this sample we find that 46% own a car. This value represents a statistic because it is a numerical summary of a sample.

- (b) Suppose the average salary of all employees in the City of Joliet is \$78,302. This value represents a parameter because it is a numerical summary of a population. Suppose a sample of 30 employees is obtained, and from this sample we find the average salary is \$75,038. This value represents a statistic because it is a numerical summary of a sample.



The methods of statistics follow a process.

CAUTION!

Many nonscientific studies are based on *convenience samples*, such as Internet surveys or phone-in polls. The results of any study performed using this type of sampling method are not reliable.

The Process of Statistics

1. *Identify the research objective.* A researcher must determine the question(s) he or she wants answered. The question(s) must clearly identify the population that is to be studied.
2. *Collect the data needed to answer the question(s) posed in (1).* Conducting research on an entire population is often difficult and expensive, so we typically look at a sample. This step is vital to the statistical process, because if the data are not collected correctly, the conclusions drawn are meaningless. Do not overlook the importance of appropriate data collection. We discuss this step in detail in Sections 1.2 through 1.6.
3. *Describe the data.* Descriptive statistics allow the researcher to obtain an overview of the data and can help determine the type of statistical methods the researcher should use. We discuss this step in detail in Chapters 2 through 4.
4. *Perform inference.* Apply the appropriate techniques to extend the results obtained from the sample to the population and report a level of reliability of the results. We discuss techniques for measuring reliability in Chapters 5 through 8 and inferential techniques in Chapters 9 through 15.

EXAMPLE 2 The Process of Statistics: Trust Your Neighbor

Pew Research conducted a poll and asked, “Do you trust all or most of your neighbors?” The following process allowed the researchers to conduct their study.

1. *Identify the Research Objective* The researchers wanted to determine the percentage of adult Americans who trust all or most of their neighbors. Therefore, the population was adult Americans.
2. *Collect the Data Needed to Answer the Question Posed in (1)* It is unreasonable to expect to survey the more than 200 million adult Americans to determine whether they trust all or most of their neighbors. So, the researchers surveyed a sample of 1628 adult Americans. Of those surveyed, 847 stated they trust all or most of their neighbors.
3. *Describe the Data* Of the 1628 individuals in the survey, 52% ($= 847/1628$) stated they trust all or most of their neighbors. This is a descriptive statistic because it is a summary of the sample data.
4. *Perform Inference* Pew Research wanted to extend the results of the survey to all adult Americans. When generalizing results from a sample to a population, the results are uncertain. To account for this uncertainty, Pew reported a 2.5% *margin of error*. This means Pew feels fairly certain (in fact, 95% certain) that the percentage of *all* adult Americans who trust all or most of their neighbors is somewhere between 49.5% ($= 52\% - 2.5\%$) and 54.5% ($= 52\% + 2.5\%$). 

NW Now Work Problem 45

③ Distinguish between Qualitative and Quantitative Variables

Once a research objective is stated, a list of the information we want to learn about the individuals must be created. **Variables** are the characteristics of the individuals within the population. For example, recently my son and I planted a tomato plant in our backyard. We collected information about the tomatoes harvested from the plant. The individuals we studied were the tomatoes. The variable that interested us was the weight of a tomato. My son noted that the tomatoes had different weights even though they came from the same plant. He discovered that variables such as weight may vary.

If variables did not vary, they would be constants, and statistical inference would not be necessary. Think about it this way: If each tomato had the same weight, then knowing the weight of one tomato would allow us to determine the weights of all tomatoes. However, the weights of the tomatoes vary. One goal of research is to learn the causes of the variability so that we can learn to grow plants that yield the best tomatoes.

Variables can be classified into two groups: *qualitative* or *quantitative*.

Definitions

Qualitative, or categorical, variables allow for classification of individuals based on some attribute or characteristic.

Quantitative variables provide numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.

Many examples in this text will include a suggested **approach**, or a way to look at and organize a problem so that it can be solved. The approach will be a suggested method of *attack* toward solving the problem. This does not mean that the approach given is the only way to solve the problem, because many problems have more than one approach leading to a correct solution.

EXAMPLE 3 Distinguishing between Qualitative and Quantitative Variables

Problem Determine whether the following variables are qualitative or quantitative.

- (a) Gender
- (b) Temperature
- (c) Number of days during the past week that a college student studied
- (d) Zip code

Approach Quantitative variables are numeric measures such that meaningful arithmetic operations can be performed on the values of the variable. Qualitative variables describe an attribute or characteristic of the individual that allows researchers to categorize the individual.

Solution

- (a) Gender is a qualitative variable because it allows a researcher to categorize the individual as male or female. Notice that arithmetic operations cannot be performed on these attributes.
- (b) Temperature is a quantitative variable because it is numeric, and operations such as addition and subtraction provide meaningful results. For example, 70°F is 10°F warmer than 60°F.
- (c) Number of days during the past week that a college student studied is a quantitative variable because it is numeric, and operations such as addition and subtraction provide meaningful results.
- (d) Zip code is a qualitative variable because it categorizes a location. Notice that, even though zip codes are numeric, adding or subtracting zip codes does not provide meaningful results.

NW Now Work Problem 11

Example 3(d) shows us that a variable may be qualitative while having numeric values. Just because the value of a variable is numeric does not mean that the variable is quantitative.

④ Distinguish between Discrete and Continuous Variables

We can further classify quantitative variables into two types: *discrete* or *continuous*.

Definitions**IN OTHER WORDS**

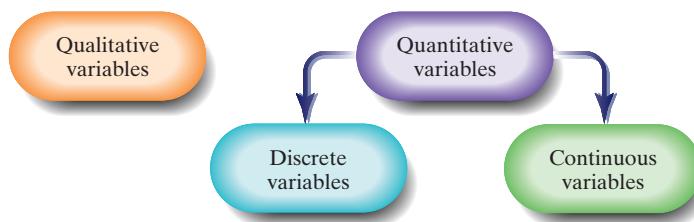
If you count to get the value of a quantitative variable, it is discrete. If you measure to get the value of a quantitative variable, it is continuous.

A **discrete variable** is a quantitative variable that has either a finite number of possible values or a countable number of possible values. The term *countable* means that the values result from counting, such as 0, 1, 2, 3, and so on. A discrete variable cannot take on every possible value between any two possible values.

A **continuous variable** is a quantitative variable that has an infinite number of possible values that are not countable. A continuous variable may take on every possible value between any two values.

Figure 2 illustrates the relationship among qualitative, quantitative, discrete, and continuous variables.

Figure 2

**EXAMPLE 4****Distinguishing between Discrete and Continuous Variables**

Problem Determine whether the quantitative variables are discrete or continuous.

- The number of heads obtained after flipping a coin five times.
- The number of cars that arrive at a McDonald's drive-thru between 12:00 P.M. and 1:00 P.M.
- The distance a 2020 Toyota Prius can travel in city driving conditions with a full tank of gas.

Approach A variable is discrete if its value results from counting. A variable is continuous if its value is measured.

Solution

- The number of heads obtained by flipping a coin five times is a discrete variable because we can count the number of heads obtained. The possible values of this discrete variable are 0, 1, 2, 3, 4, 5.
- The number of cars that arrive at a McDonald's drive-thru between 12:00 P.M. and 1:00 P.M. is a discrete variable because we find its value by counting the cars. The possible values of this discrete variable are 0, 1, 2, 3, 4, and so on. Notice that this number has no upper limit.
- The distance traveled is a continuous variable because we measure the distance (miles, feet, inches, and so on).

NW Now Work Problem 19

Continuous variables are often rounded. For example, if a certain make of car gets 24 miles per gallon (mpg) of gasoline, its miles per gallon must be greater than or equal to 23.5 and less than 24.5, or $23.5 \leq \text{mpg} < 24.5$.

The type of variable (qualitative, discrete, or continuous) dictates the methods that can be used to analyze the data.

The list of observed values for a variable is **data**. Gender is a variable; the observations male and female are data. **Qualitative data** are observations corresponding to a qualitative variable. **Quantitative data** are observations corresponding to a quantitative variable. **Discrete data** are observations corresponding to a discrete variable. **Continuous data** are observations corresponding to a continuous variable.

EXAMPLE 5**Distinguishing between Variables and Data**

Problem Table 1 presents a group of selected countries and information regarding these countries as of July, 2020. Identify the individuals, variables, and data in Table 1.

Table 1

Country	Government Type	Life Expectancy (years)	Population (in millions)
Australia	Federal parliamentary democracy	82.7	25.5
Canada	Constitutional monarchy	83.4	37.7
France	Republic	82.2	67.8
Morocco	Constitutional monarchy	73.3	35.6
Poland	Republic	78.3	38.3
Senegal	Presidential republic	63.2	15.7
United States	Federal republic	80.3	332.6

Source: CIA World Factbook

Approach An individual is an object or person for whom we wish to obtain data. The variables are the characteristics of the individuals, and the data are the specific values of the variables.

Solution The **individuals** in the study are the countries: Australia, Canada, and so on. The **variables** measured for each country are *government type*, *life expectancy*, and *population*. The variable *government type* is qualitative because it categorizes the individual. The variables *life expectancy* and *population* are quantitative.

The quantitative variable *life expectancy* is continuous because it is measured. The quantitative variable *population* is discrete because we count people. The **observations** are the data. For example, the data corresponding to the variable *life expectancy* are 82.7, 83.4, 82.2, 73.3, 78.3, 63.2, and 80.3. The following data correspond to the individual Poland: a republic government with residents whose life expectancy is 78.3 years and population is 38.3 million people. Republic is an instance of qualitative data that results from observing the value of the qualitative variable *government type*. The life expectancy of 78.3 years is an instance of quantitative data that results from observing the value of the quantitative variable *life expectancy*.

NW Now Work Problem 41

5 Determine the Level of Measurement of a Variable

Rather than classify a variable as qualitative or quantitative, we can assign a level of measurement to the variable.

Definitions

IN OTHER WORDS

The word **nominal** comes from the Latin word **nomen**, which means to name. When you see the word **ordinal**, think order.

A variable is at the **nominal level of measurement** if the values of the variable name, label, or categorize. In addition, the naming scheme does not allow for the values of the variable to be arranged in a ranked or specific order.

A variable is at the **ordinal level of measurement** if it has the properties of the nominal level of measurement, however, the naming scheme allows for the values of the variable to be arranged in a ranked or specific order.

(continued)

A variable is at the **interval level of measurement** if it has the properties of the ordinal level of measurement and the differences in the values of the variable have meaning. A value of zero does not mean the absence of the quantity. Arithmetic operations such as addition and subtraction can be performed on values of the variable.

A variable is at the **ratio level of measurement** if it has the properties of the interval level of measurement and the ratios of the values of the variable have meaning. A value of zero means the absence of the quantity. Arithmetic operations such as multiplication and division can be performed on the values of the variable.

Nominal or ordinal variables are also qualitative variables. Interval or ratio variables are also quantitative variables.

EXAMPLE 6

Determining the Level of Measurement of a Variable

Problem For each of the following variables, determine the level of measurement.

- (a) Gender
- (b) Temperature
- (c) Number of days during the past week that a college student studied
- (d) Letter grade earned in your statistics class

Approach For each variable, we ask the following: Does the variable simply categorize each individual? If so, the variable is nominal. Does the variable categorize *and* allow ranking of each value of the variable? If so, the variable is ordinal. Do differences in values of the variable have meaning, but a value of zero does not mean the absence of the quantity? If so, the variable is interval. Do ratios of values of the variable have meaning *and* there is a natural zero starting point? If so, the variable is ratio.

Solution

- (a) Gender is a variable measured at the nominal level because it only allows for categorization of male or female. Plus, it is not possible to rank gender classifications.
- (b) Temperature is a variable measured at the interval level because differences in the value of the variable make sense. For example, 70°F is 10°F warmer than 60°F. Notice that the ratio of temperatures does not represent a meaningful result. For example, 60°F is not twice as warm as 30°F. In addition, 0°F does not represent the absence of heat.
- (c) Number of days during the past week that a college student studied is measured at the ratio level, because the ratio of two values makes sense and a value of zero has meaning. For example, a student who studies four days studies twice as many days as a student who studies two days.
- (d) Letter grade is a variable measured at the ordinal level because the values of the variable can be ranked, but differences in values have no meaning. For example, an A is better than a B, but A – B has no meaning.

NW Now Work Problem 27

When classifying variables according to their level of measurement, it is extremely important that we recognize what the variable is intended to measure. For example, suppose we want to know whether cars with 4-cylinder engines get better gas mileage than cars with 6-cylinder engines. Here, engine size represents a category of data and so the variable is nominal. On the other hand, if we want to know the average number of cylinders in cars in the United States, the variable is classified as ratio (an 8-cylinder engine has twice as many cylinders as a 4-cylinder engine).



1.1 Assess Your Understanding

Vocabulary and Skill Building

1. Match each word or phrase with its definition.

Word/Phrase	Definition
(a) Statistics	I. A numerical summary of a sample.
(b) Population	II. Organizing and summarizing data through tables, graphs, and numerical summaries.
(c) Sample	III. The science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. It is also about providing a measure of confidence in any conclusions.
(d) Parameter	IV. A subset of the group of individuals that is being studied.
(e) Statistic	V. Uses methods that take results from a sample and extends them to the population, and measures the reliability of the result.
(f) Individual	VI. A person or object that is a member of the group being studied.
(g) Descriptive Statistics	VII. A numerical summary of a population.
(h) Inferential Statistics	VIII. The entire group of individuals to be studied.

2. Match each word or phrase with its definition.

Word/Phrase	Definition
(a) Discrete Variable	I. Provide numerical measures of individuals. The measures can be added or subtracted, and provide meaningful results.
(b) Data	II. Allow for classification of individuals based on some attribute or characteristic.
(c) Continuous Variable	III. The characteristics of the individuals within the population.
(d) Qualitative Variable	IV. Information that describes characteristics of an individual.
(e) Quantitative Variable	V. Has either a finite number of possible values or countable number of possible values. The values of these variables typically result from counting.
(f) Variable	VI. Has an infinite number of possible values that are not countable. The values of these variables typically result from measurement.

In Problems 3–10, determine whether the underlined value is a parameter or a statistic.

NW 3. State Government Following the 2018 national midterm election, 18% of the governors of the 50 United States were female. *Source: National Governors Association*

4. Calculus Exam The average score for a class of 28 students taking a calculus midterm exam was 72%.

5. School Bullies In a national survey of 1300 high school students (grades 9 to 12), 32% of respondents reported that someone had bullied them at school. *Source: Bureau of Justice Statistics*

6. Drug Use In a national survey on substance abuse, 13.3% of 12th graders reported using illicit drugs within the past month. *Source: National Institute on Drug Abuse*

7. Batting Average Ty Cobb is one of Major League Baseball's greatest hitters of all time, with a career batting average of .366. *Source: baseball-almanac.com*

8. Moonwalkers Only 12 men have walked on the moon. The average time these men spent on the moon was 43.92 hours. *Source: www.theguardian.com*

9. Hygiene Habits A study of 6076 adults in public rest rooms (in Atlanta, Chicago, New York City, and San Francisco) found that 23% did not wash their hands before exiting.

Source: American Society for Microbiology and the Soap and Detergent Association.

10. Public Knowledge Interviews of 100 adults 18 years of age or older, conducted nationwide, found that 44% could state the minimum age required for the office of U.S. president.

Source: Newsweek Magazine.

In Problems 11–18, classify the variable as qualitative or quantitative.

NW 11. Nation of origin

12. Number of siblings

13. Grams of carbohydrates in a doughnut

14. Number on a football player's jersey

15. Number of unpopped kernels in a bag of microwave popcorn

16. Assessed value of a house

17. Phone number

18. Student ID number

In Problems 19–26, determine whether the quantitative variable is discrete or continuous.

NW 19. Goals scored in a season by a soccer player

20. Volume of water lost each day through a leaky faucet

21. Length (in minutes) of a country song

22. Number of Sequoia trees in a randomly selected acre of Yosemite National Park

23. High temperature on a randomly selected day in Memphis, Tennessee

24. Internet connection speed in kilobytes per second

25. Points scored in an NCAA basketball game

26. Air pressure in pounds per square inch in an automobile tire

In Problems 27–34, determine the level of measurement of each variable.

NW 27. Nation of origin

28. Movie ratings of one star through five stars

29. Volume of water used by a household in a day

30. Year of birth of college students

31. Highest degree conferred (high school, bachelor's, and so on)

32. Eye color

33. Assessed value of a house

34. Time of day measured in military time

In Problems 35–40, a research objective is presented. For each, identify the population and sample in the study.

35. The Gallup Organization contacts 1028 teenagers who are 13 to 17 years of age and live in the United States and asks whether or not they had been prescribed medications for any mental disorders, such as depression or anxiety.

36. A quality-control manager randomly selects 50 bottles of Coca-Cola that were filled on October 15 to assess the calibration of the filling machine.

37. A farmer interested in the weight of his soybean crop randomly samples 100 plants and weighs the soybeans on each plant.

38. Every year the U.S. Census Bureau releases the *Current Population Report* based on a survey of 50,000 households. The goal of this report is to learn the demographic characteristics, such as income, of all households within the United States.

39. Folate and Hypertension Researchers want to determine whether or not higher folate intake is associated with a lower risk of hypertension (high blood pressure) in women (27 to 44 years of age). To make this determination, they look at 7373 cases of hypertension in these women and find that those who consume at least 1000 micrograms per day ($\mu\text{g}/\text{d}$) of total folate had a decreased risk of hypertension compared with those who consume less than 200 $\mu\text{g}/\text{d}$. *Source:* John P. Forman, MD; Eric B. Rimm, ScD; Meir J. Stampfer, MD; Gary C. Curhan, MD, ScD, "Folate Intake and the Risk of Incident Hypertension among US Women," *Journal of the American Medical Association* 293:320–329, 2005.

40. A community college notices that an increasing number of full-time students are working while attending the school. The administration randomly selects 128 students and asks how many hours per week each works.

In Problems 41 and 42, identify the individuals, variables, and data corresponding to the variables. Determine whether each variable is qualitative, continuous, or discrete.

NW 41. Driver's License Laws The following data represent driver's license laws for various states.

State	Minimum Age for Driver's License (unrestricted)	Mandatory Belt Use	Maximum Allowable Speed Limit (cars on rural interstate), mph
Alabama	17	Front	70
Colorado	17	Front	75
Indiana	18	All	70
North Carolina	16	All	70
Wisconsin	18	All	65

Source: Governors Highway Safety Association.

42. BMW Cars The following information relates to the 2019 model year product line of BMW automobiles.

Model	Body Style	Weight (lb)	Number of Seats
3 Series	Sedan	3489	4
4 Series	Coupe	3574	4
5 Series	Sedan	3790	5
7 Series	Sedan	4244	5
X3	Sport utility	4034	5
Z4	Roadster Coupe	3287	2

Source: www.motortrend.com

Applying the Concepts

43. Smoker's IQ A study was conducted in which 20,211 18-year-old Israeli male military recruits were given an exam to measure IQ. In addition, the recruits were asked to disclose their smoking status. An individual was considered a smoker if he smoked at least one cigarette per day. The goal of the study was to determine whether adolescents aged 18 to 21 who smoke have a lower IQ than nonsmokers. It was found that the average IQ of the smokers was 94, while the average IQ of the nonsmokers was 101. The researchers concluded that lower IQ individuals are more likely to choose to smoke, not that smoking makes people less intelligent.

Source: Weiser, M., Zarka, S., Werbeloff, N., Kravitz, E. and Lubin, G. (2010). "Cognitive Test Scores in Male Adolescent Cigarette Smokers Compared to Non-smokers: A Population-Based Study." *Addiction*, 105:358–363. doi: 10.1111/j.1360-0443.2009.02740.x).

- (a) What is the research objective?
- (b) What is the population being studied? What is the sample?
- (c) What are the descriptive statistics?
- (d) What are the conclusions of the study?

44. A Cure for the Common Wart A study was designed "to determine if application of duct tape is as effective as cryotherapy (liquid nitrogen applied to the wart for 10 seconds every 2 to 3 weeks) in the treatment of common warts." The researchers randomly divided 51 patients into two groups. The 26 patients in group 1 had their warts treated by applying duct tape to the wart for 6.5 days and then removing the tape for 12 hours, at which point the cycle was repeated for a maximum of 2 months. The 25 patients in group 2 had their warts treated by cryotherapy for a maximum of six treatments. Once the treatments were complete, it was determined that 85% of the patients in group 1 and 60% of the patients in group 2 had complete resolution of their warts. The researchers concluded that duct tape is significantly more effective in treating warts than cryotherapy.

Source: Dean R. Focht III, Carole Spicer, Mary P. Fairchok. "The Efficacy of Duct Tape vs. Cryotherapy in the Treatment of Verruca Vulgaris (The Common Wart)," *Archives of Pediatrics and Adolescent Medicine*, 156(10), 2002.

- (a) What is the research objective?
- (b) What is the population being studied? What is the sample?
- (c) What are the descriptive statistics?
- (d) What are the conclusions of the study?

NW 45. Government Waste Gallup News Service conducted a survey of 1017 American adults aged 18 years or older. The respondents were asked, "Of every tax dollar that goes to the federal government in Washington, D.C., do you believe 51 cents or more are wasted?" Of the 1017 individuals surveyed, 35% indicated that 51 cents or more is wasted. Gallup reported that 35% of all adult Americans 18 years or older believe the federal government wastes at least 51 cents of each dollar spent, with a margin of error of 4% and a 95% level of confidence.

- (a) What is the research objective?
- (b) What is the population?
- (c) What is the sample?
- (d) List the descriptive statistics.
- (e) What can be inferred from this survey?

46. Investment Decision The Gallup Organization conducted a survey of 1018 adults, aged 18 and older, living in the United States and asked, “If you had a thousand dollars to spend, do you think investing it in the stock market would be a good or bad idea?” Of the 1018 adults, 46% said it would be a bad idea. The Gallup Organization reported that 46% of all adults, aged 18 and older, living in the United States thought it was a bad idea to invest \$1000 in the stock market with a 4% margin of error with 95% confidence.

- (a) What is the research objective?
- (b) What is the population?
- (c) What is the sample?
- (d) List the descriptive statistics.
- (e) What can be inferred from this survey?

47. Threaded Problem: Tornado The data set

“Tornadoes_2017” located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017. For each of the following variables in the data set, indicate whether the variable is qualitative or quantitative. For those that are quantitative, indicate whether the variable is discrete or continuous.

- (a) State
- (b) F Scale (this is the Fujita scale for rating tornadoes based on wind speed, where 0 is a tornado whose wind speed is less than 73 mph; 1 is a tornado whose wind speed is 73–112 mph; up through 5, which is a tornado whose wind speed is 261–318 mph).
- (c) Fatalities
- (d) Length

48. Threaded Problem: Tornado The data set

“Tornadoes_2017” located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017. For each of the following variables in the data set, indicate the level of measurement of each variable.

- (a) State
- (b) F Scale (this is the Fujita scale for rating tornadoes based on wind speed, where 0 is a tornado whose wind speed is less than 73 mph; 1 is a tornado whose wind speed is 73–112 mph; up through 5, which is a tornado whose wind speed is 261–318 mph).
- (c) Fatalities
- (d) Length

49. What Level of Measurement? It is extremely important for a researcher to clearly define the variables in a study because this helps to determine the type of analysis that can be performed on the data. For example, if a researcher wanted to describe baseball players based on jersey number, what level of measurement would the variable *jersey number* be? Now suppose the researcher felt that certain players who were of lower caliber received higher numbers. Does the level of measurement of the variable change? If so, how?

50. Interpreting the Variable Suppose a fundraiser holds a raffle for which each person who enters the room receives a ticket numbered 1 to N , where N is the number of people at the fundraiser. The first person to arrive receives ticket number 1, the second person receives ticket number 2, and so on. Determine the level of measurement for each of the following interpretations of the variable *ticket number*.

- (a) The winning ticket number.

(b) The winning ticket number was announced as 329. An attendee noted his ticket number was 294 and stated, “I guess I arrived too early.”

(c) The winning ticket number was announced as 329. An attendee looked around the room and commented, “It doesn’t look like there are 329 people in attendance.”

51. Analyze the Article Read the newspaper article and answer the following questions:

- (a) What is the research question the study addresses?
- (b) What is the sample?
- (c) What type of variable is season in which you were born?
- (d) What can be said (in general) about individuals born in summer? Winter?
- (e) What conclusion was drawn from the study?

Season of Birth Affects Your Mood Later In Life by Nicola Fifield

Babies born in the summer are much more likely to suffer from mood swings when they grow up while those born in the winter are less likely to become irritable adults, scientists claim.

Researchers studied 400 people and matched their personality type to when in the year they were born.

They claim that people born at certain times of the year have a far greater chance of developing certain types of temperaments, which can lead to mood disorders.

The scientists, from Budapest, said this was because the seasons had an influence on certain monoamine neurotransmitters, such as dopamine and serotonin, which control mood, however more research was needed to find out why.

They discovered that the number of people with a “cyclothymic” temperament, characterized by rapid, frequent swings between sad and cheerful moods, was significantly higher in those born in the summer.

Those with a hyperthymic temperament, a tendency to be excessively positive, was significantly higher among those born in the spring and summer.

The study also found that those born in the autumn were less likely to be depressive, while those born in winter were less likely to be irritable.

Lead researcher, assistant professor Xenia Gonda, said: “Biochemical studies have shown that the season in which you are born has an influence on certain monoamine neurotransmitters, such as dopamine and serotonin, which is detectable even in adult life. This led us to believe that birth season may have a longer-lasting effect.

“Our work looked at 400 subjects and matched their birth season to personality types in later life.

“Basically, it seems that when you are born may increase or decrease your chance of developing certain mood disorders.

Professor Gonda added: “We can’t yet say anything about the mechanisms involved.

What we are now looking at is to see if there are genetic markers which are related to season of birth and mood disorder”.

The study may well provide a clue as to why some of the nation's best known personalities are good natured, while others are slightly grumpier.

The Duchess of Cambridge was born in winter, on January 9, which according to the study, means she is less likely to be irritable while Roy Keane, the famously hot-headed former Manchester United footballer, was born in August, when the scientists say people are more likely to have mood swings.

Mary Berry, the ever-cheerful presenter of the Great British Bake Off, was born in the Spring, when, according to the study, people are more likely to be excessively positive.

The study is being presented at the annual conference of the European College of Neuropsychopharmacology (ECNP) in Berlin, Germany, on Sunday.

Professor Eduard Vieta, from the ECNP, said: "Although both genetic and environmental factors are involved in one's temperament, now we know that the season at birth plays a role too."

"And the finding of "high mood" tendency (hyperthymic temperament) for those born in summer is quite intriguing." *The Telegraph, October 19, 2014*

Source: Season of Birth Affects Your Mood Later In Life by Nicola Fifield from The Telegraph. Copyright © 2014 by Telegraph Media Group Limited.

Explaining the Concepts

52. Explain the difference between a population and a sample.
53. Contrast the differences between qualitative and quantitative variables. Discuss the differences between discrete and continuous variables.
54. In your own words, define the four levels of measurement of a variable. Give an example of each.
55. Explain what is meant when we say "data vary." How does this variability affect the results of statistical analysis?
56. Explain the process of statistics.
57. The age of a person is commonly considered to be a continuous random variable. Could it be considered a discrete random variable instead? Explain.

1.2 Observational Studies versus Designed Experiments



Objectives

- ① Distinguish between an observational study and an experiment
- ② Explain the various types of observational studies

1 Distinguish between an Observational Study and an Experiment

Once a research objective is determined, the researcher develops the method for obtaining the data that can be used to answer the questions posed in our research objective. There are two methods for collecting data: *observational studies* and *designed experiments*. To see the difference between these two methods, read the following two studies.

EXAMPLE 1 Cellular Phones and Brain Tumors

Researchers wanted to determine whether there is an association between mobile phone use and brain tumors. To do so, 791,710 middle-aged women in the United Kingdom were followed over a period of 7 years. During this time, there were 1261 incidences of brain tumors. The researchers compared the women who never used a mobile phone to those who used mobile phones and found no significant difference in the incidence rate of brain tumors between the two groups.

Source: Benson, V. S. et al. "Mobile Phone Use and Risk of Brain Neoplasms and Other Cancers: Prospective Study," *International Journal of Epidemiology* 2013 Jun; 42(3): 792–802.

EXAMPLE 2 Cellular Phones and Brain Tumors

Researchers from the United States National Toxicology Program conducted a study to address the concern that radio-frequency radiation (RFR) may be associated with an increased likelihood of developing brain tumors in humans. Certainly, it is unethical

to purposely expose humans to a potential carcinogen, so rats were used instead. The researchers randomly assigned 90 rats to one of three possible groups. Each of the groups was housed in a reverberation chamber that allowed the rats to be exposed to the RFR. Group 1 rats served as a control and were not exposed to any RFR in their chamber. Group 2 rats were exposed to Global System for Mobile Communications (GSM)-modulated RFR, and Group 3 rats were exposed to Code Division Multiple Access (CDMA)-modulated RFR. GSM and CDMA are the modulations primarily used in the United States. The rats in Groups 2 and 3 were exposed to RFR using a continuous cycle of 10 minutes on (exposed) and 10 minutes off (not exposed) for a total daily exposure time of about 9 hours a day, 7 days per week for approximately two years. Each chamber was maintained on a 12-hour light/dark cycle with a temperature range of 72 degrees Fahrenheit (plus or minus 3 degrees), a humidity range of 50 plus/minus 15%, and with at least 10 air changes per hour. All the rats had the same access to food and water. The researchers found low incidences of brain tumors in rats exposed to RFR for both GSM and CDMA modulations, while there were no cases of brain tumors in the control group. However, the incidence rate was not statistically significant.

Source: M. Wyde, et al. bioRxiv 055699; doi: <https://doi.org/10.1101/055699>, June 23, 2016. “Report of Partial Findings from the National Toxicology Program Carcinogenesis Studies of Cell Phone Radiofrequency Radiation in Hsd: Sprague Dawley® SD rats (Whole Body Exposures).”



In both studies, the goal was to determine if radio frequencies from cell phones increase the risk of contracting brain tumors. Whether or not brain cancer was contracted is the *response variable*. The level of cell phone usage is the *explanatory variable*. In research, we wish to determine how varying the amount of an **explanatory variable** affects the value of a **response variable**.

What are the differences between the studies in Examples 1 and 2? Obviously, in Example 1 the study was conducted on humans, while the study in Example 2 was conducted on rats. However, there is a bigger difference. In Example 1, no attempt was made to influence the individuals in the study. The researchers simply followed the women over time to determine their use of cell phones. In other words, no attempt was made to influence the value of the explanatory variable, radio-frequency exposure (cell phone use). Because the researchers simply recorded the behavior of the participants, the study in Example 1 is an *observational study*.

Definition

An **observational study** measures the value of the response variable without attempting to influence the value of either the response or explanatory variables. That is, in an observational study, the researcher observes the behavior of the individuals without trying to influence the outcome of the study.

In the study in Example 2, the researchers obtained 90 rats and divided the rats into three groups. Each group was *intentionally* exposed to various levels of radiation. The researchers then compared the number of rats in each group that had brain tumors. Clearly, there was an attempt to influence the individuals in this study because the value of the explanatory variable (exposure to radio frequency) was manipulated to three levels. Because the researchers manipulated the value of an explanatory variable (radiation) and controlled others (temperature, food), we call the study in Example 2 a *designed experiment*.

Definition

If a researcher randomly assigns the individuals in a study to groups, intentionally manipulates the value of an explanatory variable, controls other explanatory variables at fixed values, and then records the value of the response variable for each individual, the study is a **designed experiment**.

NW Now Work Problem 9

Which Is Better? A Designed Experiment or an Observational Study?

To answer this question, let's consider another study.

EXAMPLE 3

Do Flu Shots Benefit Seniors?

Researchers wanted to determine the long-term benefits of the influenza vaccine on seniors aged 65 years and older by looking at records of over 36,000 seniors for 10 years. The seniors were divided into two groups. Group 1 were seniors who chose to get a flu vaccination shot, and group 2 were seniors who chose not to get a flu vaccination shot. After observing the seniors for 10 years, it was determined that seniors who get flu shots are 27% less likely to be hospitalized for pneumonia or influenza and 48% less likely to die from pneumonia or influenza.

Source: Kristin L. Nichol, MD, MPH, MBA, James D. Nordin, MD, MPH, David B. Nelson, PhD, John P. Mullooly, PhD, Eelko Hak, PhD. "Effectiveness of Influenza Vaccine in the Community-Dwelling Elderly," *New England Journal of Medicine* 357:1373–1381, 2007.



Wow! The results of this study sound great! All seniors should go out and get a flu shot. Right? Not necessarily. The authors were concerned about *confounding*. They were concerned that lower hospitalization and death rates may have been due to something other than the flu shot. Could it be that seniors who get flu shots are more health conscious or are able to get to the clinic more easily? Does race, income, or gender play a role in whether one might contract (and possibly die from) influenza?

Definition

Confounding in a study occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory variable and the response variable may be due to some other variable or variables not accounted for in the study.

Confounding is potentially a major problem with observational studies. Often, the cause of confounding is a *lurking variable*.

Definition

A **lurking variable** is an explanatory variable that was not considered in a study, but that affects the value of the response variable in the study. In addition, lurking variables are typically related to explanatory variables considered in the study.

In the influenza study, possible lurking variables might be age, health status, or mobility of the senior. How can we manage the effect of lurking variables? One possibility is to look at the individuals in the study to determine if they differ in any significant way. For example, it turns out in the influenza study that the seniors who elected to get a flu shot were actually *less* healthy than those who did not. The researchers also accounted for race and income. The authors identified another potential lurking variable, *functional status*, meaning the ability of the seniors to conduct day-to-day activities on their own. The authors were able to adjust their results for this variable as well.

Even after accounting for all the potential lurking variables in the study, the authors were still careful to conclude that getting an influenza shot is *associated* with a lower risk of being hospitalized or dying from influenza. The authors used the term *associated*, instead of saying the influenza shots *caused* a lower risk of death, because the study was observational.

Observational studies do not allow a researcher to claim causation, only association.

Designed experiments, on the other hand, are used whenever control of certain variables is possible and desirable. This type of research allows the researcher to identify certain cause and effect relationships among the variables in the study.

So why ever conduct an observational study if we can't claim causation? Often, it is unethical to conduct an experiment. Consider the link between smoking and lung cancer. In a designed experiment to determine if smoking causes lung cancer in humans, a researcher would divide a group of volunteers into groups. Group 1 individuals would smoke a pack of cigarettes every day for the next 10 years, and Group 2 individuals would not smoke. In addition, eating habits, sleeping habits, and exercise would be controlled so that the only difference between the two groups was smoking. After 10 years the experiment's researcher would compare the proportion of participants in the study who contract lung cancer in the smoking group to the nonsmoking group. If the two proportions differ significantly, it could be said that smoking causes cancer. This designed experiment is able to control many of the factors that might affect whether one contracts lung cancer that would not be controlled in an observational study, however, it is a very unethical study.

Other reasons exist for conducting observational studies over designed experiments. An article in support of observational studies states, "observational studies have several advantages over designed experiments, including lower cost, greater timeliness, and a broader range of patients." (Source: Kjell Benson, BA, and Arthur J. Hartz, MD, PhD. "A Comparison of Observational Studies and Randomized, Controlled Trials," *New England Journal of Medicine* 342:1878–1886, 2000.)

One final thought regarding confounding. In designed experiments, it is possible to have two explanatory variables in a study that are related to each other and related to the response variable. For example, suppose Professor Egner wanted to conduct an experiment in which she compared student success using online homework versus traditional textbook homework. To do the study, she taught her morning statistics class using the online homework and her afternoon class using traditional textbook homework. At the end of the semester, she compared the final exam scores for the online homework section to the textbook homework section. If the morning section had higher scores, could Professor Egner conclude that online homework is the cause of higher exam scores? Not necessarily. It is possible that the morning class had students who were more motivated. It is impossible to know whether the outcome was due to the online homework or to the time at which the class was taught. In this sense, we say that the time of day the class is taught is a *confounding variable*.

Definition

A **confounding variable** is an explanatory variable that was considered in a study whose effect cannot be distinguished from a second explanatory variable in the study.

The big difference between lurking variables and confounding variables is that lurking variables are not considered in the study (for example, we did not consider lifestyle in the pneumonia study) whereas confounding variables are measured in the study (for example, we measured morning versus afternoon classes).

So lurking variables are related to both the explanatory and response variables, and this relation is what creates the apparent association between the explanatory and response variable in the study. For example, lifestyle (healthy or not) is associated with the likelihood of getting an influenza shot as well as the likelihood of contracting pneumonia or influenza.

A confounding variable is a variable in a study that does not necessarily have any association with the other explanatory variable, but does have an effect on the response variable. Perhaps morning students are more motivated, and this is what led to the higher final exam scores, not the homework delivery system.

The bottom line is that both lurking variables and confounding variables can confound the results of a study, so a researcher should be mindful of their potential existence.

We will continue to look at obtaining data through various types of observational studies until Section 1.6. In Section 1.6, we will look at designed experiments.

② Explain the Various Types of Observational Studies

There are three major categories of observational studies: (1) cross-sectional studies, (2) case-control studies, and (3) cohort studies.

Cross-sectional Studies These observational studies collect information about individuals at a specific point in time or over a very short period of time.

For example, a researcher might want to assess the risk associated with smoking by looking at a group of people, determining how many are smokers, and comparing the rate of lung cancer of the smokers to the nonsmokers.

An advantage of cross-sectional studies is that they are cheap and quick to do. However, they have limitations. For our lung cancer study, individuals might develop cancer after the data are collected, so our study will not give the full picture.

Case-control Studies These studies are **retrospective**, meaning that they require individuals to look back in time or require the researcher to look at existing records. In case-control studies, individuals who have a certain characteristic may be matched with those who do not.

For example, we might match individuals who smoke with those who do not. When we say “match” individuals, we mean that we would like the individuals in the study to be as similar (homogeneous) as possible in terms of demographics and other variables that may affect the response variable. Once homogeneous groups are established, we would ask the individuals in each group how much they smoked over the past 25 years. The rate of lung cancer between the two groups would then be compared.

A disadvantage to this type of study is that it requires individuals to recall information from the past. It also requires the individuals to be truthful in their responses. An advantage of case-control studies is that they can be done relatively quickly and inexpensively.

Cohort Studies A cohort study first identifies a group of individuals to participate in the study (the cohort). The cohort is then observed over a long period of time. During this period, characteristics about the individuals are recorded and some individuals will be exposed to certain factors (not intentionally) and others will not. At the end of the study the value of the response variable is recorded for the individuals.

Typically, cohort studies require many individuals to participate over long periods of time. Because the data are collected over time, cohort studies are **prospective**. Another problem with cohort studies is that individuals tend to drop out due to the long time frame. This could lead to misleading results. That said, cohort studies are the most powerful of the observational studies.

One of the largest cohort studies is the Framingham Heart Study. In this study, more than 10,000 individuals have been monitored since 1948. The study continues to this day, with the grandchildren of the original participants taking part in the study. This cohort study is responsible for many of the breakthroughs in understanding heart disease. Its cost is in excess of \$10 million.

Some Concluding Remarks about Observational Studies versus Designed Experiments

Is a designed experiment always superior to an observational study? Not necessarily. Plus, observational studies play a role in the research process. For example, because cross-sectional and case-control observational studies are relatively inexpensive, they allow researchers to explore possible associations prior to undertaking large cohort studies or designing experiments.

Also, it is not always possible to conduct an experiment. For example, we could not conduct an experiment to investigate the perceived link between high tension wires and leukemia (on humans). Do you see why?

Census Data

Another source of data is a *census*.

Definition

A **census** is a list of all individuals in a population along with certain characteristics of each individual.

The United States conducts a census every 10 years to learn the demographic makeup of the United States. Everyone whose usual residence is within the borders of the United States must fill out a questionnaire packet. The cost of obtaining the census in 2010* was approximately \$5.4 billion; about 635,000 temporary workers were hired to assist in collecting the data.

Why is the U.S. Census so important? The results of the census are used to determine the number of representatives in the House of Representatives in each state, congressional districts, distribution of funds for government programs (such as Medicaid), and planning for the construction of schools and roads. The first census of the United States was obtained in 1790 under the direction of Thomas Jefferson. It is a constitutional mandate that a census be conducted every 10 years.

Is the United States successful in obtaining a census? Not entirely. Some individuals go uncounted due to illiteracy or language issues. People experiencing homelessness may also go uncounted. Given the political stakes that are based on the census, politicians often debate how to count these individuals. Statisticians have offered solutions to the counting problem. If you wish, go to www.census.gov and in the search box type *count homeless*. You will find many articles on the Census Bureau's attempt to count the homeless. The bottom line is that even census data can have flaws.

Obtaining Data through Web Scraping

Web scraping, or **data mining**, is the process of extracting data from the Internet. Web scraping can be used to extract data from tables on web pages and then upload the data to a file. Or, web scraping can be used to create a data set of words from an online article (that is, fetching unstructured information and transforming it into a structured format through something called *parsing* and *reformatting processes*). Web scraping can also be used to dynamically call information from websites with links. Web scraping of dynamic information is based on the fact that information on web pages is constantly changing, so some data users might want to learn information over time in an automated fashion. For example, researchers may want to learn information about prices, price changes, and sold-out items over a holiday weekend (such as the weekend after Thanksgiving).

Web scraping is becoming an important part of the data science industry. Because data is so valuable, companies scrape the web constantly to gain an edge on their competition and to learn more about their customers. Web scraping tools will search websites for information on pricing of certain products in order to do instant price comparisons. For example, companies scrape airline flight prices to find the best deals (cheapOair.com). Real estate websites scrape housing purchases to develop home price estimate algorithms (Zillow.com). Or, companies that link individuals with jobs will search the Internet for job openings and share new postings with subscribers who have posted skills that match the job (LinkedIn). Investment companies now scrape social media (primarily Twitter) to determine stockholders' sentiments to try to predict the community reaction to quarterly earnings announcements. Sports analytics firms scrape sports play-by-play information to develop ranking algorithms (Statcast). The list goes on and on.

Web scraping packages are being developed continuously within programming languages such as Python and R. Most data scientists prefer to scrape the web using Python's BeautifulSoup package. BeautifulSoup is widely known as the most advanced library for web scraping. In R, rvest and scrapeR are popular web scraping packages. It is important to realize that these tools are incredibly useful for scraping the web, but they are also limited. Without some understanding of programming languages used in building web pages, your web scraping abilities will be limited to HTML tables.

There are ethical issues associated with web scraping. After all, the scraping of data on these pages is often done without the permission of the host. A lot of dialogue is

*Costs of the 2020 census were not available at the time of printing. However, some estimates place its cost at over \$15 billion.

currently taking place in the “web capture” community surrounding the legality and ethics of web capturing. We experience many benefits as a result of web scraping (cheaper flights, the right job offers, and so on), but there are also many pitfalls (personal information is potentially sourced from a website you make a purchase from). What is the responsibility of the website host to protect your information (data)? What are your responsibilities to protect your information?

If you are interested in learning more about the methods of web scraping and the ethics surrounding the technique, type “Web Scraping” or “Web Scraping Ethics” in the search engine of your browser.

If you would like to try web scraping, consult the Student Activity Workbook that accompanies this text. There is an activity that introduces StatCrunchThis—a web scraping tool available with StatCrunch.

Downloading Data from the Web

More than ever you will find that government agencies, companies, and sports organizations regularly make data available to the public. Often, this data can be downloaded from their websites as csv (Excel) or txt (text) files. Then, the data can be uploaded into your favorite statistical analysis package (such as StatCrunch, Minitab, or R) or spreadsheet (such as Excel or Google Spreadsheet). The amount of data we have access to right now is vast and growing. For a small sample of some websites that have data available for download, go to <https://www.sullystats.com/resources>. This page is updated periodically to keep current with the immense amount of data available for analysis.



1.2 Assess Your Understanding

Vocabulary and Skill Building

1. In your own words, define explanatory variable and response variable.
 2. Match each word or phrase with its definition.

Word/Phrase	Definition
(a) Designed Experiment	<p>I. Occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory variable and the response variable may be due to some other variable not accounted for in the study.</p>
(b) Observational Study	<p>II. An explanatory variable that was considered in a study whose effect cannot be distinguished from a second explanatory variable in the study.</p>
(c) Lurking Variable	<p>III. A researcher randomly assigns the individuals in a study to groups, intentionally manipulates the value of an explanatory variable, controls other explanatory variables at fixed values, and then records the value of the response variable for each individual.</p>
(d) Confounding	<p>IV. An explanatory variable that was not considered in a study, but that affects the value of the response variable in the study. In addition, this variable is typically related to other explanatory variables in the study.</p>
(e) Confounding Variable	<p>V. A researcher measures the value of the response variable without attempting to influence the value of either the response or explanatory variables. That is, the researcher observes the behavior of individuals in the study and records the values of the explanatory and response variables.</p>

- 3.** Match each type of study to its definition.

Word	Definition
(a) Cohort Study	I. Studies that are retrospective, meaning they require the researcher to look at existing records, or the subject to recall information from the past. Individuals who have certain characteristics are matched with those who don't.
(b) Cross-sectional Study	II. Studies that follow a group of individuals over a long period of time. Characteristics of the individuals are recorded and some individuals will be exposed to certain factors (not intentionally) and others will not. Because the data are collected over time, these studies are prospective.
(c) Case-control Study	III. Studies that collect information about individuals at a specific point in time, or over a short period of time.

4. Which type of study allows the researcher to claim causation between an explanatory variable and a response variable?
 5. Given a choice, would you conduct a study using an observational study or a designed experiment? Why?
 6. The data used in the influenza study presented in Example 3 were obtained from a cohort study. What does this mean? Why is a cohort study superior to a case-control study?
 7. Explain why it would be unlikely to use a designed experiment to answer the research question posed in Example 3.
 8. What does it mean when an observational study is retrospective? What does it mean when an observational study is prospective?

In Problems 9–16, determine whether the study depicts an observational study or an experiment.

NW 9. Cancer Study The American Cancer Society is beginning a study to learn why some people never get cancer. To take part in the study, a person must be 30–65 years of age and never had cancer. The study requires that the participants fill out surveys about their health and habits and give blood samples and waist measurements. These surveys must be filled out every two years. The study is expected to last for the next 20 years.

10. Rats with cancer are divided into two groups. One group receives 5 milligrams (mg) of a medication that is thought to fight cancer, and the other receives 10 mg. After 2 years, the spread of the cancer is measured.

11. Seventh-grade students are randomly divided into two groups. One group is taught math using traditional techniques; the other is taught math using a reform method. After 1 year, each group is given an achievement test to compare proficiency.

12. Hair and Heart Disease A study in which balding men were compared with non-balding men at one point in time found that balding men were 70% more likely to have heart disease.

Source: USA Today, April 4, 2013.

13. A survey is conducted asking 400 people, “Do you prefer Coke or Pepsi?”

14. Two hundred people are asked to perform a taste test in which they drink from two randomly placed, unmarked cups and are asked which drink they prefer.

15. Sixty patients with carpal tunnel syndrome are randomly divided into two groups. One group is treated weekly with both acupuncture and an exercise regimen. The other is treated weekly with the exact same exercise regimen, but no acupuncture. After 1 year, both groups are questioned about their level of pain due to carpal tunnel syndrome.

16. Conservation agents netted 250 large-mouth bass in a lake and determined how many were carrying parasites.

Applying the Concepts

17. Happiness and Your Heart Is there an association between level of happiness and the risk of heart disease? Researchers studied 1739 people over a 10-year period and asked questions about their daily lives and the hassles they face. The researchers also determined which individuals in the study experienced any type of heart disease. After their analysis, they concluded that happy individuals are less likely to experience heart disease.

Source: European Heart Journal 31 (9):1065–1070, February 2010.

(a) What type of observational study is this? Explain.
 (b) What is the response variable? What is the explanatory variable?
 (c) In the report, the researchers stated that “the research team also hasn’t ruled out that a common factor like genetics could be causing both the emotions and the heart disease.” Use the language introduced in this section to explain what this sentence means.

18. Daily Coffee Consumption Is there an association between daily coffee consumption and the occurrence of skin cancer? Researchers asked 93,676 women to disclose their coffee-drinking habits and also determined which of the women had nonmelanoma skin cancer. The researchers concluded that consumption of six or more cups of caffeinated coffee per day was associated with a reduction in nonmelanoma skin cancer.

Source: European Journal of Cancer Prevention, 16(5): 446–452, October 2007.

- (a) What type of observational study was this? Explain.
- (b) What is the response variable in the study? What is the explanatory variable?
- (c) In their report, the researchers stated that “After adjusting for various demographic and lifestyle variables, daily consumption of six or more cups was associated with a 30% reduced prevalence of nonmelanoma skin cancer.” Why was it important to adjust for these variables?

NW 19. Television in the Bedroom Is a television (TV) in the bedroom associated with obesity? Researchers questioned 379 twelve-year-old adolescents and concluded that the body mass index (BMI) of the adolescents who had a TV in their bedroom was significantly higher than the BMI of those who did not have a TV in their bedroom.

Source: Christelle Delmas, Carine Platat, Brigitte Schweitzer, Aline Wagner, Mohamed Ouja, and Chantal Simon. “Association Between Television in Bedroom and Adiposity Throughout Adolescence,” Obesity, 15:2495–2503, 2007.

- (a) Why is this an observational study? What type of observational study is this?
- (b) What is the response variable in the study? What is the explanatory variable?
- (c) Can you think of any lurking variables that may affect the results of the study?
- (d) In the report, the researchers stated, “These results remain significant after adjustment for socioeconomic status.” What does this mean?
- (e) Can we conclude that a television in the bedroom causes a higher body mass index? Explain.

20. Get Married, Gain Weight Are young couples who marry or cohabit more likely to gain weight than those who stay single? Researchers followed 8000 men and women for 7 years. At the start of the study, none of the participants were married or living with a romantic partner. The researchers found that women who married or cohabitated during the study gained 9 pounds more than single women, and married or cohabitating men gained, on average, 6 pounds more than single men.

- (a) Why is this an observational study? What type of observational study is this?
- (b) What is the response variable in the study? What is the explanatory variable?
- (c) Identify some potential lurking variables in this study.
- (d) Can we conclude that getting married or cohabiting causes one to gain weight? Explain.

21. Midwives Researchers Sally Tracy and associates undertook a cross-sectional study looking at the method of delivery and cost of delivery for first-time “low risk” mothers under three delivery scenarios:

- (1) Caseload midwifery
- (2) Standard hospital care
- (3) Private obstetric care

The results of the study revealed that 58.5% of all births with midwifery were vaginal deliveries compared with 48.2% of standard hospital births and 30.8% of private obstetric care. In addition, the costs of delivery from midwifery was \$3903.78 compared with \$5279.23 for standard hospital care and \$5413.69 for private obstetric care.

Source: Sally K Tracy, Alec Welsh, Bev Hall, Donna Hartz, Anne Lainchbury, Andrew Bisits, Jan White, and Mark Tracy “Caseload midwifery compared to standard or private obstetric care for first time mothers in a public teaching hospital in Australia: a cross sectional study of cost and birth outcomes” BMC Pregnancy and Childbirth 2014, 14:46.

- (a) Why is this a cross-sectional observational study?
- (b) Name the explanatory variable in the study.
- (c) Name the two response variables in the study and determine whether each is qualitative or quantitative.

22. Web Page Design Magnum, LLC, is a web page design firm that has two designs for an online hardware store. To determine which is the more effective design, Magnum uses one page in the Denver area and a second page in the Miami area. For each visit, Magnum records the amount of time visiting the site and the amount spent by the visitor.

- (a) What is the explanatory variable in this study? Is it qualitative or quantitative?
- (b) What are the two response variables? For each response variable, state whether it is qualitative or quantitative.
- (c) Explain how confounding might be an issue with this study.

23. Analyze the Article Write a summary of the following opinion. The opinion was posted at abcnews.com. Include the type of study conducted, possible lurking variables, and conclusions. What is the message of the author of the article?

Power Lines and Cancer—To Move or Not to Move

New Research May Cause More Fear Than Warranted, One Physician Explains

OPINION by JOSEPH MOORE, M.D.

A recent study out of Switzerland indicates there might be an increased risk of certain blood cancers in people with prolonged exposure to electromagnetic fields, like those generated from high-voltage power lines.

If you live in a house near one of these high-voltage power lines, a study like this one might make you wonder whether you should move.

But based on what we know now, I don't think that's necessary. We can never say there is no risk, but we can say that the risk appears to be extremely small.

“Scare Science”

The results of studies like this add a bit more to our knowledge of potential harmful environmental exposures, but they should also be seen in conjunction with the results of hundreds of studies that have gone before. It cannot be seen as a definitive call to action in and of itself.

The current study followed more than 20,000 Swiss railway workers over a period of 30 years. True, that represents a lot of people over a long period of time.

However, the problem with many epidemiological studies, like this one, is that it is difficult to have an absolute control group of people to compare results with. The researchers compared the incidence of different cancers of workers with a high amount of electromagnetic field exposure to those workers with lower exposures.

These studies aren't like those that have identified definitive links between an exposure and a disease—like those involving smoking and lung cancer. In those studies, we can actually measure the damage done to lung tissue as a direct result of smoking. But usually it's very difficult for the conclusions of an epidemiological study to rise to the level of controlled studies in determining public policy.

Remember the recent scare about coffee and increased risk of pancreatic cancer? Or the always-simmering issue of cell phone use and brain tumors?

As far as I can tell, none of us have turned in our cell phones. In our own minds, we've decided that any links to cell phone use and brain cancer have not been proven definitively. While we can't say that there is absolutely no risk in using cell phones, individuals have determined on their own that the potential risks appear to be quite small and are outweighed by the benefits.

Findings Shouldn't Lead to Fear

As a society, we should continue to investigate these and other related exposures to try to prove one way or another whether they are disease-causing. If we don't continue to study, we won't find out. It's that simple.

When findings like these come out, and I'm sure there will be more in the future, I would advise people not to lose their heads. Remain calm. You should take the results as we scientists do—as intriguing pieces of data about a problem we will eventually learn more about, either positively or negatively, in the future. It should not necessarily alter what we do right now.

What we can do is take actions that we know will reduce our chances of developing cancer.

Stop smoking and avoid passive smoke. It is the leading cause of cancer that individuals have control over.

Whenever you go outside, put on sunscreen or cover up.

Eat a healthy diet and stay physically active.

Make sure you get tested or screened. Procedures like colonoscopies, mammograms, pap smears and prostate exams can catch the early signs of cancer, when the chances of successfully treating them are the best.

Taking the actions above will go much farther in reducing your risks for cancer than moving away from power lines or throwing away your cell phone.

Dr. Joseph Moore is a medical oncologist at Duke University Comprehensive Cancer Center.

Source: Copyright © by Joseph Moore.

24. Cellular Phones Researchers wanted to determine whether there is an association between mobile phone use and body mass index. To do so, 105,028 men and women aged 18 years or over from the United Kingdom were recruited and their cell-phone use behavior was studied (number of calls per day, number of hours per week, year cell phone was first used) along with other variables (amount of exercise, body mass index) of the individuals. The researchers found a strong positive association between duration of phone calls on a cell phone and body mass index (that is, as the duration of phone calls increases, body mass index tends to increase as well).

Source: Mireille B. Toledano, Rachel B. Smith, Irene Chang, Margaret Douglass, and Paul Elliott, “Cohort Profile: UK COSMOS—a UK cohort for study of environment and health,” *International Journal of Epidemiology*, 46(3):775–787, June 1, 2017, <https://doi.org/10.1093/ije/dyw203>

- (a) What type of observational study is this?

- (b) Many studies involving cell phones look for a link between cell-phone usage and negative health outcomes (such as stroke or cancer) due to radio-frequency exposure. The following quote is from the article: “Obesity is associated with health outcomes such as stroke and cancers, which are of interest in relation to radio frequency exposure, and therefore is potential for confounding.” Explain what this means.

25. A Flawed Retrospective Study In an infamous study, researchers suggested that left-handed individuals died younger than right-handed individuals. In the study, researchers identified 987 individuals who died in 1990 and then used historical records to determine birth year as well as whether the individual was right-handed or not. They found that individuals who were right-handed lived 75 years, on average, while those who were not right-handed (left-handed or ambidextrous) lived 66 years, on average. Explain the flaw in this retrospective study and point out the potential dangers in retrospective studies. *Hint:* In the early 1900s individuals were often pressured to become right-handed at an early age. This pressure subsided, and the percentage of individuals born around 1950 or later that are left-handed is around 10%–12%, the norm.

26. Putting It Together: Passive Smoke? The following abstract appears in *The New England Journal of Medicine*:

BACKGROUND. The relation between passive smoking and lung cancer is of great public health importance. Some previous studies have suggested that exposure to environmental tobacco smoke in the household can cause lung cancer, but others have found no effect. Smoking by the spouse has been the most commonly used measure of this exposure.

METHODS. In order to determine whether lung cancer is associated with exposure to tobacco smoke within the household, we conducted a case-control study of 191 patients with lung cancer who had never smoked and an equal number of persons without lung cancer who had never smoked. Lifetime residential histories including information on exposure to environmental tobacco smoke were compiled and analyzed. Exposure was measured in terms of “smoker-years,” determined by multiplying

the number of years in each residence by the number of smokers in the household.

RESULTS. Household exposure to 25 or more smoker-years during childhood and adolescence doubled the risk of lung cancer. Approximately 15 percent of the control subjects who had never smoked reported this level of exposure. Household exposure of less than 25 smoker-years during childhood and adolescence did not increase the risk of lung cancer. Exposure to a spouse’s smoking, which constituted less than one third of total household exposure on average, was not associated with an increase in risk.

CONCLUSIONS. The possibility of recall bias and other methodologic problems may influence the results of case-control studies of environmental tobacco smoke. Nonetheless, our findings regarding exposure during early life suggest that approximately 17 percent of lung cancers among nonsmokers can be attributed to high levels of exposure to cigarette smoke during childhood and adolescence.

- (a) What is the research objective?
 - (b) What makes this study a case-control study? Why is this a retrospective study?
 - (c) What is the response variable in the study? Is it qualitative or quantitative?
 - (d) What is the explanatory variable in the study? Is it qualitative or quantitative?
 - (e) Can you identify any lurking variables that may have affected this study?
 - (f) What is the conclusion of the study? Can we conclude that exposure to smoke in the household causes lung cancer?
 - (g) Would it be possible to design an experiment to answer the research question in part (a)? Explain.
27. Name three ways that web scraping can be used to obtain data.
28. Discuss the ethics behind scraping data from the Internet. In particular, answer the following questions. What is the responsibility of the website host to protect your information (data)? What are your responsibilities to protect your information? For assistance, type “Web Scraping” or “Web Scraping Ethics” in the search engine of your browser.

1.3 Simple Random Sampling

Objective ① Obtain a simple random sample



Sampling

Besides the observational studies that we looked at in Section 1.2, observational studies can also be conducted by administering a survey. When administering a survey, the researcher must first identify the population that is to be targeted. For example, the Gallup Organization regularly surveys Americans about various pop-culture and political issues. Often, the population of interest is Americans aged 18 years or older. Of course, the Gallup Organization cannot survey *all* adult Americans (there are over 200 million), so instead the group typically surveys a *random sample* of about 1000 adult Americans.

Definition

Random sampling is the process of using chance to select individuals from a population to be included in the sample.

For the results of a survey to be reliable, the characteristics of the individuals in the sample must be representative of the characteristics of the individuals in the population. The key to obtaining a sample representative of a population is to let *chance* or *randomness* play a role in dictating which individuals are in the sample, rather than convenience. **If convenience is used to obtain a sample, the results of the survey are meaningless.**

Suppose that Gallup wants to know the proportion of adult Americans who consider themselves to be baseball fans. If Gallup obtained a sample by standing outside of Fenway Park (home of the Boston Red Sox professional baseball team), the survey results are not likely to be reliable. Why? Clearly, the individuals in the sample do not accurately reflect the makeup of the entire population. As another example, suppose you wanted to learn the proportion of students on your campus who work. It might be convenient to survey the students in your statistics class, but do these students represent the overall student body? Does the proportion of freshmen, sophomores, juniors, and seniors in your class mirror the proportion of freshmen, sophomores, juniors, and seniors on campus? Does the proportion of males and females in your class resemble the proportion of males and females across campus? Probably not. For this reason, the convenient sample is not representative of the population, which means any results reported from your survey are misleading.

We will discuss four basic sampling techniques: *simple random sampling*, *stratified sampling*, *systematic sampling*, and *cluster sampling*. These sampling methods are designed so that any selection biases introduced (knowingly or unknowingly) by the surveyor during the selection process are eliminated. In other words, the surveyor does not have a choice as to which individuals are in the study. We will discuss simple random sampling now and the remaining three types of sampling in Section 1.4.

1 Obtain a Simple Random Sample

The most basic sample survey design is *simple random sampling*.

Definition

A sample of size n from a population of size N is obtained through **simple random sampling** if every possible sample of size n has an equally likely chance of occurring. The sample is then called a **simple random sample**.

IN OTHER WORDS

Simple random sampling is like selecting names from a hat.

The number of individuals in the sample is always less than the number of individuals in the population.

EXAMPLE 1

Illustrating Simple Random Sampling

Problem Sophia has four tickets to a concert. Six of her friends, Yolanda, Michael, Kevin, Marissa, Annie, and Katie, have all expressed an interest in going to the concert. Sophia decides to randomly select three of her six friends to attend the concert.

- List all possible samples of size $n = 3$ from the population of size $N = 6$. Once an individual is chosen, he or she cannot be chosen again.
- Comment on the likelihood of the sample containing Michael, Kevin, and Marissa.

Approach List all possible combinations of three people chosen from the six. Remember, in simple random sampling, each sample of size 3 is equally likely to occur.

Solution

- The possible samples of size 3 are listed in Table 2.

Table 2

Yolanda, Michael, Kevin	Yolanda, Michael, Marissa	Yolanda, Michael, Annie	Yolanda, Michael, Katie
Yolanda, Kevin, Marissa	Yolanda, Kevin, Annie	Yolanda, Kevin, Katie	Yolanda, Marissa, Annie
Yolanda, Marissa, Katie	Yolanda, Annie, Katie	Michael, Kevin, Marissa	Michael, Kevin, Annie
Michael, Kevin, Katie	Michael, Marissa, Annie	Michael, Marissa, Katie	Michael, Annie, Katie
Kevin, Marissa, Annie	Kevin, Marissa, Katie	Kevin, Annie, Katie	Marissa, Annie, Katie

NW Now Work Problem 7

From Table 2, we see that there are 20 possible samples of size 3 from the population of size 6. The term *sample* means the individuals in the sample.

- (b) Only 1 of the 20 possible samples contains Michael, Kevin, and Marissa, so there is a 1 in 20 chance that the simple random sample will contain these three. In fact, all the samples of size 3 have a 1 in 20 chance of occurring.

IN OTHER WORDS

A frame lists all the individuals in a population. For example, a list of all registered voters in a particular precinct might be a frame.

Obtaining a Simple Random Sample

The results of Example 1 leave one question unanswered: How do we select the individuals in a simple random sample? We could write the names of the individuals in the population on different sheets of paper and then select names from a hat. Often, however, the size of the population is so large that performing simple random sampling in this fashion is not practical. Instead, each individual in the population is assigned a unique number between 1 and N , where N is the size of the population. Then n distinct random numbers from this list are selected, where n represents the size of the sample. To number the individuals in the population, we need a **frame**—a list of all the individuals within the population.

EXAMPLE 2

Obtaining a Simple Random Sample Using a Table of Random Numbers

Problem The accounting firm of Senese and Associates has grown. To make sure their clients are still satisfied with the services they are receiving, the company decides to send a survey out to a simple random sample of 5 of its 30 clients.

Approach

Step 1 The clients must be listed (the frame) and numbered from 01 to 30.

Step 2 Five unique numbers will be randomly selected. The clients corresponding to the numbers are sent a survey. This process is called *sampling without replacement*. In a **sample without replacement**, an individual who is selected is removed from the population and cannot be chosen again. In a **sample with replacement**, a selected individual is placed back into the population and could be chosen a second time. We use sampling without replacement so that we don't select the same client twice.

Solution

Step 1 Table 3 shows the list of clients. We arrange them in alphabetical order, although this is not necessary, and number them from 01 to 30.

Table 3

01. ABC Electric	11. Fox Studios	21. R&Q Realty
02. Brassil Construction	12. Haynes Hauling	22. Ritter Engineering
03. Bridal Zone	13. House of Hair	23. Simplex Forms
04. Casey's Glass House	14. John's Bakery	24. Spruce Landscaping
05. Chicago Locksmith	15. Logistics Management, Inc.	25. Thors, Robert DDS
06. DeSoto Painting	16. Lucky Larry's Bistro	26. Travel Zone
07. Dino Jump	17. Moe's Exterminating	27. Ultimate Electric
08. Euro Car Care	18. Nick's Tavern	28. Venetian Gardens Restaurant
09. Farrell's Antiques	19. Orion Bowling	29. Walker Insurance
10. First Fifth Bank	20. Precise Plumbing	30. Worldwide Wireless

Step 2 A table of random numbers can be used to select the individuals to be in the sample. See Table 4 on the next page.* We pick a starting place in the table by closing

*Each digit is in its own column. The digits are displayed in groups of five for ease of reading. The digits in row 1 are 893922321274483, and so on. The first digit, 8, is in column 1; the second digit, 9, is in column 2; the ninth digit, 1, is in column 9.

(continued)

Table 4

Column 4

Row Number	Column Number									
	01–05	06–10	11–15	16–20	21–25	26–30	31–35	36–40	41–45	46–50
01	89392	23212	74483	36590	25956	36544	68518	40805	09980	00467
02	61458	17639	96252	95649	73727	33912	72896	66218	52341	97141
03	11452	74197	81962	48433	90360	26480	73231	37740	26628	44690
04	27575	04429	31308	02241	01698	19191	18948	78871	36030	23980
05	36829	59109	88976	46845	28329	47460	88944	08264	00843	84592
06	81902	93458	42161	26099	09419	89073	82849	09160	61845	40906
07	59761	55212	33360	68751	86737	79743	85262	31887	37879	17525
08	46827	25906	64708	20307	78423	15910	86548	08763	47050	18513
09	24040	66449	32353	83668	13874	86741	81312	54185	78824	00718
10	98144	96372	50277	15571	82261	66628	31457	00377	63423	55141
11	14228	17930	30118	00438	49666	65189	62869	31304	17117	71489
12	55366	51057	90065	14791	62426	02957	85518	28822	30588	32798
13	96101	30646	35526	90389	73634	79304	96635	06626	94683	16696
14	38152	55474	30153	26525	83647	31988	82182	98377	33802	80471
15	85007	18416	24661	95581	45868	15662	28906	36392	07617	50248
16	85544	15890	80011	18160	33468	84106	40603	01315	74664	20553
17	10446	20699	98370	17684	16932	80449	92654	02084	19985	59321
18	67237	45509	17638	65115	29757	80705	82686	48565	72612	61760
19	23026	89817	05403	82209	30573	47501	00135	33955	50250	72592
20	67411	58542	18678	46491	13219	84084	27783	34508	55158	78742

We skip 52 because it is larger than 30.

our eyes and placing a finger on it. This method accomplishes the goal of being random. Suppose we start in column 4, row 13. Because our data have two digits, we select two-digit numbers from the table using columns 4 and 5. We select numbers between 01 and 30, inclusive, and skip 00, numbers greater than 30, and numbers already selected.

The first number in the list is 01, so the client corresponding to 01 will receive a survey. Reading down, the next number in the list is 52, which is greater than 30, so we skip it. Continuing down the list, the following numbers are selected from the list:

01, 07, 26, 11, 23

We display each of the random numbers used to select the individuals in the sample in boldface type in Table 4 to help you to understand where they came from. The clients corresponding to these numbers are

ABC Electric, Dino Jump, Travel Zone, Fox Studios, Simplex Forms



EXAMPLE 3 Obtaining a Simple Random Sample Using Technology

Problem Find a simple random sample of five clients for the problem presented in Example 2.

Approach The approach is similar to that given in Example 2.

Step 1 Obtain the frame and assign the clients numbers from 01 to 30.

Step 2 Randomly select five numbers using a random number generator. To do this, we must first set the **seed**. The **seed** is an initial point for the generator to start creating

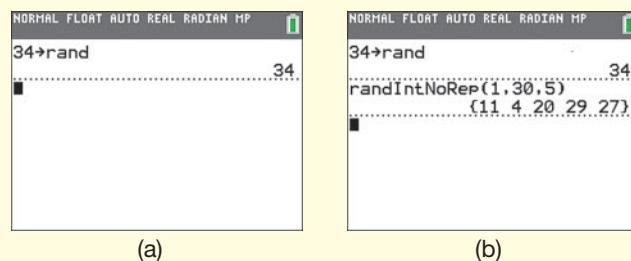
random numbers—like selecting the initial point in the table of random numbers. The seed can be any nonzero number. Statistical software such as StatCrunch, Minitab, or Excel can be used to generate random numbers, but we will use a TI-84 Plus C graphing calculator. The steps for obtaining random numbers using StatCrunch, Minitab, Excel, and the TI-83/84 Plus/84C graphing calculator can be found in the Technology Step-by-Step shown below.

Solution

Step 1 Table 3 on page 25 shows the list of clients and numbers corresponding to the clients.

Step 2 Figure 3(a) shows the seed set at 34 on a TI-84 Plus C graphing calculator. Now we can generate a list of random numbers, which are shown in Figure 3(b).

Figure 3



The following numbers are generated by the calculator:

11, 4, 20, 29, 27

The clients corresponding to these numbers are the clients to be surveyed: Fox Studios, Casey's Glass House, Precise Plumbing, Walker Insurance, and Ultimate Electric.

Using Technology

If you are using a different statistical package or type of calculator, the random numbers generated will likely be different. This does not mean you are wrong. There is no such thing as a wrong random sample as long as the correct procedures are followed.

NW Now Work Problem 11

CAUTION!

Random-number generators are not truly random, because they are programs, and programs do not act “randomly.” The seed dictates the random numbers that are generated.

Notice an important difference in the solutions of Examples 2 and 3. Because both samples were obtained randomly, they resulted in different individuals in the sample! For this reason, each sample will likely result in different descriptive statistics. Any inference based on each sample *may* result in different conclusions regarding the population. This is the nature of statistics. Inferences based on samples will vary because the individuals in different samples vary.

Technology Step-by-Step

Obtaining a Simple Random Sample

TI-83/84 Plus

- Enter any nonzero number (the seed) on the HOME screen.
- Press the STO ▶ button.
- Press the MATH button.
- Highlight the PRB menu and select 1: rand.
- From the HOME screen press ENTER.
- Press the MATH button. Highlight the PRB menu and select 5: randInt(.
- With randInt(on the HOME screen, enter 1, N), where N is the population size. For example, if N = 500, enter the following:

$\text{randInt}(1,500)$

Press ENTER to obtain the first individual in the sample. Continue pressing ENTER until the desired sample size is obtained.

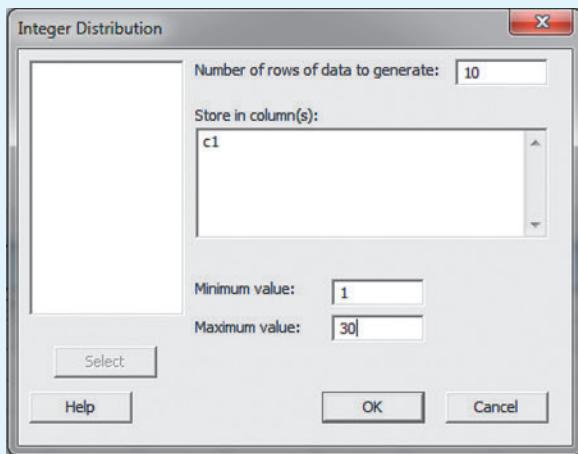
TI-84 Plus C

- Enter any nonzero number (the seed) on the HOME screen.
- Press the STO ▶ button.
- Press the MATH button.
- Highlight the PROB menu and select 1: rand.
- From the HOME screen press ENTER.
- Press the MATH button.
- Highlight the PROB menu and select 8: randIntNoRep(.

8. Type in the values for lower, upper, and n .
9. Highlight Paste. Press ENTER. Press ENTER a second time from the HOME screen.

Minitab

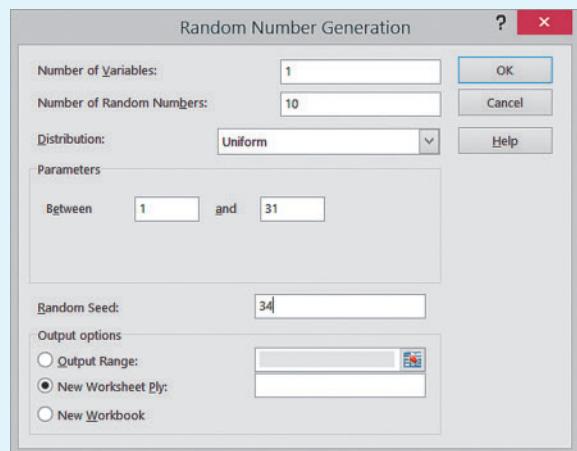
1. Select the **Calc** menu and highlight **Set Base . . .**
2. Enter any seed number you desire. Note that it is not necessary to set the seed, because Minitab uses the time of day in seconds to set the seed.
3. Select the **Calc** menu, highlight **Random Data**, and select **Integer . . .**
4. Fill in the following window with the appropriate values. To obtain a simple random sample for the situation in Example 2, we would enter the following:



Generate 10 rows of data (instead of 5) in case any of the random numbers repeat. Select **OK**, and the random numbers will appear in column 1 (C1) in the spreadsheet.

Excel

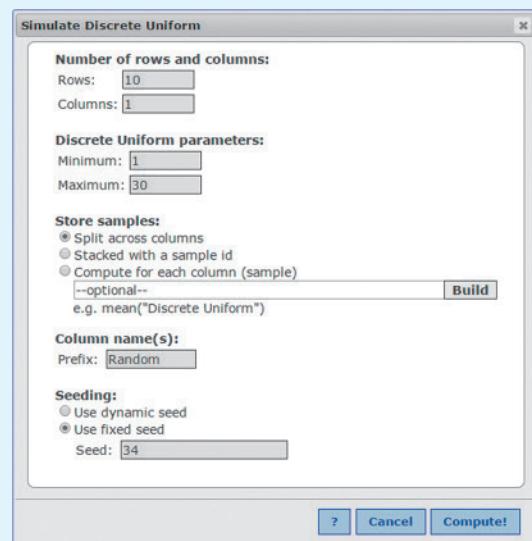
1. Be sure the Data Analysis ToolPak is activated. This is done by selecting **File** and then **Options**. Select **Add-Ins**. Under **Manage**, select **Excel Add-Ins**. Click **Go . . .**. Check the **Analysis ToolPak** box. Click **OK**.
2. Select **Data** and then **Data Analysis**. Highlight **Random Number Generation** and select **OK**.
3. Fill in the window with the appropriate values. To obtain a simple random sample for the situation in Example 2, fill in the window as shown in the next column. Generate 10 rows of data (instead of 5) in case any of the random numbers repeat. Notice also that the parameter is between 1 and 31, so any value greater than or equal to 1 and less than or equal to 31 is possible. In the unlikely event that 31 appears, simply ignore it. Select **OK** and



the random numbers will appear in column 1 (A1) in the spreadsheet. Ignore any values to the right of the decimal place.

StatCrunch

1. Select **Data**, highlight **Simulate**, then select **Discrete Uniform**.
2. Fill in the window with the appropriate values. To obtain a simple random sample for the situation in Example 2, enter the values shown in the figure. Generate 10 rows of data (instead of 5) in case any of the random numbers repeat. Click **Compute!**, and the random numbers will appear in the spreadsheet. *Note:* You could also select the dynamic seed radio button, if you like, to set the seed.



1.3 Assess Your Understanding

Vocabulary and Skill Building

1. What is a frame?
2. Define simple random sampling.

3. What does it mean when sampling is done without replacement?
4. What is random sampling? Why is it used and how does it compare with convenience sampling?

- 5. Literature** As part of a college literature course, students must read three classic works of literature from the provided list. Obtain a simple random sample of size 3 from this list. Write a short description of the process you used to generate your sample.

Pride and Prejudice	The Scarlet Letter
As I Lay Dying	The Jungle
Death of a Salesman	Huckleberry Finn
The Sun Also Rises	Crime and Punishment
A Tale of Two Cities	

- 6. Team Captains** A coach must select two players to serve as captains. He wants to randomly select two players to be the captains. Obtain a simple random sample of size 2 from the following list: Mady, Breanne, Evin, Tori, Emily, Clair, Caty, Jory, Payton, Jordyn. Write a short description of the process you used to generate your sample.

- NW 7. Course Selection** A student entering a doctoral program in educational psychology is required to select two courses from the list of courses provided as part of his or her program.

EPR 616, *Research in Child Development*
 EPR 630, *Educational Research Planning and Interpretation*
 EPR 631, *Nonparametric Statistics*
 EPR 632, *Methods of Multivariate Analysis*
 EPR 645, *Theory of Measurement*
 EPR 649, *Fieldwork Methods in Educational Research*
 EPR 650, *Interpretive Methods in Educational Research*

- (a) List all possible two-course selections.
 (b) Comment on the likelihood that the pair of courses EPR 630 and EPR 645 will be selected.

- 8. Merit Badge Requirements** To complete the Citizenship in the World merit badge, one must select two of the following seven organizations and describe their role in the world.

Source: Boy Scouts of America

1. The United Nations
2. The World Court
3. World Organization of the Scout Movement
4. The World Health Organization
5. Amnesty International
6. The International Committee of the Red Cross
7. CARE

- (a) List all possible pairs of organizations.
 (b) Comment on the likelihood that the pair “The United Nations” and “Amnesty International” will be selected.

Applying the Concepts

- 9. Sampling the Faculty** A community college employs 87 full-time faculty members. To gain the faculty’s opinions about an upcoming building project, the college president wishes to obtain a simple random sample that will consist of 9 faculty members. He numbers the faculty from 1 to 87.

- (a) Using Table I from Appendix A, the president closes his eyes and drops his ink pen on the table. It points to the digit in row 5, column 22. Using this position as the starting point and proceeding downward, determine the numbers for the 9 faculty members who will be included in the sample.
 (b) If the president uses technology, determine the numbers for the 9 faculty members who will be included in the sample.

- 10. Sampling the Students** The same community college from Problem 9 has 7656 students currently enrolled in classes. To gain the students’ opinions about an upcoming building project, the college president wishes to obtain a simple random sample of 20 students. He numbers the students from 1 to 7656.

- (a) Using Table I from Appendix A, the president closes his eyes and drops his ink pen on the table. It points to the digit in row 11, column 32. Using this position as the starting point and proceeding downward, determine the numbers for the 20 students who will be included in the sample.

- (b) If the president uses technology, determine the numbers for the 20 students who will be included in the sample.

- NW 11. Obtaining a Simple Random Sample** Suppose the president  of the United States wants to hold a forum with the governors of ten states. Use the frame shown to answer parts (a) and (b).

1. Alabama	18. Louisiana	35. Ohio
2. Alaska	19. Maine	36. Oklahoma
3. Arizona	20. Maryland	37. Oregon
4. Arkansas	21. Massachusetts	38. Pennsylvania
5. California	22. Michigan	39. Rhode Island
6. Colorado	23. Minnesota	40. South Carolina
7. Connecticut	24. Mississippi	41. South Dakota
8. Delaware	25. Missouri	42. Tennessee
9. Florida	26. Montana	43. Texas
10. Georgia	27. Nebraska	44. Utah
11. Hawaii	28. Nevada	45. Vermont
12. Idaho	29. New Hampshire	46. Virginia
13. Illinois	30. New Jersey	47. Washington
14. Indiana	31. New Mexico	48. West Virginia
15. Iowa	32. New York	49. Wisconsin
16. Kansas	33. North Carolina	50. Wyoming
17. Kentucky	34. North Dakota	

- (a) Obtain a simple random sample of size 10 using Table I in Appendix A, a graphing calculator, or computer software.
 (b) Obtain a second simple random sample of size 10 using Table I in Appendix A, a graphing calculator, or computer software.

- DAT 12. Obtaining a Simple Random Sample** The following table lists the 45 presidents of the United States.

1. Washington	16. Lincoln	31. Hoover
2. J. Adams	17. A. Johnson	32. F. D. Roosevelt
3. Jefferson	18. Grant	33. Truman
4. Madison	19. Hayes	34. Eisenhower
5. Monroe	20. Garfield	35. Kennedy
6. J. Q. Adams	21. Arthur	36. L. B. Johnson
7. Jackson	22. Cleveland	37. Nixon
8. Van Buren	23. B. Harrison	38. Ford
9. W. H. Harrison	24. Cleveland	39. Carter
10. Tyler	25. McKinley	40. Reagan
11. Polk	26. T. Roosevelt	41. G. H. Bush
12. Taylor	27. Taft	42. Clinton
13. Fillmore	28. Wilson	43. G. W. Bush
14. Pierce	29. Harding	44. Obama
15. Buchanan	30. Coolidge	45. Trump

- (a) Obtain a simple random sample of size 8 using Table I in Appendix A, a graphing calculator, or computer software.
 (b) Obtain a second simple random sample of size 8 using Table I in Appendix A, a graphing calculator, or computer software.

13. Obtaining a Simple Random Sample The president of the student government wants to conduct a survey to determine the student body's opinion regarding student services. The administration provides you with a list of the names and phone numbers of the 19,935 registered students.

(a) Discuss the procedure you would follow to obtain a simple random sample of 25 students.

(b) Obtain this sample.

14. Obtaining a Simple Random Sample The mayor of Justice, Illinois, asks you to poll the residents of the village and provides you with a list of the names and phone numbers of the 5832 residents of the village.

(a) Discuss the procedure you would follow to obtain a simple random sample of 20 residents.

(b) Obtain this sample.

DATA 15. Future Government Club The Future Government Club wants to sponsor a panel discussion on the upcoming national election. The club wants four of its members to lead the panel discussion. Obtain a simple random sample of size 4 from the below table. Write a short description of the process you used to generate your sample.

Blouin	Fallenbuchel	Niemeyer	Rice
Bolden	Grajewski	Nolan	Salihar
Bolt	Haydra	Ochs	Tate
Carter	Keating	Opacian	Thompson
Cooper	Khouri	Pawlak	Trudeau
Debold	Lukens	Pechtold	Washington
De Young	May	Ramirez	Wright
Engler	Motola	Redmond	Zenkel

DATA 16. Worker Morale The owner of a private food store is concerned about employee morale. She decides to survey the employees to learn about work environment and job satisfaction. Obtain a simple random sample of size 5 from the names in the given table. Write a short description of the process you used to generate your sample.

Archer	Foushi	Kemp	Oliver
Bolcerek	Gow	Lathus	Orsini
Bryant	Grove	Lindsey	Salazar
Carlisle	Hall	Massie	Ullrich
Cole	Hills	McGuffin	Vaneck
Dimas	Houston	Musa	Weber
Ellison	Kats	Nickas	Zavodny
Everhart			

DATA 17. Chicago High Schools Open the data set 1_3_17 from www.pearsonhighered.com/sullivanstats. The data set represents a list of every high school in the city of Chicago. Suppose you wish to conduct a survey of all the students enrolled for a simple random sample of 8 high schools in the city of Chicago. Record the name of the 8 high schools (individuals) selected. Write a description of the process you used to generate your sample.

1.4 Other Effective Sampling Methods



- Objectives**
- ① Obtain a stratified sample
 - ② Obtain a systematic sample
 - ③ Obtain a cluster sample

The goal of sampling is to obtain as much information as possible about the population at the least cost. Remember, we are using the word *cost* in a general sense. Cost includes monetary outlays, time, and other resources. With this goal in mind, it may be advantageous to use sampling techniques other than simple random sampling.

1 Obtain a Stratified Sample

Under certain circumstances, *stratified sampling* provides more information about the population for less cost than simple random sampling.

Definition

A **stratified sample** is obtained by separating the population into nonoverlapping groups called *strata* and then obtaining a simple random sample from each stratum. The individuals within each stratum should be homogeneous (or similar) in some way.

IN OTHER WORDS

Stratum is singular, while **strata** is plural. The word **strata** means divisions. So a stratified sample is a simple random sample of different divisions of the population.

For example, suppose Congress was considering a bill that abolishes estate taxes. In an effort to determine the opinion of her constituency, a senator asks a pollster to conduct a survey within her state. The pollster may divide the population of registered voters within the state into three strata: Republican, Democrat, and Independent. This grouping makes sense because the members within each of the three parties may have the same opinion regarding estate taxes, but opinions between parties may differ. The main criterion in performing a stratified sample is that each group (stratum) must have a common attribute that results in the individuals being similar within the stratum.

An advantage of stratified sampling over simple random sampling is that it may allow fewer individuals to be surveyed while obtaining the same or more information. This result occurs because individuals within each subgroup have similar characteristics, so opinions within the group are not as likely to vary much from one individual to the next. In addition, a stratified sample guarantees that each stratum is represented in the sample.

EXAMPLE 1**Obtaining a Stratified Sample**

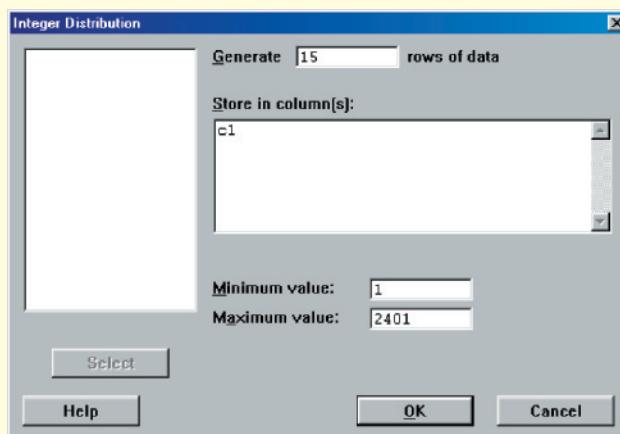
Problem The president of DePaul University wants to conduct a survey to determine the community's opinion regarding campus safety. The president divides the DePaul community into three groups: resident students, nonresident (commuting) students, and staff (including faculty) so that he can obtain a stratified sample. Suppose there are 6204 resident students, 13,304 nonresident students, and 2401 staff, for a total of 21,909 individuals in the population. The president wants to obtain a sample of size 100, with the number of individuals selected from each stratum weighted by the population size. So resident students make up $6204/21,909 = 28\%$ of the sample, nonresident students account for 61% of the sample, and staff constitute 11% of the sample. A sample of size 100 requires a stratified sample of $0.28(100) = 28$ resident students, $0.61(100) = 61$ nonresident students, and $0.11(100) = 11$ staff.

Approach To obtain the stratified sample, conduct a simple random sample within each group. That is, obtain a simple random sample of 28 resident students (from the 6204 resident students), a simple random sample of 61 nonresident students, and a simple random sample of 11 staff. Be sure to use a different seed for each stratum.

Solution Using Minitab, with the seed set to 4032 and the values shown in Figure 4, we obtain the following sample of staff:

240, 630, 847, 190, 2096, 705, 2320, 323, 701, 471, 744

Figure 4

**CAUTION!**

Do not use the same seed (or starting point in Table I) for each stratum in a stratified sample, because we want the simple random samples within each stratum to be independent of each other.

Repeat this procedure for the resident and nonresident students using a different seed.



An advantage of stratified sampling over simple random sampling is that the researcher can determine characteristics within each stratum. This allows an analysis to be performed on each stratum to see if any significant differences among them exist. For example, we could analyze the data obtained in Example 1 to see if there is a difference in the opinions of students versus staff.

② Obtain a Systematic Sample

In both simple random sampling and stratified sampling, a frame—a list of the individuals in the population being studied—must exist. Therefore, these sampling techniques require some preliminary work before obtaining the sample. A sampling technique that does not require a frame is *systematic sampling*.

Definition

A **systematic sample** is obtained by selecting every k th individual from the population. The first individual selected corresponds to a random number between 1 and k .

Because systematic sampling does not require a frame, it is a useful technique when you cannot obtain a list of the individuals in the population.

To obtain a systematic sample, select a number k , randomly select a number between 1 and k and survey that individual, then survey every k th individual thereafter. For example, we might decide to survey every $k = 8$ th individual. Now randomly select a number between 1 and 8, such as 5. This means we survey the 5th, $5 + 8 = 13$ th, $13 + 8 = 21$ st, $21 + 8 = 29$ th, and so on, individuals until we reach the desired sample size.

EXAMPLE 2

Obtaining a Systematic Sample without a Frame

Problem The manager of Kroger Food Stores wants to measure the satisfaction of the store's customers. Design a sampling technique that can be used to obtain a sample of 40 customers.

Approach A frame of Kroger customers would be difficult, if not impossible, to obtain. Therefore, it is reasonable to use systematic sampling by surveying every k th customer who leaves the store.

Solution The manager decides to obtain a systematic sample by surveying every 7th customer. He randomly determines a number between 1 and 7, say 5. He then surveys the 5th customer exiting the store and every 7th customer thereafter, until he has a sample of 40 customers. The survey will include customers 5, 12, 19, . . . , 278.*

But how do we select the value of k ? If the size of the population is unknown, there is no mathematical way to determine k . The value of k must be small enough to achieve our desired sample size, and large enough to obtain a sample that is representative of the population.

To clarify this point, let's revisit Example 2. If k is too large, say 30, we will survey every 30th shopper, starting with the 5th. A sample of size 40 would require that 1175 shoppers visit Kroger on that day. If Kroger does not have 1175 shoppers, the desired sample size will not be achieved. On the other hand, if k is too small, say 4, the store would survey the 5th, 9th, . . . , 161st shopper. The 161st shopper might exit the store at 3 P.M., so our survey would not include any of the evening shoppers. This sample is not representative of *all* Kroger patrons! An estimate of the size of the population would help to determine an appropriate value for k .

*Because we are surveying 40 customers, the first individual surveyed is the 5th, the second is the $5 + 7 = 12$ th, the third is the $5 + (2)7 = 19$ th, and so on, until we reach the 40th, which is the $5 + (39)7 = 278$ th shopper.

To determine the value of k when the size of the population, N , is known is relatively straightforward. Suppose the population size is $N = 20,325$ and we desire a sample of size $n = 100$. To guarantee that individuals are selected evenly from both the beginning and the end of the population (such as early and late shoppers), compute N/n and round down to the nearest integer. For example, $20,325/100 = 203.25$, so $k = 203$. Then randomly select a number between 1 and 203 and select every 203rd individual thereafter. So, if we randomly selected 90 as the starting point, we would survey the 90th, 293rd, 496th, . . . , 20,187th individuals.

We summarize the procedure as follows:

Steps in Systematic Sampling

1. If possible, approximate the population size, N .
2. Determine the sample size desired, n .
3. Compute $\frac{N}{n}$ and round down to the nearest integer. This value is k .
4. Randomly select a number between 1 and k . Call this number p .
5. The sample will consist of the following individuals:

$$p, p + k, p + 2k, \dots, p + (n - 1)k$$

Because systematic sampling does not require a frame, it typically provides more information for a given cost than does simple random sampling. In addition, systematic sampling is easier to employ, so there is less likelihood of interviewer error occurring, such as selecting the wrong individual to be surveyed.

NW Now Work Problem 27

3 Obtain a Cluster Sample

A fourth sampling method is called *cluster sampling*. Like the previous three sampling methods, this method has benefits under certain circumstances.

Definition

A **cluster sample** is obtained by selecting all individuals within a randomly selected collection or group of individuals.

IN OTHER WORDS

Imagine a parking lot. Each subsection of the lot could be a cluster (Section F-4, for example).

Suppose a school administrator wants to learn the characteristics of students enrolled in online classes. Rather than obtaining a simple random sample based on the frame of all students enrolled in online classes, the administrator could treat each online class as a cluster and then obtain a simple random sample of these clusters. The administrator would then survey *all* the students in the selected clusters. This reduces the number of classes that get surveyed.

EXAMPLE 3

Obtaining a Cluster Sample

Problem A sociologist wants to gather data regarding household income within the city of Boston. Obtain a sample using cluster sampling.

Approach The city of Boston can be set up so that each city block is a cluster. Once the city blocks have been identified, obtain a simple random sample of the city blocks and survey all households on the blocks selected.

Solution Suppose there are 10,493 city blocks in Boston. First, the sociologist must number the blocks from 1 to 10,493. Suppose the sociologist has enough time and money to survey 20 clusters (city blocks). The sociologist should obtain a simple

(continued)

CAUTION!

Stratified and cluster samples are different. In a stratified sample, we divide the population into two or more homogeneous groups. Then we obtain a simple random sample from each group. In a cluster sample, we divide the population into groups, obtain a simple random sample of some of the groups, and survey *all* individuals in the selected groups.

random sample of 20 numbers between 1 and 10,493 and survey all households from the clusters selected. Cluster sampling is a good choice in this example because it reduces the travel time to households that is likely to occur with both simple random sampling and stratified sampling. In addition, there is no need to obtain a frame of all the households with cluster sampling. The only frame needed is one that provides information regarding city blocks.



The following are a few of the questions that arise in cluster sampling:

- How do I cluster the population?
- How many clusters do I sample?
- How many individuals should be in each cluster?

First, we must determine whether the individuals within the proposed cluster are homogeneous (similar individuals) or heterogeneous (dissimilar individuals). In Example 3, city blocks tend to have similar households. Survey responses from houses on the same city block are likely to be similar. This results in duplicate information. We conclude that if the clusters have homogeneous individuals it is better to have more clusters with fewer individuals in each cluster.

What if the cluster is heterogeneous? Under this circumstance, the heterogeneity of the cluster likely resembles the heterogeneity of the population. In other words, each cluster is a scaled-down representation of the overall population. For example, a quality-control manager might use shipping boxes that contain 100 light bulbs as a cluster, since the rate of defects within the cluster would resemble the rate of defects in the population, assuming the bulbs are randomly placed in the box. Thus, when each cluster is heterogeneous, fewer clusters with more individuals in each cluster are appropriate.

NW Now Work Problem 13

Convenience Sampling

In the four sampling techniques just presented, the individuals are selected randomly. Inappropriate sampling methods are those in which the individuals are not randomly selected.

Have you ever watched a talk show where the host asks listeners to reply to a poll through Twitter? This is a nonscientific data collection method, so the results of any analysis are suspect because the data was obtained using a *convenience sample*.

Definition

A **convenience sample** is a sample in which the individuals are easily obtained and not based on randomness.

CAUTION!

Studies that use convenience sampling generally have results that are suspect. The results should be looked on with extreme skepticism.

The most popular of the many types of convenience samples are those in which the individuals in the sample are **self-selected** (the individuals themselves decide to participate in a survey). These are also called **voluntary response** samples. One example of self-selected sampling is phone-in polling; a radio personality will ask his or her listeners to phone the station to submit their opinions. Another example is the use of the Internet to conduct surveys. For example, a television news show will present a story regarding a certain topic and ask its viewers to “tell us what you think” by completing a questionnaire online or submitting an opinion using a hash-tag(#) on Twitter. Both of these samples are poor designs because the individuals who decide to be in the sample generally have strong opinions about the topic. A more typical individual in the population will not bother logging on to a computer or using Twitter to complete a survey. Any inference made regarding the population from this type of sample should be made with extreme caution.

Convenience samples yield unreliable results because the individuals participating in the survey are not chosen using random sampling. Instead, the interviewer or

participant selects who is in the survey. Would an interviewer select an ornery individual? Of course not! Therefore, the sample is likely not to be representative of the population.

Multistage Sampling

In practice, most large-scale surveys obtain samples using a combination of the techniques just presented.

As an example of multistage sampling, consider Nielsen Media Research. Nielsen randomly selects households and monitors the television programs these households are watching through a People Meter. The meter is an electronic box placed on each TV within the household. The People Meter measures what program is being watched and who is watching it. Nielsen selects the households with the use of a two-stage sampling process.

Stage 1 Using U.S. Census data, Nielsen divides the country into geographic areas (strata). The strata are typically city blocks in urban areas and geographic regions in rural areas. About 6000 strata are randomly selected.

Stage 2 Nielsen sends representatives to the selected strata and lists the households within the strata. The households are then randomly selected through a simple random sample.

Nielsen sells the information obtained to television stations and companies. These results are used to help determine prices for commercials.

As another example of multistage sampling, consider the sample used by the Census Bureau for the Current Population Survey. This survey requires five stages of sampling:

Stage 1 Stratified sample

Stage 2 Cluster sample

Stage 3 Stratified sample

Stage 4 Cluster sample

Stage 5 Systematic sample

This survey is very important because it is used to obtain demographic estimates of the United States in noncensus years. Details about the Census Bureau's sampling method can be found in *The Current Population Survey: Design and Methodology*, Technical Paper No. 40.

Sample Size Considerations

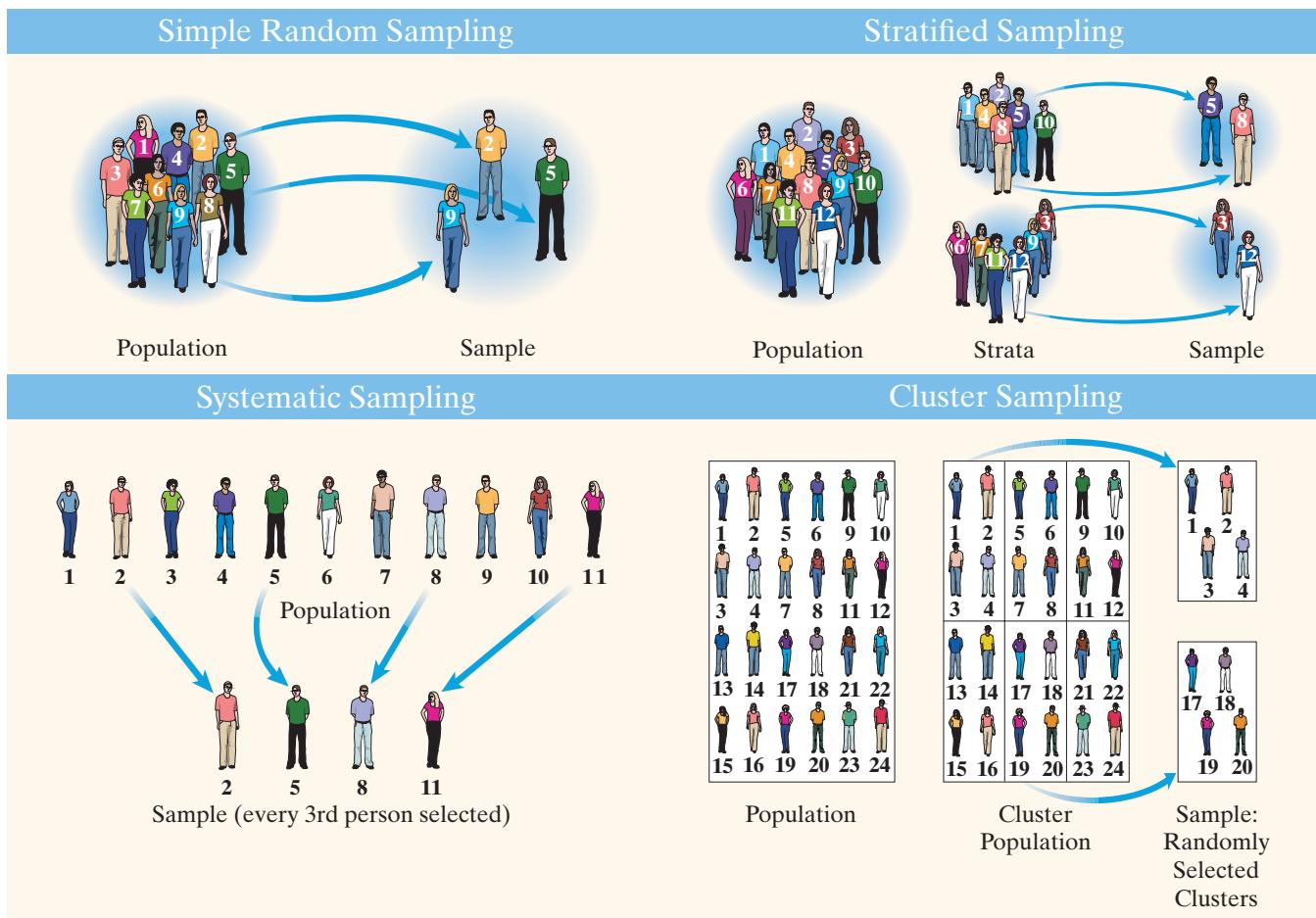
Throughout our discussion of sampling, we did not mention how to determine the sample size. Researchers need to know how many individuals they must survey to draw conclusions about the population within some predetermined margin of error. They must find a balance between the reliability of the results and the cost of obtaining these results. Time and money determine the level of confidence researchers will place on the conclusions drawn from the sample data. The more time and money researchers have available, the more accurate the results of the statistical inference.

In Sections 9.1 and 9.2, we discuss techniques for determining the sample size required to estimate characteristics regarding the population within some margin of error. (For a detailed discussion of sample size considerations, consult a text on sampling techniques such as *Elements of Sampling Theory and Methods* by Z. Govindarajulu, Pearson, 1999.)

Summary

Figure 5 on the next page illustrates the four sampling techniques presented.

Figure 5



1.4 Assess Your Understanding

Vocabulary and Skill Building

- Describe a circumstance in which stratified sampling would be an appropriate sampling method?
- Which sampling method does not require a frame?
- Why are convenience samples ill advised?
- A(n) _____ sample is obtained by dividing the population into groups and selecting all individuals from within a random sample of the groups.
- A(n) _____ sample is obtained by dividing the population into homogeneous groups and randomly selecting individuals from each group.
- True or False:** When taking a systematic random sample of size n , every group of size n from the population has the same chance of being selected.
- True or False:** A simple random sample is always preferred because it obtains the same information as other sampling plans but requires a smaller sample size.
- True or False:** When conducting a cluster sample, it is better to have fewer clusters with more individuals when the clusters are heterogeneous.
- True or False:** Inferences based on voluntary response samples are generally not reliable.

- True or False:** When obtaining a stratified sample, the number of individuals included within each stratum must be equal.

In Problems 11–22, identify the type of sampling used.

- To estimate the percentage of defects in a recent manufacturing batch, a quality-control manager at Intel selects every 8th chip that comes off the assembly line starting with the 3rd until she obtains a sample of 140 chips.
- To determine the prevalence of human growth hormone (HGH) use among high school varsity baseball players, the State Athletic Commission randomly selects 50 high schools. All members of the selected high schools' varsity baseball teams are tested for HGH.
- NW** To determine customer opinion of its boarding policy, Southwest Airlines randomly selects 60 flights during a certain week and surveys all passengers on the flights.
- A member of Congress wishes to determine her constituency's opinion regarding estate taxes. She divides her constituency into three income classes: low-income households, middle-income households, and upper-income households. She then takes a simple random sample of households from each income class.

15. In an effort to identify whether an advertising campaign has been effective, a marketing firm conducts a nationwide poll by randomly selecting individuals from a list of known users of the product.

16. A radio station asks its listeners to call in their opinion regarding the use of U.S. forces in peacekeeping missions.

17. A farmer divides his orchard into 50 subsections, randomly selects 4, and samples all the trees within the 4 subsections to approximate the yield of his orchard.

18. A college official divides the student population into five classes: freshman, sophomore, junior, senior, and graduate student. The official takes a simple random sample from each class and asks the members' opinions regarding student services.

19. A survey regarding download time on a certain website is administered on the Internet by a market research firm to anyone who would like to take it.

20. The presider of a guest-lecture series at a university stands outside the auditorium before a lecture begins and hands every fifth person who arrives, beginning with the third, a speaker evaluation survey to be completed and returned at the end of the program.

21. To determine his DSL Internet connection speed, Shawn divides up the day into four parts: morning, midday, evening, and late night. He then measures his Internet connection speed at 5 randomly selected times during each part of the day.

22. 24 Hour Fitness wants to administer a satisfaction survey to its current members. Using its membership roster, the club randomly selects 40 club members and asks them about their level of satisfaction with the club.

23. A salesperson obtained a systematic sample of size 20 from a list of 500 clients. To do so, he randomly selected a number from 1 to 25, obtaining the number 16. He included in the sample the 16th client on the list and every 25th client thereafter. List the numbers that correspond to the 20 clients selected.

24. A quality-control expert wishes to obtain a cluster sample by selecting 10 of 795 clusters. She numbers the clusters from 1 to 795. Using Table I from Appendix A, she closes her eyes and drops a pencil on the table. It points to the digit in row 8, column 38. Using this position as the starting point and proceeding downward, determine the numbers for the 10 clusters selected.

Applying the Concepts

NW 25. Stratified Sampling The Future Government Club wants to sponsor a panel discussion on the upcoming national election. The club wants to have four of its members lead the panel discussion. To be fair, however, the panel should consist of two Democrats and two Republicans. From the list of current members of the club, obtain a stratified sample of two Democrats and two Republicans to serve on the panel.

Democrats

Bolden	Fallenbucel	Motola	Ramirez
Bolt	Haydra	Nolan	Tate
Carter	Khouri	Opacian	Washington
Debold	Lukens	Pawlak	Wright

Republicans

Blouin	Grajewski	Ochs	Salihar
Cooper	Keating	Pechtold	Thompson
De Young	May	Redmond	Trudeau
Engler	Niemeyer	Rice	Zenkel

26. Stratified Sampling The owner of a private food store is concerned about employee morale. She decides to survey the managers and hourly employees to see if she can learn about work environment and job satisfaction. From the list of workers at the store, obtain a stratified sample of two managers and four hourly employees to survey.

Managers		Hourly Employees		
Carlisle	Oliver	Archer	Foushi	Massie
Hills	Orsini	Bolcerek	Gow	Musa
Kats	Ullrich	Bryant	Grove	Nickas
Lindsey	McGuffin	Cole	Hall	Salazar
		Dimas	Houston	Vaneck
		Ellison	Kemp	Weber
		Everhart	Lathus	Zavodny

NW 27. Systematic Sample The human resource department at a certain company wants to conduct a survey regarding worker morale. The department has an alphabetical list of all 4502 employees at the company and wants to conduct a systematic sample.

(a) Determine k if the sample size is 50.

(b) Determine the individuals who will be administered the survey. More than one answer is possible.

28. Systematic Sample To predict the outcome of a county election, a newspaper obtains a list of all 945,035 registered voters in the county and wants to conduct a systematic sample.

(a) Determine k if the sample size is 130.

(b) Determine the individuals who will be administered the survey. More than one answer is possible.

29. Which Method? The mathematics department at a university wishes to administer a survey to a sample of students taking college algebra. The department is offering 32 sections of college algebra, similar in class size and makeup, with a total of 1280 students. They would like the sample size to be roughly 10% of the population of college algebra students this semester. How might the department obtain a simple random sample? A stratified sample? A cluster sample? Which method do you think is best in this situation?

30. Good Sampling Method? To obtain students' opinions about proposed changes to course registration procedures, the administration of a small college asked for faculty volunteers who were willing to administer a survey in one of their classes. Twenty-three faculty members volunteered. Each faculty member gave the survey to all the students in one course of their choosing. Would this sampling method be considered a cluster sample? Why or why not?

31. Sample Design The city of Naperville is considering the construction of a new commuter rail station. The city wishes to survey the residents of the city to obtain their opinion regarding the use of tax dollars for this purpose. Design a sampling method to obtain the individuals in the sample. Be sure to support your choice.

32. Sample Design A school board at a local community college is considering raising the student services fees. The board wants to obtain the opinion of the student body before proceeding. Design a sampling method to obtain the individuals in the sample. Be sure to support your choice.

33. Sample Design Target wants to open a new store in the village of Lockport. Before construction, Target's marketers want to obtain some demographic information regarding the area under consideration. Design a sampling method to obtain the individuals in the sample. Be sure to support your choice.

34. Sample Design The county sheriff wants to determine if a certain highway has a high proportion of speeders traveling on it. Design a sampling method to obtain the individuals in the sample. Be sure to support your choice.

35. Sample Design A pharmaceutical company wants to conduct a survey of 30 individuals who have high cholesterol. The company has obtained a list from doctors throughout the country of 6600 individuals who are known to have high cholesterol. Design a sampling method to obtain the individuals in the sample. Be sure to support your choice.

36. Sample Design A marketing executive for Coca-Cola, Inc., wants to identify television shows that people in the Boston area who typically drink Coke are watching. The executive has a list of all households in the Boston area. Design a sampling method to obtain the individuals in the sample. Be sure to support your choice.

37. Putting It Together: Comparing Sampling Methods

Suppose a political strategist wants to get a sense of how American adults aged 18 years or older feel about health care and health insurance.

- (a) In a political poll, what would be a good frame to use for obtaining a sample?
- (b) Explain why simple random sampling may not guarantee that the sample has an accurate representation of registered Democrats, registered Republicans, and registered Independents.

(c) How can stratified sampling guarantee this representation?

38. Putting It Together: Thinking about Randomness What is random sampling? Why is it necessary for a sample to be obtained randomly rather than conveniently? Will randomness guarantee that a sample will provide accurate information about the population? Explain.

39. Research the origins of the Gallup Poll and the current sampling method the organization uses. Report your findings to the class.

40. Research the sampling methods used by a market research firm in your neighborhood. Report your findings to the class. The report should include the types of sampling methods used, number of stages, and sample size.

1.5 Bias in Sampling



Objective ① Explain the sources of bias in sampling

1 Explain the Sources of Bias in Sampling

So far we have looked at *how* to obtain samples, but not at some of the problems that inevitably arise in sampling. Remember, the goal of sampling is to obtain information about a population through a sample.

Definition

If the results of the sample are not representative of the population, then the sample has **bias**.

IN OTHER WORDS

The word **bias** could mean to give preference to selecting some individuals over others; it could also mean that certain responses are more likely to occur in the sample than in the population.

There are three sources of bias in sampling:

1. Sampling bias
2. Nonresponse bias
3. Response bias

Sampling Bias

Sampling bias means that the technique used to obtain the individuals in the sample tends to favor one part of the population over another. Any convenience sample has sampling bias because the individuals are not chosen through a random sample.

Sampling bias also results due to **undercoverage**, which occurs when the proportion of one segment of the population is lower in a sample than it is in the population. Undercoverage can result if the frame used to obtain the sample is incomplete or not representative of the population. Some frames, such as the list of all registered voters,

may seem easy to obtain; but even this frame may be incomplete since people who recently registered to vote may not be on the published list of registered voters.

Sampling bias can lead to incorrect predictions. For example, the magazine *Literary Digest* predicted that Alfred M. Landon would defeat Franklin D. Roosevelt in the 1936 presidential election. The *Literary Digest* conducted a poll based on a list of its subscribers, telephone directories, and automobile owners. On the basis of the results, the *Literary Digest* predicted that Landon would win with 57% of the popular vote. However, Roosevelt won the election with about 62% of the popular vote. This election took place during the height of the Great Depression. In 1936, most subscribers to the magazine, households with telephones, and automobile owners were Republican, the party of Landon. Therefore, the choice of the frame used to conduct the survey led to an incorrect prediction due to sampling bias. Essentially, there was undercoverage of Democrats.

It is difficult to gain access to a *complete* list of individuals in a population. For example, public-opinion polls often use random digit dialing (RDD) telephone surveys, which implies that the frame is all households with telephones (landlines or cell phones). This method of sampling excludes households without telephones, as well as homeless people. If these people differ in some way from people with telephones or homes, the results of the sample may not be valid.

Nonresponse Bias

Nonresponse bias exists when individuals selected to be in the sample who do not respond to the survey have different opinions from those who do. Nonresponse can occur because individuals selected for the sample do not wish to respond or the interviewer was unable to contact them.

All surveys will suffer from nonresponse. The federal government's Current Population Survey has a response rate of about 92%, but it varies depending on the age of the individual. For example, the response rate for 20- to 29-year-olds is 85%, and for individuals 70 and older, it is 99%. Response rates in random digit dialing telephone surveys are typically around 70%, e-mail survey response rates hover around 40%, and mail surveys can have response rates as high as 60%.

Nonresponse bias can be controlled using callbacks. For example, if a mailed questionnaire was not returned, a callback might mean phoning the individual to conduct the survey. If an individual was not at home, a callback might mean returning to the home at other times in the day.

Another method to improve nonresponse is using rewards, such as cash payments for completing a questionnaire, or incentives such as a cover letter that states that the responses to the questionnaire will determine future policy. For example, I received \$1 with a survey regarding my satisfaction with a recent purchase. The \$1 "payment" was meant to make me feel guilty enough to fill out the questionnaire. As another example, a city may send out questionnaires to households and state in a cover letter that the responses to the questionnaire will be used to decide pending issues within the city.

Let's consider the *Literary Digest* poll again. The *Literary Digest* mailed out more than 10 million questionnaires and 2.3 million people responded. The rather low response rate (23%) contributed to the incorrect prediction. After all, Roosevelt was the incumbent president and only those who were unhappy with his administration were likely to respond. In the same election, the 35-year-old George Gallup predicted that Roosevelt would win the election in his survey involving 50,000 people.



Response Bias

Response bias exists when the answers on a survey do not reflect the true feelings of the respondent. Response bias can occur in a number of ways.

Interviewer Error A trained interviewer is essential to obtain accurate information from a survey. A skilled interviewer can elicit responses from individuals and make the interviewee feel comfortable enough to give truthful responses. For example, a good interviewer can obtain truthful answers to questions as sensitive as "Have you ever

cheated on your taxes?” Do not be quick to trust surveys conducted by poorly trained interviewers. Do not trust survey results if the sponsor has a vested interest in the results of the survey. Would you trust a survey conducted by a car dealer that reports 90% of customers say they would buy another car from the dealer?

Misrepresented Answers Some survey questions result in responses that misrepresent facts or are flat-out lies. For example, a survey of recent college graduates may find that self-reported salaries are inflated. Also, people may overestimate their abilities. For example, ask people how many push-ups they can do in 1 minute, and then ask them to do the push-ups. How accurate were they?

CAUTION!

The wording of questions can significantly affect the responses and, therefore, the validity of a study. A great source for phrasing survey questions is Saris, W. E. and Gallhofer, I. N. (2007) *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ: Wiley.

Wording of Questions The way a question is worded can lead to response bias in a survey, so questions must always be asked in balanced form. For example, the “yes/no” question

Do you oppose the reduction of estate taxes?

should be written

Do you favor or oppose the reduction of estate taxes?

The second question is balanced. Do you see the difference? Consider the following report based on studies from Schuman and Presser (*Questions and Answers in Attitude Surveys*, 1981, p. 277), who asked the following two questions:

- (A) Do you think the United States should forbid public speeches against democracy?
- (B) Do you think the United States should allow public speeches against democracy?

For those respondents presented with question A, 21.4% gave “yes” responses, while for those given question B, 47.8% gave “no” responses. The conclusion you may arrive at is that most people are not willing to forbid something, but more people are willing not to allow something. These results illustrate how wording a question can alter a survey’s outcome.

Another consideration in wording a question is not to be vague. The question “How much do you study?” is too vague. Does the researcher mean how much do I study for all my classes or just for statistics? Does the researcher mean per day or per week? The question should be written “How many hours do you study statistics each week?”

Ordering of Questions or Words Many surveys will rearrange the order of the questions within a questionnaire so that responses are not affected by prior questions. Consider an example from Schuman and Presser in which the following two questions were asked:

- (A) Do you think the United States should let Communist newspaper reporters from other countries come in here and send back to their papers the news as they see it?
- (B) Do you think a Communist country such as Russia should let American newspaper reporters come in and send back to America the news as they see it?

For surveys conducted in 1980 in which the questions appeared in the order (A, B), 54.7% of respondents answered “yes” to A and 63.7% answered “yes” to B. If the questions were ordered (B, A), then 74.6% answered “yes” to A and 81.9% answered “yes” to B. When Americans are first asked if U.S. reporters should be allowed to report Communist news, they are more likely to agree that Communists should be allowed to report American news. Questions should be rearranged as much as possible to help reduce effects of this type.

Pollsters will also rearrange words within a question. For example, the Gallup Organization routinely asks the following question of adults aged 18 years or older:

Do you [rotated: approve (or) disapprove] of the job “the current president” is doing as president?

The words *approve* and *disapprove* are rotated to remove the effect that may occur by writing the word *approve* first in the question.

Type of Question One of the first considerations in designing a question is determining whether the question should be *open* or *closed*.

An **open question** allows the respondent to choose his or her response:

What is the most important problem facing America's youth today?

A **closed question** requires the respondent to choose from a list of predetermined responses:

What is the most important problem facing America's youth today?

- (a) Drugs
- (b) Violence
- (c) Single-parent homes
- (d) Promiscuity
- (e) Peer pressure

In closed questions, the possible responses should be rearranged because respondents are likely to choose early choices in a list rather than later choices.

An open question should be phrased so that the responses are similar. (You don't want a wide variety of responses.) This allows for easy analysis of the responses. Closed questions limit the number of respondent choices and, therefore, the results are much easier to analyze. The limited choices, however, do not always include a respondent's desired choice. In that case, the respondent will have to choose a secondary answer or skip the question.

Survey designers recommend conducting pretest surveys with open questions and then using the most popular answers as the choices on closed-question surveys. Another issue to consider in the closed-question design is the number of possible responses. The option "no opinion" should be omitted, because this option does not allow for meaningful analysis. The goal is to limit the number of choices in a closed question without forcing respondents to choose an option they do not prefer, which would make the survey have response bias.

Data-entry Error Although not technically a result of response bias, data-entry error will lead to results that are not representative of the population. Once data are collected, the results may need to be entered into a computer, which could result in input errors. Or, a respondent may make a data entry error. For example, 39 may be entered as 93. It is imperative that data be checked for accuracy. In this text, we present some suggestions for checking for data error.

Can a Census Have Bias?

The discussion so far has focused on bias in samples, but bias can also occur when conducting a census. A question on a census form could be misunderstood, thereby leading to response bias in the results. We also mentioned that it is often difficult to contact each individual in a population. For example, the U.S. Census Bureau is challenged to count each homeless person in the country, so the census data published by the U.S. government likely suffers from nonresponse bias.

Sampling Error versus Nonsampling Error Nonresponse bias, response bias, and data-entry errors are types of *nonsampling error*. However, whenever a sample is used to learn information about a population, there will inevitably also be *sampling error*.

Definitions

IN OTHER WORDS

We can think of sampling error as error that results from using a subset of the population to describe characteristics of the population. Nonsampling error is error that results from obtaining and recording the information collected.

Nonsampling errors result from undercoverage, nonresponse bias, response bias, or data-entry error. Such errors could also be present in a complete census of the population. **Sampling error** results from using a sample to estimate information about a population. This type of error occurs because a sample gives incomplete information about a population.

By incomplete information, we mean that the individuals in the sample cannot reveal all the information about the population. Suppose we wanted to determine the average age of the students enrolled in an introductory statistics course. To do this, we obtain a simple random sample of four students and ask them to write their age on a sheet of paper and turn it in. The average age of these four students is 23.25 years.

Assume that no students lied about their age or misunderstood the question, and the sampling was done appropriately. If the actual average age of all 30 students in the class (the population) is 22.91 years, then the sampling error is $23.25 - 22.91 = 0.34$ year. Now suppose the same survey is conducted, but this time one student lies about his age. Then the results of the survey will also have nonsampling error.



1.5 Assess Your Understanding

Vocabulary and Skill Building

- What is a closed question? What is an open question? Discuss the advantages and disadvantages of each type of question.
- What does it mean when a part of the population is underrepresented?
- Match each word or phrase to the definition.

Word/Phrase	Definition
(a) Bias	I. When the techniques used to select individuals to be in the sample favor one part of the population over another.
(b) Sampling Bias	II. When the answers on a survey do not reflect the true feelings of the respondent.
(c) Nonresponse Bias	III. The results of the sample are not representative of the population.
(d) Response Bias	IV. When the individuals selected to be in the sample who do not respond to the survey have different opinions from those who do respond.

- Distinguish between nonsampling error and sampling error.

In Problems 5–16, the survey has bias. (a) Determine the type of bias. (b) Suggest a remedy.

- A retail store manager wants to conduct a study regarding the shopping habits of his customers. He selects the first 60 customers who enter his store on a Saturday morning.
 - The village of Oak Lawn wishes to conduct a study regarding the income level of households within the village. The village manager selects 10 homes in the southwest corner of the village and sends an interviewer to the homes to determine household income.
 - An antigen advocate wants to estimate the percentage of people who favor stricter gun laws. He conducts a nationwide survey of 1203 randomly selected adults 18 years old and older. The interviewer asks the respondents, “Do you favor harsher penalties for individuals who sell guns illegally?”
 - Suppose you are conducting a survey regarding the sleeping habits of students. From a list of registered students, you obtain a simple random sample of 150 students. One survey question is “How much sleep do you get?”
 - A polling organization conducts a study to estimate the percentage of households that speaks a foreign language as the primary language. It mails a questionnaire to 1023 randomly selected households throughout the United States and asks the head of household if a foreign language is the primary language spoken in the home. Of the 1023 households selected, 12 responded.
 - Cold Stone Creamery is considering opening a new store in O’Fallon. Before opening, the company wants to know the percentage of households in O’Fallon that regularly visit an ice cream shop. The market researcher obtains a list of households in O’Fallon, randomly selects 150, and mails a questionnaire that asks about ice cream eating habits and flavor preferences. Of the 150 questionnaires mailed, 4 are returned.
 - A newspaper article reported, “The *Cosmopolitan* magazine survey of more than 5000 Australian women aged 18–34 found about 42 percent considered themselves overweight or obese.”
- Source: Herald Sun, September 9, 2007.*
- A health teacher wants to research the weight of college students. She obtains the weights for all the students in her 9 A.M. class by looking at their driver’s licenses or state IDs.
 - A magazine is conducting a study on the effects of infidelity in a marriage. The editors randomly select 400 women whose husbands were unfaithful and ask, “Do you believe a marriage can survive when the husband destroys the trust that must exist between husband and wife?”
 - A textbook publisher wants to determine what percentage of college professors either require or recommend that their students purchase textbook packages with supplemental materials. The publisher sends surveys by e-mail to a random sample of 320 faculty members who have registered with its website. The publisher reports that 80% of college professors require or recommend that their students purchase some type of textbook package.
 - Suppose you are conducting a survey regarding illicit drug use among teenagers in the Baltimore school district. You obtain a cluster sample of 12 schools within the district and sample all sophomore students in the randomly selected schools. The survey is administered by the teachers.
 - To determine the public’s opinion of the police department, the police chief obtains a cluster sample of 15 census tracts within his jurisdiction and samples all households in the randomly selected tracts. Uniformed police officers go door to door to conduct the survey.

cream shop. The market researcher obtains a list of households in O’Fallon, randomly selects 150, and mails a questionnaire that asks about ice cream eating habits and flavor preferences. Of the 150 questionnaires mailed, 4 are returned.

- A newspaper article reported, “The *Cosmopolitan* magazine survey of more than 5000 Australian women aged 18–34 found about 42 percent considered themselves overweight or obese.”

Source: Herald Sun, September 9, 2007.

- A health teacher wants to research the weight of college students. She obtains the weights for all the students in her 9 A.M. class by looking at their driver’s licenses or state IDs.

- A magazine is conducting a study on the effects of infidelity in a marriage. The editors randomly select 400 women whose husbands were unfaithful and ask, “Do you believe a marriage can survive when the husband destroys the trust that must exist between husband and wife?”

- A textbook publisher wants to determine what percentage of college professors either require or recommend that their students purchase textbook packages with supplemental materials. The publisher sends surveys by e-mail to a random sample of 320 faculty members who have registered with its website. The publisher reports that 80% of college professors require or recommend that their students purchase some type of textbook package.

- Suppose you are conducting a survey regarding illicit drug use among teenagers in the Baltimore school district. You obtain a cluster sample of 12 schools within the district and sample all sophomore students in the randomly selected schools. The survey is administered by the teachers.

- To determine the public’s opinion of the police department, the police chief obtains a cluster sample of 15 census tracts within his jurisdiction and samples all households in the randomly selected tracts. Uniformed police officers go door to door to conduct the survey.

Applying the Concepts

- Response Rates** Surveys tend to suffer from low response rates. Based on past experience, a researcher determines that the typical response rate for an e-mail survey is 40%. She wishes to obtain a sample of 300 respondents, so she e-mails the survey to 1500 randomly selected e-mail addresses. Assuming the response rate for her survey is 40%, will the respondents form an unbiased sample? Explain.

- Delivery Format** The General Social Survey asked, “About how often did you have sex in the past 12 months?” About 47% of respondents indicated they had sex at least once a week. In an internet survey for a marriage and family wellness center, respondents were asked, “How often do you and your partner have sex (on average)?” About 31% of respondents indicated they had sex with their partner at least once a week. Explain how the delivery method for such a question could result in biased responses.

19. Order of the Questions Consider the following two questions:

- (a) Currently, social media companies, such as Facebook, profit by selling user data. Do you believe the government should regulate the ability of social media companies to sell user data?
- (b) Would you be willing to pay for social media services if your user data was your personal property that the media company could not sell?

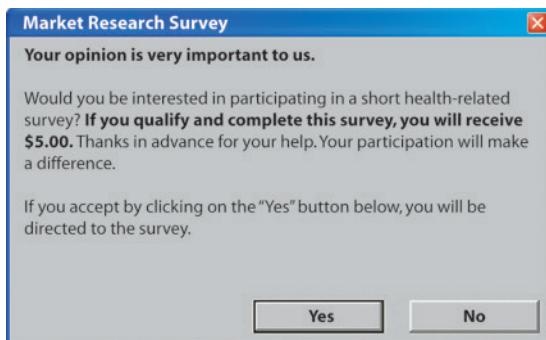
Do you think the order in which the questions are asked will affect the survey results? If so, what can the pollster do to alleviate this response bias?

20. Order of the Questions Consider the following two questions:

- (a) Do you believe that the government should or should not be allowed to prohibit individuals from expressing their religious beliefs at their place of employment?
- (b) Do you believe that the government should or should not be allowed to prohibit teachers from expressing their religious beliefs in public school classrooms?

Do you think the order in which the questions are asked will affect the survey results? If so, what can the pollster do to alleviate this response bias? Discuss the choice of the word *prohibit* in the survey questions.

21. Improving Response Rates Suppose you are reading an article at psychcentral.com and the following text appears in a pop-up window:



What tactic is the company using to increase the response rate for its survey?

22. Rotating Choices Consider this question from a recent Gallup poll:

Which of the following approaches to solving the nation's energy problems do you think the U.S. should follow right now—[ROTATED: emphasize production of more oil, gas and coal supplies (or) emphasize more conservation by consumers of existing energy supplies]?

Why is it important to rotate the two choices presented in the question?

23. Sampling Strategies Many national polls are based on random-digit dialing (RDD). In this method, a computer randomly generates a phone number in the hopes of reaching an individual at his/her residence. In registration-based sampling (RBS), voter files are used to obtain a random sample. Researchers at Pew Research conducted an opinion poll using both RDD and RBS in order to compare the results. *Source:* “Comparing Survey Sampling Strategies: Random-Digit Dial vs. Voter Files,” Kennedy, Courtney et al., *Pew Research*, Oct. 9, 2018.

(a) Voter files are typically used for election surveys (to find likely voters). What type of bias would you expect would exist in using a voter file to obtain a random sample of all adult Americans aged 18 years or older?

(b) Do you think the type of bias that exists when using RBS also exists for RDD? Which method do you think would suffer more from this type of bias? Explain.

(c) Which method of sampling do you think has the lower response rate?

(d) Which method of sampling do you think oversampled individuals who identified with or leaned Republican?

24. Robocalls According to Martin Boon of ICM Limited, a polling firm in Britain, in 1995 it took 3000 to 4000 phone calls to obtain a sample of size 2000. Today, it takes over 30,000 calls. To reduce costs, more polling is done using robocalls and Internet-based polling. Robocalling to cellular telephones is illegal. How do these methods potentially lead to nonsampling error? What types of nonsampling error do you think this may lead to?

25. Don't Call Me! The Telephone Consumer Protection Act (TCPA) allows consumers to put themselves on a do-not-call registry. If a number is on the registry, commercial telemarketers are not allowed to call you. Do you believe this has affected the ability of surveyors to obtain accurate polling results? If so, how?

26. Current Population Survey In the federal government's Current Population Survey, the response rate for 20- to 29-year-olds is 85%; for individuals at least 70 years of age it is 99%. Why do you think this is?

27. Exit Polling in the 2016 Election During every election, pollsters conduct exit interviews to help determine which candidate people voted for. During the 2016 presidential election, exit polls under-sampled the number of voters called “Democrat white working-class voters” and over-sampled the number of voters called “white college-educated voters” (who tend to be socially liberal). Research the role nonsampling error played in the exit polls for the 2016 presidential election. Some sources of your research should include Edison Research (the firm that conducts exit polls) and the *New York Times*.

28. Write Your Own Survey Develop a survey that you could administer using online survey tools such as StatCrunch, surveymonkey.com, or polldaddy.com. Administer the survey. Did the responses accurately reflect the goals of each question? What types of nonsampling error did you encounter in the survey? If you invited individuals to take the survey via an e-mail, what type of response rate did you obtain? What approach did you take to increase response rate?

29. Wording Survey Questions In the early 1990s, Gallup asked Americans whether they supported the United States bombing Serbian forces in Bosnia. In this survey, 35% of respondents supported the idea. The very same day, ABC News asked whether Americans would support the United States, along with its allies in Europe, bombing Serbian forces in Bosnia. In this survey, 65% supported the idea. Explain the difference in the wording of the question. What does this suggest?

30. Wording Survey Questions Write a survey question that contains strong wording and one that contains tempered wording. Post each question in an online survey site such as StatCrunch, surveymonkey.com, or polldaddy.com. Administer the survey to at least 25 different people for each question. How does the wording affect the response?

31. Order in Survey Questions Write two questions that could have different responses, depending on the order in which the questions are presented. Or write a single question such that the order in which words are presented could affect the response. Administer the survey to at least 25 different people for each question. Did the results differ?

32. Informed Opinions People often respond to survey questions without any knowledge of the subject matter. A common example of this is the discussion on banning dihydrogen monoxide. The Centers for Disease Control (CDC) reports that there were 1423 deaths due to asbestos in 2005, but over 3443 deaths were attributed to dihydrogen monoxide in 2007. Articles and websites such as www.dhmo.org tell how this substance is widely used despite the dangers associated with it. Many people have joined the cause to ban this substance without realizing that dihydrogen monoxide is simply water (H_2O). Their eagerness to protect the environment or their fear of seeming uninformed may be part of the problem. Put together a survey that asks individuals whether dihydrogen monoxide should or should not be banned. Give the survey to 20 randomly selected students around campus and report your results to the class. An example survey might look like the following:

Dihydrogen monoxide is colorless, odorless, and kills thousands of people every year. Most of these deaths are caused by accidental inhalation, but the dangers of dihydrogen monoxide do not stop there. Prolonged exposure to its solid form can severely damage skin tissue. Symptoms of ingestion can include excessive sweating and urination and possibly a bloated feeling, nausea, vomiting, and body electrolyte imbalance. Dihydrogen monoxide is a major component of acid rain and can cause corrosion after coming in contact with certain metals.

Do you believe that the government should or should not ban the use of dihydrogen monoxide?

33. Name two biases that led to the *Literary Digest* making an incorrect prediction in the presidential election of 1936.

34. Research on George Gallup Research the polling done by George Gallup in the 1936 presidential election. Write a report on your findings and include information about the sampling technique and sample size. Next, research the polling done by Gallup for the 1948 presidential election. Did Gallup accurately predict the outcome of the election? What lessons were learned by Gallup?

35. The Challenge in Polling One of the challenges in polling for elections is deciding who to include in your frame and who might actually turn out to vote.

- (a) Suppose you were asked to conduct a poll for a senatorial election. Explain how you might design your sample. In your explanation include a discussion of the difference between “registered voters” and “likely voters.” What role would stratification play in your sampling?
- (b) Voter turnout is different for presidential election cycles (2012, 2016, 2020, and so on) versus non-presidential election cycles (2014, 2018, 2022, and so on). Explain the role election cycle plays in voter turnout and explain how this may affect your sampling methodology.
- (c) During the 2014 election, Nate Silver of FiveThirtyEight said “the pre-election polling averages (not the FiveThirtyEight forecasts, which also account for other factors) in the 10 most competitive Senate races had a 6-percentage point Democratic bias as compared to the votes counted in each state so far.” Explain what this means and explain how this would have impacted polling results compared with actual results.

Explaining the Concepts

- 36.** Why is it rare for frames to be completely accurate?
- 37.** What are some solutions to nonresponse?
- 38.** Discuss the benefits of having trained interviewers.
- 39.** What are the advantages of having a presurvey when constructing a questionnaire that has closed questions?
- 40.** Discuss the pros and cons of telephone interviews that take place during dinner time in the early evening.
- 41.** Why is a high response rate desired? How would a low response rate affect survey results?
- 42.** Discuss why the order of questions or choices within a questionnaire are important in sample surveys.
- 43.** Suppose a survey asks, “How many hours do you study?” Explain how this could be interpreted in more than one way. Suggest a way in which the question could be improved.
- 44.** Discuss a possible advantage of offering rewards or incentives to increase response rates. Are there any disadvantages?

1.6 The Design of Experiments



Objectives

- ① Describe the characteristics of an experiment
- ② Explain the steps in designing an experiment
- ③ Explain the completely randomized design
- ④ Explain the matched-pairs design

The major theme of this chapter has been data collection. Section 1.2 briefly discussed the idea of an experiment, but the main focus was on observational studies. Sections 1.3 through 1.5 focused on sampling and surveys. In this section, we further develop the idea of collecting data through an experiment.

1 Describe the Characteristics of an Experiment

Remember, in an observational study, if an association exists between an explanatory variable and response variable the researcher cannot claim causality. To demonstrate how changes in the explanatory variable *cause* changes in the response variable, the researcher needs to conduct an *experiment*.

Definitions

An **experiment** is a controlled study conducted to determine the effect varying one or more explanatory variables or **factors** has on a response variable. Any combination of the values of the factors is called a **treatment**.

Historical Note

Sir Ronald Fisher, often called the Father of Modern Statistics, was born in England on February 17, 1890. He received a BA in astronomy from Cambridge University in 1912. In 1914, he took a position teaching mathematics and physics at a high school. He did this to help serve his country during World War I. (He was rejected by the army because of his poor eyesight.) In 1919, Fisher took a job as a statistician at Rothamsted Experimental Station, where he was involved in agricultural research. In 1933, Fisher became Galton Professor of Eugenics at Cambridge University, where he studied Rh blood groups. In 1943 he was appointed to the Balfour Chair of Genetics at Cambridge. He was knighted by Queen Elizabeth in 1952. Fisher retired in 1957 and died in Adelaide, Australia, on July 29, 1962. One of his famous quotations is “To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of.”



In an experiment, the **experimental unit** is a person, object, or some other well-defined item upon which a treatment is applied. We often refer to the experimental unit as a **subject** when he or she is a person. The subject is analogous to the individual in a survey.

The goal in an experiment is to determine the effect various treatments have on the response variable. For example, we might want to determine whether a new treatment is superior to an existing treatment (or no treatment at all). To make this determination, experiments require a *control group*. A **control group** serves as a baseline treatment that can be used to compare it to other treatments. For example, a researcher in education might want to determine if students who do their homework using an online homework system do better on an exam than those who do their homework from the text. The students doing the text homework might serve as the control group (since this is the currently accepted practice). The factor is the type of homework. There are two treatments: online homework and text homework.

A second method for defining the control group is through the use of a *placebo*. A **placebo** is a treatment that looks just like the “real” treatments in the study. A placebo could be an innocuous medication, such as a sugar tablet, that looks, tastes, and smells like the experimental medication. A placebo might also be a procedure that follows the same steps as the experimental procedure, but leaves out a key intervention. For example, a procedure called vertebroplasty, where medical cement is pumped into a spine fracture, was tested through a designed experiment. All subjects went through a surgery to repair the spine, but only half received the medical cement.

An interesting outcome occurred in the vertebroplasty experiment. A subject in the placebo group found that the procedure resulted in complete abatement of the back pain even though she did not receive the medical cement! This type of phenomenon in an experiment is referred to as the **placebo effect**. Jo Marchant explores the placebo effect in her book titled *Cure*. Marchant suggests that placebo treatments can have measurable effects. For example, in patients with Parkinson’s disease placebos caused an increase of the neurotransmitter dopamine. A study of 459 migraine sufferers found that the placebo effect accounted for about 60% of the benefit of the drug Maxalt. Of course, the placebo effect will not account for improvements in someone with a tumor or replace insulin for someone with diabetes. However, the Maxalt study suggests that remedies for pain, nausea, or depression rely extensively on the placebo effect.

In an experiment, it is important that each group be treated the same way. It is also important that individuals do not adjust their behavior because of the treatment they are receiving. For this reason, many experiments use a technique called *blinding*. **Blinding** refers to nondisclosure of the treatment an experimental unit is receiving. There are two types of blinding: *single blinding* and *double blinding*.

Definitions

In **single-blind** experiments, the experimental unit (or subject) does not know which treatment he or she is receiving. In **double-blind** experiments, neither the experimental unit nor the researcher in contact with the experimental unit knows which treatment the experimental unit is receiving.

EXAMPLE 1 The Characteristics of an Experiment

Problem Lipitor™ is a cholesterol-lowering drug made by Pfizer. In the Collaborative Atorvastatin Diabetes Study (CARDS), the effect of Lipitor on cardiovascular disease was assessed in 2838 subjects, ages 40 to 75, with type 2 diabetes, without prior history of cardiovascular disease. In this placebo-controlled, double-blind experiment, subjects were randomly allocated to either Lipitor 10 mg daily (1428) or placebo (1410) and were followed for 4 years. The response variable was the occurrence of any major cardiovascular event.

Lipitor significantly reduced the rate of major cardiovascular events (83 events in the Lipitor group versus 127 events in the placebo group). There were 61 deaths in the Lipitor group versus 82 deaths in the placebo group.

- (a) What does it mean for the experiment to be placebo-controlled?
- (b) What does it mean for the experiment to be double-blind?
- (c) What is the population for which this study applies? What is the sample?
- (d) What are the treatments?
- (e) What is the response variable? Is it qualitative or quantitative?

Approach Apply the definitions just presented.

Solution

- (a) The placebo is a medication that looks, smells, and tastes like Lipitor. The placebo (control) group serves as a baseline against which to compare the results from the group receiving Lipitor. The placebo is also used because people tend to behave differently when they are in a study. By having a placebo control group, the effect of this is neutralized.
- (b) Since the experiment is double-blind, the subjects, as well as the individual monitoring the subjects, do not know whether the subjects are receiving Lipitor or the placebo. The experiment is double-blind so that the subjects receiving the medication do not behave differently from those receiving the placebo and so the individual monitoring the subjects does not treat those in the Lipitor group differently from those in the placebo group.
- (c) The population is individuals from 40 to 75 years of age with type 2 diabetes without a prior history of cardiovascular disease. The sample is the 2838 subjects in the study.
- (d) The treatments are 10 mg of Lipitor or a placebo daily.
- (e) The response variable is whether the subject had any major cardiovascular event, such as a stroke, or not. It is a qualitative variable.

NW Now Work Problem 7



② Explain the Steps in Designing an Experiment

To **design** an experiment means to describe the overall plan in conducting the experiment. Conducting an experiment requires a series of steps.

Steps in Designing an Experiment

Step 1 Identify the Problem to Be Solved. The statement of the problem should be as explicit as possible and should provide the experimenter with direction. The statement must also identify the response variable and the population to be studied. Often, the statement is referred to as the *claim*.

Step 2 Determine the Factors That Affect the Response Variable. The factors are usually identified by an expert in the field of study. In identifying the factors, ask, “What things affect the value of the response variable?” After the factors are identified, determine which factors to fix at some predetermined level, which to manipulate, and which to leave uncontrolled.

Step 3 Determine the Number of Experimental Units. As a general rule, choose as many experimental units as time and money allow. Techniques (such as those discussed in Sections 9.1 and 9.2) exist for determining sample size, provided certain information is available.

Step 4 Determine the Level of Each Factor. There are two ways to deal with the factors: control or randomize.

1. Control: There are two ways to control the factors.

- (a) Set the level of a factor at one value throughout the experiment (if you are *not* interested in its effect on the response variable).
- (b) Set the level of a factor at various levels (if you are interested in its effect on the response variable). The combinations of the levels of all varied factors constitute the treatments in the experiment.

2. Randomize: Randomly assign the experimental units to treatment groups. Because it is difficult, if not impossible, to identify all factors in an experiment, randomly assigning experimental units to treatment groups mutes the effect of variation attributable to factors (explanatory variables) not controlled.

Step 5 Conduct the Experiment.

- (a) **Replication** occurs when each treatment is applied to more than one experimental unit. Using more than one experimental unit for each treatment ensures the effect of a treatment is not due to some characteristic of a single experimental unit. It is a good idea to assign an equal number of experimental units to each treatment.
- (b) Collect and process the data. Measure the value of the response variable for each replication. Then organize the results. The idea is that the value of the response variable for each treatment group is the *same* before the experiment because of randomization. Then any *difference* in the value of the response variable among the different treatment groups is a result of differences in the level of the treatment.

Step 6 Test the Claim. This is the subject of inferential statistics. **Inferential statistics** is a process in which generalizations about a population are made on the basis of results obtained from a sample. Provide a statement regarding the level of confidence in the generalization. Methods of inferential statistics are presented in Chapters 9 through 15.

Well-designed studies will account for the potential of confounding. Recall that confounding occurs when the effects of two or more explanatory variables are not separated.

Consider the Lipitor study from Example 1. The population to which this study applies is individuals 40 to 75 years of age who have type 2 diabetes and no history of cardiovascular disease. In Step 4 of the steps to design an experiment, the researcher controls any factors (explanatory variables) that may play a role in the occurrence of a cardiovascular event (the response variable) but that are not of interest to the researcher. For example, the researcher should have the subjects on the same diet and the same exercise regimen throughout the study because these variables are known to play a role in whether an individual has a cardiovascular event, or not, but are not of interest to the researcher. That is, the variables diet and exercise are controlled and set at one level for all subjects in the study. The variable of interest to the researcher, however, is controlled and set at different levels. So, the variable *drug* is controlled and set at two levels: Lipitor or placebo. Some variables that may affect the value of the response variable cannot be controlled, such as family history of cardiovascular events. Well-designed experiments deal with explanatory variables that are not controlled through randomization. The thinking is that randomization “evens out” the effect of these variables among the various treatment groups.

Suppose, in our study, when we randomized the subjects, we ended up with the younger individuals in the Lipitor group. After analyzing the data, let’s say we found a lower incidence rate of cardiovascular events in the Lipitor group. We would not know whether this outcome was a result of the drug or the age of the subjects (since the typical age for a cardiovascular event is the late sixties). In this case, age of the subjects is a confounding variable. Of course, in a well-designed study, the researcher would determine, before starting the treatment, if the average age of the subjects in the two groups is similar.

③ Explain the Completely Randomized Design

The steps just given apply to any type of designed experiment. We now concentrate on the simplest type of experiment.

Definition

A **completely randomized design** is one in which each experimental unit is randomly assigned to a treatment.

EXAMPLE 2

A Completely Randomized Design

Problem A farmer wishes to determine the optimal level of a new fertilizer on his soybean crop. Design an experiment that will assist him.

Approach Follow the steps for designing an experiment.

Solution

Step 1 The farmer wants to identify the optimal level of fertilizer for growing soybeans. We define *optimal* as the level that maximizes yield. So the response variable will be crop yield.

Step 2 Some factors that affect crop yield are fertilizer, precipitation, sunlight, method of tilling the soil, type of soil, plant, and temperature.

Step 3 In this experiment, we will plant 60 soybean plants (experimental units).

Step 4 List the factors and their levels.

IN OTHER WORDS

The various levels of the factor are the treatments in a completely randomized design.

Figure 6



See Figure 6.

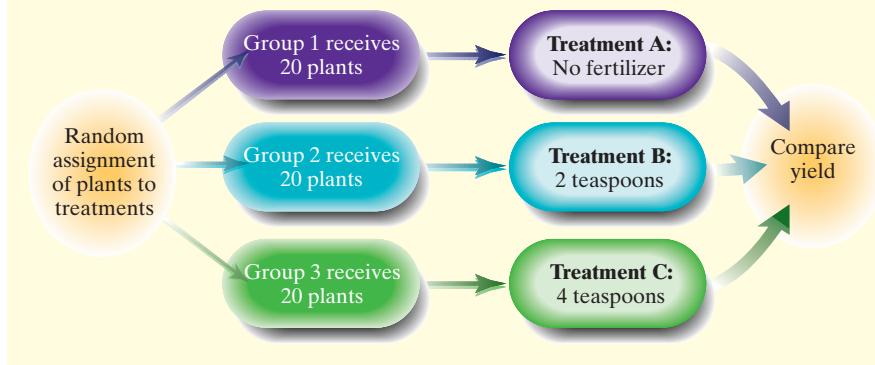
- **Fertilizer.** This factor will be controlled and set at three levels. We wish to measure the effect of varying the level of this variable on the response variable, yield. We will set the treatments (level of fertilizer) as follows:
 - Treatment A: 20 soybean plants receive no fertilizer.
 - Treatment B: 20 soybean plants receive 2 teaspoons of fertilizer per gallon of water every 2 weeks.
 - Treatment C: 20 soybean plants receive 4 teaspoons of fertilizer per gallon of water every 2 weeks.

Step 5

- (a) Randomly assign each plant to a treatment group. First, number the plants from 1 to 60 and randomly generate 20 numbers. The plants corresponding to these numbers get treatment A. Next, number the remaining plants 1 to 40 and randomly generate 20 numbers. The plants corresponding to these numbers get treatment B. The remaining plants get treatment C. Now till the soil, plant the soybean plants, and fertilize according to the schedule prescribed.
- (b) At the end of the growing season, determine the crop yield for each plant.

Step 6 Determine any differences in yield among the three treatment groups. Figure 7 illustrates the experimental design.

Figure 7



Example 2 is a completely randomized design because the experimental units (the plants) were randomly assigned to the treatments. It is the most popular experimental design because of its simplicity, but it is not always the best. We discuss inferential procedures for the completely randomized design with two treatments and quantitative response variable in Section 11.3 and with three or more treatments in Section 13.1. We discuss inferential procedures for the completely randomized design for a qualitative response variable in Sections 11.1 and 12.2.

NW Now Work Problem 9

④ Explain the Matched-Pairs Design

Another type of experimental design is called a *matched-pairs design*.

Definition

A **matched-pairs design** is an experimental design in which the experimental units are paired up. The pairs are selected so that they are related in some way (that is, the same person before and after a treatment, twins, husband and wife, same geographical location, and so on). There are only two levels of treatment in a matched-pairs design.

In matched-pairs design, one matched individual will receive one treatment and the other receives a different treatment. The matched pair is randomly assigned to the treatment using a coin flip or a random-number generator. We then look at the difference in the results of each matched pair. One common type of matched-pairs design is to measure a response variable on an experimental unit before and after a treatment is applied. In this case, the individual is matched against itself. These experiments are sometimes called before-after or pretest-posttest experiments.

EXAMPLE 3

A Matched-Pairs Design



Problem An educational psychologist wants to determine whether listening to music has an effect on a student's ability to learn. Design an experiment to help the psychologist answer the question.

Approach Use a matched-pairs design by matching students according to IQ and gender (just in case gender plays a role in learning with music).

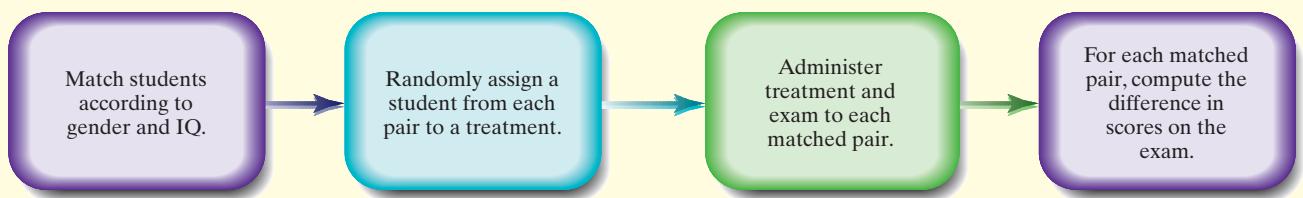
Solution Match students according to IQ and gender. For example, match two females with IQs in the 110 to 115 range.

For each pair of students, flip a coin to determine which student is assigned the treatment of a quiet room or a room with music playing in the background.

(continued)

Each student will be given a statistics textbook and asked to study Section 1.1. After 2 hours, the students will enter a testing center and take a short quiz on the material in the section. Compute the difference in the scores of each matched pair. Any differences in scores will be attributed to the treatment. Figure 8 illustrates the design.

Figure 8

**NW Now Work Problem 11**

We discuss statistical inference for the matched-pairs design for a qualitative response variable in Section 12.3 and quantitative response variable in Section 11.2.

One note about the relation between a designed experiment and simple random sampling: It is often the case that the experimental units selected to participate in a study are not randomly selected. This is because we often need the experimental units to have some common trait, such as high blood pressure. For this reason, participants in experiments are recruited or volunteer to be in a study. However, once we have the experimental units, we use simple random sampling to assign them to treatment groups. With random assignment we assume that the participants are similar at the start of the experiment. Because the treatment is the only difference between the groups, we can say the treatment *caused* the difference observed in the response variable.



1.6 Assess Your Understanding

Vocabulary

1. Define the following:

- (a) Experimental unit
- (b) Treatment
- (c) Response variable
- (d) Factor
- (e) Placebo
- (f) Confounding
- (g) Blinding

2. What is replication in an experiment?

3. Explain the difference between a single-blind and a double-blind experiment.

4. A(n) _____ design is one in which each experimental unit is randomly assigned to a treatment. A(n) _____ design is one in which the experimental units are paired up.

5. **True or False:** Observational studies are used to determine causality between an explanatory variable and response variable.

6. **True or False:** Generally, the goal of an experiment is to determine the effect that treatments will have on the response variable.

Applying the Concepts

NW 7. Chew Your Food Researchers wanted to determine the association between number of times one chews food and food consumption. They identified 45 individuals who were 18 to 45 years of age. First, the researchers determined a baseline for number of chews before swallowing food. Next, each participant attended three sessions to eat pizza for lunch until comfortably full by chewing each portion of food 100%, 150%, and 200% of their baseline number of chews before swallowing. Food intake for each of the three chewing treatments was then measured. It was found that food consumption was reduced significantly, by 9.5% and 14.8%, respectively, for the 150% and 200% number of chews compared to the baseline.



Source: Yong Zhu and James H. Hollis. "Increasing the Number of Chews before Swallowing Reduces Meal Size in Normal-Weight, Overweight, and Obese Adults," *Journal of the Academy of Nutrition and Dietetics*, 11 November 2013.

- (a) What is the research objective of the study?
- (b) What is the response variable in this study? Is it quantitative or qualitative?
- (c) What is the explanatory variable in this study? Is it quantitative or qualitative?
- (d) Who are the experimental units?

- (e) How is control used in this study?
- (f) Each individual chewed 100%, 150%, and 200% of their baseline number of chews before swallowing. This is referred to as a *repeated-measures study* since the same participants were exposed to each treatment. The order in which chewing took place (100% versus 150% versus 200%) was determined randomly. Explain why this is important.

8. Alcohol Dependence To determine if topiramate is a safe and effective treatment for alcohol dependence, researchers conducted a 14-week trial of 371 men and women aged 18 to 65 years diagnosed with alcohol dependence. In this double-blind, randomized, placebo-controlled experiment, subjects were randomly given either 300 milligrams (mg) of topiramate (183 subjects) or a placebo (188 subjects) daily, along with a weekly compliance enhancement intervention. The variable used to determine the effectiveness of the treatment was self-reported percentage of heavy drinking days. Results indicated that topiramate was more effective than placebo at reducing the percentage of heavy drinking days. The researchers concluded that topiramate is a promising treatment for alcohol dependence.

Source: Bankole A. Johnson, Norman Rosenthal, et al. "Topiramate for Treating Alcohol Dependence: A Randomized Controlled Trial," *Journal of the American Medical Association*, 298(14):1641–1651, 2007.

- (a) What does it mean for the experiment to be placebo-controlled?
- (b) What does it mean for the experiment to be double-blind? Why do you think it is necessary for the experiment to be double-blind?
- (c) What does it mean for the experiment to be randomized?
- (d) What is the population for which this study applies? What is the sample?
- (e) What are the treatments?
- (f) What is the response variable?

NW 9. School Psychology A school psychologist wants to test the effectiveness of a new method for teaching reading. She recruits 500 first-grade students in District 203 and randomly divides them into two groups. Group 1 is taught by means of the new method, while group 2 is taught by traditional methods. The same teacher is assigned to teach both groups. At the end of the year, an achievement test is administered and the results of the two groups are compared.

- (a) What is the response variable in this experiment?
- (b) Think of some of the factors in the study. How are they controlled?
- (c) What are the treatments? How many treatments are there?
- (d) How are the factors that are not controlled dealt with?
- (e) Which group serves as the control group?
- (f) What type of experimental design is this?
- (g) Identify the subjects.
- (h) Explain how time of day the course is taught could potentially confound the results of the study. Can anything be done to eliminate the effect of this confounding variable?
- (i) Draw a diagram similar to Figure 7 or 8 to illustrate the design.

10. Pharmacy A pharmaceutical company has developed an experimental drug meant to relieve symptoms associated with

the common cold. The company identifies 300 adult males 25 to 29 years old who have a common cold and randomly divides them into two groups. Group 1 is given the experimental drug, while group 2 is given a placebo. After 1 week of treatment, the subjects report whether they still have cold symptoms, or not.

- (a) What is the response variable in this experiment?
- (b) Think of some of the factors in the study. How are they controlled?
- (c) What are the treatments? How many treatments are there?
- (d) How are the factors that are not controlled dealt with?
- (e) What type of experimental design is this?
- (f) Identify the subjects.
- (g) Draw a diagram similar to Figure 7 or 8 to illustrate the design.

NW 11. Whiter Teeth An ad for Crest Whitestrips Premium claims that the strips will whiten teeth in 7 days and the results will last for 12 months. A researcher who wishes to test this claim studies 20 sets of identical twins. Within each set of twins, one is randomly selected to use Crest Whitestrips Premium in addition to regular brushing and flossing, while the other just brushes and flosses. Whiteness of teeth is measured at the beginning of the study, after 7 days, and every month thereafter for 12 months.

- (a) What type of experimental design is this?
- (b) What is the response variable in this experiment?
- (c) What are the treatments?
- (d) What are other factors (controlled or uncontrolled) that could affect the response variable?
- (e) What might be an advantage of using identical twins as subjects in this experiment?

12. Assessment To help assess student learning in her developmental math courses, a mathematics professor at a community college implemented pre- and posttests for her students. A knowledge-gained score was obtained by taking the difference of the two test scores.

- (a) What type of experimental design is this?
- (b) What is the response variable in this experiment?
- (c) What is the treatment?

13. Automatic Social Behavior The principle of ideomotor action suggests that the act of thinking about a behavior increases the tendency to engage in that behavior. This is sometimes referred to as "priming." In one such study, 30 male and female undergraduate students were randomly assigned to either "an elderly prime condition or a neutral prime condition." The individual assigning the students did not know which group the student was assigned to. The priming manipulation was a scrambled-sentence task, which the subjects believed was for a language proficiency test. Half the students were assigned words such as *Florida, old, forgetful, wrinkle, knits, or alone* (this group is called the elderly priming condition). The other students were assigned words not associated with being elderly (call this the neutral priming condition). For each replication of the study, the subject was given five words and asked to construct a sentence using four of the five words. After the sentence was created, the subject was thanked and told the elevator was down the hall. In the hall, the amount of time it took the subject to travel 9.75 meters was determined. The subjects in the elderly priming

condition had a mean of 8.28 seconds to travel the 9.75 meters while the subjects in the neutral priming condition had a mean of 7.30 seconds to travel the 9.75 meters. The travel time of the elderly priming condition subjects was statistically significantly higher than that of the neutral priming condition. It is also interesting to note that upon exit interviews the subjects were told the nature of the experiment, and none of the elderly priming condition subjects believed the words had an impact on their behavior. *Source:* John A. Bargh, Mark Chen, and Lara Burrows. "Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action." *Journal of Personality and Social Psychology*, 71(2): 230–244, 1996.

- (a) What type of experimental design is this?
- (b) What is the response variable? Is it qualitative or quantitative?
- (c) What is the treatment? How many treatments are there?
- (d) Who are the subjects?
- (e) What is the role of blinding in this study?
- (f) What is the conclusion of the study?

14. Insomnia Researchers wanted to test the effectiveness of a new cognitive behavioral therapy (CBT) compared with both an older behavioral treatment and a placebo therapy for treating insomnia. They identified 75 adults with insomnia. Patients were randomly assigned to one of three treatment groups. Twenty-five patients were randomly assigned to receive CBT (sleep education, stimulus control, and time-in-bed restrictions), another 25 received muscle relaxation training (RT), and the final 25 received a placebo treatment. Treatment lasted 6 weeks, with follow-up conducted at 6 months. To measure the effectiveness of the treatment, researchers used wake time after sleep onset (WASO). CBT produced larger improvements than did RT or placebo treatment. For example, the CBT-treated patients achieved an average 54% reduction in their WASO, whereas RT-treated and placebo-treated patients, respectively, achieved only 16% and 12% reductions in this measure. Results suggest that CBT treatment leads to significant sleep improvements within 6 weeks, and these improvements appear to endure through 6 months of follow-up.

Source: Jack D. Edinger, PhD; William K. Wohlgemuth, PhD; Rodney A. Radtke, MD; Gail R. Marsh, PhD; Ruth E. Quillian, PhD. "Cognitive Behavioral Therapy for Treatment of Chronic Primary Insomnia," *Journal of the American Medical Association* 285:1856–1864, 2001.

- (a) What type of experimental design is this?
- (b) What is the population being studied?
- (c) What is the response variable in this study?
- (d) What are the treatments?
- (e) Identify the experimental units.
- (f) Draw a diagram similar to Figure 7 or 8 to illustrate the design.

15. The Memory Drug? Researchers wanted to evaluate whether ginkgo, an over-the-counter herb marketed as enhancing memory, improves memory in elderly adults as measured by objective tests. To do this, they recruited 98 men and 132 women older than 60 years and in good health. Participants were randomly assigned to receive ginkgo, 40 milligrams (mg) 3 times per day, or a matching placebo. The measure of memory improvement was determined by a standardized test of learning and memory. After 6 weeks of treatment, the data indicated that ginkgo did not increase

performance on standard tests of learning, memory, attention, and concentration. These data suggest that, when taken following the manufacturer's instructions, ginkgo provides no measurable increase in memory or related cognitive function to adults with healthy cognitive function.



Source: Paul R. Solomon et al. "Ginkgo for Memory Enhancement," *Journal of the American Medical Association* 288:835–840, 2002.

- (a) What type of experimental design is this?
- (b) What is the population being studied?
- (c) What is the response variable in this study?
- (d) What is the factor that is set to predetermined levels? What are the treatments?
- (e) Identify the experimental units.
- (f) What is the control group in this study?
- (g) Draw a diagram similar to Figure 7 or 8 to illustrate the design.

16. Shrinking Stomach? Researchers wanted to determine whether the stomach shrinks as a result of dieting. To do this, they randomly divided 23 obese patients into two groups. The 14 individuals in the experimental group were placed on a diet that allowed them to consume 2508 kilojoules (kJ) per day for 4 weeks. The 9 subjects in the control group ate as they normally would. To assess the size of the stomach, a latex gastric balloon was inserted into each subject's stomach and filled with the maximum amount of water that could be tolerated by the patient. The volume of water was compared to the volume that could be tolerated at the beginning of the study. The experimental subjects experienced a 27% reduction in gastric capacity, while the subjects in the control group experienced no change in gastric capacity. It was concluded that a reduction in gastric capacity occurs after a restricted diet.

Source: A. Geliebter et al. "Reduced Stomach Capacity in Obese Subjects after Dieting," *American Journal of Clinical Nutrition* 63(2):170–173, 1996.

- (a) What type of experimental design is this?
- (b) What is the population that the results of this experiment apply to?
- (c) What is the response variable in this study? Is it qualitative or quantitative?
- (d) What are the treatments?
- (e) Identify the experimental units.
- (f) Draw a diagram similar to Figure 7 or 8 to illustrate the design.

17. Platelet-Rich Plasma Does a platelet-rich plasma (PRP) injection into the scalp promote hair growth? Researchers identified 30 female patients with female pattern hair loss. The patients ranged in age from 20 to 45 years. For each patient, two areas with hair loss were identified. A coin toss was used to

decide which area received the PRP injection and which received a saline injection. Injections were given to each patient at one-week intervals. After 6 months the change in the patients' hair density (number of hairs per square centimeter) and hair diameter (millimeters) was measured. The mean difference in hair density (PCP minus saline) was 59.47 hairs/cm² and the mean difference in hair thickness was 0.08 mm. *Source:* A. A. Tawfik and M. A. R. Osman, "The Effect of Autologous Activated Platelet-Rich Plasma Injection on Female Pattern Hair-Loss: A Randomized Placebo-Controlled Study," *J Cosmet Dermatol.* 17:47–53, 2018, <https://doi.org/10.1111/jocd.12357>

- (a) What type of experimental design is this?
- (b) What is the population that is being studied?
- (c) What are the response variables in this study?
- (d) What is the treatment?
- (e) Who are the experimental units?
- (f) What role does randomization play in this study?
- (g) Draw a diagram similar to Figure 7 or 8 to illustrate the design.

18. Dominant Hand Professor Andy Neill wanted to determine if the reaction time of people differs in their dominant hand versus their nondominant hand. To do this, he recruited 15 students. Each student was asked to hold a yardstick between the index finger and thumb. The student was asked to open the hand, release the yardstick, and then catch the yardstick between the index finger and thumb. The distance that the yardstick fell served as a measure of reaction time. A coin flip was used to determine whether the student would use their dominant hand first or the nondominant hand. Results indicated that the reaction time in the dominant hand was less than that of the nondominant hand.

- (a) What type of experimental design is this?
- (b) What is the response variable in this study?
- (c) What is the treatment?
- (d) Identify the experimental units.
- (e) Why did Professor Neill use a coin flip to determine whether the student should begin with the dominant hand or the nondominant hand?
- (f) Draw a diagram similar to Figure 7 or 8 to illustrate the design.

19. Drug Effectiveness A pharmaceutical company wants to test the effectiveness of an experimental drug meant to reduce high cholesterol. The researcher at the pharmaceutical company has decided to test the effectiveness of the drug through a completely randomized design. She has obtained 20 volunteers with high cholesterol: Ann, John, Michael, Kevin, Marissa, Christina, Eddie, Shannon, Julia, Randy, Sue, Tom, Wanda, Roger, Laurie, Rick, Kim, Joe, Colleen, and Bill. Number the volunteers from 1 to 20. Use a random-number generator to randomly assign 10 of the volunteers to the experimental group. The remaining volunteers will go into the control group. List the individuals in each group.

20. Effects of Alcohol A researcher has recruited 20 volunteers to participate in a study. The researcher wishes to measure the effect of alcohol on an individual's reaction time. The 20 volunteers are randomly divided into two groups. Group 1 serves as a control group in which participants drink four 1-ounce cups of a liquid that looks, smells, and tastes like alcohol in 15-minute increments. Group 2 serves as an experimental group in which participants drink four 1-ounce cups of 80-proof alcohol in 15-minute increments. After drinking the last 1-ounce cup, the

participants sit for 20 minutes. After the 20-minute resting period, the reaction time to a stimulus is measured.

- (a) What type of experimental design is this?
 - (b) Use Table I in Appendix A or a random-number generator to divide the 20 volunteers into groups 1 and 2 by assigning the volunteers a number between 1 and 20. Then randomly select 10 numbers between 1 and 20. The individuals corresponding to these numbers will go into group 1.
- 21. Green Tea** You wonder whether green tea lowers cholesterol.
- (a) To research the claim that green tea lowers LDL (so-called bad) cholesterol, you ask a random sample of individuals to divulge whether they are regular green tea users or not. You also obtain their LDL cholesterol levels. Finally, you compare the LDL cholesterol levels of the green tea drinkers to those of the non-green tea drinkers. Explain why this is an observational study.
 - (b) Name some lurking variables that might exist in the study.
 - (c) Suppose, instead of surveying individuals regarding their tea-drinking habits, you decide to conduct a designed experiment. You identify 120 volunteers to participate in the study and decided on three levels of the treatment: a placebo, one cup of green tea daily, two cups of green tea daily. The experiment is to run for one year. The response variable will be the change in LDL cholesterol for each subject from the beginning of the study to the end. What type of experimental design is this?
 - (d) Explain how you would use blinding in this experiment.
 - (e) What is the treatment? How many levels is it set at?
 - (f) What factors might you attempt to control and fix at one level in this experiment?
 - (g) Explain how to use randomization in this experiment. How does randomization neutralize those variables that are not controlled?
 - (h) Suppose you assigned 40 subjects to each of the three treatment groups. In addition, you decided to control the variable exercise by having each subject perform 150 minutes of cardiovascular exercise each week by walking on a treadmill. However, the 40 subjects in the placebo group decided they did not want to walk on the treadmill and skipped the weekly exercise. Explain how exercise is now a confounding variable.
- 22. Priming for Healthy Food** Does alerting shoppers at a grocery store regarding the healthiness (or lack thereof) of energy-dense snack foods change the shopping habits of overweight individuals? To answer this question, researchers randomly gave 42 overweight shoppers a recipe flyer that either contained health information or did not contain the health information. This type of intervention is referred to as *priming*. To determine purchases, the receipts of the participants were reviewed. Results of the study found that shoppers primed with the health- and diet-related words on the recipe bought significantly (almost 75%) fewer unhealthy snacks than those without the primes.
- Source:* E. K. Papies and associates, "Using Health Primes to Reduce Unhealthy Snack Purchases Among Overweight Consumers in a Grocery Store," *International Journal of Obesity* (2013), 1–6.
- (a) What is the research objective?
 - (b) Who are the subjects?
 - (c) Explain why blinding is not possible in this study.
 - (d) What is the explanatory variable in the study?
 - (e) The response variable was number of unhealthy snacks purchased. Is this quantitative or qualitative?

(f) Another factor in the study was weight status (normal weight vs. overweight). Suppose all the normal weight subjects were given the flyer with the prime and overweight subjects were given the flyer without the prime. Explain how confounding would play a role in the study.

23. Batteries An engineer wants to determine the effect of temperature on battery voltage. In particular, he is interested in determining if there is a significant difference in the voltage of the batteries when exposed to temperatures of 90°F, 70°F, and 50°F. Help the engineer design the experiment. Include a diagram to illustrate your design.

24. Tire Design An engineer has just developed a new tire design. However, before going into production, the tire company wants to determine if the new tire reduces braking distance on a car traveling 60 miles per hour compared with radial tires. Design an experiment to help the engineer determine if the new tire reduces braking distance.

25. Designing an Experiment Researchers want to know if there is a link between hypertension (high blood pressure) and consumption of salt. Past studies have indicated that the consumption of fruits and vegetables offsets the negative impact of salt consumption. It is also known that there is quite a bit of person-to-person variability in the ability of the body to process and eliminate salt. However, no method exists for identifying individuals who have a higher ability to process salt. The U.S. Department of Agriculture recommends that daily intake of salt should not exceed 2400 milligrams (mg). The researchers want to keep the design simple, so they choose to conduct their study using a completely randomized design.

(a) What is the response variable in the study?

(b) Name three factors that have been identified.

(c) For each factor identified, determine whether the variable can be controlled or cannot be controlled. If a factor cannot be controlled, what should be done to reduce variability in the response variable?

(d) How many treatments would you recommend? Why?

26. Search a newspaper, magazine, or other periodical that describes an experiment. Identify the population, experimental unit, response variable, treatment, factors, and their levels.

27. Research the *placebo effect* and the *Hawthorne effect*. Write a paragraph that describes how each affects the outcome of an experiment.

28. Coke or Pepsi Design an experiment where the goal is to determine whether people prefer Coke or Pepsi. In the design, be sure to identify the response variable, the role of blinding, and randomization.

Explaining the Concepts

29. The Dictator Game In their book *SuperFreakonomics*, authors Steven Levitt and Stephen Dubner describe the research of behavioral economist John List. List recruited customers and dealers at a baseball-card show to participate in an experiment in which the customer would state how much he was willing to pay for a single baseball card. The prices ranged from \$4 (lowball) to \$50 (premium card). The dealer would then give the customer a card that was supposed to correspond to the offer price. In this setting, the dealer could certainly give the buyer a card worth less than the offer price, but this rarely happened. The card received by the buyer was

close in value to the price offered. Next, List went to the trading floor at the show and again recruited customers. But this time the customers approached dealers at their booth. The dealers did not know they were being watched. The scenario went something like this: as the customer approached the dealer's booth, he would say, "Please give me the best Derek Jeter card you can for \$20." In this scenario, the dealers consistently ripped off the customers by giving them cards worth much less than the offer price. In fact, the dealers who were the worst offenders were the same dealers who refused to participate in List's study. Do you believe that individuals who volunteer for experiments are scientific do-gooders? That is, do you believe that in designed experiments subjects strive to meet the expectations of the researcher? In addition, do you believe that results of experiments may suffer because many experiments require individuals to volunteer, and individuals who are not do-gooders do not volunteer for studies? Now, explain why control groups are needed in designed experiments and the role they can play in neutralizing the impact of scientific do-gooders.

30. Ebola In October 2014, there was an Ebola breakout in West Africa. At the time, there was no vaccine for the virus, however, there were some experimental drugs that had not yet been approved for humans. Because the spread of the disease was reaching an epidemic, there were calls to initiate randomized trials of an experimental drug on human subjects right away.

(a) Discuss how you would go about designing a randomized trial to assess the efficacy of an experimental Ebola vaccine.

(b) Doctors Without Borders was on the record, prior to any randomized trial, as saying that trials in which subjects are assigned to a control group are unethical. Discuss the ethics behind a randomized trial of a potential life-saving vaccine to test its efficacy while an epidemic is raging.

31. Covid-19 In May, 2020, the biotechnology company Novavax began human trials of its Covid-19 vaccine. The Phase I trial was a placebo-controlled, observer-blinded design. Most clinical trials go through five phases: Phase 0 (Pre-clinical), Phase I (Safety), Phase II (Efficacy), Phase III (Comparison), and Phase IV (Surveillance). Research the five phases of a clinical trial. Discuss the ethics involved in using control groups for a potential life-saving vaccine.

32. The Placebo Effect Research the link between the release of dopamine and the placebo effect.

33. What is the role of randomization in a designed experiment? If you were conducting a completely randomized design with three treatments and 90 experimental units, describe how you would randomly assign the experimental units to the treatments.

34. Treatments may be a combination of factors rather than a single factor. For example, suppose we want to investigate the role of diet and drugs in weight loss. Suppose we have three diet plans (a saturated-fat diet, the Mediterranean diet, or the U.S. National Cholesterol Education Plan, or NCEP-1 Diet) along with a new experimental weight-loss drug versus a placebo. So, diet has three treatment levels and drug has two treatment levels. Determine the six treatments that consider all possible combinations of these two factors.



Chapter 1 Review

Summary

We defined statistics as a science in which data are collected, organized, summarized, and analyzed to infer characteristics regarding a population. Statistics also provides a measure of confidence in the conclusions that are drawn. Descriptive statistics consists of organizing and summarizing information, while inferential statistics consists of drawing conclusions about a population based on results obtained from a sample. The population is a collection of individuals about which information is desired and the sample is a subset of the population.

Data are the observations of a variable. Data can be either qualitative or quantitative. Quantitative data are either discrete or continuous.

Data can be obtained from four sources: a census, an existing source, an observational study, or a designed experiment. A census will list all the individuals in the population, along with certain characteristics. Due to the cost of obtaining a census, most researchers opt

for obtaining a sample. Web-scraping is the process of extracting data from the Internet and is a relatively new method for obtaining data. In addition, many organizations and governments are now making their data available for download from the web. In observational studies, the response variable is measured without attempting to influence its value, so the explanatory variable is not manipulated. Designed experiments are used when control of the individuals in the study is desired to isolate the effect of a certain treatment on a response variable.

We introduced five sampling methods: simple random sampling, stratified sampling, systematic sampling, cluster sampling, and convenience sampling. All the sampling methods, except for convenience sampling, allow for unbiased statistical inference to be made. Convenience sampling typically leads to an unrepresentative sample and biased results.

Vocabulary

Statistics (p. 3)
Data (pp. 3, 8)
Population (p. 5)
Individual (p. 5)
Sample (p. 5)
Statistic (p. 5)
Descriptive statistics (p. 5)
Inferential statistics (pp. 5, 47)
Parameter (p. 5)
Variable (p. 6)
Qualitative or categorical variable (p. 7)
Quantitative variable (p. 7)
Discrete variable (p. 8)
Continuous variable (p. 8)
Qualitative data (p. 8)
Quantitative data (p. 8)
Discrete data (p. 8)
Continuous data (p. 8)
Nominal level of measurement (p. 9)
Ordinal level of measurement (p. 9)
Interval level of measurement (p. 10)
Ratio level of measurement (p. 10)
Explanatory variable (p. 15)

Response variable (p. 15)
Observational study (p. 15)
Designed experiment (p. 15)
Confounding (p. 16)
Lurking variable (p. 16)
Confounding variable (p. 17)
Retrospective (p. 18)
Prospective (p. 18)
Census (p. 19)
Web scraping (p. 19)
Data mining (p. 19)
Random sampling (p. 23)
Simple random sampling (p. 24)
Simple random sample (p. 24)
Frame (p. 25)
Sampling without replacement (p. 25)
Sampling with replacement (p. 25)
Seed (p. 26)
Stratified sample (p. 30)
Systematic sample (p. 32)
Cluster sample (p. 33)
Convenience sample (p. 34)
Self-selected (p. 34)
Voluntary response (p. 34)

Bias (p. 38)
Sampling bias (p. 38)
Undercoverage (p. 38)
Nonresponse bias (p. 39)
Response bias (p. 39)
Open question (p. 41)
Closed question (p. 41)
Nonsampling error (p. 41)
Sampling error (p. 41)
Experiment (p. 45)
Factors (p. 45)
Treatment (p. 45)
Experimental unit (p. 45)
Subject (p. 45)
Control group (p. 45)
Placebo (p. 45)
Placebo effect (p. 45)
Blinding (p. 45)
Single-blind (p. 45)
Double-blind (p. 45)
Design (p. 46)
Replication (p. 47)
Completely randomized design (p. 48)
Matched-pairs design (p. 49)

Objectives

Section	You should be able to . . .	Example(s)	Review Exercises
1.1	1 Define statistics and statistical thinking (p. 3)	pp. 3–4	
	2 Explain the process of statistics (p. 4)	2	3, 10, 11
	3 Distinguish between qualitative and quantitative variables (p. 6)	3	7–9
	4 Distinguish between discrete and continuous variables (p. 7)	4, 5	7, 8
	5 Determine the level of measurement of a variable (p. 9)	6	12–15
1.2	1 Distinguish between an observational study and an experiment (p. 14)	1–3	16, 17, 27(b)
	2 Explain the various types of observational studies (p. 18)	pp. 18–20	2, 18
1.3	1 Obtain a simple random sample (p. 24)	1–3	24, 26
1.4	1 Obtain a stratified sample (p. 30)	1	21
	2 Obtain a systematic sample (p. 32)	2	22, 25
	3 Obtain a cluster sample (p. 33)	3	20
1.5	1 Explain the sources of bias in sampling (p. 38)	pp. 38–42	4, 5
1.6	1 Describe the characteristics of an experiment (p. 45)	1	1(f)
	2 Explain the steps in designing an experiment (p. 46)	pp. 46–47	6
	3 Explain the completely randomized design (p. 48)	2	27, 30, 31
	4 Explain the matched-pairs design (p. 49)	3	28, 31

Review Exercises

1. Define each of the following.

- (a) Response variable
- (b) Variable
- (c) Qualitative variable
- (d) Quantitative variable
- (e) Observational study
- (f) Designed experiment
- (g) Confounding
- (h) Lurking variable

2. List and describe the three major types of observational studies.

3. What is meant by the *process of statistics*?

4. List and explain the three sources of bias in sampling. Provide some methods that might be used to minimize bias in sampling.

5. Distinguish between sampling and nonsampling error.

6. Explain the steps in designing an experiment.

In Problems 7–9, classify the variable as qualitative or quantitative. If the variable is quantitative, state whether it is discrete or continuous.

7. Number of new automobiles sold at a dealership on a given day

8. Weight in carats of an uncut diamond

9. Brand name of a pair of running shoes

In Problems 10 and 11, determine whether the underlined value is a parameter or a statistic.

10. In a survey of 1011 people age 50 or older, 73% agreed with the statement “I believe in life after death.”

Source: Bill Newcott. “Is There Life after Death?,” *AARP Magazine*, Sept./Oct. 2007.

11. **Completion Rate** In the 2019 NCAA Football Championship Game, quarterback Trevor Lawrence completed 62.5% of his passes for a total of 347 yards.

In Problems 12–15, determine the level of measurement of each variable.

12. Birth year

13. Marital status

14. Stock rating (strong buy, buy, hold, sell, strong sell)

15. Number of siblings

In Problems 16 and 17, determine whether the study depicts an observational study or a designed experiment.

16. A parent group examines 25 randomly selected PG-13 movies and 25 randomly selected PG movies, records the number of sexual innuendos and curse words that occur in each, and then compares the number of sexual innuendos and curse words between the two movie ratings.

17. A sample of 504 patients in early stages of Alzheimer’s disease is divided into two groups. One group receives an experimental drug; the other receives a placebo. The advance of the disease in the patients from the two groups is tracked at 1-month intervals over the next year.

- 18.** Read the following description of an observational study and determine whether it is a cross-sectional, a case-control, or a cohort study. Explain your choice.

The Cancer Prevention Study II (CPS-II) examines the relationship among environmental and lifestyle factors of cancer cases by tracking approximately 1.2 million men and women. Study participants completed an initial study questionnaire in 1982 providing information on a range of lifestyle factors, such as diet, alcohol and tobacco use, occupation, medical history, and family cancer history. These data have been examined extensively in relation to cancer mortality. The vital status of study participants is updated biennially.

Source: American Cancer Society.

In Problems 19–22, determine the type of sampling used.

- 19.** On election day, a pollster for Fox News positions herself outside a polling place near her home and asks the first 50 voters leaving the facility to complete a survey.
- 20.** An Internet service provider randomly selects 15 residential blocks from a large city and surveys every household in these 15 blocks to determine the number that would use a high-speed Internet service.
- 21.** Thirty-five sophomores, 22 juniors, and 35 seniors are randomly selected to participate in a study from 574 sophomores, 462 juniors, and 532 seniors at a certain high school.
- 22.** Officers for the Department of Motor Vehicles pull aside every 40th tractor trailer passing through a weigh station, starting with the 12th, for an emissions test.
- 23.** Each of the following surveys has bias. Determine the type of bias and suggest a remedy.
- A politician sends a survey about tax issues to a random sample of subscribers to a literary magazine.
 - An interviewer with little foreign language knowledge is sent to an area where her language is not commonly spoken.
 - A data-entry clerk mistypes survey results into his computer.
- 24. Obtaining a Simple Random Sample** The mayor of a town wants to conduct personal interviews with small business owners to determine if there is anything he could do to help improve business conditions. The following list gives the names of the companies in the town. Obtain a simple random sample of size 5 from the companies in the town.

Allied Tube and Conduit	Lighthouse Financial	Senese's Winery
Bechstien Construction Co.	Mill Creek Animal Clinic	Skyline Laboratory
Cizer Trucking Co.	Nancy's Flowers	Solus, Maria, DDS
D & M Welding	Norm's Jewelry	Trust Lock and Key
Grace Cleaning Service	Papoose Children's Center	Ultimate Carpet
Jiffy Lube	Plaza Inn Motel	Waterfront Tavern
Levin, Thomas, MD	Risky Business Security	WPA Pharmacy

- 25. Obtaining a Systematic Sample** A quality-control engineer wants to be sure that bolts coming off an assembly line are within prescribed tolerances. He wants to conduct a systematic sample by selecting every 9th bolt to come off the assembly line. The machine produces 30,000 bolts per day, and the engineer wants a sample of 32 bolts. Which bolts will be sampled?

- 26. Obtaining a Simple Random Sample** Based on the Military Standard 105E (ANSI/ASQC Z1.4, ISO 2859) Tables, a lot of 91 to 150 items with an acceptable quality level (AQL) of 1% and a normal inspection plan would require a sample of size 13 to be inspected for defects. If the sample contains no defects, the entire lot is accepted. Otherwise, the entire lot is rejected. A shipment of 100 night-vision goggles is received and must be inspected. Discuss the procedure you would follow to obtain a simple random sample of 13 goggles to inspect.

- 27. Tooth-Whitening Gum** Smoking and drinking coffee have a tendency to stain teeth. In an effort to determine the ability of chewing gum to remove stains on teeth, researchers conducted an experiment in which 64 bovine incisors (teeth) were stained with natural pigments such as coffee for 10 days. Each tooth was randomly assigned to one of four treatments: gum A, gum B, gum C, or saliva. Each tooth group was placed into a device that simulated a human chewing gum. The temperature of the device was maintained at body temperature and the tooth was in the device for 20 minutes. The process was repeated six times (for a total of 120 minutes of chewing). The researcher conducting the experiment did not know which treatment was being applied to each experimental unit. Upon removing a tooth from the chewing apparatus, the color change was measured using a spectrophotometer. The percentage of stain removed by each treatment after 120 minutes is as follows: Gum A, 47.6%; Gum B, 45.2%, Gum C, 21.4%, Saliva, 2.1%.

The researchers concluded that gums A and B removed significantly more stain than gum C or saliva. In addition, gum C removed significantly more stain than saliva.

Source: Michael Moore et al. "In Vitro Tooth Whitening Effect of Two Medicated Chewing Gums Compared to a Whitening Gum and Saliva," *BioMed Central Oral Health* 8:23, 2008.

- Identify the research objective.
- Is this an observational study or designed experiment? Why?
- If observational, what type of observational study is this? If an experiment, what type of experimental design is this?
- What is the response variable?
- What is the explanatory variable? Is it qualitative or quantitative?
- Identify the experimental units.
- State a factor that could affect the value of the response variable that is fixed at a set level.
- What is the conclusion of the study?

28. Reaction Time Researchers wanted to assess the effect of low alcohol consumption on reaction time in seniors, believing that even low levels of alcohol consumption can impair the ability to walk, thereby increasing the likelihood of falling. They identified 13 healthy seniors who were not heavy consumers of alcohol. The experiment took place in late afternoon. Each subject was instructed to have a light lunch and not to drink any caffeinated drinks in the 4 hours prior to arriving at the lab. The seniors were asked to walk on a treadmill on which an obstacle would appear randomly. The reaction time was measured by determining the time it took the senior to lift his or her foot upon the appearance of the obstacle. First, each senior walked the treadmill after consuming a drink consisting of water mixed with orange juice with the scent and taste of vodka. The senior was then asked to drink two additional drinks (40% vodka mixed with orange juice). The senior then walked on the treadmill again. The average response time increased by 19 milliseconds after the alcohol treatment. The researchers concluded that response times are significantly delayed even for low levels of alcohol consumption.

*Source: Judith Hegeman et al. “Even Low Alcohol Concentrations Affect Obstacle Avoidance Reactions in Healthy Senior Individuals,” *BMC Research Notes* 3:243, 2010.*

- (a) What type of experimental design is this?
- (b) What is the response variable in this experiment? Is it quantitative or qualitative?
- (c) What is the treatment?
- (d) What factors were controlled and set at a fixed level in this experiment?
- (e) Can you think of any factors that may affect reaction to alcohol that were not controlled?
- (f) Why do you think the researchers used a drink that had the scent and taste of vodka to serve as the treatment for a baseline measure?
- (g) What was the conclusion of the study? To whom does this conclusion apply?

29. Multiple Choice A common tip for taking multiple-choice tests is to always pick (b) or (c) if you are unsure. The idea is that instructors tend to feel the answer is more hidden if it is surrounded by distractor answers. An astute statistics instructor is aware of this and decides to use a table of random digits to select which choice will be the correct answer. If each question has five choices, use Table I in Appendix A or a random-number

generator to determine the correct answers for a 20-question multiple-choice exam.

30. Humor in Advertising A marketing research firm wants to know whether information presented in a commercial is better recalled when presented using humor or serious commentary by adults between 18 and 35 years of age. They will use an exam that asks questions of 50 subjects about information presented in the ad. The response variable will be percentage of information recalled. Create a completely randomized design to answer the question. Be sure to include a diagram to illustrate your design.

31. Describe what is meant by a matched-pairs design. Contrast this experimental design with a completely randomized design.

32. Internet Search Go to an online science magazine such as *Science Daily* (www.sciencedaily.com) or an open source online medical journal such as BioMed Central (www.biomedcentral.com) and identify an article that includes statistical research.

- (a) Was the study you selected a designed experiment or an observational study?
- (b) What was the research objective?
- (c) What was the response variable in the study?
- (d) Summarize the conclusions of the study.

33. Cell Phones Many newspaper articles discuss the dangers of teens texting while driving. Suppose you are a journalist and want to chime in on the discussion. However, you want your article to be more compelling, so you decide to conduct an experiment with one hundred 16- to 19-year-old volunteers. Design an experiment that will assess the dangers of texting while driving. Decide on the type of experiment (completely randomized, matched-pairs, or other), a response variable, the explanatory variables, and any controls that will be imposed. Also, explain how you are going to obtain the data without potentially harming your subjects. Write an article that presents the experiment to your readers so that they know what to anticipate in your follow-up article. Remember, your article is for the general public, so be sure to clearly explain the various facets of your experiment.

34. What is the role of randomization in a designed experiment? If you were conducting a completely randomized design with four treatments and 100 experimental units, describe how you would randomly assign the experimental units to the treatments.



Chapter Test

1. List the four components that comprise the definition of statistics.
2. What is meant by the *process of statistics*?

In Problems 3–5, determine if the variable is qualitative or quantitative. If the variable is quantitative, determine if it is discrete or continuous. State the level of measurement of the variable.

3. Time to complete the 500-meter race in speed skating.
4. Video game rating system by the Entertainment Software Rating Board (EC, E, E10+, T, M, AO, RP)
5. The number of surface imperfections on a camera lens.

In Problems 6 and 7, determine whether the study depicts an observational study or a designed experiment. Identify the response variable in each case.

6. A random sample of 30 digital cameras is selected and divided into two groups. One group uses a brand-name battery, while the other uses a generic plain-label battery. All variables besides battery type are controlled. Pictures are taken under identical conditions and the battery life of the two groups is compared.
7. A pollster asks 800 adult Americans whether the gap between the rich and poor will grow or shrink over the next 30 years.
8. Contrast the three major types of observational studies in terms of the time frame when the data are collected.
9. Compare and contrast observational studies and designed experiments. Which study allows a researcher to claim causality?
10. Explain why it is important to use a control group and blinding in an experiment.

11. List the steps required to conduct an experiment.
12. A cellular phone company is looking for ways to improve customer satisfaction. They want to select a simple random sample of four stores from their 15 franchises in which to conduct customer satisfaction surveys. Discuss the procedure you would use, and then use the procedure to select a simple random sample of size $n = 4$. The locations are as follows:

Afton	Ballwin	Chesterfield	Clayton	Deer Creek
Ellisville	Farmington	Fenton	Ladue	Lake St. Louis
O'Fallon	Pevely	Shrewsbury	Troy	Warrenton

13. A congresswoman wants to survey her constituency regarding public policy. She asks one of her staff members to obtain a sample of residents of the district. The frame she has available lists 9012 Democrats, 8302 Republicans, and 3012 Independents. Obtain a stratified random sample of 8 Democrats, 7 Republicans, and 3 Independents. Be sure to discuss the procedure used.

14. A farmer has a 500-acre orchard in Florida. Each acre is subdivided into blocks of 5. Altogether, there are 2500 blocks of trees on the farm. After a frost, he wants to get an idea of the extent of the damage. Obtain a sample of 10 blocks of trees using a cluster sample. Be sure to discuss the procedure used.

15. A casino manager wants to inspect a sample of 14 slot machines in his casino for quality-control purposes. There are 600 sequentially numbered slot machines operating in the casino. Obtain a systematic sample of 14 slot machines. Be sure to discuss how you obtained the sample.

16. Describe what is meant by an experiment that has a completely randomized design.

17. Each of the following surveys has bias. Identify the type of bias.

- (a) A television survey that gives 900 phone numbers for viewers to call with their vote. Each call costs \$2.00.
- (b) An employer distributes a survey to her 450 employees asking them how many hours each week, on average, they surf the Internet during business hours. Three of the employees complete the survey.
- (c) A question on a survey asks, “Do you favor or oppose a minor increase in property tax to ensure fair salaries for teachers and properly equipped school buildings?”
- (d) A researcher conducting a poll about national politics sends a survey to a random sample of subscribers to *Time* magazine.

18. **Shapely Glasses** Does the shape of a glass play a role in determining the amount of time it takes to finish the drink? Researchers identified 159 male and female self-professed social drinkers. One week the subjects were given a 12 ounce beer with either a straight glass or a curved glass. A week later the subjects were given a 12 ounce beer in the other glass. The first week, subjects were given a glass and asked to consume the drink at their own pace while watching television. The time to complete the drink was measured. During the second week, the subjects were given the other shaped glass and asked to complete a computer task. Again, the time to complete the drink was measured. The type of glass given in the first week was determined randomly. The researchers found that the time to complete the drink was significantly faster for the curved glass.

*Source: A. S. Attwood, N. E. Scott-Samuel, G. Stothart, M. R. Munafò (2012) “Glass Shape Influences Consumption Rate for Alcoholic Beverages.” *PLoS ONE* 7(8): e43007. doi:10.1371/journal.pone.0043007*

- (a) What type of experimental design is this?
- (b) Who are the subjects?
- (c) What is the treatment?
- (d) What is the response variable? Is it qualitative or quantitative?
- (e) Explain the role randomization plays in this experiment.
- (f) Draw a figure that illustrates the design.

19. Nucryst Pharmaceuticals, Inc., announced the results of its first human trial of NPI 32101, a topical form of its skin ointment. A total of 225 patients diagnosed with skin irritations were randomly divided into three groups as part of a double-blind, placebo-controlled study to test the effectiveness of the new topical cream. The first group received a 0.5% cream, the second group received a 1.0% cream, and the third group received a placebo. Groups were treated twice daily for a 6-week period.

Source: www.nucryst.com

- (a) What type of experimental design is this?
- (b) What is the factor that is set to predetermined levels? What are the treatments?
- (c) What does it mean for this study to be double-blind?
- (d) What is the control group for this study?
- (e) Identify the experimental units.
- (f) Draw a diagram to illustrate the design.

20. Researchers Katherine Tucker and associates wanted to determine whether consumption of cola is associated with lower bone mineral density. They looked at 1125 men and 1413 women in the Framingham Osteoporosis Study, which is a cohort that began in 1971. The first examination in this study began between 1971 and 1975, with participants returning for an examination every 4 years. Based on results of questionnaires, the researchers were able to determine

cola consumption on a weekly basis. Analysis of the results indicated that women who consumed at least one cola per day (on average) had a bone mineral density that was significantly lower at the femoral neck than those who consumed less than one cola per day. The researchers did not find this relation in men.

Source: “Colas, but not other carbonated beverages, are associated with low bone mineral density in older women: The Framingham Osteoporosis Study,” *American Journal of Clinical Nutrition* 84: 936–942, 2006.

- (a) Why is this a cohort study?
 - (b) What is the response variable in this study? What is the explanatory variable?
 - (c) Is the response variable qualitative or quantitative?
 - (d) The following appears in the article: “Variables that could potentially confound the relation between carbonated beverage consumption and bone mineral density were obtained from information collected (in the questionnaire).” What does this mean?
 - (e) Can you think of any lurking variables that should be accounted for?
 - (f) What are the conclusions of the study? Does increased cola consumption cause a lower bone mineral density?
- 21.** Explain the difference between a lurking variable and a confounding variable.

Making an Informed Decision

What College Should I Attend?

One of the most difficult tasks of surveying is phrasing questions so that they are not misunderstood. In addition, questions must be phrased so that the researcher obtains answers that allow for meaningful analysis. We wish to create a questionnaire that can be used to make an informed decision about what college to attend. In addition, we want to develop a plan for determining how well a particular college educates its students.

1. Using the school you are currently attending, determine a sampling plan to obtain a random sample of 20 students.
2. Develop a questionnaire that can be administered to the students. The questions in the survey should obtain demographic information about the student (age, gender) as well as questions that pertain to why they chose the school. Also, develop questions that address the students' satisfaction with their school choice. Finally, develop questions that relate to things you consider important in a college environment.

Administer the survey either by paper or use an online survey program such as Poll Monkey, Google, or StatCrunch.

3. Summarize your findings. Based on your findings, would you choose the school for yourself?
4. A second gauge in determining college choice is how well the school educates its students. Certainly, some schools have higher-caliber students as incoming freshmen than others, but you do not want academic ability upon entering the school to factor in your decision. Rather, you want to measure how much the college has increased its students' skill set. Design an experiment that would allow you to measure the response variable “increase in academic skill set” for a particular college. Provide detail for controls in obtaining this information.





Descriptive Statistics

Remember, statistics is a process. The first chapter (Part 1) dealt with the first two steps in the statistical process: (1) identify the research objective and (2) collect the data needed to answer the questions in the research objective. The next three chapters deal with organizing, summarizing, and presenting the data collected. This step in the process is called *descriptive statistics*.

CHAPTER 2 Summarizing Data in Tables and Graphs

CHAPTER 3 Numerically Summarizing Data

CHAPTER 4 Describing the Relation between Two Variables



Summarizing Data in Tables and Graphs

Outline

- 2.1** Organizing Qualitative Data
- 2.2** Organizing Quantitative Data
- 2.3** Graphical Misrepresentations of Data

Making an Informed Decision



Suppose you work for the school newspaper. Your editor approaches you with a special reporting assignment. Your task is to write an article that describes the “typical” student at your school, complete with supporting information. How are you going to do this assignment? See the Decisions project on page 106.

Putting It Together

Chapter 1 discussed how to identify the research objective and collect data. We learned that data can be obtained from either observational studies or designed experiments. When data are obtained, they are referred to as **raw data**.

The purpose of this chapter is to learn how to organize raw data into a meaningful form so that we can understand what the data are telling us. The first step in determining how to organize raw data is to determine whether the data are qualitative or quantitative.

2.1 Organizing Qualitative Data



Preparing for This Section Before getting started, review the following:

- Qualitative data (Section 1.1, pp. 6–8)
- Level of measurement (Section 1.1, pp. 9–10)

Objectives

- Organize qualitative data in tables
- Construct bar graphs
- Construct pie charts

In this section we will concentrate on tabular and graphical summaries of qualitative data. In Section 2.2 we discuss methods for summarizing quantitative data.

① Organize Qualitative Data in Tables

Recall that qualitative (or categorical) data provide measures that categorize or classify an individual. When raw qualitative data are collected, we often first determine the number of individuals within each category.

Definition

A **frequency distribution** lists each category of data and the number of occurrences for each category of data.

EXAMPLE 1 Organizing Qualitative Data into a Frequency Distribution

Problem A physical therapist wants to determine types of rehabilitation required by her patients. To do so, she obtains a simple random sample of 30 of her patients and records the body part requiring rehabilitation. See Table 1. Construct a frequency distribution of location of injury.

Approach To construct a frequency distribution, create a list of the body parts (categories) and tally each occurrence. Finally, add up the number of tallies to determine the frequency.

Solution Table 2 shows that the back is the most common body part requiring rehabilitation, with a total of 12.

CAUTION!

The data in Table 2 are still qualitative. The frequency simply represents the count of each category.

Table 1

Back	Back	Hand
Wrist	Back	Groin
Elbow	Back	Back
Back	Shoulder	Shoulder
Hip	Knee	Hip
Neck	Knee	Knee
Shoulder	Shoulder	Back
Back	Back	Back
Knee	Knee	Back
Hand	Back	Wrist

Source: Krystal Catton, student at Joliet Junior College.

Table 2

Body Part	Tally	Frequency
Back		12
Wrist		2
Elbow		1
Hip		2
Shoulder		4
Knee		5
Hand		2
Groin		1
Neck		1



In any frequency distribution, it is a good idea to add up the frequency column to make sure that it equals the number of observations. In Example 1, the frequency column adds up to 30, as it should.

Often, we want to know the *relative frequency* of the categories, rather than the frequency.

Definitions

IN OTHER WORDS

A frequency distribution shows the number of observations that belong in each category. A relative frequency distribution shows the proportion of observations that belong in each category.

The **relative frequency** is the proportion (or percent) of observations within a category and is found using the formula

$$\text{Relative frequency} = \frac{\text{frequency}}{\text{sum of all frequencies}} \quad (1)$$

A **relative frequency distribution** lists each category of data together with the relative frequency.

EXAMPLE 2

Constructing a Relative Frequency Distribution of Qualitative Data

Problem Using the summarized data in Table 2, construct a relative frequency distribution.

Approach Add all the frequencies, and then use Formula (1) to compute the relative frequency of each category of data.

Solution The sum of all the values in the frequency column in Table 2 is 30.

Now compute the relative frequency of each category. For example, the relative frequency of the category *Back* is $12/30 = 0.4$. The relative frequency distribution is shown in Table 3.

Table 3

Body Part	Frequency	Relative Frequency
Back	12	$\frac{12}{30} = 0.4$
Wrist	2	$\frac{2}{30} \approx 0.0667$
Elbow	1	0.0333
Hip	2	0.0667
Shoulder	4	0.1333
Knee	5	0.1667
Hand	2	0.0667
Groin	1	0.0333
Neck	1	0.0333
Total	30	1

Using Technology

Some statistical spreadsheets such as Minitab and StatCrunch have a command that will construct a frequency and relative frequency distribution of raw qualitative data.

From the distribution, the most common body part for rehabilitation is the back.



It is a good idea to add up the relative frequencies to be sure they sum to 1. In fraction form, the sum should be exactly 1. In decimal form, the sum may differ slightly from 1 due to rounding.

2 Construct Bar Graphs

Once raw data are organized in a table, we can create graphs. Just as “a picture is worth a thousand words,” pictures of data result in a more powerful message than tables. Try the following exercise: Open a newspaper and look at a table and a graph. Study each.

Now put the paper away and close your eyes. What do you see in your mind's eye? Can you recall information more easily from the table or the graph? In general, people are more likely to recall information obtained from a graph than they are from a table.

A common device for graphically representing qualitative data is a bar graph.

Definition

A **bar graph** is constructed by labeling each category of data on either the horizontal or vertical axis and the frequency or relative frequency of the category on the other axis. Rectangles of equal width are drawn for each category. The height of each rectangle represents the category's frequency or relative frequency.

EXAMPLE 3

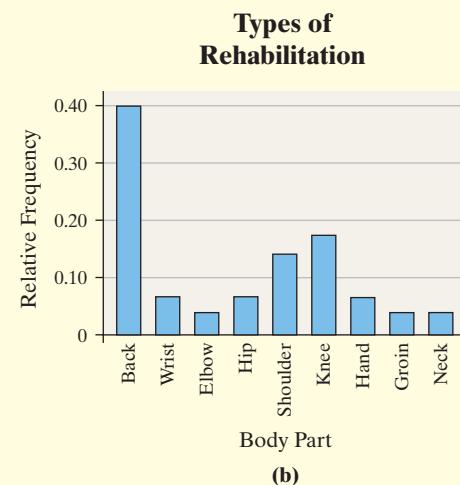
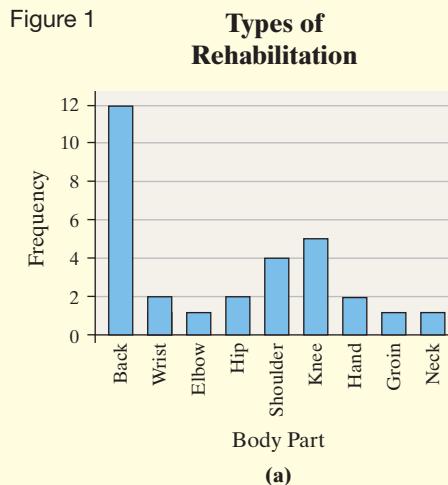
Constructing a Frequency and a Relative Frequency Bar Graph

Problem Use the data summarized in Table 3 to construct a frequency bar graph and a relative frequency bar graph.

Approach Use a horizontal axis to indicate the categories of the data (body parts) and a vertical axis to represent the frequency or relative frequency. Draw rectangles of equal width to the height that is the frequency or relative frequency for each category. The bars do not touch each other.

Solution Figure 1(a) shows the frequency bar graph, and Figure 1(b) shows the relative frequency bar graph.

Figure 1



CAUTION!

Graphs that start the scale at some value other than 0 or have bars with unequal widths, bars with different colors, or three-dimensional bars can misrepresent the data.

EXAMPLE 4

Constructing a Frequency or Relative Frequency Bar Graph Using Technology

Problem Use a statistical spreadsheet to construct a frequency or relative frequency bar graph for the data in Example 1.

Approach We will use Excel to construct the frequency and relative frequency bar graphs. The steps for constructing the graphs using Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on pages 69–70. **Note:** TI-graphing calculators cannot draw frequency or relative frequency bar graphs.

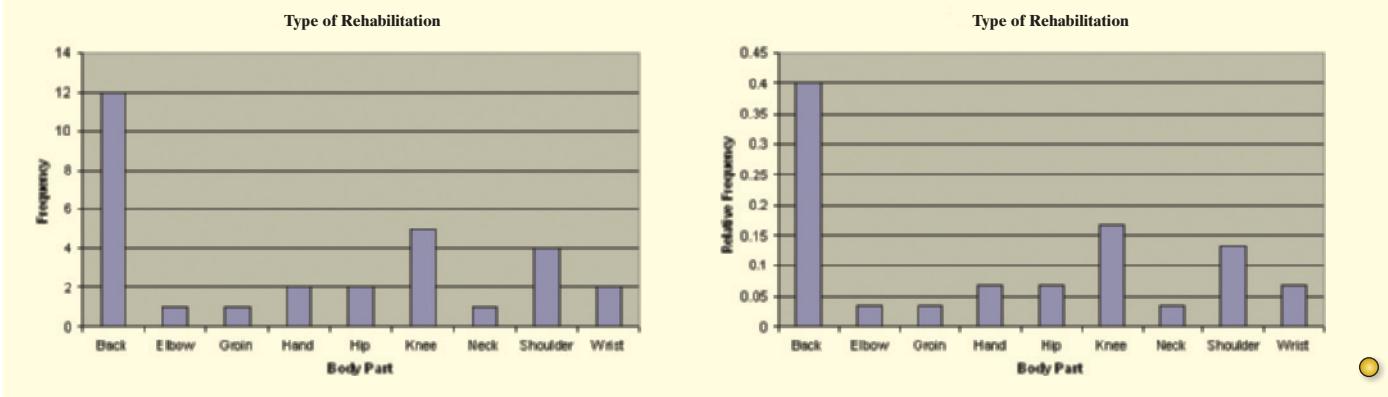
Solution Figure 2 on the following page shows the frequency and relative frequency bar graphs obtained from Excel.

(continued)

Using Technology

The graphs obtained from a different statistical package may differ from those in Figure 2. Some packages use the word *count* in place of *frequency* or *percent* in place of *relative frequency*.

Figure 2

**NW Now Work Problems 21(c)–(d)**

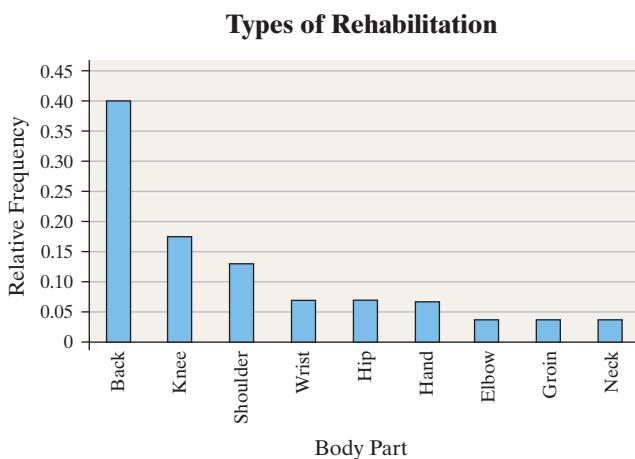
In bar graphs, the order of the categories usually does not matter. However, bar graphs that have categories arranged in decreasing order of frequency help prioritize categories for decision-making purposes in areas such as quality control, human resources, and marketing.

Definition

A **Pareto chart** is a bar graph whose bars are drawn in decreasing order of frequency or relative frequency.

Figure 3 illustrates a relative frequency Pareto chart for the data in Table 3.

Figure 3

**Side-by-Side Bar Graphs**

Suppose we want to know whether more people are finishing college today than in 1990. We could draw a **side-by-side bar graph** to compare the data for the two different years. Data sets should be compared by using relative frequencies, because different sample or population sizes make comparisons using frequencies difficult or misleading.

EXAMPLE 5**Comparing Two Data Sets**

Problem The data in Table 4 represent the educational attainment in 1990 and 2017 of adults 25 years and older who are residents of the United States. The data are in thousands. So 39,344 represents 39,344,000.

- Draw a side-by-side relative frequency bar graph of the data.
- Make some general conclusions based on the graph.

Table 4

Educational Attainment	1990	2017
Not a high school graduate	39,344	26,582
High school diploma	47,643	60,032
Some college, no degree	29,780	45,110
Associate's degree	9,792	18,761
Bachelor's degree	20,833	43,585
Graduate or professional degree	11,478	27,181
Totals	158,870	221,251

Source: U.S. Census Bureau.

Approach First, determine the relative frequencies of each category for each year. To construct side-by-side bar graphs, draw two bars for each category of data, one for 1990, the other for 2017.

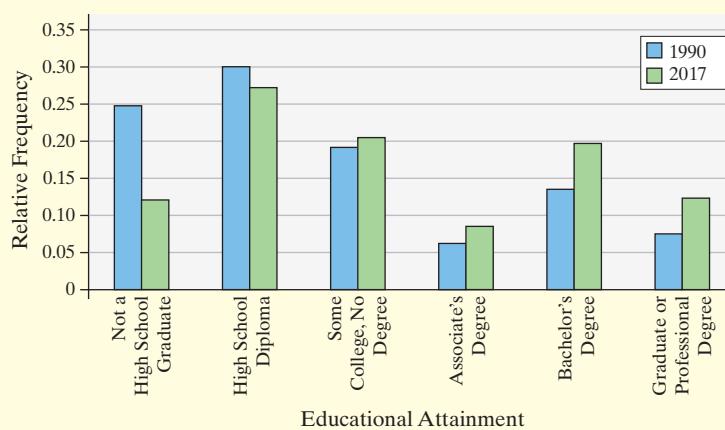
Solution

(a) Table 5 shows the relative frequency for each category. The side-by-side bar graph is shown in Figure 4.

Table 5

Educational Attainment	1990	2017
Not a high school graduate	0.2476	0.1201
High school diploma	0.2999	0.2713
Some college, no degree	0.1874	0.2039
Associate's degree	0.0616	0.0848
Bachelor's degree	0.1311	0.1970
Graduate or professional degree	0.0722	0.1229

Figure 4

Educational Attainment in 1990 versus 2017

- (b) The relative frequency of adults who are not high school graduates in 2017 is about half that of 1990. In 2017, a much higher proportion of the adult population has at least a bachelor's degree. However, the proportion of the population with a bachelor's degree has not doubled (as the frequencies in Table 4 might suggest). An overall conclusion is that adult Americans are more educated in 2017 than they were in 1990.

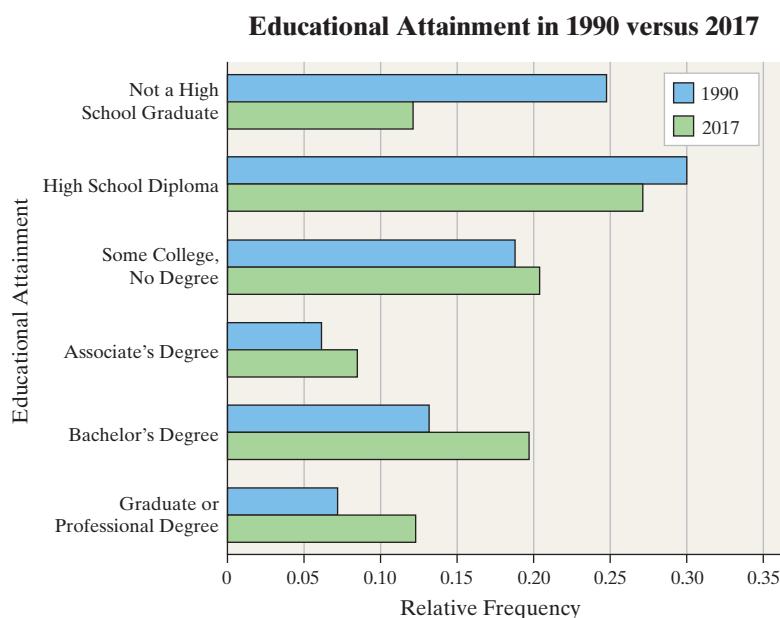
NW Now Work Problem 17



Horizontal Bars

Bar graphs may also be drawn with horizontal bars. Horizontal bars are preferable when category names are lengthy. For example, Figure 5 on the next page uses horizontal bars to display the same data as in Figure 4.

Figure 5



③ Construct Pie Charts

Pie charts are typically used to present the relative frequency of qualitative data. In most cases the data are nominal, but ordinal data can also be displayed in a pie chart.

Definition

A **pie chart** is a circle divided into sectors. Each sector represents a category of data. The area of each sector is proportional to the frequency of the category.

EXAMPLE 6

Constructing a Pie Chart

Table 6

Educational Attainment	Frequency
Not a high school graduate	26,582
High school diploma	60,032
Some college, no degree	45,110
Associate's degree	18,761
Bachelor's degree	43,585
Graduate or professional degree	27,181
Total	221,251

Using Technology

Most statistical spreadsheets are capable of drawing pie charts. See the Technology Step-by-Step on pages 69–70 for instructions on drawing pie charts using Minitab, Excel, and StatCrunch. The TI-83 and TI-84 Plus graphing calculators do not draw pie charts.

Problem The data presented in Table 6 represent the educational attainment of residents of the United States 25 years or older in 2017, based on data obtained from the U.S. Census Bureau. The data are in thousands. Construct a pie chart of the data.

Approach The pie chart will have one part or sector corresponding to each category of data. The area of each sector is proportional to the frequency of each category. For example, from Table 5, the proportion of all U.S. residents 25 years or older who are not high school graduates is 0.1201. The category “not a high school graduate” will make up 12.01% of the area of the pie chart. Since a circle has 360 degrees, the degree measure of the sector for this category will be $(0.1201)360^\circ \approx 43^\circ$. Use a protractor to measure each angle.

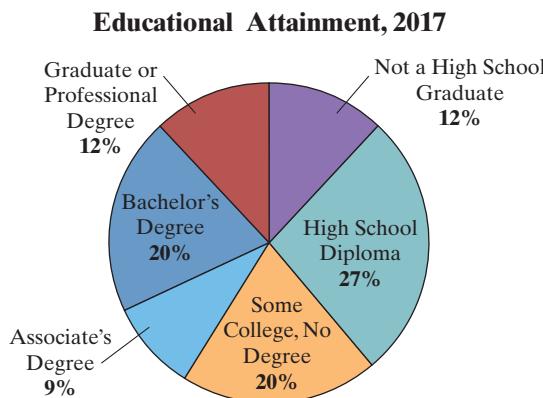
Solution Use the same approach for the remaining categories to obtain Table 7.

Table 7

Educational Attainment	Frequency	Relative Frequency	Degree Measure of Each Sector
Not a high school graduate	26,582	0.1201	43
High school diploma	60,032	0.2713	98
Some college, no degree	45,110	0.2039	73
Associate's degree	18,761	0.0848	31
Bachelor's degree	43,585	0.1970	71
Graduate or professional degree	27,181	0.1229	44

To construct a pie chart by hand, use a protractor to approximate the angles for each sector. See Figure 6.

Figure 6



NW Now Work Problem 21(e)

To make a pie chart, we need all the categories of the variable under consideration. For example, using Example 1, we could create a bar graph that lists the proportion of patients requiring back, shoulder, or knee rehabilitation, but it would not make sense to construct a pie chart for this situation. Do you see why? Only 70% of the data would be represented.

When should a bar graph or a pie chart be used? Pie charts are useful for showing the division of all possible values of a qualitative variable into its parts. However, because angles are often hard to judge in pie charts, they are not as useful in comparing two specific values of the qualitative variable. Instead the emphasis is on comparing the part to the whole. Bar graphs are useful when we want to compare the different parts, not necessarily the parts to the whole. For example, to get the “big picture” regarding educational attainment in 2017, a pie chart is a good visual summary. However, to compare bachelor’s degrees to high school diplomas, a bar graph is a better visual summary. Since bars are easier to draw and compare, some practitioners forgo pie charts in favor of Pareto charts when comparing parts to the whole.

Technology Step-by-Step

Drawing Bar Graphs and Pie Charts

TI-83/84 Plus

The TI-83 or TI-84 Plus does not have the ability to draw bar graphs or pie charts.

Minitab

Frequency or Relative Frequency Distributions from Raw Data

1. Enter the raw data in C1.
2. Select **Stat** and highlight **Tables** and select **Tally Individual Variables** . . .
3. Fill in the window with appropriate values. In the “Variables” box, enter C1. Check “counts” for a frequency distribution and/or “percents” for a relative frequency distribution. Click OK.

Bar Graphs from Summarized Data

1. Enter the categories in C1 and the frequency or relative frequency in C2.
2. Select **Graph** and then **Bar Chart** . . .
3. In the “Bars represent” pull-down menu, select “Values from a table” and highlight “Simple.” Press OK.
4. Fill in the window with the appropriate values. In the “Graph variables” box, enter C2. In the “Categorical variable” box, enter C1. By pressing Labels, you can add a title to the graph. Click OK to obtain the bar graph.

Bar Graphs from Raw Data

1. Enter the raw data in C1.
2. Select **Graph** and then **Bar Chart** . . .
3. In the “Bars represent” pull-down menu, select “Counts of unique values” and highlight “Simple.” Press OK.
4. Fill in the window with the appropriate values. In the “Categorical variables” box, enter C1. By pressing Labels, you can add a title to the graph. Click OK to obtain the bar graph.

Pie Chart from Raw or Summarized Data

1. If the data are summarized in a table, enter the categories in C1 and the frequency or relative frequency in C2. If the data are raw, enter the data in C1.

2. Select **Graph** and then **Pie Chart** . . .

3. Fill in the window with the appropriate values. If the data are summarized, click the “Chart values from a table” radio button; if the data are raw, click the “Chart counts of unique values” radio button. For summarized data, enter C1 in the “Categorical variables” box and C2 in the “Summary variables” box. If the data are raw, enter C1 in the “Categorical variables” box. By pressing Labels, you can add a title to the graph. Click OK to obtain the pie chart.

Excel

Bar Graphs from Summarized Data

1. Enter the categories in column A and the frequency or relative frequency in column B.
2. Highlight the data to be graphed.
3. Select the Insert menu. Click the “column” or “bar” chart type. Select the chart type in the upper-left corner.
4. Click the “+” to enter axes labels and chart title.

Bar Graphs from Raw Data Using XLSTAT

1. Load the XLSTAT plug-in.
2. Enter the data into a spreadsheet.
3. Select XLSTAT > Describing Data > Descriptive Statistics.
4. General tab: Fill out the box as follows:
 - Qualitative data: Check the box and highlight the data cell range to be analyzed (such as A1:A40).
 - Range/Sheet/Workbook: Select the Sheet option.
 - Sample labels: Check the box if the first row of data contains a label.
5. Options tab: Check the Charts box.
6. Outputs tab: Under “Qualitative Data” select Categories, Frequency per category, and Rel. frequency per category (%).
7. Charts (2) tab: Check the box for Bar charts. Under “Values used” choose either Frequencies or Relative frequencies, whichever is preferred. Click OK.
8. Manipulate the chart options (title, axes titles, colors, etc.) as desired.

Pie Charts from Summarized Data

1. Enter the categories in column A and the frequencies in column B.
2. Highlight the data to be graphed.
3. Select the Insert menu and click the “pie” chart type. Select the pie chart in the upper-left corner.
4. Click the “+” to enter labels, chart title, and legend.

Pie Chart from Raw Data

Follow the same steps as were given in Bar Graphs from Raw Data, except check the box for Pie charts rather than Bar charts.

StatCrunch

Frequency or Relative Frequency Distributions from Raw Data

1. If necessary, enter the raw data into the spreadsheet. Name the column.
2. Select **Stat**, highlight **Tables**, and select **Frequency**.
3. Click on the variable you wish to summarize. Click the type of table you want. If you want both Frequency and Relative frequency, highlight Frequency, then press Ctrl (or Command on an Apple) and select Relative frequency. Click Compute!.

Bar Graphs from Summarized Data

1. If necessary, enter the summarized data into the spreadsheet. Name the variable and frequency (or relative frequency) columns.
2. Select **Graph**, highlight **Bar Plot**, then select **With Summary**.
3. Select the “Categories in:” variable and “Counts in:” variable. Choose the type of bar graph (frequency or relative frequency). Enter labels for the X-axis and Y-axis. Enter a title for the graph. Click Compute!.

Bar Graphs from Raw Data

1. If necessary, enter the raw data into the spreadsheet. Name the column variable.

2. Select **Graph**, highlight **Bar Plot**, then highlight **With Data**.
3. Click on the column name of the variable you wish to summarize. Leave the grouping option as “Split bars.” Choose the type of bar graph (frequency or relative frequency). Enter labels for the X-axis and Y-axis. Enter a title for the graph. Click Compute!.

Side-by-Side Bar Graphs from Summarized Data

1. If necessary, enter the summarized data into the spreadsheet. Name the columns.
2. Select **Graph**, highlight **Chart**, then select **Columns**.
3. Select the column variables that contain the frequency or relative frequency of each category. Select the column of the variable that has the row labels. Choose the display you would like (vertical or horizontal split bars). Click Compute!.

Pie Chart from Summarized Data

1. If necessary, enter the summarized data into the spreadsheet. Name the column.
2. Select **Graph**, highlight **Pie Chart**, then select **With Summary**.
3. Select the “Categories in:” variable and “Counts in:” variable. Choose the display you would like. Enter a title for the graph. Click Compute!.

Pie Chart from Raw Data

1. If necessary, enter the raw data into the spreadsheet. Name the column variable.
2. Select **Graph**, highlight **Pie Chart**, then select **With Data**.
3. Click on the column name of the variable you wish to summarize. Choose the display you would like. Enter a title for the graph. Click Compute!.



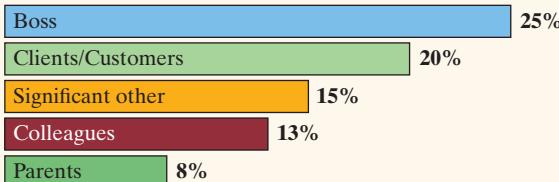
2.1 Assess Your Understanding

Vocabulary and Skill Building

1. Define raw data in your own words.
2. A frequency distribution lists the _____ of occurrences of each category of data, while a relative frequency distribution lists the _____ of occurrences of each category of data.
3. In a relative frequency distribution, what should the relative frequencies add up to?
4. What is a bar graph? What is a Pareto chart?
5. **Selling Yourself** This *USA Today*-type chart shows the top responses to the question, “Who’s the most difficult to ‘sell yourself’ to?”

Who’s the Most Difficult to “Sell Yourself” to?

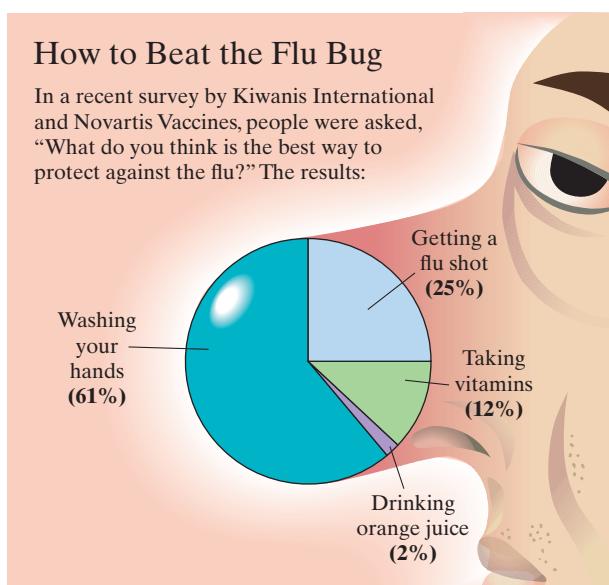
Top Responses:



Source: Sandler Training Survey of 1,053 adults.

- (a) What was the most common response?
 (b) The survey is of 1053 adults. How many stated their significant other is most difficult to “sell yourself” to?
 (c) Explain why this graphic cannot be displayed as a pie chart.

6. Flu Season The pie chart shown, the type we see in *USA Today*, depicts the approaches people use to avoid getting the flu.

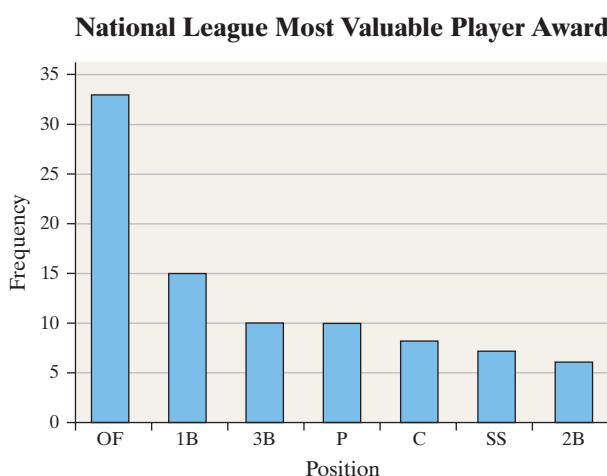


Source: Kiwanis International and Novartis Vaccines.

- (a) What is the most common approach? What percentage of the population chooses this method?
 (b) What is the least used approach? What percentage of the population chooses this method?
 (c) What percentage of the population thinks flu shots are the best way to beat the flu?

7. Most Valuable Player The following Pareto chart shows the position played by the most valuable player (MVP) in the National League since 1931.

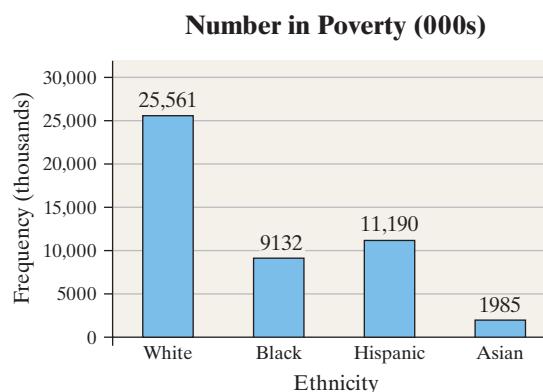
Source: <http://www.baseball-almanac.com/>



- (a) Which position had the most MVPs?
 (b) How many MVPs played first base (1B)?

- (c) How many more MVPs played first base (1B) than third base (3B)?
 (d) There are three outfield (OF) positions (left field, center field, right field). Given this, how might the graph be misleading?

8. Poverty The U.S. Census Bureau uses money income thresholds to define poverty. For example, in 2018 the poverty threshold for a family of four with two children was \$25,100. The bar graph represents the number of people living in poverty (in thousands) in the United States in 2017, by ethnicity.

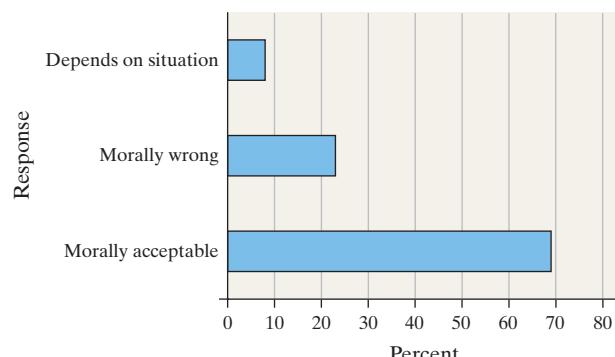


Source: U.S. Census Bureau.

- (a) How many whites were living in poverty in 2017?
 (b) Of the impoverished listed, what percent were Hispanic?
 (c) How might this graph be misleading?

9. Divorce The following graph represents the results of a survey, in which a random sample of adult Americans was asked, “Please tell me whether you personally believe that in general divorce is morally acceptable or morally wrong.”

Opinion Regarding Divorce

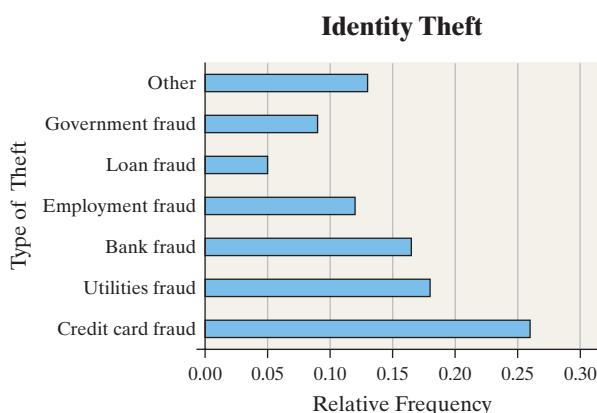


Source: Gallup.

- (a) What percent of the respondents believe divorce is morally acceptable?
 (b) If there are 240 million adults in America, how many believe that divorce is morally wrong?
 (c) If Gallup claimed that the results of the survey indicate that 8% of adult Americans believe that divorce is acceptable in certain situations, would you say this statement is descriptive or inferential? Why?

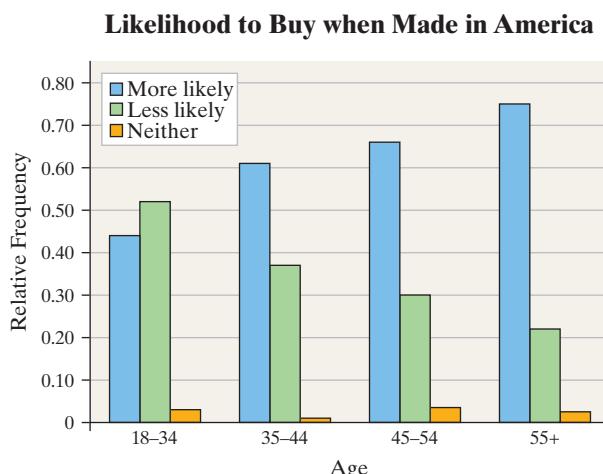
10. Identity Theft Identity fraud occurs when someone else’s personal information is used to open credit card accounts, apply for a job, receive benefits, and so on. The following

relative frequency bar graph represents the various types of identity theft based on a study conducted by the Federal Trade Commission.



Source: Federal Trade Commission.

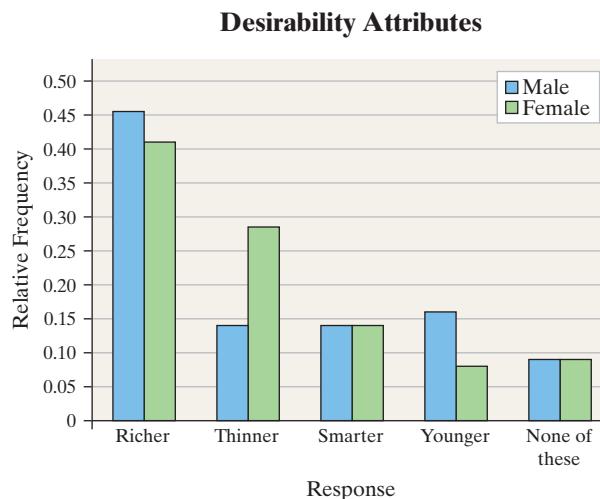
- (a) Approximately what percentage of identity theft was loan fraud (such as applying for a loan in someone else's name)?
 - (b) If there were 10 million cases of identity fraud in a recent year, how many were credit card fraud (someone uses someone else's credit card to make a purchase)?
- 11. Made in America** A random sample of 2163 adults (aged 18 and over) was asked, "When you see an ad emphasizing that a product is 'Made in America,' are you more likely to buy it, less likely to buy it, or neither more nor less likely to buy it?" The results of the survey are presented in the side-by-side bar graph.



Source: Harris Interactive.

- (a) What proportion of 18- to 34-year-old respondents are more likely to buy when made in America? What proportion of 35- to 44-year-old respondents are more likely to buy when made in America?
- (b) Which age group has the greatest proportion who are more likely to buy when made in America?
- (c) Which age group has a majority of respondents who are less likely to buy when made in America?
- (d) What is the apparent association between age and likelihood to buy when made in America?

- 12. Desirability Attributes** A random sample of 2163 adults (aged 18 and over) was asked, "Given a choice of the following, which one would you most want to be?" The results of the survey are presented in the side-by-side bar graph.



Source: Harris Interactive.

- (a) What proportion of males would like to be richer? What proportion of females would like to be richer?
- (b) Which attribute do females desire more than males?
- (c) Which attribute do males prefer over females two-to-one?
- (d) Which attribute do males and females desire in equal proportion?

Applying the Concepts

- 13. College Survey** In a national survey conducted by the Centers for Disease Control to determine health-risk behaviors among college students, college students were asked, "How often do you wear a seat belt when riding in a car driven by someone else?" The frequencies were as follows:

Response	Frequency
Never	125
Rarely	324
Sometimes	552
Most of the time	1257
Always	2518

- (a) Construct a relative frequency distribution.
 - (b) What percentage of respondents answered "Always"?
 - (c) What percentage of respondents answered "Never" or "Rarely"?
 - (d) Construct a frequency bar graph.
 - (e) Construct a relative frequency bar graph.
 - (f) Construct a pie chart.
 - (g) Suppose that a representative from the Centers for Disease Control says, "52.7% of college students surveyed always wear a seat belt." Is this a descriptive or inferential statement?
- 14. College Survey** In a national survey conducted by the Centers for Disease Control to determine health-risk behaviors among college students, college students were asked, "How

often do you wear a seat belt when driving a car?" The frequencies were as follows:

Response	Frequency
I do not drive a car	249
Never	118
Rarely	249
Sometimes	345
Most of the time	716
Always	3093

- (a) Construct a relative frequency distribution.
- (b) What percentage of respondents answered "Always"?
- (c) What percentage of respondents answered "Never" or "Rarely"?
- (d) Construct a frequency bar graph.
- (e) Construct a relative frequency bar graph.
- (f) Construct a pie chart.
- (g) Compute the relative frequencies of "Never," "Rarely," "Sometimes," "Most of the time," and "Always," excluding those that do not drive. Compare with those in Problem 13. What might you conclude?
- (h) Suppose that a representative from the Centers for Disease Control says, "2.5% of the college students in this survey responded that they never wear a seat belt." Is this a descriptive or inferential statement?

DATA 15. Use the Internet? The Gallup organization conducted a survey in which 1025 randomly sampled adult Americans were asked, "How much time, if at all, do you personally spend using the Internet—more than 1 hour a day, up to 1 hour a day, a few times a week, a few times a month or less, or never?" The results of the survey were as follows:

Response	Frequency
More than 1 hour a day	377
Up to 1 hour a day	192
A few times a week	132
A few times a month or less	81
Never	243

- (a) Construct a relative frequency distribution.
- (b) What proportion of those surveyed never use the Internet?
- (c) Construct a frequency bar graph.
- (d) Construct a relative frequency bar graph.
- (e) Construct a pie chart.
- (f) A local news broadcast reported that 37% of adult Americans use the Internet more than 1 hour a day. What is wrong with this statement?

DATA 16. Dining Out A sample of 521 adults was asked, "How often do you dine out?" The results of the survey are given in the table in the next column.

- (a) Construct a relative frequency distribution.
- (b) What proportion of those surveyed dine out once or twice a week?
- (c) Construct a frequency bar graph.
- (d) Construct a relative frequency bar graph.

Response	Frequency
Several times a week	103
Once or twice a week	204
A few times a month	130
Very rarely	79
Never	5

Source: www.amylovesit.com

NW 17. Online Shopping A random sample of college students was asked, "What social platform most influences your online shopping?" Results are shown below.

Platform	Males	Females
Facebook	326	222
Instagram	380	535
Snapchat	56	29
Twitter	122	57
None	678	271

Source: www.Whatsgoodly.com

- (a) Construct a relative frequency distribution for males.
- (b) Construct a relative frequency distribution for females.
- (c) Construct a side-by-side relative frequency bar graph.
- (d) Compare the online shopping habits of males and females.

DATA 18. Texting A survey of U.S. adults and teens (ages 12–17) was administered by Pew Research, to determine the number of texts sent in a single day.

Number of Texts	Adults	Teens
None	173	13
1–10	978	138
11–20	249	69
21–50	249	113
51–100	134	113
101+	153	181

Source: Pew Internet.

- (a) Construct a relative frequency distribution for adults.
- (b) Construct a relative frequency distribution for teens.
- (c) Construct a side-by-side relative frequency bar graph.
- (d) Compare the texting habits of adults and teens.

DATA 19. Dream Job A survey of adult men and women asked, "Which one of the following jobs would you most like to have?" The results of the survey are shown in the table.

Response	Men	Women
Professional athlete	40	18
Actor/actress	26	37
President of the United States	13	13
Rock star	13	13
Not sure	7	19

Source: Marist Poll.

- (a) Construct a relative frequency distribution for men and women.
- (b) Construct a side-by-side relative frequency bar graph.
- (c) What are the apparent differences in gender as it pertains to this question?

- DATA** **20. Car Color** A survey of 100 randomly selected autos in the luxury car segment and 100 randomly selected autos in the sports car segment that were recently purchased yielded the following colors.

Color	Number of Luxury Cars	Number of Sports Cars
White	25	10
Black	22	15
Silver	16	18
Gray	12	15
Blue	7	13
Red	7	15
Gold	6	5
Green	3	2
Brown	2	7

Source: Based on results from www.infoplease.com

- (a) Construct a relative frequency distribution for each car type.
- (b) Draw a side-by-side relative frequency bar graph.
- (c) Compare the colors for the two car types. Make a conjecture about the reasons for the differences.

- NW 21. Walt Disney Stock** The table shows the movement of Walt Disney stock for 30 randomly selected trading days. “Up” means the stock price increased in value for the day, “Down” means the stock price decreased in value for the day, and “No Change” means the stock price closed at the same price it closed for the previous day.

Down	Up	Up	Down	Down	Up
Down	Up	Down	Up	Down	Up
Down	Down	Up	Up	Up	Up
Down	Down	Down	Up	Down	Up
No Change	Up	Down	Down	No Change	Down

Source: Yahoo!Finance.

- (a) Construct a frequency distribution.
- (b) Construct a relative frequency distribution.
- (c) Construct a frequency bar graph.
- (d) Construct a relative frequency bar graph.
- (e) Construct a pie chart.

- DATA 22. Favorite Day to Eat Out** A survey was conducted by Wakefield Research in which participants were asked to disclose their favorite night to order takeout for dinner. The following data are based on their results.

Thursday	Saturday	Friday	Friday	Sunday
Wednesday	Saturday	Friday	Tuesday	Friday
Saturday	Monday	Friday	Friday	Sunday
Friday	Tuesday	Wednesday	Saturday	Friday
Wednesday	Monday	Wednesday	Wednesday	Friday
Friday	Wednesday	Thursday	Tuesday	Friday
Tuesday	Saturday	Friday	Tuesday	Friday
Saturday	Saturday	Saturday	Sunday	Friday

Source: Based on results from Wakefield Research.

- (a) Construct a frequency distribution.
- (b) Construct a relative frequency distribution.

- (c) If you own a restaurant, which day would you purchase an advertisement in the local newspaper? Are there any days you would avoid purchasing advertising space?
- (d) Construct a frequency bar graph.
- (e) Construct a relative frequency bar graph.
- (f) Construct a pie chart.

- DATA 23. Online Groceries** The following data represent the day of the week an order was placed for groceries using the online grocery delivery service Instacart.

Wednesday	Saturday	Monday	Wednesday	Wednesday	Wednesday
Friday	Saturday	Saturday	Tuesday	Tuesday	Monday
Thursday	Tuesday	Sunday	Friday	Tuesday	Sunday
Friday	Monday	Friday	Thursday	Saturday	Saturday
Tuesday	Monday	Wednesday	Monday	Tuesday	Sunday
Saturday	Sunday	Sunday	Saturday	Sunday	Wednesday
Thursday	Tuesday	Saturday	Sunday	Sunday	Saturday

Source: www.instacart.com

- (a) Construct a frequency and relative frequency distribution.
- (b) If you own an Instacart franchise, what day would you want to have the most drivers available to deliver groceries?
- (c) Construct a relative frequency bar graph.
- (d) Construct a pie chart.

- DATA 24. Blood Type** A phlebotomist draws the blood of a random sample of 50 patients and determines their blood types as shown:

O	O	A	A	O
B	O	B	A	O
AB	B	A	B	AB
O	O	A	A	O
AB	O	A	B	A
O	A	A	O	A
O	A	O	AB	A
O	B	A	A	O
O	O	O	A	O
O	A	O	A	O

- (a) Construct a frequency distribution.
- (b) Construct a relative frequency distribution.
- (c) According to the data, which blood type is most common?
- (d) According to the data, which blood type is least common?
- (e) Use the results of the sample to conjecture the percentage of the population that has type O blood. Is this an example of descriptive or inferential statistics?
- (f) Contact a local hospital and ask them the percentage of the population that is blood type O. Why might the results differ?
- (g) Draw a frequency bar graph.
- (h) Draw a relative frequency bar graph.
- (i) Draw a pie chart.

- DATA 25. Threaded Problem: Tornado** The data set “Tornadoes_2017” located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017. F scale is a qualitative variable that categorizes tornadoes by their wind speed. The

table below shows the F scale rating from the National Oceanic and Atmospheric Administration.

F Scale	Wind Speed (mph)
F0	< 73
F1	73–112
F2	113–157
F3	158–206
F4	207–260
F5	261–318

- (a) Construct a frequency and relative frequency distribution of F scale. **Note:** An entry of –9 means the F scale was not reported for that tornado. How many tornadoes did not have their F scale reported?
- (b) Draw a relative frequency bar graph of F scale.
- (c) Construct a pie chart of F scale.
- (d) Create a graphical summary that will display the month that had the most tornadoes in 2017.
- (e) Which state had the most tornadoes in 2017?

26. Highest Elevation The following data represent the land area and highest elevation for each of the seven continents.

Continent	Land Area (square miles)	Highest Elevation (feet)
Africa	11,608,000	19,340
Antarctica	5,100,000	16,066
Asia	17,212,000	29,035
Australia	3,132,000	7,310
Europe	3,837,000	18,510
North America	9,449,000	20,320
South America	6,879,000	22,834

Source: www.infoplease.com

- (a) Would it make sense to draw a pie chart for land area? Why? If so, draw a pie chart.
- (b) Would it make sense to draw a pie chart for the highest elevation? Why? If so, draw a pie chart.

DATA 27. StatCrunch Survey Choose a qualitative variable from the Sullivan StatCrunch Survey I data set at www.pearsonhighered.com/sullivanstats and summarize the variable.

DATA 28. StatCrunch Survey Choose a qualitative variable from the Sullivan StatCrunch Survey I data set at www.pearsonhighered.com/sullivanstats and summarize the variable by gender.

29. Putting It Together: Online Homework Keeping students engaged in the learning process greatly increases their chance of success in a course. Traditional lecture-based math instruction has given way to a more student-engaged approach where students interact with the teacher in class and receive immediate feedback to their responses. The teacher presence allows students, when incorrect in a response, to be guided through a solution and then immediately be given a similar problem to attempt.

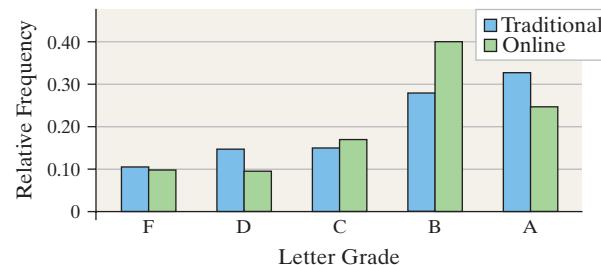
A researcher conducted a study to investigate whether an online homework system using an attempt—feedback—reattempt approach improved student learning over traditional pencil-and-paper homework. The online homework system was designed to increase student engagement outside class, something commonly missing in traditional pencil-and-paper assignments, ultimately leading to increased learning.

The study was conducted using two first-semester calculus classes taught by the researcher in a single semester. One class was assigned traditional homework and the other was assigned online homework that used the attempt—feedback—reattempt approach. The summaries are based on data from the study.

- (a) What is the research objective?
- (b) Is this study an observational study or experiment?
- (c) Give an example of how the researcher attempted to control variables in the study.
- (d) Explain why assigning homework type to entirely separate classes can confound the conclusions of the study.
- (e) For the data in the table, (i) identify the variables, (ii) indicate whether the variables are qualitative or quantitative, and (iii) for each quantitative variable, indicate whether the variable is discrete or continuous.

	Prior College Experience		No Prior College Experience	
	Traditional	Online	Traditional	Online
Number of students	10	9	23	18
Average age	22.8	19.4	18.1	18.1
Average exam score	84.5	68.9	79.4	80.6

Grades Earned on Exams (No Prior College Experience)



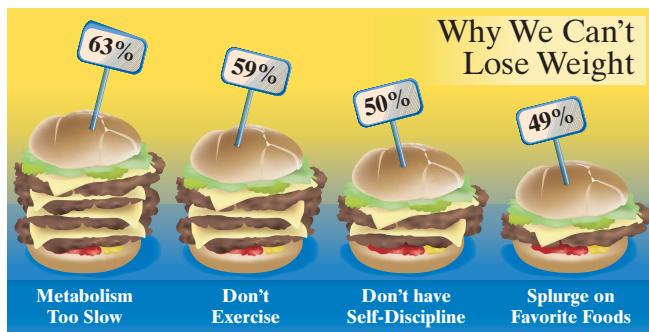
Source: *Journal of Computers in Mathematics and Science Teaching* 26(1):55–73, 2007.

- (f) What type of variable is letter grade? What level of measurement is letter grade? Do you think presenting the data in a table from A to F would be a better representation of the data than presenting it in a graph?
- (g) What type of graph is displayed?
- (h) Could the data in the graph be presented in a pie chart? If so, what is the “whole”? If not, why not?
- (i) Considering the students with no prior college experience, how might the table and the graph generate conflicting conclusions?

Explaining the Concepts

30. When should relative frequencies be used when comparing two data sets? Why?
31. Suppose you need to summarize ordinal data in a bar graph. How would you arrange the categories of data on the horizontal axis? Is it possible to make the order of the data apparent in a pie chart?
32. Describe the circumstances in which a bar graph is preferable to a pie chart. When is a pie chart preferred over a bar graph? Are there circumstances in which a pie chart cannot be drawn, but a bar graph could be drawn? What are these circumstances?

33. Consider the information in the chart shown below, which is in the *USA Today* style of graph. Could the information provided be organized into a pie chart? Why or why not?



2.2 Organizing Quantitative Data



Preparing for This Section Before getting started, review the following:

- Quantitative variable (Section 1.1, p. 7)
- Discrete variable (Section 1.1, pp. 7–9)
- Continuous variable (Section 1.1, pp. 7–9)

Objectives

- ① Organize discrete data in tables
- ② Construct histograms of discrete data
- ③ Organize continuous data in tables
- ④ Construct histograms of continuous data
- ⑤ Draw dot plots
- ⑥ Identify the shape of a distribution
- ⑦ Draw time-series graphs

In summarizing quantitative data, first determine whether the data are discrete or continuous. If the data are discrete with relatively few different values of the variable, then the categories of data (called **classes**) will be the observations (as in qualitative data). If the data are discrete, but with many different values of the variable or if the data are continuous, then categories of data (the *classes*) must be created using intervals of numbers. We will first present the techniques for organizing discrete quantitative data when there are relatively few different values and then proceed to organizing continuous quantitative data.

1 Organize Discrete Data in Tables

Use the values of the discrete variable to create the classes when the number of distinct data values is small.

EXAMPLE 1

Constructing Frequency and Relative Frequency Distributions from Discrete Data

Problem The manager of a Wendy's fast-food restaurant wants to know the typical number of customers who arrive during the lunch hour. The data in Table 8 represent the number of customers who arrive at Wendy's for 40 randomly selected 15-minute



intervals of time during lunch. For example, during one 15-minute interval, seven customers arrived. Construct a frequency and relative frequency distribution of the data.

Table 8**Number of Arrivals at Wendy's**

7	6	6	6	4	6	2	6
5	6	6	11	4	5	7	6
2	7	1	2	4	8	2	6
6	5	5	3	7	5	4	6
2	2	9	7	5	9	8	5

Approach The number of people arriving could be 0, 1, 2, 3, Table 8 shows there are 11 categories of data from this study: 1, 2, 3, . . . , 11. Tally the number of observations for each category, count each tally, and create the frequency and relative frequency distributions.

Solution The two distributions are shown in Table 9.

Table 9

Number of Customers	Tally	Frequency	Relative Frequency
1		1	$\frac{1}{40} = 0.025$
2		6	0.15
3		1	0.025
4		4	0.1
5		7	0.175
6		11	0.275
7		5	0.125
8		2	0.05
9		2	0.05
10		0	0.0
11		1	0.025

On the basis of the relative frequencies, 27.5% of the 15-minute intervals had 6 customers arrive at Wendy's during the lunch hour.

NW Now Work Problems 25(a)–(e)

② Construct Histograms of Discrete Data

The *histogram*, a graph used to present quantitative data, is similar to the bar graph.

Definition

A **histogram** is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same and the rectangles touch each other.

EXAMPLE 2

Drawing a Histogram for Discrete Data

Problem Construct a frequency histogram and a relative frequency histogram of the number of customers arriving at Wendy's using the data in Table 9.

(continued)

CAUTION!

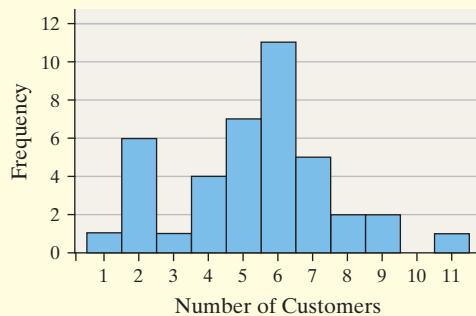
The rectangles in histograms touch, but the rectangles in bar graphs do not touch.

Approach The value of each category of data (number of customers) is on the horizontal axis and the frequency or relative frequency is on the vertical axis. Draw rectangles of equal width centered at the value of each category. For example, the first rectangle is centered at 1. For the frequency histogram, the height of the rectangle is the frequency of the category; for the relative frequency histogram, the height of the rectangle is the relative frequency of the category. Remember, the rectangles touch.

Solution Figures 7(a) and (b) show the frequency and relative frequency histograms, respectively.

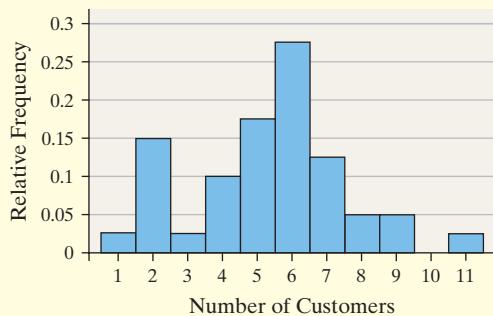
Figure 7

Arrivals at Wendy's



(a)

Arrivals at Wendy's



(b)

NW Now Work Problems 25(f)–(g)



3 Organize Continuous Data in Tables

Classes are categories into which data are grouped. When a data set consists of a large number of different discrete data values or when a data set consists of continuous data, we create classes by using intervals of numbers.

Table 10 is a typical frequency distribution created from continuous data. The data represent the number of mothers in the United States, ages 15–54, who had their fourth child in 2016.

Notice that the data are categorized, or grouped, by intervals of numbers. Each interval represents a class. For example, the first class is 15- to 19-year-old U.S. mothers who had their fourth child. We read this interval as follows: “The number of U.S. mothers, ages 15–19, who had their fourth child in 2016 was 443.” There are eight classes in the table, each with a **lower class limit** (the smallest value within the class) and an **upper class limit** (the largest value within the class). The lower class limit for the first class in Table 10 is 15; the upper class limit is 19. The **class width** is the difference between consecutive lower class limits. In Table 10 the class width is $20 - 15 = 5$.

The data in Table 10 are continuous. So the class 15–19 actually represents 15–19.999..., or 15 up to every value less than 20. Some will report the interval 15–19 as 15– < 20 (read, “fifteen to less than 20”).

Notice that the classes in Table 10 do not overlap. This is necessary to avoid confusion as to which class a data value belongs. Notice also that the class widths are equal for all classes.

One exception to the requirement of equal class widths occurs in open-ended tables. A table is **open ended** if the first class has no lower class limit or the last class has no upper class limit. The data in Table 11 represent the number of residents of the United States who have earned a Bachelor’s degree as of 2017. The last class in the table, “65 and over,” is open-ended.

Table 10

Age	Number
15–19	443
20–24	25,328
25–29	84,217
30–34	101,070
35–39	63,461
40–44	14,558
45–49	867
50–54	84

Source: National Vital Statistics Reports, Vol. 67, No. 1.

Table 11

Age	Number (000s)
25–44	30,758
45–64	25,972
65 and over	14,036

Source: U.S. Census Bureau.

EXAMPLE 3**Organizing Continuous Data into a Frequency and Relative Frequency Distribution****IN OTHER WORDS**

For qualitative and many discrete data, the classes are formed by using the data. For continuous data, the classes are formed by using intervals of numbers, such as 30–39.

CAUTION!

Watch out for tables with classes that overlap, such as a first class of 20–30 and a second class of 30–40.

NOTE

These data are given to the nearest penny, so the classes are also to the nearest penny. If data are expressed to a whole number (such as 211, 151, and so on), the classes would also be expressed to a whole number as in

50–74
75–99
100–124

Using Technology

Many technologies have a “sort” feature that makes tallying data by hand much easier.

Problem The data in Table 12 represent the total fine including late penalties, in dollars, for a simple random sample of 50 parking and camera violations in the City of New York. Construct a frequency and relative frequency distribution of the data.

Table 12**Parking and Camera Violation Fines**

211.09	209.29	150.56	271.20	210.20	125.76	190.72	187.23	229.18	125.00
105.00	256.69	262.99	256.70	322.61	243.80	236.91	260.00	162.46	65.00
227.59	224.35	223.99	155.40	193.42	151.46	127.37	140.06	216.19	210.25
105.23	125.21	208.40	208.39	207.79	124.85	124.67	148.02	206.29	183.16
147.70	206.01	123.41	204.49	134.66	167.65	121.25	131.76	120.53	143.52

Source: NYC OpenData.

Approach To construct a frequency distribution, first create classes of equal width. Table 12 has 50 observations that range from 65 to 322.61, so we decide to create the classes such that the lower class limit of the first class is 50 and the class width is 25. There is nothing magical about the choice of 25 as the class width. We could have selected a class width of 20, 50, or any other class width. Choose a class width that will nicely summarize the data. If the choice of class width does not accomplish this, try another. The second class has a lower class limit of $50 + 25 = 75$. The classes cannot overlap, so the upper class limit of the first class is 74.99. Continuing in this fashion, we obtain the following classes:

50–74.99
75–99.99
⋮
300–324.99

This represents 11 classes. Tally the number of observations in each class, count the tallies, and create the frequency distribution. Divide the frequency of each class by 50, the number of observations, to obtain the relative frequency.

Solution Tally the data as shown in the second column of Table 13. The third column shows the frequency of each class. From the frequency distribution, we conclude that a fine between \$200 and \$224.99 occurs with the most frequency. The fourth column shows the relative frequency of each class. So, 26% of the fines are between \$200 and \$224.99.

Table 13

Class (Amount of Fine)	Tally	Frequency	Relative Frequency
50–74.99		1	1/50 = 0.02
75–99.99		0	0/50 = 0
100–124.99		7	7/50 = 0.14
125–149.99		10	10/50 = 0.20
150–174.99		5	5/50 = 0.10
175–199.99		4	4/50 = 0.08
200–224.99		13	13/50 = 0.26
225–249.99		4	4/50 = 0.08
250–274.99		5	5/50 = 0.10
275–299.99		0	0/50 = 0
300–324.99		1	1/50 = 0.02

Historical Note

Florence Nightingale was born in Italy on May 12, 1820. She was named after the city of her birth. Nightingale was educated by her father, who attended Cambridge University. Between 1849 and 1851, she studied nursing throughout Europe. In 1854, she was asked to oversee the introduction of female nurses into the military hospitals in Turkey. While there, she greatly improved the mortality rate of wounded soldiers. She collected data and invented graphs (the polar area diagram), tables, and charts to show that improving sanitary conditions would lead to decreased mortality rates. In 1869, Nightingale founded the Nightingale School Home for Nurses. After a long and eventful life as a reformer of health care and contributor to graphics in statistics, Florence Nightingale died on August 13, 1910.

**IN OTHER WORDS**

Creating the classes for summarizing continuous data is an art form. There is no such thing as *the* correct frequency distribution. However, there can be less desirable frequency distributions. The larger the class width, the fewer classes a frequency distribution will have.

IN OTHER WORDS

Rounding *up* is different from rounding *off*. For example, 6.2 rounded *up* would be 7, while 6.2 rounded *off* would be 6.

Though formulas and procedures exist for creating frequency distributions from raw data, they do not necessarily provide better summaries. There is no one correct frequency distribution for a particular set of data. However, some frequency distributions can better illustrate patterns within the data than others. So constructing frequency distributions is somewhat of an art form. Use the distribution that seems to provide the best overall summary of the data.

The goal in constructing a frequency distribution is to reveal interesting features of the data, but we also typically want the number of classes to be between 5 and 20. When the data set is small, we usually want fewer classes. When the data set is large, we usually want more classes. Why do you think this is reasonable?

Although there is no “right” frequency distribution, there are bad ones. Use the following guidelines to help determine an appropriate lower class limit of the first class and class width.

Guidelines for Determining the Lower Class Limit of the First Class and Class Width**Choosing the Lower Class Limit of the First Class**

Choose the smallest observation in the data set or a convenient number slightly lower than the smallest observation in the data set. For example, in Table 12, the smallest observation is 65. A convenient lower class limit of the first class is 50 or 60.

Determining the Class Width

- Decide on the number of classes. Generally, there should be between 5 and 20 classes. The smaller the data set, the fewer classes you should have. For example, we chose 11 classes for the data in Table 12.
- Determine the class width by computing

$$\text{Class width} \approx \frac{\text{largest data value} - \text{smallest data value}}{\text{number of classes}}$$

Round this value *up* to a convenient number. For example, using the data in

Table 12, we obtain class width $\approx \frac{322.61 - 65}{11} = 23.419$. Use 25 as the class

width because this is an easy number to work with. Rounding up may result in fewer classes than were originally intended.

Applying these guidelines, to the fine data, we would end up with the frequency distribution shown in Table 13.

4

Construct Histograms of Continuous Data

We are now ready to draw histograms of continuous data.

EXAMPLE 4**Drawing a Histogram of Continuous Data**

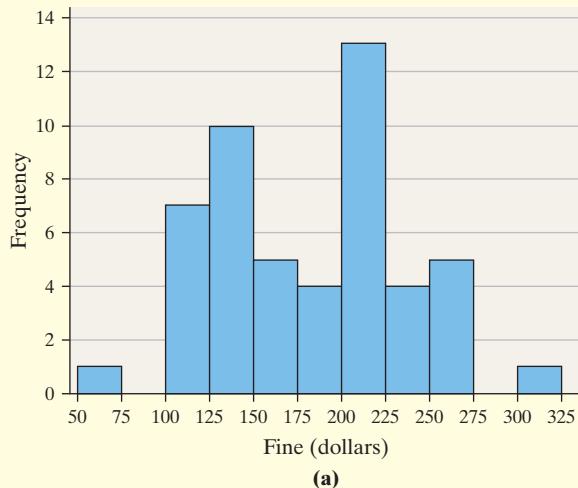
Problem Construct a frequency and relative frequency histogram of the fine data discussed in Example 3.

Approach To draw the frequency histogram, use the frequency distribution in Table 13. First, label the lower class limits of each class on the horizontal axis. Then, for each class, draw a rectangle whose width is the class width and whose height is the frequency. For the relative frequency histogram, the height of the rectangle is the relative frequency.

Solution Figures 8(a) and (b) show the frequency and relative frequency histograms, respectively.

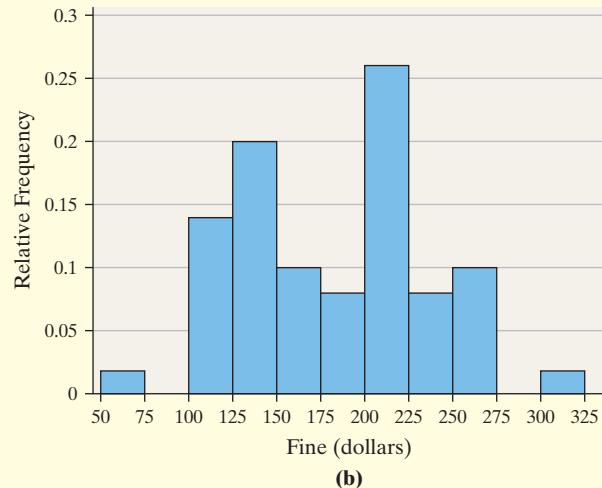
Figure 8

Fines for Parking and Camera Violations in New York City



(a)

Fines for Parking and Camera Violations in New York City



(b)

EXAMPLE 5 Drawing a Histogram for Continuous Data Using Technology

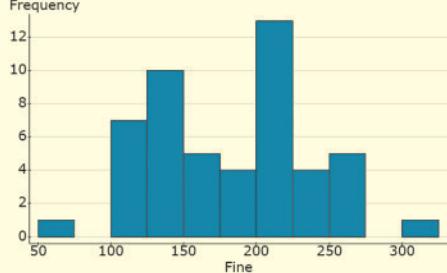
Problem Construct a frequency and relative frequency histogram of the fine data discussed in Example 3.

Approach We will use StatCrunch to construct the frequency and relative frequency histograms. The steps for constructing the graphs using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch, are given in the Technology Step-by-Step on pages 84–85.

Solution Figures 9(a) and (b) show the frequency and relative frequency histograms, respectively, obtained from StatCrunch.

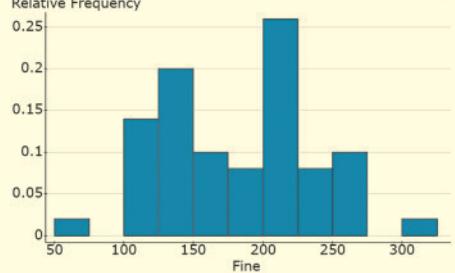
Figure 9

Fines for Parking and Camera Violations in New York City



(a)

Fines for Parking and Camera Violations in New York City



(b)

NW Now Work Problems 27(c)–(d)

Using technology to construct histograms is a convenient and efficient way to explore patterns in data using different class widths.

5 Draw Dot Plots

A **dot plot** is drawn by placing each observation horizontally in increasing order and placing a dot above the observation each time it is observed.

EXAMPLE 6

Drawing a Dot Plot

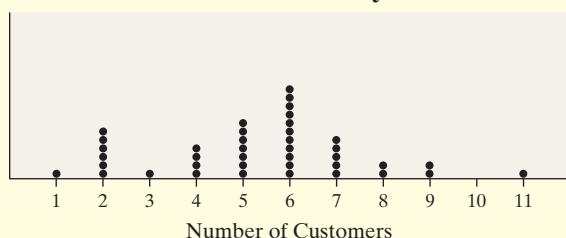
Problem Draw a dot plot for the number of arrivals at Wendy's data from Table 8 on page 77.

Approach The smallest observation in the data set is 1 and the largest is 11. Write the numbers 1 through 11 horizontally. For each observation, place a dot above the value of the observation.

Solution Figure 10 shows the dot plot.

Figure 10

Arrivals at Wendy's



NW Now Work Problem 35

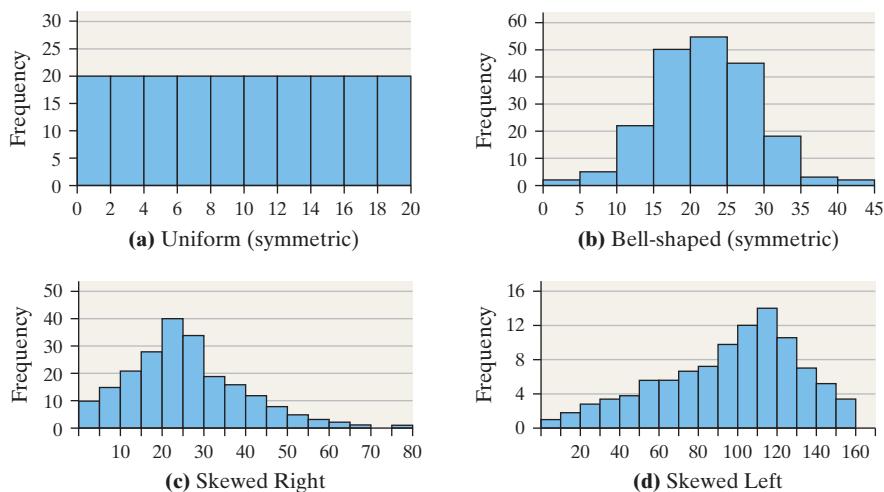


6 Identify the Shape of a Distribution

One way that a variable is described is through the shape of its distribution. Distribution shapes are typically classified as symmetric, skewed left, or skewed right. Figure 11 displays various histograms and the shape of the distribution.

Figures 11(a) and (b) show symmetric distributions. They are symmetric because, if we split the histogram down the middle, the right and left sides are almost mirror images. Figure 11(a) is a **uniform distribution** because the frequency of each value of the variable is evenly spread out across the values of the variable. Figure 11(b) displays a **bell-shaped distribution** because the highest frequency occurs in the middle and frequencies tail off to the left and right of the middle. That is, the graph looks like the profile of a bell. The distribution in Figure 11(c) is **skewed right**. Notice that the tail to the right of the peak is longer than the tail to the left of the peak. Finally, Figure 11(d) illustrates a distribution that is **skewed left**, because the tail to the left of the peak is longer than the tail to the right of the peak.

Figure 11



CAUTION!

We do not describe qualitative data as skewed left, skewed right, or uniform.

CAUTION!

It is important to recognize that data will not always exhibit behavior that perfectly matches any of the shapes given in Figure 11. To identify the shape of a distribution, some flexibility is required. In addition, people may disagree on the shape, since identifying shape is subjective.

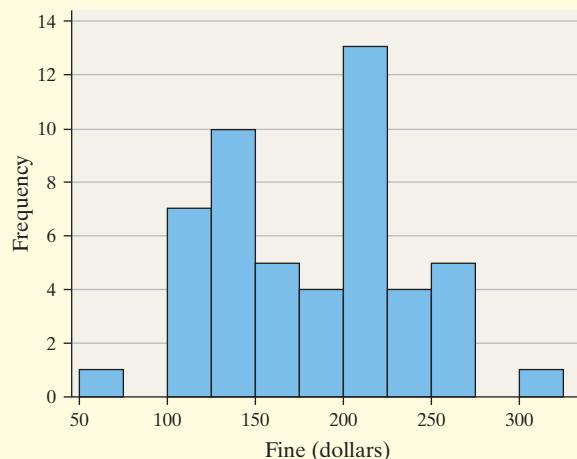
EXAMPLE 7 Identifying the Shape of a Distribution

Problem Figure 12 displays the histogram obtained in Example 4 for the fine paid for parking and camera violations in New York City. Describe the shape of the distribution.

Approach We compare the shape of the distribution displayed in Figure 12 with those in Figure 11.

Solution Since the histogram looks most like Figure 11(b), the distribution is fairly symmetric. However, it is not bell-shaped.

Figure 12

Fines for Parking and Camera Violations in New York City


NW Now Work Problem 27(e)

7 Draw Time-Series Graphs

If the value of a variable is measured at different points in time, the data are referred to as **time-series data**. The closing price of Cisco Systems stock at the end of each month for the past 12 years is an example of time-series data.

Definition

A **time-series plot** is obtained by plotting the time in which a variable is measured on the horizontal axis and the corresponding value of the variable on the vertical axis. Line segments are then drawn connecting the points.

Time-series plots are very useful in identifying trends in the data over time.

EXAMPLE 8 Drawing a Time-Series Plot

Problem The Partisan Conflict Index (PCI) tracks the degree of political disagreement among U.S. politicians in the federal government. It is found by measuring the frequency of newspaper articles reporting disagreement in a given month. Higher values of the index suggest greater conflict among political parties, Congress, and the President. The data in Table 14 on the following page represents the PCI in October from 2005 to 2019. Construct a time-series plot of the data. In what year was the index highest?

Table 14

Year	Partisan Conflict Index (PCI)
2005	74.39
2006	80.16
2007	88.88
2008	93.00
2009	90.54
2010	147.52
2011	129.75
2012	147.55
2013	252.10
2014	137.58
2015	146.35
2016	155.96
2017	175.36
2018	163.96
2019	109.02

Source: Federal Reserve Bank of Philadelphia

Using Technology

Statistical spreadsheets, such as StatCrunch, Excel, or Minitab, and certain graphing calculators, such as the TI-83 or TI-84 Plus, can create time-series graphs.

NW Now Work Problem 37

Approach

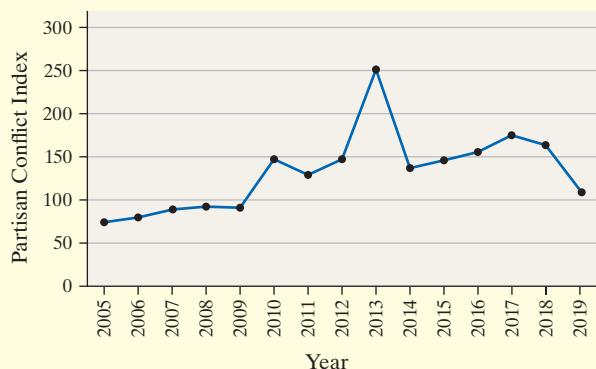
Step 1 Plot points for each year, with the date on the horizontal axis and the Partisan Conflict Index on the vertical axis.

Step 2 Connect the points with line segments.

Solution Figure 13 shows the time-series plot. Notice the jump in the PCI just prior to the November, 2010 midterm elections and the huge spike in October, 2013 the year the PCI was highest. During October, 2013, the federal government shut down because Congress could not agree on a budget, largely due to a standoff over the Affordable Care Act (aka Obamacare).

Figure 13

Partisan Conflict Index in the United States Federal Government



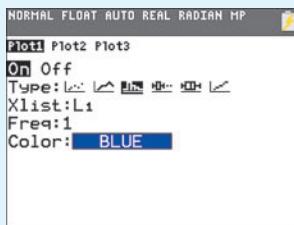
Technology Step-by-Step

Drawing Histograms, Dot Plots, and Time-Series Plots

TI-83/84 Plus

Histograms

- Enter the raw data in L1 by pressing STAT and selecting 1: Edit.
- Press 2nd Y = to access the StatPlot menu. Select 1: Plot1.
- Place the cursor on “ON” and press ENTER.
- Place the cursor on the histogram icon (see the figure) and press ENTER. Press 2nd QUIT to exit the Plot1 menu.



- Press WINDOW. Set Xmin to the lower class limit of the first class. Set Xmax to the lower class limit of the

class following the class containing the largest value. For example, if the first class is 0–9, set Xmin to 0. If the class width is 10 and the last class is 90–99, set Xmax to 100. Set Xscl to the class width. Set Ymin to 0. Set Ymax to a value larger than the frequency of the class with the highest frequency.

- Press GRAPH.

Helpful Hints: To determine each class frequency, press TRACE and use the arrow keys to scroll through each class. If you decrease the value of Ymin to a value such as –5, you can see the values displayed on the screen easier. The TI graphing calculators do not draw dot plots.

Time-Series Plots

- Enter the values for the x-axis in L1 and the values for the y-axis in L2 by pressing STAT and selecting 1: Edit.
- Press 2nd Y = to access the StatPlot menu. Select 1: Plot1.
- Place the cursor on “ON” and press ENTER.
- Place the cursor on the line plot icon and press ENTER. Press 2nd QUIT to exit the Plot1 menu.

5. Press WINDOW. Set Xmin to the smallest value in L1 and set Xmax to the largest value in L1. Set Ymin to the smallest value in L2 and set Ymax to the largest value in L2. Press GRAPH.

Minitab

Histograms

1. Enter the raw data in C1.
2. Select the **Graph** menu and then **Histogram . . .**
3. Highlight the “simple” icon and press OK.
4. Put the cursor in the “Graph variables” box. Highlight C1 and press Select. Click SCALE and select the Y-Scale Type tab. For a frequency histogram, click the frequency radio button. For a relative frequency histogram, click the percent radio button. Click OK twice.

Note: To adjust the class width and to change the labels on the horizontal axis to the lower class limit, double-click inside one of the bars in the histogram. Select the “binning” tab in the window that opens. Click the cutpoint button and the midpoint/cutpoint positions radio button. In the midpoint/cutpoint box, enter the lower class limits of each class. Click OK.

Dot Plots

1. Enter the raw data in C1.
2. Select the **Graph** menu and then **Dotplot . . .**
3. Highlight the “simple” icon and press OK.
4. Put the cursor in the “Graph variables” box. Highlight C1 and press Select. Click OK.

Time-Series Plots

1. Enter the values for the *x*-axis in C1 and the values for the *y*-axis in C2.
2. Select the **Graph** menu and then **Scatterplot . . .**
3. Choose the “With Connect Line” plot. Click OK.
4. Choose the Y variable from C2 and the X variable from C1. Click OK.

Excel

Histograms

1. Load the XLSTAT Add-in.
2. Enter the raw data in column A.
3. Select XLSTAT. Click Describing data, then select Histograms.
4. With the cursor in the Data cell, highlight the data in Column A.
5. Click either the Continuous or Discrete radio button.
6. Click the Options tab. Decide on either a certain number of intervals or enter your own lower class limits. To enter your own intervals, enter the lower class limits in Column B.
7. Click the Charts tab. Choose either Frequency or Relative Frequency. Click OK.

Time-Series Plots

1. Enter the values for the *x*-axis in column A and the values for the *y*-axis in column B.
2. Highlight the column containing the *y*-values. Select Insert > Line Chart Charts > Line with Markers.
3. Select Design > Select Data. In the Select Data Source dialog box, select Horizontal (Category) Axis Labels | Edit.

4. In the Axis Labels dialog box highlight the *x*-values click OK.
5. In the Select Data Source dialog box click OK.

Manipulate the chart options (title, axes titles, colors, etc.) as desired.

StatCrunch

Frequency and Relative Frequency Distributions of Discrete Data

1. Enter the raw data into the spreadsheet. Name the column variable.
2. Select **Stat**, highlight **Tables**, and select **Frequency**.
3. Click on the variable you wish to summarize. Click the Type of table you want. If you want both Frequency and Relative frequency, highlight Frequency, then press Ctrl (or Command on an Apple) and select Relative frequency. Click Compute!.

Frequency and Relative Frequency Distributions of Continuous Data

1. If necessary, enter the raw data into the spreadsheet. Name the column variable.
2. Select **Data** and then **Bin**.
3. Click the variable you wish to summarize. Click the “Use fixed width bins” radio button. Enter the lower class limit of the first class in the “Start at:” cell. Enter the class width in the “Bin width:” cell. Leave the “Include left endpoint” radio button selected. Click Compute!.
4. Select **Stat** and highlight **Tables**, then **Frequency**.
5. Choose the Bin(column name) variable. Under Type:, select Frequency and Relative frequency. Click Compute!.

Histograms

1. If necessary, enter the raw data into the spreadsheet. Name the column variable.
2. Select **Graph** and then **Histogram**.
3. Click on the variable you wish to summarize. Choose the type of histogram (frequency or relative frequency). You have the option of choosing a lower class limit for the first class by entering a value in the cell marked “Bins: Start at:” You have the option of choosing a class width by entering a value in the cell marked “Bins: Width:” Enter labels for the X-axis and Y-axis. Enter a title for the graph. Click Compute!.

Dot Plots

1. If necessary, enter the raw data into the spreadsheet. Name the column variable.
2. Select **Graph** and then **Dotplot**.
3. Click on the variable you wish to summarize. Enter labels for the X-axis and Y-axis. Enter a title for the graph. Click Compute!.

Time-Series Plots

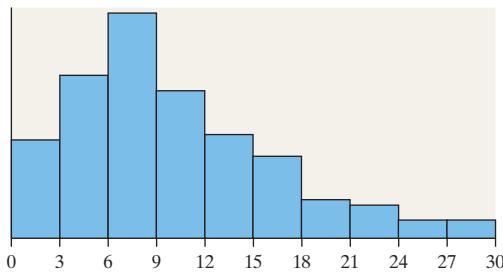
1. Enter the values for the *x*-axis in var1 and the values for the *y*-axis in var2. Title the columns.
2. Select **Graph** and highlight **Scatter Plot**.
3. The values in var1 are the X variable. The values in var2 are the Y variable. Highlight “Points and Lines” in the Display window. Label the axes and title the graph. Click Compute!.



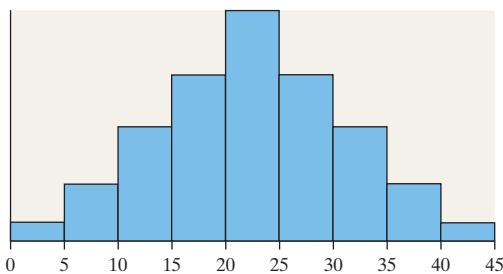
2.2 Assess Your Understanding

Vocabulary and Skill Building

1. The categories by which data are grouped are called _____.
2. The _____ class limit is the smallest value within the class and the _____ class limit is the largest value within the class.
3. The _____ is the difference between consecutive lower class limits.
4. What does it mean if a distribution is said to be “skewed left”?
5. *True or False:* There is not one particular frequency distribution that is correct, but there are frequency distributions that are less desirable than others.
6. *True or False:* Suppose the first class of a frequency distribution is 0–9.9 and the second class is 10–19.9. Then, the class width is 9.9.
7. *True or False:* The shape of the distribution shown is best classified as skewed left.

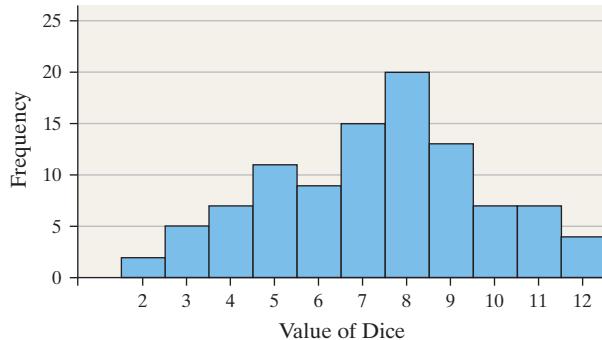


8. *True or False:* The shape of the distribution shown is best classified as uniform.



9. **Rolling the Dice** An experiment was conducted in which two fair dice were thrown 100 times. The sum of the pips showing on the dice was then recorded. The following frequency histogram gives the results.

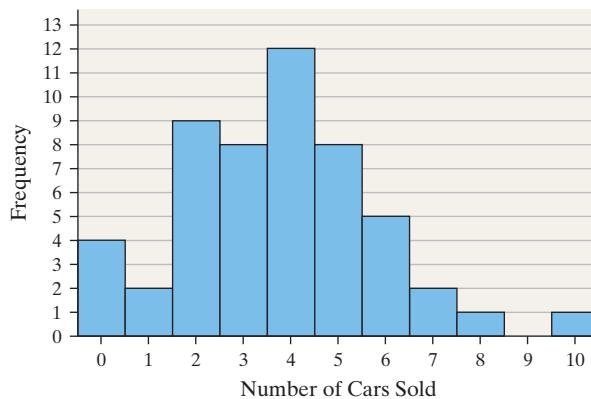
Sum of Two Dice



- (a) What was the most frequent outcome of the experiment?
- (b) What was the least frequent?
- (c) How many times did we observe a 7?
- (d) How many more 5's were observed than 4's?
- (e) Determine the percentage of time a 7 was observed.
- (f) Describe the shape of the distribution.

10. **Car Sales** A car salesman records the number of cars he sold each week for the past year. The following frequency histogram shows the results.

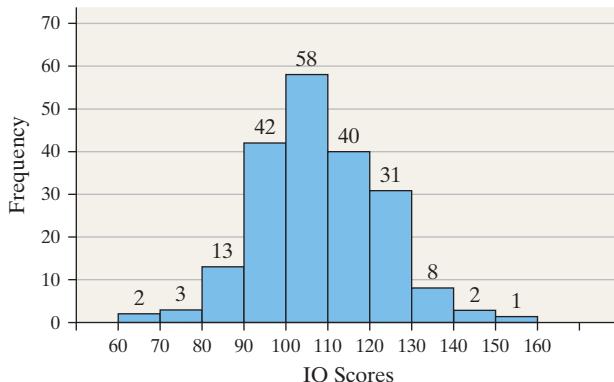
Cars Sold per Week



- (a) What is the most frequent number of cars sold in a week?
- (b) For how many weeks were two cars sold?
- (c) Determine the percentage of time two cars were sold.
- (d) Describe the shape of the distribution.

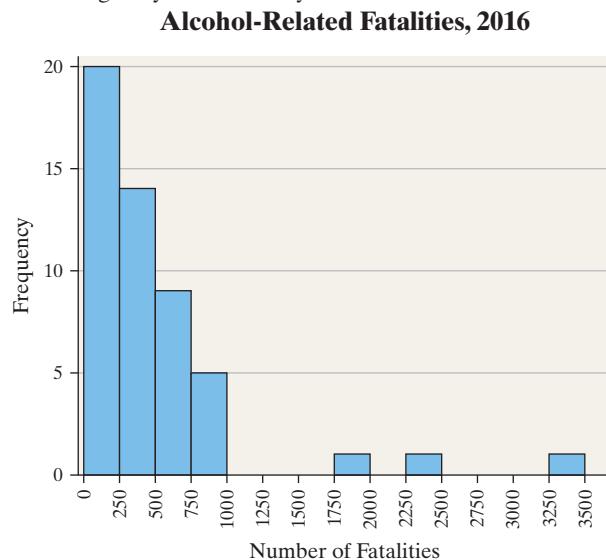
11. **IQ Scores** The following frequency histogram represents the IQ scores of a random sample of seventh-grade students. IQs are measured to the nearest whole number. The frequency of each class is labeled above each rectangle.

IQs of 7th-Grade Students



- (a) How many students were sampled?
- (b) Determine the class width.
- (c) Identify the classes and their frequencies.
- (d) Which class has the highest frequency?
- (e) Which class has the lowest frequency?
- (f) What percent of students had an IQ of at least 130?
- (g) Did any students have an IQ of 165?

- 12. Alcohol-Related Traffic Fatalities** The frequency histogram represents the number of alcohol-related traffic fatalities by state (including Washington, D.C.) in 2016 according to the National Highway Traffic Safety Administration.

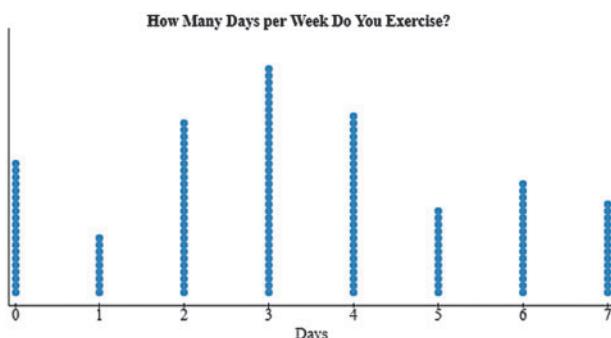


- (a) Determine the class width.
- (b) Identify the classes.
- (c) Which class has the highest frequency?
- (d) Describe the shape of the distribution.
- (e) A reporter writes the following statement: “According to the data, Texas had 3341 alcohol-related deaths, while Vermont had only 62. So the roads in Vermont are much safer.” Explain what is wrong with this statement and how a fair comparison can be made between alcohol-related traffic fatalities in Texas versus Vermont.

In Problems 13 and 14, for each variable presented, state whether you would expect a histogram of the data to be bell-shaped, uniform, skewed left, or skewed right. Justify your reasoning.

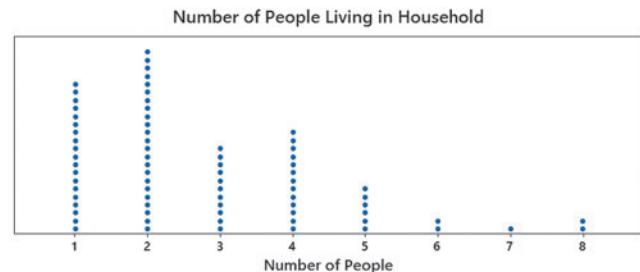
13. (a) Annual household incomes in the United States
 (b) Scores on a standardized exam such as the SAT
 (c) Number of people living in a household
 (d) Ages of patients diagnosed with Alzheimer’s disease
14. (a) Number of alcoholic drinks consumed per week
 (b) Ages of students in a public school district
 (c) Ages of hearing-aid patients
 (d) Heights of full-grown men

- 15. Exercise** A student conducted a survey asking a random sample of adults, “How many days per week do you exercise?”. The results of the survey are displayed in the dot plot drawn in StatCrunch shown below.



- (a) What is the most popular number of days to exercise?
- (b) What is the least popular number of days to exercise?
- (c) How many of the individuals surveyed indicated that they exercise seven days per week?
- (d) How many individuals do not exercise?
- (e) A total of 160 individuals were surveyed, what percent of the individuals surveyed exercise seven days per week? What percent of individuals surveyed do not exercise?

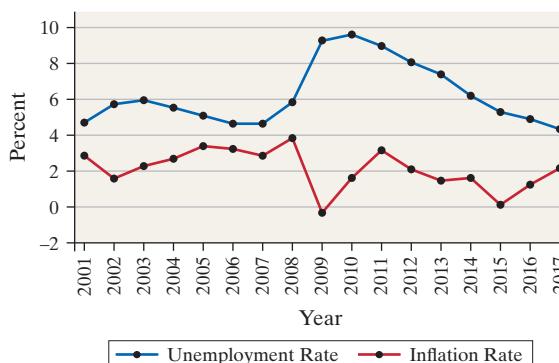
- 16. People in Household** The following dot plot drawn in Minitab shows the number of people living in a household for a random sample of households in the United States.



- (a) If you randomly selected a household, what is the most likely number of people living in the household?
- (b) If you randomly selected a household, what is the least likely number of people living in the household?
- (c) How many households in the sample had five people living in the household?
- (d) How many more households have four people living in the household than three?

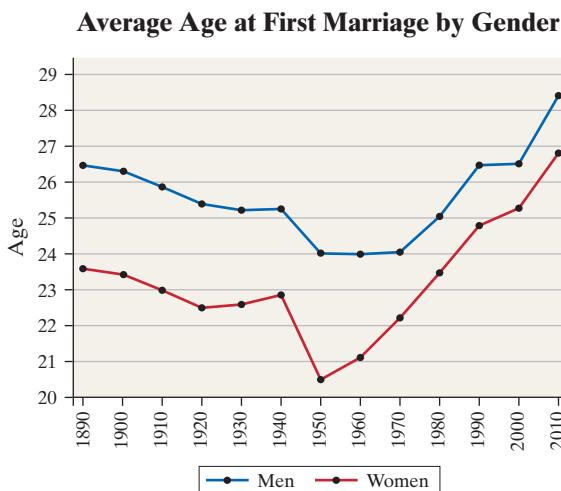
- 17. Misery Index** The following time-series plot shows the annual unemployment and inflation rates for the years 2001 through 2017. Source: www.miseryindex.us

Unemployment and Inflation Rates



- (a) Estimate the unemployment rate in 2012.
- (b) In what year was the unemployment rate highest?
- (c) In what year was the inflation rate highest?
- (d) In what year were the unemployment rate and inflation rate furthest apart?
- (e) The misery index is defined as the sum of the unemployment rate and the inflation rate. According to the misery index, which year was more “miserable”, 2008 or 2011?

- 18. Age at First Marriage** The following time-series plot shows the average age at which individuals first marry by gender for each year of the census since 1890.



- (a) To the nearest year, what was the average age of a man who first married in 1980?
- (b) To the nearest year, what was the average age of a woman who first married in 1960?
- (c) In which year was the difference in the average age of men and women at which they first married the largest? Approximately, what is the age difference?
- (d) In which year was the difference in the average age of men and women at which they first married the least?

Applying the Concepts

- DATA 19. Predicting School Enrollment** To predict future enrollment in a school district, fifty households within the district were sampled, and asked to disclose the number of children under the age of five living in the household. The results of the survey are presented in the following table.

Number of Children under 5	Number of Households
0	16
1	18
2	12
3	3
4	1

- (a) Construct a relative frequency distribution of the data.
- (b) What percentage of households has two children under the age of 5?
- (c) What percentage of households has one or two children under the age of 5?

- DATA 20. Free Throws** In an experiment, a researcher asks a basketball player to record the number of free throws she shoots until she misses. The experiment is repeated 50 times. The following table lists the distribution of the number of free throws attempted until a miss is recorded.

Number of Free Throws until a Miss	Frequency
1	16
2	11
3	9
4	7
5	2
6	3
7	0
8	1
9	0
10	1

- (a) Construct a relative frequency distribution of the data.
- (b) What percentage of the time did she first miss on her fourth free throw?
- (c) What percentage of the time did she first miss the tenth free throw?
- (d) What percentage of the time did she make at least five in a row?

In Problems 21 and 22, find (a) the number of classes, (b) the class limits, and (c) the class width.

- 21. Sprint Speed** The following frequency distribution represents the sprint speed (in feet per second) of all players in Major League Baseball during the 2018 baseball season.

Speed (ft/sec)	Number of Players
22–23.9	18
24–25.9	104
26–27.9	253
28–29.9	166
30–31.9	8

Source: Statcast.

- 22. Earthquakes** The following data represent the magnitude of earthquakes worldwide in October 2018.

Magnitude	Number
0–0.9	3145
1.0–1.9	4145
2.0–2.9	1264
3.0–3.9	241
4.0–4.9	770
5.0–5.9	130
6.0–6.9	14

Source: U.S. Geological Survey, Earthquake Hazards Program.

In Problems 23 and 24, construct (a) a relative frequency distribution, (b) a frequency histogram, and (c) a relative frequency histogram for the given data. Then answer the questions that follow.

- 23. Use the data in Problem 21. What percentage of players had a sprint speed between 24 and 25.9 ft/sec? What percentage of players had a sprint speed less than 24 ft/sec?
- 24. Use the data in Problem 22. What percentage of earthquakes registered 4.0 to 4.9? What percentage of earthquakes registered less than 5.0?

- NW DATA 25. Televisions in the Household** A researcher with A.C. Nielsen wanted to determine the number of televisions in households. He conducts a survey of 40 randomly selected households and obtains the following data.

1	1	4	2	3	3	5	1
1	2	2	4	1	1	0	3
1	2	2	1	3	1	1	3
2	3	2	2	1	2	3	2
1	2	2	2	2	1	3	1

Source: Based on data from the U.S. Department of Energy.

- (a) Are these data discrete or continuous? Explain.
- (b) Construct a frequency distribution of the data.
- (c) Construct a relative frequency distribution of the data.
- (d) What percentage of households in the survey have three televisions?
- (e) What percentage of households in the survey have four or more televisions?
- (f) Construct a frequency histogram of the data.
- (g) Construct a relative frequency histogram of the data.
- (h) Describe the shape of the distribution.

- DATA 26. Waiting** The data below represent the number of customers waiting for a table at 6:00 P.M. for 40 consecutive Saturdays at Bobak's Restaurant.

11	5	11	3	6	8	6	7
4	5	13	9	6	4	14	11
13	10	9	6	8	10	9	5
10	8	7	3	8	8	7	8
7	9	10	4	8	6	11	8

- (a) Are these data discrete or continuous? Explain.
- (b) Construct a frequency distribution of the data.
- (c) Construct a relative frequency distribution of the data.
- (d) What percentage of the Saturdays had 10 or more customers waiting for a table at 6:00 P.M.?
- (e) What percentage of the Saturdays had five or fewer customers waiting for a table at 6:00 P.M.?
- (f) Construct a frequency histogram of the data.
- (g) Construct a relative frequency histogram of the data.
- (h) Describe the shape of the distribution.

- NW DATA 27. Gini Index** The Gini Index is a measure of how evenly income is distributed within a country, ranging from 0 to 100. An index of 0 suggests income is distributed with perfect equality. The higher the number, the worse the income inequality. The data below represent the Gini Index for a random sample of countries. **Note:** The United States has a Gini Index of 45 and Sweden has the lowest Gini Index.

23.0	27.0	30.0	32.0	34.0	36.5	38.5	40.0	41.9	45.0	47.2	50.4	55.1
23.8	27.2	30.3	32.1	34.1	36.7	39.0	40.0	42.4	45.3	47.4	50.8	57.7
24.3	27.4	30.6	32.6	34.2	36.8	39.0	40.1	42.5	45.3	47.5	50.9	58.5
24.7	28.0	30.7	32.7	34.4	36.8	39.0	40.2	43.2	45.5	47.7	51.9	59.2
24.8	28.0	30.9	32.8	34.5	36.8	39.0	40.5	43.7	45.6	47.8	51.9	59.7
25.0	28.2	30.9	33.0	35.2	37.6	39.2	40.8	44.3	45.8	48.3	52.1	61.3
26.0	28.2	31.0	33.2	35.3	37.6	39.4	40.9	44.5	46.0	49.0	53.0	62.9
26.0	28.9	31.3	33.2	35.5	37.6	39.4	41.3	44.6	46.2	50.1	53.2	63.0
26.0	29.0	31.9	33.4	36.2	37.7	39.5	41.5	44.6	46.8	50.2	53.6	63.1
26.3	29.6	31.9	33.7	36.2	37.9	39.7	41.7	44.8	46.9	50.3	53.7	63.2
26.8	30.0	32.0	33.9	36.5	38.0							

Source: CIA World Factbook.

With a first class having a lower class limit of 20 and a class width of 5:

- (a) Construct a frequency distribution.
- (b) Construct a relative frequency distribution.
- (c) Construct a frequency histogram of the data.
- (d) Construct a relative frequency histogram of the data.
- (e) Describe the shape of the distribution.
- (f) Repeat parts (a)–(e) using a class width of 10.
- (g) Does one frequency distribution provide a better summary of the data than the other? Explain.

- DATA 28. Average Income** The following data represent the median household income (in dollars) for the 50 states and the District of Columbia in 2017.

\$51,113	\$83,382	\$57,872	\$43,441	\$62,447	\$54,971	\$45,392
\$72,231	\$53,681	\$51,348	\$56,885	\$50,343	\$56,894	\$63,451
\$61,125	\$57,016	\$43,903	\$59,087	\$59,886	\$55,240	\$57,837
\$48,829	\$73,575	\$51,664	\$59,619	\$59,768	\$59,295	
\$69,759	\$60,208	\$81,084	\$56,550	\$55,006	\$71,319	
\$74,172	\$64,609	\$73,227	\$74,801	\$64,610	\$63,805	
\$72,780	\$58,873	\$57,700	\$72,997	\$63,173	\$71,293	
\$62,318	\$63,481	\$71,920	\$47,855	\$66,390	\$75,418	

Source: Federal Reserve Bank of St. Louis.

With the first class having a lower class limit of 40,000 and a class width of 5000:

- (a) Construct a frequency distribution.
- (b) Construct a relative frequency distribution.
- (c) Construct a frequency histogram of the data.
- (d) Construct a relative frequency histogram of the data.
- (e) Describe the shape of the distribution.
- (f) Repeat parts (a)–(e) using a class width of 10,000.
- (g) Does one frequency distribution provide a better summary of the data than the other? Explain.

- DATA 29. Exit Velocity** The following data represent the exit velocity (in miles per hour) for a simple random sample of 50 homeruns hit during the 2018 Major League Baseball season.

111.0	102.1	105.8	100.5	104.6
104.3	101.8	101.4	105.9	104.0
108.7	104.6	105.9	107.1	103.6
101.0	100.6	104.0	97.6	102.7
105.5	99.3	106.6	90.9	112.2
104.2	108.3	100.8	105.1	99.7
99.9	102.9	93.8	107.3	100.4
104.1	103.4	96.4	103.4	94.8
101.2	106.5	107.3	99.5	100.7
102.8	103.2	103.0	109.3	104.7

Source: Statcast.

With a first class having a lower class limit of 90 and a class width of 4:

- (a) Construct a relative frequency distribution.
- (b) Construct a relative frequency histogram.
- (c) Describe the shape of the distribution.
- (d) Repeat parts (a)–(c) using a class width of 2.5.
- (e) Does one distribution provide a better summary of the data than the other? Explain.

- DATA** **30. Cigarette Tax Rates** The table shows the tax, in dollars, on a pack of cigarettes in each of the 50 states and Washington, DC, as of January 2015. **Note:** The state with the lowest tax is Missouri and the state with the highest tax is New York.

0.425	1.339	0.60	0.17	0.45	1.53	2.52
2.00	0.37	0.36	1.70	0.44	0.62	0.60
2.00	3.20	2.00	0.64	1.25	1.41	2.90
1.15	0.57	2.00	0.80	1.03	1.70	
0.87	1.98	3.51	1.78	1.31	2.75	
0.84	0.995	2.00	2.70	1.60	0.30	
3.40	1.36	3.43	1.66	3.50	3.025	
1.60	0.79	0.68	4.35	0.57	0.55	

Source: Tax Foundation.

With a first class having a lower class limit of 0 and a class width of 0.50:

- (a) Construct a frequency distribution.
- (b) Construct a relative frequency distribution.
- (c) Construct a frequency histogram of the data.
- (d) Construct a relative frequency histogram of the data.
- (e) Describe the shape of the distribution.
- (f) Repeat parts (a)–(e) using a class width of 1.
- (g) Does one frequency distribution provide a better summary of the data than the other? Explain.

- DATA** **31. Student Loans** The following data represent the default rate on student loans for a random sample of 40 colleges or universities in the United States.

12.5	2.6	7.0	23.6	1.7	18.5	4.0	7.4
6.2	0.0	3.7	17.8	5.4	4.1	23.0	2.0
0.0	16.1	3.3	22.4	0.0	6.9	7.5	20.3
0.0	9.0	18.4	10.3	2.7	20.6	8.1	8.8
10.6	12.5	27.3	7.0	5.4	16.2	8.6	10.7

Source: United States Department of Education.

- (a) If seven classes are to be formed, choose an appropriate lower class limit for the first class and a class width.
- (b) Draw a relative frequency histogram of the data.
- (c) Describe the shape of the distribution.

- DATA** **32. Volume of Altria Group Stock** The volume of a stock is the number of shares traded on a given day. The following data, in millions, so that 6.42 represents 6,420,000 shares traded, represent the volume of Altria Group stock traded for a random sample of 35 trading days in 2018.

6.42	23.59	18.91	7.85	7.76
8.51	9.05	14.83	14.43	8.55
6.37	10.30	10.16	10.90	11.20
13.57	9.13	7.83	15.32	14.05
7.84	7.88	17.10	16.58	7.68
7.69	10.22	10.49	8.41	7.85
10.94	20.15	8.97	15.39	8.32

Source: TD Ameritrade.

- (a) If six classes are to be formed, choose an appropriate lower class limit for the first class and a class width.
- (b) Construct a frequency distribution.
- (c) Construct a relative frequency distribution.
- (d) Construct a frequency histogram of the data.
- (e) Construct a relative frequency histogram of the data.
- (f) Describe the shape of the distribution.

- DATA** **33. Return on Investment** Payscale.com tracks the graduates of all institutions of higher education and determines the return on investment (ROI) of the school's graduates. ROI can be thought of as the investment return on the expenses associated with attending college. Go to www.pearsonhighered.com/sullivanstats and download the file 2_2_33. Construct a relative frequency histogram of the data. Comment on what you notice based on the graphical summary.

- DATA** **34. Sullivan Survey** What is the ideal number of children to have? This question was asked on the Sullivan Statistics Survey I. Draw a dot plot of the variable "Children" from the SullivanStatsSurveyI data set at www.pearsonhighered.com/sullivanstats. Now draw a dot plot of the variable "Children" from the survey data set by "gender." Do there appear to be any differences in the ideal number of children for males and females? Is there a better graph that could be drawn to make the comparison easier?

- NW** **35. Televisions in the Household** Draw a dot plot of the televisions per household data from Problem 25.

- 36. Waiting** Draw a dot plot of the waiting data from Problem 26.

- NW** **37. Federal Debt** The following data represent the total federal debt (in trillions of dollars) of the United States from 2000 to 2018.

Year	Debt	Year	Debt
2000	5.7	2010	13.6
2001	5.8	2011	14.8
2002	6.2	2012	16.1
2003	6.8	2013	16.7
2004	7.4	2014	17.8
2005	7.9	2015	18.2
2006	8.5	2016	19.6
2007	9.0	2017	20.2
2008	10.0	2018	21.5
2009	11.9	2019	22.7

Source: TreasuryDirect.gov

- (a) Construct a time-series plot of the data.

- (b) Percentage change may be found using the formula:

$$\text{Percentage change} = \frac{P_2 - P_1}{P_1}.$$

For example, the

percentage change in debt from 2000 to 2001 was

$$\text{Percentage change} = \frac{5.8 - 5.7}{5.7} = 0.018 = 1.8\%.$$

That

is, the debt increased 1.8% from 2000 to 2001. What was the percentage change in debt from 2008 to 2009?

Have there been any years when the debt decreased (since 2000)?

- DATA 38. Hurricanes** The following data represent the number of hurricanes in the Atlantic Ocean for the years 2000 to 2017.

Year	Hurricanes	Year	Hurricanes
2000	8	2009	3
2001	9	2010	12
2002	4	2011	7
2003	7	2012	10
2004	9	2013	2
2005	15	2014	6
2006	5	2015	4
2007	6	2016	7
2008	8	2017	10

Source: Stormfax Weather Almanac.

- (a) Construct a time-series plot of the data.
 (b) See Problem 37(b). What was percentage change from 2015 to 2016? What was the percentage change from 2016 to 2017?

- DATA 39. Births per Woman** Open the data set 2_2_39, which represents the number of births per woman between the ages of 15 and 44 for the years 1960 to 2016 in the United States. Construct a time-series plot and comment on any trends. In what year was the number of births per woman lowest?

- DATA 40. Life Expectancy** Open the data set 2_2_40, which represents the life expectancy of individuals born in the given year for the years 1960 to 2016 in the United States. Construct a time-series plot and comment on any trends. In what year was the life expectancy highest?

- 41. Threaded Problem: Tornado** The data set “Tornadoes_2017” located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a) Draw a relative frequency histogram of the length of all the tornadoes using a lower class limit of 0 and a class width of 5.
 (b) What is the shape of the distribution?
 (c) What is the relative frequency of tornadoes between 5 and 9.999 miles in length?
 (d) Draw a relative frequency histogram of the length of tornadoes in Texas using a lower class limit of 0 and a class width of 5. **Hint:** If you are using StatCrunch, enter “State = TX” in the Where: box of the histogram dialogue window. What is the relative frequency of tornadoes between 35 and 39.999 miles in length? Round your answer to three decimal places.
 (e) Draw a dot plot of the number of fatalities of all tornadoes in 2017. How many tornadoes resulted in four or more fatalities?
 (f) The column “NumberStates” represents the number of states the tornado traveled through. How many tornadoes traveled through two states?

- DATA 42. Putting It Together: Time Viewing a Web Page** Nielsen/NetRatings measures the amount of time an individual spends viewing a specific web page. The following data represent the amount of time, in seconds, a random sample of 40 surfers spent viewing a web page. Decide on an appropriate graphical summary and create the graphical summary. Write a few sentences that describe the data. Be sure to include in your description any interesting features the data may exhibit.

19	185	4	104	23
27	73	27	12	16
15	27	51	10	40
111	83	27	31	5
9	75	11	48	65
86	51	69	257	14
20	45	19	13	81
26	42	156	12	114

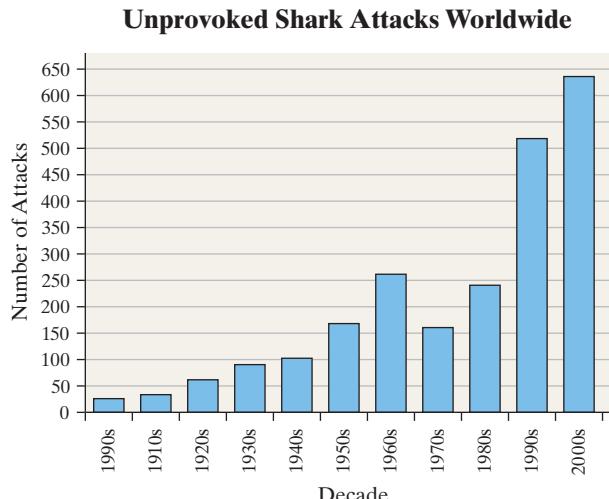
Source: Based on information provided by Nielsen/NetRatings.

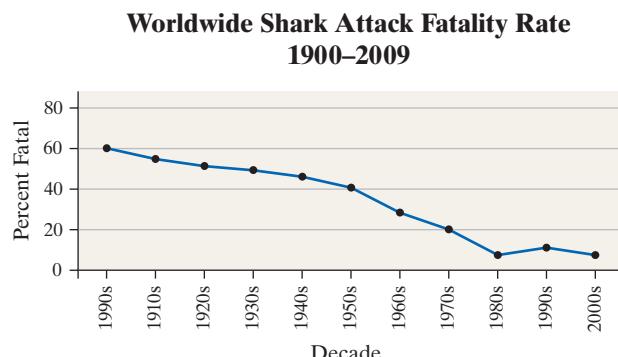
- DATA 43. Putting It Together: Red Light Cameras** Chicago has installed cameras at various intersections throughout the city. The camera photographs the license plate of any car engaging in a moving violation (such as driving through a red light or failure to completely stop prior to turning on red). Open the data set 2_2_43, which represents the number of violations recorded by all cameras on October 17, 2018. The data set is located at www.pearsonhighered.com/sullivanstats.

- (a) What type of data is “number of violations”?
 (b) Is this data from a population or sample? Explain.
 (c) How many red light cameras are there in the city of Chicago?
 (d) Draw a relative frequency histogram of the data using a lower class limit of the first class of 0 and a class width of 5. Describe the shape of the distribution.
 (e) Draw a dot plot of the data. Do you prefer the dot plot or the histogram? Explain.
 (f) Were there any cameras that did not record any violations on October 17, 2018? If so, how many?

- 44. Putting It Together: Which Graphical Summary?** Suppose you just obtained data from a survey in which you learned the following information about 50 individuals: age, income, marital status, number of vehicles in household. For each variable, explain the type of graphical summary you might be able to draw to provide a visual summary of the data.

- 45. Putting It Together: Shark!** The graphic below and the graphic on the following page represent the number of reported shark attacks and fatality rate worldwide since 1900. Write a report about the trends in the graphs. In your report discuss the apparent contradiction between the increase in shark attacks, but the decrease in fatality rate.





Source: Florida Museum of Natural History.

Explaining the Concepts

46. Why shouldn't classes overlap when summarizing continuous data in a frequency or relative frequency distribution?

47. Is there such a thing as the correct choice for a class width? Is there such a thing as a poor choice for a class width? Explain your reasoning.

48. Describe the situations in which it is preferable to use relative frequencies over frequencies when summarizing quantitative data.

49. **StatCrunch** Choose any data set that has at least 50 observations of a quantitative variable. In StatCrunch, open the “Histogram with Sliders” applet. Adjust the bin width and starting point of the histogram. Can the choice of bin width (class width) affect the shape of the histogram? Explain.

50. Sketch four histograms—one skewed right, one skewed left, one bell-shaped, and one uniform. Label each histogram according to its shape. What makes a histogram skewed left? Skewed right? Symmetric?

51. What type of variable is required when drawing a time-series plot? Why do we draw time-series plots?

2.3 Graphical Misrepresentations of Data



Objective ① Describe what can make a graph misleading or deceptive

① Describe What Can Make a Graph Misleading or Deceptive

Statistics: The only science that enables different experts using the same figures to draw different conclusions.—EVAN ESAR

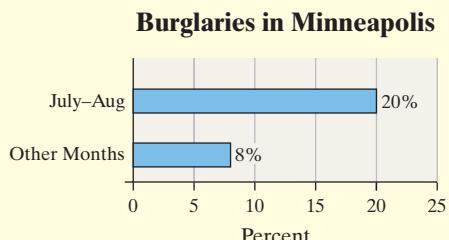
Statistics often gets a bad rap for having the ability to manipulate data to support any position. One method of distorting the truth is through graphics. We mentioned in Section 2.1 how visual displays send more powerful messages than raw data or even tables of data. Since graphics are so powerful, care must be taken in constructing graphics and in interpreting their messages. Graphics may *mislead* or *deceive*. We will call graphs misleading if they unintentionally create an incorrect impression. We consider graphs deceptive if they purposely create an incorrect impression. In either case, a reader's incorrect impression can have serious consequences. Therefore, it is important to be able to recognize misleading and deceptive graphs.

The most common graphical misrepresentations of data involve the scale of the graph, an inconsistent scale, or a misplaced origin. Increments between tick marks should be constant, and scales for comparative graphs should be the same. Also, because readers usually assume that the baseline, or zero point, is at the bottom of the graph, a graph that begins at a higher or lower value can be misleading.

EXAMPLE 1 Misrepresentation of Data

Problem A home security company located in Minneapolis, Minnesota, develops a summer ad campaign with the slogan “When you leave for vacation, burglars leave for work.” According to the city of Minneapolis, roughly 20% of home burglaries occur during the peak vacation months of July and August. The advertisement contains the graphic shown in Figure 14. Explain what is wrong with the graphic.

Figure 14



Approach Look for any characteristics that may mislead a reader, such as inconsistent scales or poorly defined categories.

Solution Consider how the categories of data are defined. The sum of the percentages (the relative frequencies) over all 12 months should be 1. Because $10(0.08) + 0.20 = 1$, it is clear that the bar for Other Months represents an average percent for each month, while the bar for July—August represents the average percent for the months July and August combined. The unsuspecting reader is misled into thinking that July and August each have a burglary rate of 20%.

Figure 15 gives a better picture of the burglary distribution. The increase during the month of July is not as dramatic as the bar graph in Figure 14 implies and August actually has fewer burglaries than September or October. In fact, Figure 14 would be considered deceitful because the security company is intentionally trying to convince consumers that July and August are much higher burglary months.

Figure 15



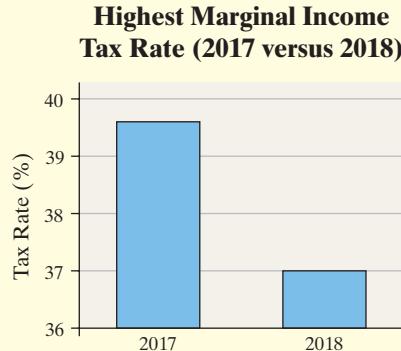
NW Now Work Problem 5



EXAMPLE 2 Misrepresentation of Data by Manipulating the Vertical Scale

Problem A national news organization developed the graphic shown in Figure 16 to illustrate the change in the highest marginal tax rate effective January 1, 2018. Why might this graph be considered misleading?

Figure 16

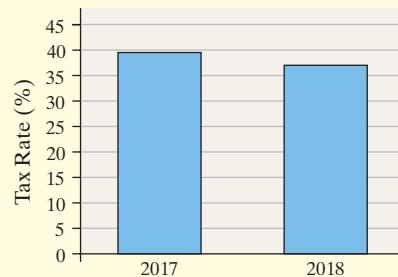


Approach We need to find any characteristics that may mislead a reader, such as manipulation of the vertical scale.

Solution The graph in Figure 16 may lead the reader to believe that marginal tax rates are almost three times as high in 2017 as in 2018 since the height of the bar for 2017 is almost three times the length of the bar for 2018 when in fact the difference is only 2.6 percentage points (because the tax rate in 2017 is 39.6% and the tax rate in 2018 is 37%). The reason for this incorrect conclusion is due to the fact that the vertical scale does not begin at 0. A graph that does not distort the difference in tax rates is given in Figure 17. Notice that the decrease in the tax rate is still apparent in the graph without the distortion as to the size of the decrease.

Figure 17

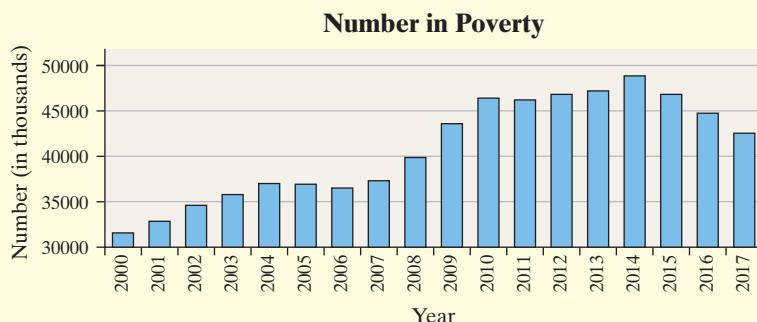
Highest Marginal Income Tax Rate (2017 versus 2018)



EXAMPLE 3 Misrepresentation of Data by Manipulating the Vertical Scale

Problem The time-series graph in Figure 18 depicts the number of residents in the United States living in poverty. Why might this graph be considered misrepresentative?

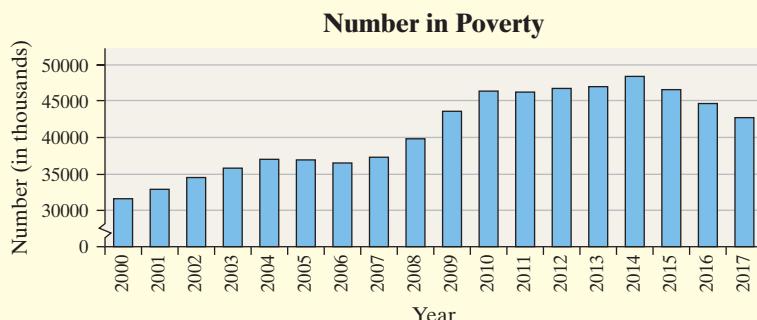
Figure 18



Approach Look for any characteristics that may mislead a reader, such as manipulation of the vertical scale.

Solution The graph may mislead readers to believe that the number in poverty has more than doubled from 2007 to 2014 because the bar for 2014 is more than twice the length of the bar from 2007. Notice that the vertical axis begins at 30,000 instead of 0. This type of scaling is common when the smallest observed data value is a rather large number. It is not necessarily done intentionally to confuse or mislead the reader. Often, the main purpose in graphs (particularly time-series graphs) is to discover a trend, rather than the actual differences in the data. However, the author of the graph should clearly indicate that the graph does not begin at 0 by including the symbol ↴ in the vertical scale. This symbol indicates that the scale has been truncated and the graph has a gap. See Figure 19.

Figure 19



Although the data do stand out in Figure 19, it is better to use a time-series plot when displaying time series data (rather than a bar plot). In addition, it is better to use the percent of the population in poverty rather than the number living in poverty. This is due to the fact that increases in poverty may be due to increases in the population as well as a deterioration of the economy. Figure 20 shows a time-series plot of the percentage of U.S. residents living in poverty. The lack of bars allows us to focus on the trend in the data, rather than the relative size (or area) of the bars.

Figure 20

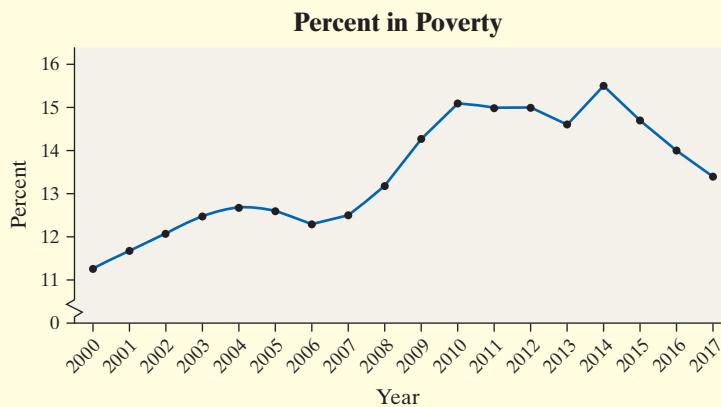
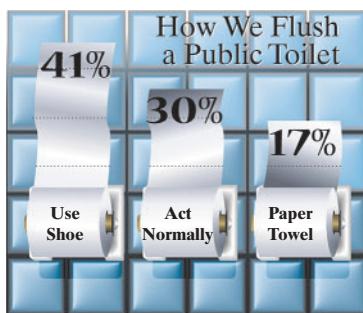
**NW Now Work Problem 3**

Figure 21

**NW Now Work Problem 9**

EXAMPLE 4 Misrepresentation of Data

Problem The bar graph illustrated in Figure 21 is a *USA Today*-type graph. A survey was conducted by Impulse Research for Quilted Northern Confidential in which individuals were asked how they would flush a toilet when the facilities are not sanitary. What's wrong with the graphic?

Approach Compare the vertical scales of each bar to see if they accurately depict the given percentages.

Solution First, it is unclear whether the bars include the roll of toilet paper or not. In either case, the roll corresponding to “use shoe” should be $2.4 (= 41/17)$ times taller than the roll corresponding to “paper towel.” If we include the roll of toilet paper, then the bar corresponding to “use shoe” is less than double the height of “paper towel.” If we do not include the roll of toilet paper, then the bar corresponding to “use shoe” is almost exactly double the height of the bar corresponding to “paper towel.” The vertical scaling is incorrect.

Newspapers, magazines, and websites often go for a “wow” factor when displaying graphs. The graph designer may be more interested in catching the reader’s eye than making the data stand out. The two most commonly used tactics are 3-D graphs and pictograms (graphs that use pictures to represent the data). The use of 3-D effects is strongly discouraged, because such graphs are often difficult to read, add little value to the graph, and distract the reader from the data.

When comparing bars, our eyes are really comparing the *areas* of the bars. That is why we emphasized that the bars or classes should be of the same width. Uniform width ensures that the area of the bar is proportional to its height, so we can compare the heights of the bars. However, when we use two-dimensional pictures in place of bars, it is not possible to obtain a uniform width. To avoid distorting the picture when values increase or decrease, both the height and width of the picture must be adjusted. This often leads to misleading graphs.

EXAMPLE 5 Misleading Graphs

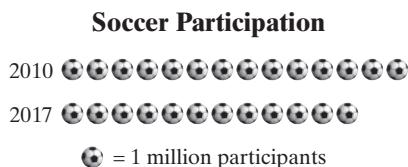
Figure 22

Soccer Participation



NW Now Work Problem 13

Figure 23



Problem Soccer seems to be losing popularity as a sport in the United States. In 2010, there were approximately 14 million participants in the United States. By 2017 this number had dropped to 12 million. To illustrate this decrease, we could create a graphic like the one shown in Figure 22. Describe how the graph may be misleading.

Source: U.S. Census Bureau; National Sporting Goods Association.

Approach Look for characteristics of the graph that seem to manipulate the facts, such as an incorrect depiction of the size of the graphics.

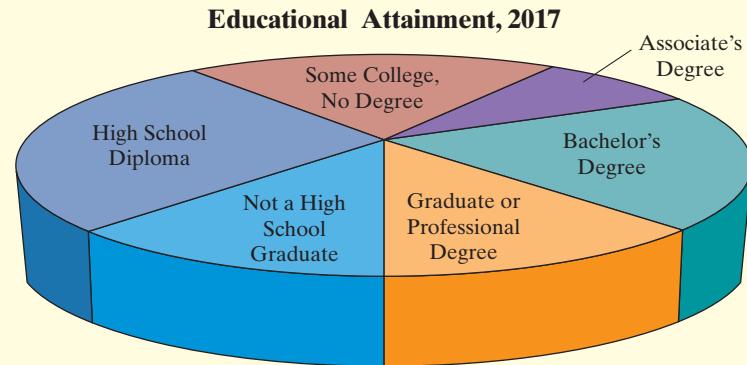
Solution The graph on the left of the figure has an area that is more than four times the area of the graph on the right of the figure. While the number of participants is given in the problem statement, they are not included in the graph, which makes the reader rely on the graphic alone to compare soccer participation in the two years. There was a 14% decrease in participation from 2010 to 2017, not more than 300% as indicated by the graphic. To be correct, the graph for 2017 should have an area that is only 14% less than the area of the graph for 2010. Adding the data values to the graphic would help reduce the chance of misinterpretation due to the oversized graph.



EXAMPLE 6 Misrepresentations of Data: Three-Dimensional Scale

Problem Figure 24 represents the educational attainment (level of education) in 2017 of adults 25 years and older who are U.S. residents. Why might this graph be considered misrepresentative.

Figure 24



Approach Find any characteristics that may mislead a reader, such as overemphasis on one category of data.

Solution Three-dimensional pie charts tend to overstate the significance of categories closer to the reader. For example, in this graph the category "Graduate or Professional Degree" looks about the same size as the size of the category "High School Diploma." Yet, in the raw data shown in Table 15 there are 27,181,000 individuals who have a Graduate or Professional Degree and 60,032,000 individuals who have a High School Diploma.

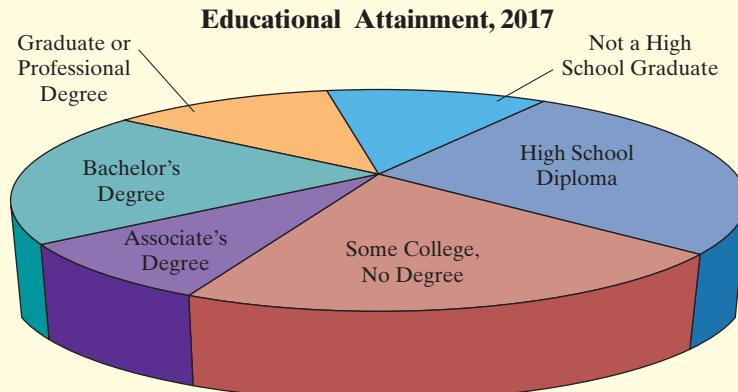
In fact, if we rotate the pie chart to bring the category "Some College, No Degree" to the front, we obtain the graph shown in Figure 25. Now, "Some College,

Table 15

Educational Attainment	Frequency (000s)
Not a High School Graduate	26,582
High School Diploma	60,032
Some College, No Degree	45,110
Associate's Degree	18,761
Bachelor's Degree	43,585
Graduate or Professional Degree	27,181

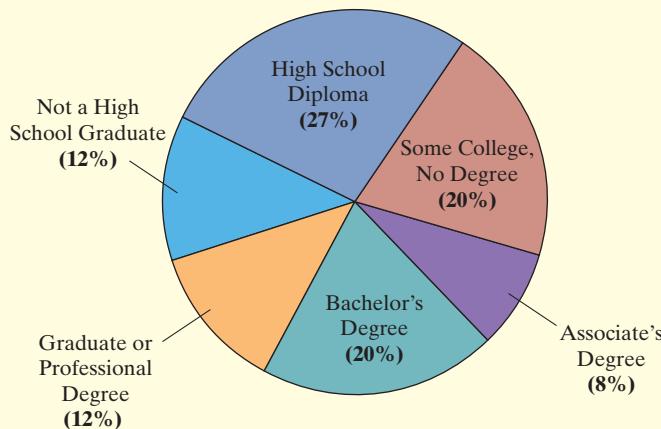
Source: U.S. Census Bureau.

Figure 25



"No Degree" appears to be the largest piece of the pie. Clearly, the three-dimensional representation distorts the number of individuals in each category, so these types of graphs should be avoided. Instead, stick to two-dimensional graphs with clear labels as shown in Figure 26.

Figure 26

Educational Attainment, 2017

The material presented in this section is by no means all-inclusive. There are many ways to create graphs that mislead. Two popular texts written about ways that graphs mislead or deceive are *How to Lie with Statistics* (W.W. Norton & Company, Inc., 1982) by Darrell Huff and *The Visual Display of Quantitative Information* (Graphics Press, 2001) by Edward Tufte.

We conclude this section with some guidelines for constructing good graphics.

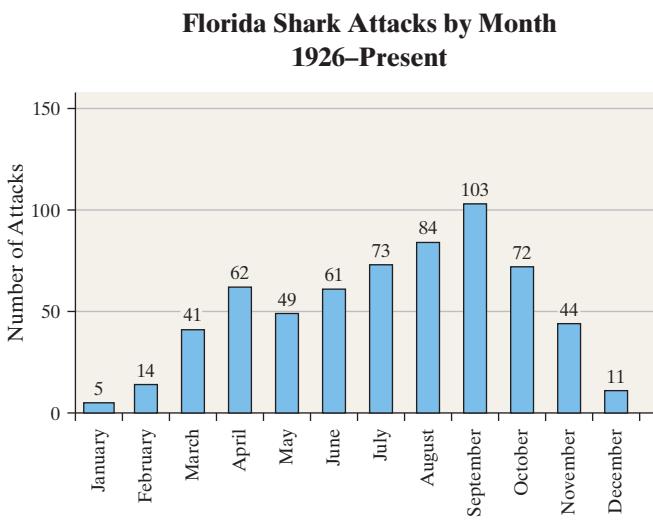
- Title and label the graphic axes clearly, providing explanations if needed. Include units of measurement and a data source when appropriate.
- Avoid distortion. Never lie about the data.
- Minimize the amount of white space in the graph. Use the available space to let the data stand out. If you truncate the scales, clearly indicate this to the reader.
- Avoid clutter, such as excessive gridlines and unnecessary backgrounds or pictures. Don't distract the reader.
- Avoid three dimensions. Three-dimensional charts may look nice, but they distract the reader and often lead to misinterpretation of the graphic.
- Do not use more than one design in the same graphic. Sometimes graphs use a different design in one portion of the graph to draw attention to that area. Don't try to force the reader to any specific part of the graph. Let the data speak for themselves.
- Avoid relative graphs that are devoid of data or scales.



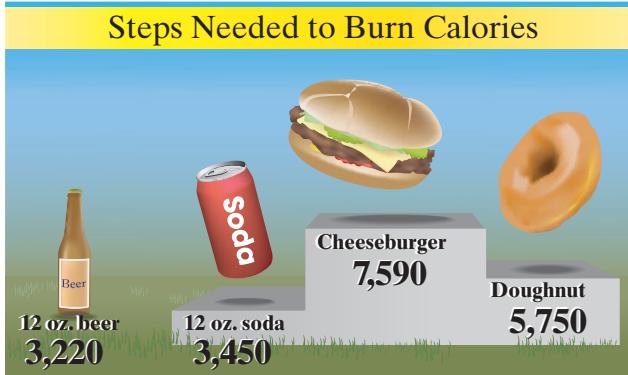
2.3 Assess Your Understanding

Applying the Concepts

- 1. Shark Attacks** Explain how the following graph is misleading.

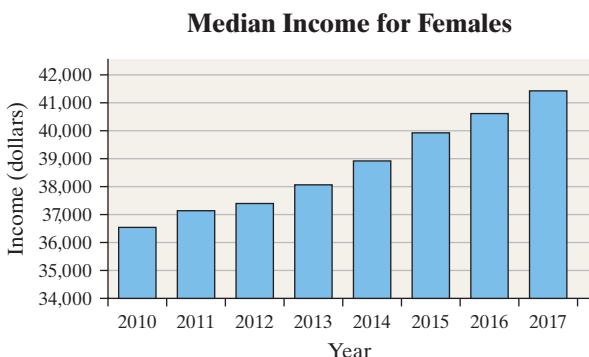


- 2. Burning Calories** The following is a *USA Today*-type graph.



- (a) Explain how it is misleading.
 (b) What could be done to improve the graphic?

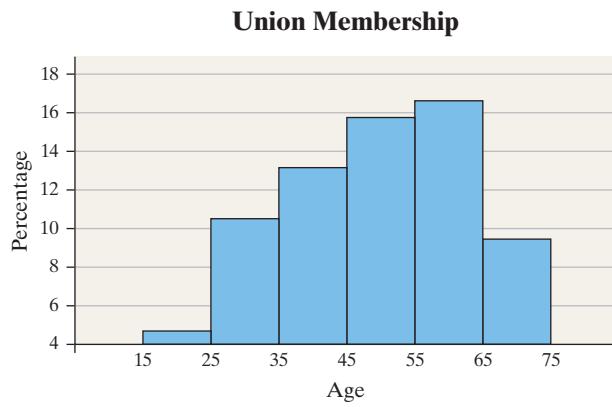
- NW 3. Median Earnings** The graph shows the median income for females from 2010 to 2017 in constant 2017 dollars.



Source: U.S. Census Bureau.

- (a) How is the graph misleading? What does the graph seem to convey?
 (b) Redraw the graph so that it is not misleading. What does the new graph seem to convey?

- 4. Union Membership** The following relative frequency histogram represents the proportion of employed people aged 15–74 years old who are members of a union.

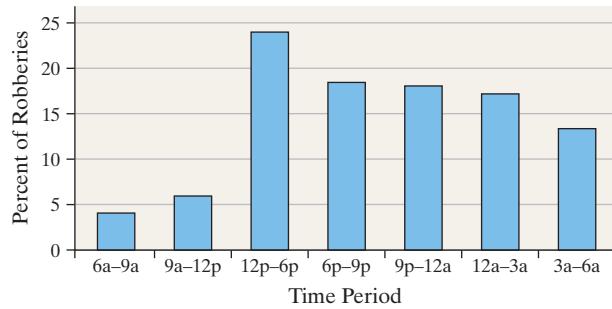


- (a) Describe how this graph is misleading. What might a reader conclude from the graph?

- (b) Redraw the histogram so that it is not misleading.

- NW 5. Robberies** A newspaper article claimed that the afternoon hours were the worst in terms of robberies and provided the following graph in support of this claim. Explain how this graph is misleading.

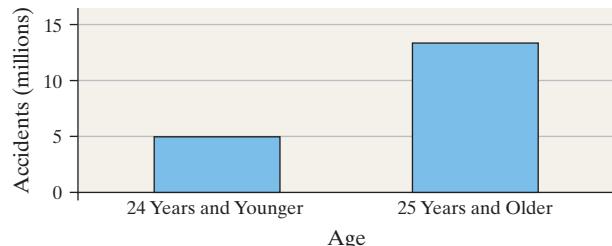
Hourly Crime Distribution (Robbery)



Source: U.S. Statistical Abstract.

- 6. Car Accidents** An article in a student newspaper claims that younger drivers are safer than older drivers and provides the following graph to support the claim. Explain how this graph is misleading.

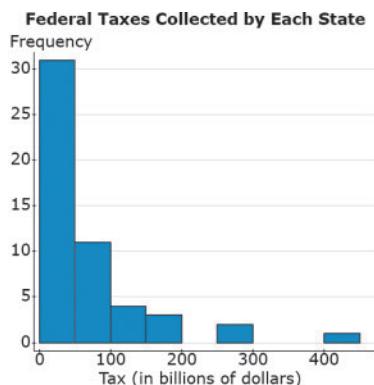
Number of Motor Vehicle Accidents



Source: U.S. Statistical Abstract.

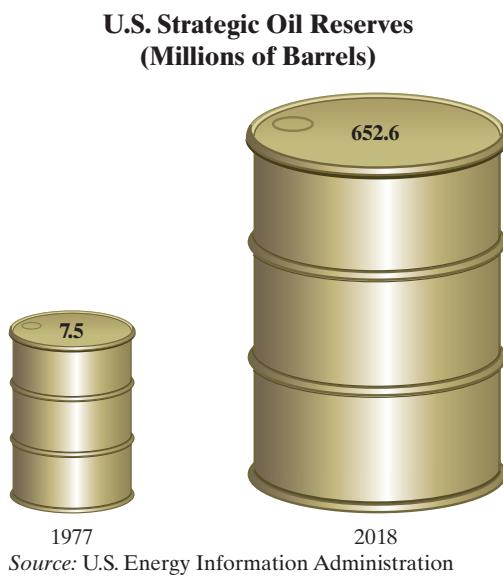
7. Tax Revenue The following histogram drawn in StatCrunch represents the total tax collected by the Internal Revenue Service for each state plus Washington, DC and Puerto Rico. Explain why the graph is misleading.

Note: Puerto Rico paid the least in tax (\$3.5 million), while California paid the most (\$405.9 million).



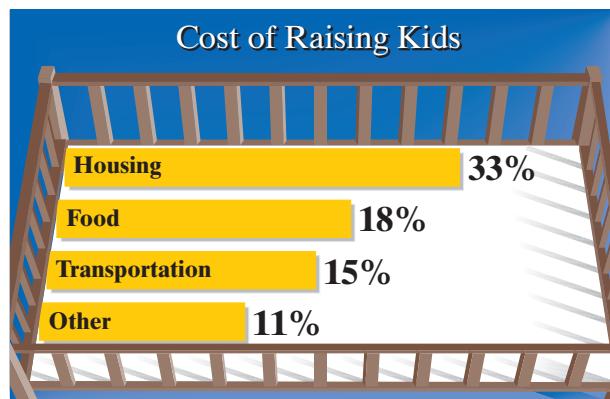
8. You Explain It! Oil Reserves The U.S. Strategic Oil Reserve is a government-owned stockpile of crude oil. It was established after the oil embargo in the mid-1970s and is meant to serve as a national defense fuel reserve, as well as to offset reductions in commercial oil supplies that would threaten the U.S. economy. The graphic depicts oil reserves in 1977 and 2018.

- (a) How many times larger should the graphic for 2018 be than the 1977 graphic (to the nearest whole number)?
- (b) The United States imported approximately 10.14 million barrels of oil per day in 2018. At that rate, assuming no change in U.S. oil production, how long would the U.S. strategic oil reserve last if no oil were imported?

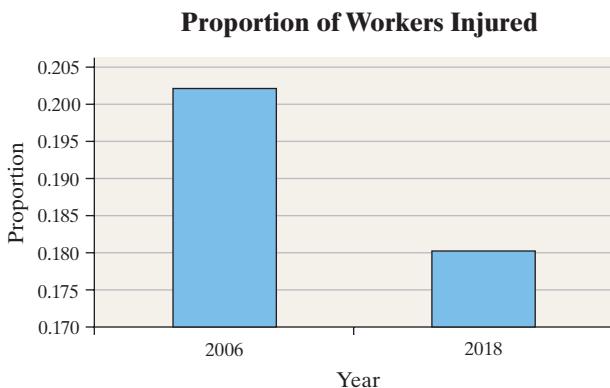


NW 9. Cost of Kids The *USA Today*-type graph shown in the next column is based on data from the Department of Agriculture. It represents the percentage of income a middle-class family will spend on their children.

- (a) How is the graphic misleading?
- (b) What could be done to improve the graphic?



10. Worker Injury The safety manager at Klutz Enterprises provides the following graph to the plant manager and claims that the rate of worker injuries has been reduced by 67% over a 12-year period. Does the graph support his claim? Explain.



DATA 11. Health Care Expenditures The following data represent health care expenditures per capita (per person) and as a percentage of the U.S. gross domestic product (GDP) from 2007 to 2016. Gross domestic product is the total value of all goods and services created during the course of the year.

Health Care		
Year	per Capita	Percent of GDP
2007	7627	5.5
2008	7897	3.5
2009	8143	3.1
2010	8412	3.3
2011	8644	2.8
2012	8924	3.2
2013	9121	2.2
2014	9515	4.3
2015	9994	5.0
2016	10,348	3.5

Source: Center for Medicare and Medicaid Services, Office of the Actuary.

- (a) Construct a time-series plot that a politician would create to support the position that health care expenditures are increasing and must be slowed.
- (b) Construct a time-series plot that the health care industry would create to refute the opinion of the politician. Is your graph convincing?
- (c) Explain how different measures may be used to support two completely different positions.

12. Gas Hike The average per gallon price for regular unleaded gasoline in the United States rose from \$1.46 in 2001 to \$2.77 in 2018. *Source:* U.S. Energy Information Administration.

- (a) Construct a graphic that is not misleading to depict this situation.
- (b) Construct a misleading graphic that makes it appear the average price roughly quadrupled between 2001 and 2018.

NW 13. Overweight Between 1980 and 2016, the percent of adults in the United States who were overweight more than doubled from 15% to 40%.

Source: Centers for Disease Control and Prevention.

- (a) Construct a graphic that is not misleading to depict this situation.
- (b) Construct a misleading graphic that makes it appear that the percent of overweight adults has more than quadrupled between 1980 and 2016.

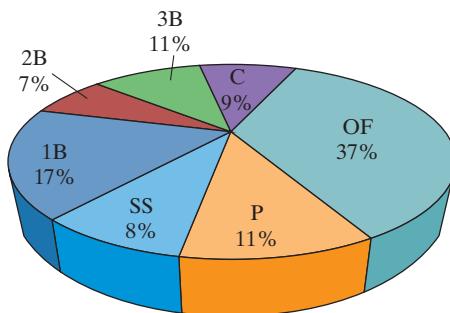
14. Ideal Family Size The following *USA Today*-type graphic illustrates the ideal family size (total children) based on a survey of adult Americans.



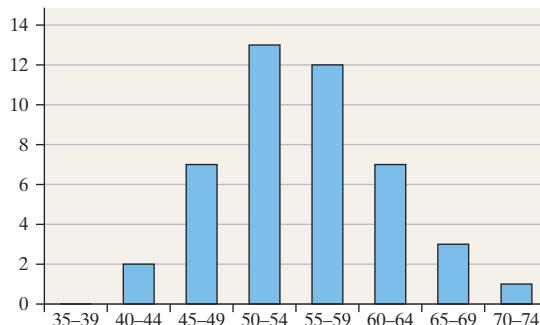
- (a) What type of graphic is being displayed?
- (b) Describe any problems with the graphic.
- (c) Construct a graphic that is not misleading and makes the data stand out.

15. National League Baseball MVP The following pie chart displays the position played by the most valuable player (MVP) in the National League of Major League Baseball from 1931 through 2018. Explain how the graphic is misleading. What should be done to improve the graphic?

National League MVP [1931–2018]



16. Inauguration Day The intention of the following graph is to display the ages of U.S. presidents on Inauguration Day. Explain any problems you see with the graph.



Chapter 2 Review

Summary

In this chapter, we learned how to graphically display raw data. Raw data is data that has not yet been organized in any form. It is the data we typically obtain from an observational study or designed experiment. A simple example of raw data would be Male, Female, Female, Male, Female, and so on.

Raw data may be organized into tables. The first step in organizing any data is to determine whether the data are qualitative or quantitative.

Qualitative data may be organized into tables by using the data itself to form categories. For example, if the raw data is for the variable gender, then the categories would be Female and Male. The tables formed are either

frequency tables or relative frequency tables. The types of graphs drawn from qualitative data include bar graphs, side-by-side bar graphs, and pie charts.

Quantitative data may also be organized into tables. For discrete data with only a few outcomes, the outcomes form the categories of data. However, for discrete data with a variety of outcomes, or for continuous data, the categories of data are formed using classes. Quantitative data can be organized into frequency or relative frequency tables. Quantitative data may also be represented graphically in histograms, dot plots, and time-series plots.

Finally, in creating graphs, care must be taken not to draw a graph that misleads or deceives the reader.

Vocabulary

Raw data (p. 62)	Pie chart (p. 68)	Uniform distribution (p. 82)
Frequency distribution (p. 63)	Classes (p. 76)	Bell-shaped distribution (p. 82)
Relative frequency (p. 64)	Histogram (p. 77)	Skewed right (p. 82)
Relative frequency distribution (p. 64)	Lower and upper class limits (p. 78)	Skewed left (p. 82)
Bar graph (p. 65)	Class width (p. 78)	Time-series data (p. 83)
Pareto chart (p. 66)	Open ended (p. 78)	Time series plot (p. 83)
Side-by-side bar graph (p. 66)	Dot plot (p. 82)	

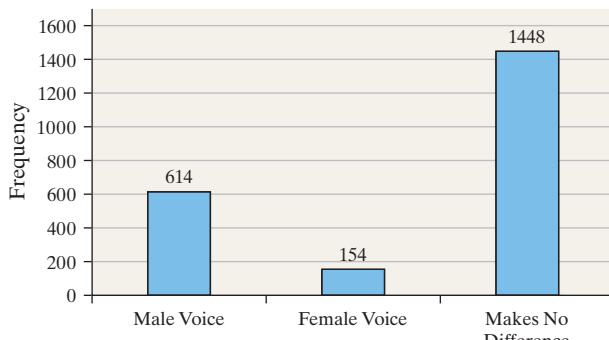
Objectives

Section	You should be able to ...	Example(s)	Review Exercises
2.1	1 Organize qualitative data in tables (p. 63) 2 Construct bar graphs (p. 64) 3 Construct pie charts (p. 68)	1, 2 3 through 5 6	2(a), 4(a) and (b) 2(b), 4(c) 2(c), 4(d)
2.2	1 Organize discrete data in tables (p. 76) 2 Construct histograms of discrete data (p. 77) 3 Organize continuous data in tables (p. 78) 4 Construct histograms of continuous data (p. 80) 5 Draw dot plots (p. 82) 6 Identify the shape of a distribution (p. 82) 7 Draw time-series graphs (p. 83)	1 2 3 4, 5 6 7 8	5(a) and (b) 5(c) and (d) 3(a), 6(a) and (b), 7(a) and (b) 3(b) and (c), 6(c) and (d), 7(c) and (d) 5(g) 3(b), 5(c), 6(e), 7(c), 8 10
2.3	1 Describe what can make a graph misleading or deceptive (p. 92)	1 through 6	9(c), 10, 11, 12

Review Exercises

- 1. Effective Commercial** Harris Interactive conducted a poll of U.S. adults and asked, “When there is a voiceover in a commercial, which type of voice is more likely to sell me a car?” Results of the survey are in the bar graph.

Convincing Voice in Purchasing a Car



- (a) How many participants were in the survey?
 (b) What is the relative frequency of the respondents who indicated that it made no difference which voice they heard?
 (c) Redraw the graph as a Pareto chart.

- (d) Research automotive commercials. Do you believe that auto manufacturers use the results of this survey when developing their commercials?

- 2. Weapons Used in Homicide** The following frequency distribution represents the cause of death in homicides for a random sample of homicides.

Type of Weapon	Frequency
Firearms	844
Knives or cutting instruments	149
Personal weapons (hands, fists, etc.)	69
Other weapon or not stated	162

Source: FBI, Uniform Crime Reports.

- (a) Construct a relative frequency distribution.
 (b) Construct a relative frequency bar graph.
 (c) Construct a pie chart.
3. Live Births The frequency distribution on the following page represents the number of live births (in thousands) in the United States in 2016 by age of mother.

Age of Mother (years)	Births (thousands)
10–14	2
15–19	194
20–24	765
25–29	1124
30–34	1092
35–39	555
40–44	115
45–49	8
50–54	1

Source: National Center for Health Statistics.

- (a) Construct a relative frequency distribution.
- (b) Construct a frequency histogram. Describe the shape of the distribution.
- (c) Construct a relative frequency histogram.
- (d) What percentage of live births were to mothers aged 20 to 24?
- (e) What percentage of live births were to mothers of age 30 or older?

DATA 4. Political Affiliation One hundred randomly selected registered voters in the city of Naperville were asked their political affiliation: Democrat (D), Republican (R), or Independent (I). The results of the survey are shown below.

D	R	D	R	D	R	D	D	R	D
R	D	D	D	R	R	D	D	D	D
R	R	I	I	D	R	D	R	R	R
I	D	D	R	I	I	R	D	R	R
D	I	R	D	D	D	D	I	I	R
R	I	R	R	I	D	D	D	D	R
D	I	I	D	D	R	R	R	R	D
D	R	R	R	D	D	I	I	D	D
D	D	I	D	R	I	D	D	D	D
R	R	R	R	R	D	R	D	R	D

- (a) Construct a frequency distribution of the data.
- (b) Construct a relative frequency distribution of the data.
- (c) Construct a relative frequency bar graph of the data.
- (d) Construct a pie chart of the data.
- (e) What appears to be the most common political affiliation in Naperville?

DATA 5. Family Size A random sample of 60 couples married for seven years were asked to give the number of children they have. The results of the survey are as follows:

0	2	3	1	3	2
3	2	3	5	3	2
1	1	3	2	3	2
4	3	3	0	4	2
2	0	1	0	3	4
0	3	3	2	1	2
2	3	3	1	3	2
4	3	2	2	0	2
0	1	4	2	2	3
4	3	3	2	4	3

- (a) Construct a frequency distribution of the data.
- (b) Construct a relative frequency distribution of the data.

- (c) Construct a frequency histogram of the data. Describe the shape of the distribution.
- (d) Construct a relative frequency histogram of the data.
- (e) What percentage of couples married seven years has two children?
- (f) What percentage of couples married seven years has at least two children?
- (g) Draw a dot plot of the data.

DATA 6. Home Ownership Rates The table shows the home ownership rate in each of the 50 states and Washington, DC, in 2017.

69.5	71.1	69.8	62.2	56.3	67.9	66.2	66.8
64.0	40.9	70.8	74.1	74.5	70.7	62.7	70.2
66.5	64.4	68.2	71.4	66.1	63.4	69.7	
64.2	64.0	68.6	72.0	66.5	70.2	70.8	
55.2	59.6	65.4	71.3	50.5	63.1	65.8	
65.5	69.6	70.9	65.6	64.3	69.5	64.9	
66.7	64.8	67.4	68.3	62.4	69.2	74.8	

Source: U.S. Census Bureau.

Note: The state with the highest home ownership rate is West Virginia and the lowest is Washington, DC.

With a lower class limit of the first class of 45 and a class width of 5:

- (a) Construct a frequency distribution.
- (b) Construct a relative frequency distribution.
- (c) Construct a frequency histogram of the data.
- (d) Construct a relative frequency histogram of the data.
- (e) Describe the shape of the distribution.
- (f) Repeat parts (a)–(e) using a lower class limit for the first class of 40 and a class width of 10.
- (g) Does one frequency distribution provide a better summary of the data than the other? Explain.

DATA 7. Diameter of a Cookie The following data represent the diameter (in inches) of a random sample of 34 Keebler Chips Deluxe™ Chocolate Chip Cookies.

2.3414	2.3010	2.2850	2.3015	2.2850	2.3019	2.2400
2.3005	2.2630	2.2853	2.3360	2.3696	2.3300	2.3290
2.2303	2.2600	2.2409	2.2020	2.3223	2.2851	2.2382
2.2438	2.3255	2.2597	2.3020	2.2658	2.2752	2.2256
2.2611	2.3006	2.2011	2.2790	2.2425	2.3003	

Source: Trina S. McNamara, student at Joliet Junior College.

- (a) Construct a frequency distribution.
- (b) Construct a relative frequency distribution.
- (c) Construct a frequency histogram. Describe the shape of the distribution.
- (d) Construct a relative frequency histogram.

DATA 8. Time Online The data on the following page represent the average number of hours per week that a random sample of 40 college students spend online. The data are based on the ECAR Study of Undergraduate Students and Information Technology. Construct a stem-and-leaf diagram of the data, and comment on the shape of the distribution.

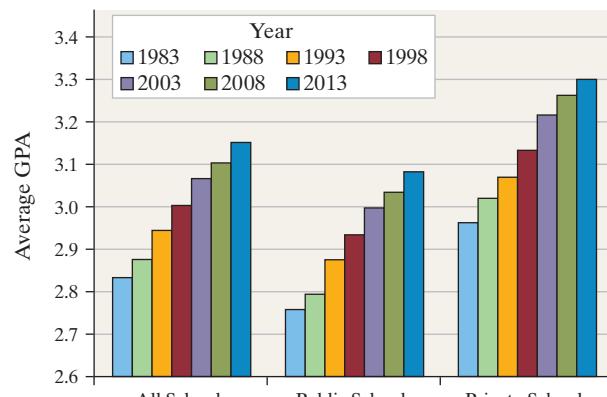
18.9	14.0	24.4	17.4	13.7	16.5	14.8	20.8
22.9	22.2	13.4	18.8	15.1	21.9	21.1	14.7
18.6	18.0	21.1	15.6	16.6	20.6	17.3	17.9
15.2	16.4	14.5	17.1	25.7	17.4	18.8	17.1
13.6	20.1	15.3	19.2	23.4	14.5	18.6	23.8

9. Grade Inflation The side-by-side bar graph to the right shows the average grade point average for the academic years beginning in 1983, 1988, 1993, 1998, 2003, 2008, and 2013, for colleges and universities.

- (a) Does the graph suggest that grade inflation is a problem in colleges?
- (b) In 1998, a grade of A became the most popular grade nationwide. Determine the percentage increase in GPAs for public schools from 1998 (2.92) to 2013 (3.07). Determine the percentage increase in GPAs for private schools from 1998 (3.11) to 2013 (3.30). Which type of institution appears to have the higher inflation?

(c) Do you believe the graph is misleading? Explain.

Recent GPA Trends Nationwide Four-Year Colleges and Universities



Source: gradeinflation.com



10. Income Distribution The following data represent the percentage of total adjusted gross income (AGI) earned and percentage of tax paid by various income classes. The top 1% represents the percentage of total AGI earned and tax paid by those whose income is higher than 99% of all earners. The bottom 50% represents the percentage of total AGI earned and tax paid by those whose income is in the bottom 50% of all income earners. For example, in 2006, 22.06% of all income earned in the United States was earned by those in the top 1% of all income earners, while 12.51% of all income earned in the United States was earned by those in the bottom 50% of income earners.

Year	Income Share of Top 1% of Earners	Income Share of Bottom 50% of Earners	Income Tax Share of Top 1% of Earners	Income Tax Share of Bottom 50% of Earners
2006	22.06	12.51	39.89	2.99
2007	22.83	12.26	40.41	2.89
2008	20.00	12.75	38.02	2.70
2009	17.21	11.88	36.34	2.46
2010	18.87	11.74	37.38	2.36
2011	18.70	11.55	35.06	2.89
2012	21.86	11.10	38.09	2.78
2013	19.04	11.49	37.80	2.78
2014	20.58	11.27	39.48	2.75
2015	20.65	11.28	39.04	2.83
2016	19.72	11.59	37.32	3.04

Source: The Tax Foundation.

- (a) Use the data to make an argument that adjusted gross incomes are diverging among Americans.
- (b) Use the data to make an argument that, while incomes are diverging between the top 1% and bottom 50%, the total taxes paid as a share of income are also diverging.

11. Misleading Graphs In 2017, the average earnings of a high school graduate were \$30,624. At \$52,484, the average earnings of a recipient of a bachelor's degree were about 71% higher.

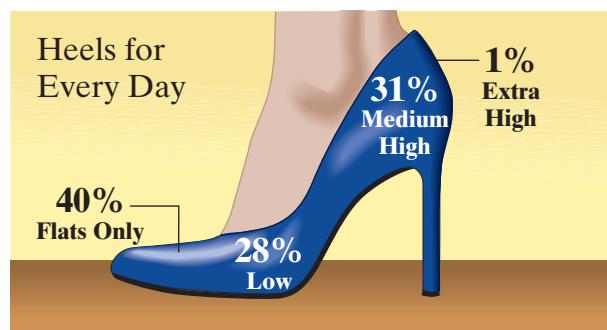
Source: U.S. Census Bureau.

- (a) Construct a misleading graph that a college recruiter might create to convince high school students that they should attend college.

- (b) Construct a graph that does not mislead.

12. High Heels The graphic to the right is a *USA Today*-type graph displaying women's preference for shoes.

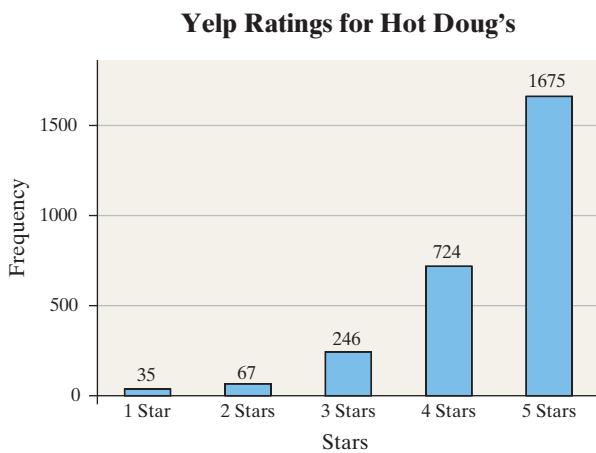
- (a) Which type of shoe is preferred the most? The least?
- (b) How is the graph misleading?





Chapter Test

1. The graph shows the ratings on Yelp for Hot Doug's Restaurant.



- (a) Which was the most popular rating for Hot Doug's?
 (b) How many ratings were posted on Hot Doug's?
 (c) How many more 5 Star ratings are there than 4 Star ratings?
 (d) What percentage of ratings are 5 Star ratings?
 (e) Is it appropriate to describe the shape of the distribution as skewed left? Why or why not?
2. A random sample of 1005 adult Americans was asked, "How would you prefer to pay for new road construction?" Results of the survey are below.

Response	Frequency
New tolls	412
Increase gas tax	181
No new roads	412

Source: HNTB Corporation.

- (a) Construct a relative frequency distribution.
 (b) What percent of the respondents indicated they would like to see an increase in gas taxes?
 (c) Construct a frequency bar graph.
 (d) Construct a relative frequency bar graph.
 (e) Construct a pie chart.
- DATA** 3. Interested in knowing the educational background of its customers the Metra Train Company contracted a marketing firm to conduct a survey asking 50 randomly selected commuters at the train station to disclose their educational attainment. The results shown in the column to the right were obtained.
- (a) Construct a frequency distribution of the data.
 (b) Construct a relative frequency distribution of the data.
 (c) Construct a relative frequency bar graph of the data.
 (d) Construct a pie chart of the data.
 (e) What is the most common educational level of a commuter?

No high school diploma	Some college
High school graduate	No high school diploma
Some college	No high school diploma
No high school diploma	High school graduate
Associate's degree	High school graduate
Bachelor's degree	High school graduate
Bachelor's degree	High school graduate
High school graduate	Associate's degree
Associate's degree	High school graduate
Some college	No high school diploma
Some college	High school graduate
High school graduate	Advanced degree
High school graduate	High school graduate
Bachelor's degree	Bachelor's degree
High school graduate	Some college
Bachelor's degree	Advanced degree
Advanced degree	No high school diploma
Bachelor's degree	High school graduate
Bachelor's degree	No high school diploma
Associate's degree	Some college
Advanced degree	Some college
High school graduate	No high school diploma
Bachelor's degree	No high school diploma
Some college	Some college
High school graduate	High school graduate

- DATA** 4. The following data represent the number of cars that arrived at a McDonald's drive-through between 11:50 A.M. and 12:00 noon each Wednesday for the past 50 weeks:

1	7	3	8	2	3	8	2	6	3
6	5	6	4	3	4	3	8	1	2
5	3	6	3	3	4	3	2	1	2
4	4	9	3	5	2	3	5	5	5
2	5	6	1	7	1	5	3	8	4

- (a) Construct a frequency distribution of the data.
 (b) Construct a relative frequency distribution of the data.
 (c) Construct a frequency histogram of the data. Describe the shape of the distribution.
 (d) Construct a relative frequency histogram of the data.
 (e) What percentage of weeks did exactly three cars arrive between 11:50 A.M. and 12:00 noon?
 (f) What percentage of weeks did three or more cars arrive between 11:50 A.M. and 12:00 noon?
 (g) Draw a dot plot of the data.

- DATA** 5. Dr. Paul Oswiecimski randomly selects 40 of his 20- to 29-year-old patients and obtains the following data regarding their serum HDL cholesterol:

70	56	48	48	53	52	66	48
36	49	28	35	58	62	45	60
38	73	45	51	56	51	46	39
56	32	44	60	51	44	63	50
46	69	53	70	33	54	55	52

Source: Paul Oswiecimski.

- (a) Construct a frequency distribution.
 (b) Construct a relative frequency distribution.
 (c) Construct a frequency histogram of the data.
 (d) Construct a relative frequency histogram of the data.
 (e) Describe the shape of the distribution.
- DATA** 6. The following data represent the time (in minutes) students spent working their Section 1.1 homework from Sullivan's College Algebra course (based on time logged into MyLabMath). Draw a stem-and-leaf diagram of the data and comment on the shape of the distribution.

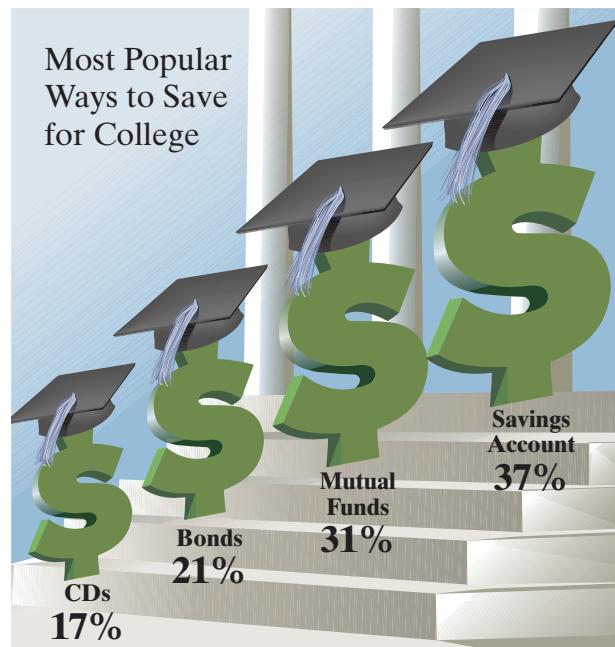
46	47	110	56	71	109
63	91	111	93	125	78
85	108	73	118	70	89
99	45	73	125	96	109
110	61	40	52	103	126

- DATA** 7. The data in the next column shows birth rate and per capita income (in thousands of 2012 dollars) from 2005 through 2017. Draw a time-series plot for both birth rate and per capita income. Comment on any trends.

Year	Birth Rate (births per 1000 women age 15–44)	Per Capita Income (thousands of 2012 dollars)
2005	14.0	36.5
2006	14.3	37.6
2007	14.3	38.1
2008	14.0	38.1
2009	13.5	37.7
2010	13.0	38.2
2011	12.7	38.8
2012	12.6	40.0
2013	12.4	39.0
2014	12.5	40.3
2015	12.4	41.6
2016	12.2	42.0
2017	11.8	42.8

Source: Centers for Disease Control and U.S. Census Bureau.

8. The following is a *USA Today*-type graph. Do you think the graph is misleading? Why? If you think it is misleading, what might be done to improve the graph?



Making an Informed Decision

Tables or Graphs?

You work for the school newspaper. Your editor approaches you with a special reporting assignment. Your task is to write an article that describes the “typical” student at your school, complete with supporting information. To write this article, you have to survey at least 40 students and ask them to respond to a questionnaire. The editor would like to have at least two qualitative and two quantitative variables that describe the typical student. The results of the survey will be presented in your article, but you are unsure whether you should present tabular or graphical summaries, so you decide to perform the following experiment.

1. Develop a questionnaire that results in obtaining the values of two qualitative and two quantitative variables. Administer the questionnaire to at least 40 students on your campus.
2. Summarize the data in both tabular and graphical form.
3. Select 20 individuals. (They don’t have to be students at your school.) Give the tabular summaries to 10 individuals and the graphical summaries to the other 10. Ask each individual to study the table or graph for 5 seconds. After 1 minute, give a questionnaire that asks various questions regarding the

information contained in the table or graph. For example, if you summarize age data, ask the individual which age group has the highest frequency. Record the number of correct answers for each individual.



Which summary results in a higher percentage of correct answers, the tables or the graphs? Write a report that discusses your findings.

4. Now use the data collected from the questionnaire to create some misleading graphs. Again, select 20 individuals. Give 10 individuals the misleading graphs and 10 individuals the correct graphs. Ask each individual to study each graph for 5 seconds. After 1 minute has elapsed, give a questionnaire that asks various questions regarding the information contained in the graphs. Record the number of correct answers for each individual. Did the misleading graphs mislead? Write a report that discusses your findings.

Note: Be sure to check with your school’s administration regarding privacy laws and policies regarding studies involving human subjects.

3

Numerically Summarizing Data

Outline

- 3.1** Measures of Central Tendency
- 3.2** Measures of Dispersion
- 3.3** Measures of Central Tendency and Dispersion from Grouped Data
- 3.4** Measures of Position and Outliers
- 3.5** The Five-Number Summary and Boxplots

Making an Informed Decision



Suppose that you are in the market for a used car. To make an informed decision regarding your purchase, you decide to collect as much information as possible. What information is important in helping you make this decision? See the Decisions project on page 169.

Putting It Together

When we look at a distribution of data, we should consider three characteristics of the distribution: shape, center, and spread. In the last chapter, we discussed methods for organizing raw data into tables and graphs. These graphs (such as the histogram) allow us to identify the shape of the distribution: symmetric (in particular, bell shaped or uniform), skewed right, or skewed left.

The center and spread are numerical summaries of the data. The center of a data set is commonly called the **average**. There are many ways to describe the average value of a distribution. In addition, there are many ways to measure the spread of a distribution. The most appropriate measure of center and spread depends on the distribution's shape.

Once these three characteristics of the distribution are known, we can analyze the data for interesting features, including unusual data values, called **outliers**.

3.1 Measures of Central Tendency



Preparing for This Section Before getting started, review the following:

- Population versus sample (Section 1.1, p. 5)
- Parameter versus statistic (Section 1.1, p. 5)
- Quantitative data (Section 1.1, p. 8)
- Qualitative data (Section 1.1, p. 8)
- Simple random sampling (Section 1.3, pp. 24–27)

Objectives

- ① Determine the arithmetic mean of a variable from raw data
- ② Determine the median of a variable from raw data
- ③ Explain what it means for a statistic to be resistant
- ④ Determine the mode of a variable from raw data

A measure of central tendency numerically describes the average or typical data value. We hear the word *average* in the news all the time:

- The average miles per gallon of gasoline of the 2020 Chevrolet Corvette Z06 in highway driving is 22.
- According to the U.S. Census Bureau, the national average commute time to work in 2018 was 27.1 minutes.
- According to the U.S. Census Bureau, the average household income in 2019 was \$63,030.
- The average American woman is 5'4" tall and weighs 142 pounds.

CAUTION!

Whenever you hear the word *average*, be aware that the word may not always be referring to the mean. One average could be used to support one position, while another average could be used to support a different position.

In this chapter, we discuss the three most widely-used measures of central tendency: the *mean*, the *median*, and the *mode*. In the media (newspapers, blogs, and so on), *average* usually refers to the mean. But beware: some reporters use *average* to refer to the median or mode. As we shall see, these three measures of central tendency can give very different results!

1 Determine the Arithmetic Mean of a Variable from Raw Data

In everyday language, the word *average* often represents the arithmetic mean. To compute the arithmetic mean of a set of data, the data must be quantitative.

Definitions

The **arithmetic mean** of a variable is computed by adding all the values of the variable in the data set and dividing by the number of observations. The **population arithmetic mean**, μ (pronounced “mew”), is computed using all the individuals in a population. The population mean is a parameter.

The **sample arithmetic mean**, \bar{x} (pronounced “x-bar”), is computed using sample data. The sample mean is a statistic.

While other types of means exist (see Problems 39 and 40), the arithmetic mean is generally referred to as the **mean**. We will follow this practice for the remainder of the text.

We usually use Greek letters to represent parameters and Roman letters (such as x or s) to represent statistics. The formulas for computing population and sample means follow:

IN OTHER WORDS

To find the mean of a set of data, add up all the observations and divide by the number of observations.

If x_1, x_2, \dots, x_N are the N observations of a variable from a population, then the population mean, μ , is

$$\mu = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum x_i}{N} \quad (1)$$

If x_1, x_2, \dots, x_n are n observations of a variable from a sample, then the sample mean, \bar{x} , is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n} \quad (2)$$

Note that N represents the size of the population, and n represents the size of the sample. The symbol Σ (the Greek letter capital sigma) tells us to add the terms. The subscript i shows that the various values are distinct and does not serve as a mathematical operation. For example, x_1 is the first data value, x_2 is the second, and so on.

EXAMPLE 1

Computing a Population Mean and a Sample Mean

Table 1

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

Problem The data in Table 1 represent the first exam score of 10 students enrolled in Introductory Statistics. Treat the 10 students as a population.

- (a) Compute the population mean.
- (b) Find a simple random sample of size $n = 4$ students.
- (c) Compute the sample mean of the sample found in part (b).

Approach

- (a) To compute the population mean, add all the data values (test scores) and divide by the number of individuals in the population.
- (b) Recall from Section 1.3 that we can use Table I in Appendix A, a calculator with a random-number generator, or computer software to obtain simple random samples. We will use a TI-84 Plus C graphing calculator.
- (c) Find the sample mean by adding the data values corresponding to the individuals in the sample and then dividing by $n = 4$, the sample size.

Solution

- (a) Compute the population mean by adding the scores of all 10 students:

$$\begin{aligned}\sum x_i &= x_1 + x_2 + x_3 + \cdots + x_{10} \\ &= 82 + 77 + 90 + 71 + 62 + 68 + 74 + 84 + 94 + 88 \\ &= 790\end{aligned}$$

Divide this result by 10, the number of students in the class.

$$\mu = \frac{\sum x_i}{N} = \frac{790}{10} = 79$$

Although it was not necessary in this problem, we will agree to round the mean to one more decimal place than that in the raw data.

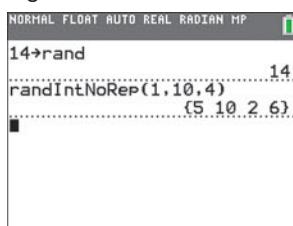
- (b) To find a simple random sample of size $n = 4$ from a population of size $N = 10$, we will use the TI-84 Plus C random-number generator with a seed of 14. (Recall that the seed gives the calculator its starting point to generate the list of random numbers.) Figure 1 shows the students in the sample: Jennifer (62), Juan (88), Ryanne (77), and Dave (68).
- (c) Compute the sample mean by adding the scores of the four students:

$$\begin{aligned}\sum x_i &= x_1 + x_2 + x_3 + x_4 \\ &= 62 + 88 + 77 + 68 \\ &= 295\end{aligned}$$

Divide this result by 4, the number of individuals in the sample.

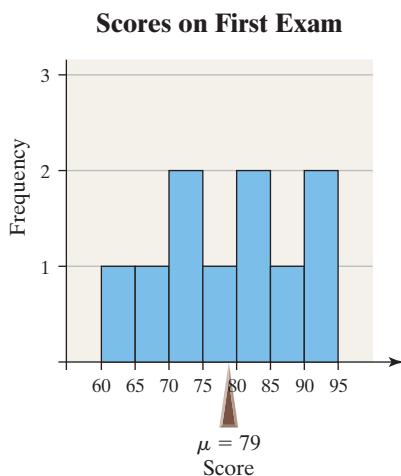
$$\bar{x} = \frac{\sum x_i}{n} = \frac{295}{4} = 73.8 \quad \text{Round to the nearest tenth.}$$

Figure 1



It helps to think of the mean of a data set as the center of gravity. In other words, the mean is the value such that a histogram of the data is perfectly balanced, with equal weight on each side of the mean. Figure 2 shows a histogram of the data in Table 1 with the mean labeled. The histogram balances at $\mu = 79$.

Figure 2



② Determine the Median of a Variable from Raw Data

A second measure of central tendency is the median. To compute the median of a set of data, the data must be quantitative.

Definition

The **median** of a variable is the value that lies in the middle of the data when arranged in ascending order. We use M to represent the median.

IN OTHER WORDS

To help remember the idea behind the median, think of the median of a highway; it divides the highway in half. So the median divides the data in half, with at most half the data below the median and at most half above it.

Steps in Finding the Median of a Data Set

Step 1 Arrange the data in ascending order.

Step 2 Determine the number of observations, n .

Step 3 Determine the observation in the middle of the data set.

- If the number of observations is odd, then the median is the data value exactly in the middle of the data set. That is, the median is the observation that lies in the $\frac{n+1}{2}$ position.
- If the number of observations is even, then the median is the mean of the two middle observations in the data set. That is, the median is the mean of the observations that lie in the $\frac{n}{2}$ position and the $\frac{n}{2} + 1$ position.

EXAMPLE 2

Determining the Median of a Data Set with an Odd Number of Observations

Problem The data in Table 2 represent the length (in seconds) of a random sample of songs released in the 1970s. Find the median length of the songs.

Approach Follow the steps listed above.

Table 2

Song Name	Length
“Sister Golden Hair”	201
“Black Water”	257
“Free Bird”	284
“The Hustle”	208
“Southern Nights”	179
“Stayin’ Alive”	222
“We Are Family”	217
“Heart of Glass”	206
“My Sharona”	240

Solution**Step 1** Arrange the data in ascending order:

$$179, 201, 206, 208, 217, 222, 240, 257, 284$$

Step 2 There are $n = 9$ observations.**Step 3** Since n is odd, the median, M , is the observation exactly in the middle of the data set, 217 seconds (the $\frac{n+1}{2} = \frac{9+1}{2} = 5$ th data value). We list the data in ascending order and show the median in blue.

$$179, 201, 206, 208, \textcolor{blue}{217}, 222, 240, 257, 284$$

Notice there are four observations on each side of the median. **EXAMPLE 3****Determining the Median of a Data Set with an Even Number of Observations****Problem** Find the median score of the data in Table 1 on page 109.**Approach** Follow the steps listed on the previous page.**Solution****Step 1** Arrange the data in ascending order:

$$62, 68, 71, 74, 77, 82, 84, 88, 90, 94$$

Step 2 There are $n = 10$ observations.**Step 3** Because n is even, the median is the mean of the two middle observations, the fifth ($\frac{n}{2} = \frac{10}{2} = 5$) and sixth ($\frac{n}{2} + 1 = \frac{10}{2} + 1 = 6$) observations with the data written in ascending order. So the median is the mean of 77 and 82:

$$M = \frac{77 + 82}{2} = 79.5$$

Notice that there are five observations on each side of the median.

$$62, 68, 71, 74, 77, 82, 84, 88, 90, 94$$

\uparrow
 $M = 79.5$

Notice that 50% (or half) of the students scored less than 79.5 and 50% (or half) of the students scored above 79.5. **NW** Now compute the median of the data in Problem 15 by hand**EXAMPLE 4****Finding the Mean and Median Using Technology****Problem** Use statistical software or a calculator to determine the population mean and median of the student test score data in Table 1 on page 109.**Approach** We will use StatCrunch to obtain the mean and median. The steps for obtaining measures of central tendency using the TI-83/84 Plus graphing calculator, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on page 116.**Solution** Figure 3 shows the output obtained from StatCrunch. The results agree with the “by hand” solution from Examples 1 and 3. 

Figure 3

Summary statistics:			
Column	n	Mean	Median
Scores	10	79	79.5

③ Explain What It Means for a Statistic to Be Resistant

Which measure of central tendency is better to use—the mean or the median? It depends.

EXAMPLE 5

Comparing the Mean and the Median

Problem Yolanda wants to know how much time she typically spends on her cell phone. She goes to her phone's website and records the call length for a random sample of 12 calls, shown in Table 3. Find the mean and median length of a cell phone call. Which measure of central tendency better describes the length of a typical phone call?

Table 3

1	7	4	1
2	4	3	48
3	5	3	6

Source: Yolanda Sullivan's cell phone records

Approach We will find the mean and median using Minitab. To help judge which is the better measure of central tendency, we will also draw a dot plot of the data.

Solution Figure 4 indicates that the mean call length is $\bar{x} = 7.3$ minutes and the median call length is 3.5 minutes. Figure 5 shows a dot plot of the data using Minitab.

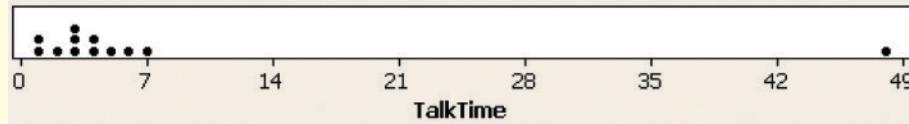
Figure 4

Descriptive Statistics: TalkTime

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
TalkTime	12	0	7.25	3.74	12.96	1.00	2.25	3.50	5.75	48.00

Figure 5



Which measure of central tendency do you think better describes the typical call length? Since only one phone call is longer than the mean, the mean is not representative of the typical call length. So the median is the better measure of central tendency for this data set.



Look back at the data in Table 3. Suppose Yolanda's 48-minute call was actually a 5-minute call. Then the mean call length would be 3.7 minutes and the median call length would still be 3.5 minutes. So one extreme observation (48 minutes) can cause the mean to increase substantially, but have no effect on the median. In other words, the mean is sensitive to extreme values while the median is not. In fact, if Yolanda's 48-minute call had actually been 148 minutes long, the median would still be 3.5 minutes, but the mean would increase to 15.6 minutes. The median is based on the value of the middle observation, so the value of the largest observation does not affect its computation.

Definition

A numerical summary of data is said to be **resistant** if extreme observations (very large or small) relative to the data do not affect its value substantially.

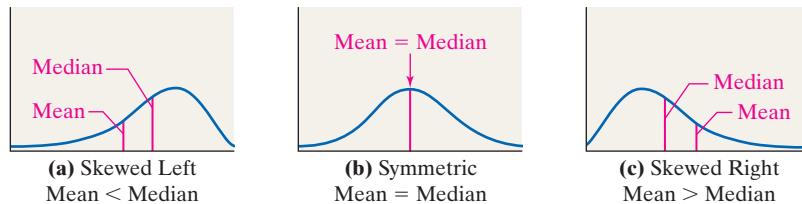
So the median is resistant, while the mean is not resistant.

When data are skewed, there are extreme values in the tail, which tend to pull the mean in the direction of the tail. For example, in skewed-right distributions, there are large observations in the right tail. These observations increase the value of the mean, but have little effect on the median. Similarly, if a distribution is skewed left, the mean tends to be smaller than the median. In symmetric distributions, the mean and the median are close in value. We summarize these ideas in Table 4 and Figure 6.

Table 4**Relation between the Mean, Median, and Distribution Shape**

Distribution Shape	Mean versus Median
Skewed left	Mean substantially smaller than median
Symmetric	Mean roughly equal to median
Skewed right	Mean substantially larger than median

Figure 6
Mean or median versus skewness



A word of caution is in order. The relation between the mean, median, and skewness are guidelines. The guidelines tend to hold up well for continuous data, but when the data are discrete, the rules can be easily violated. See Problem 42.*

You may be asking yourself, “Why would I ever compute the mean?” After all, the mean and median are close in value for symmetric data, and the median is the better measure of central tendency for skewed data. The reason we compute the mean is that much of the statistical inference that we perform is based on the mean. We will have more to say about this in Chapter 8. Plus, the mean uses all the data, while the median relies only on the position of the data.

EXAMPLE 6**Describing the Shape of a Distribution****Table 5**

5.8	7.4	9.2	7.0	8.5	7.6
7.9	7.8	7.9	7.7	9.0	7.1
8.7	7.2	6.1	7.2	7.1	7.2
7.9	5.9	7.0	7.8	7.2	7.5
7.3	6.4	7.4	8.2	9.1	7.3
9.4	6.8	7.0	8.1	8.0	7.5
7.3	6.9	6.9	6.4	7.8	8.7
7.1	7.0	7.0	7.4	8.2	7.2
7.6	6.7				

Problem The data in Table 5 represent the birth weights (in pounds) of 50 randomly sampled babies.

- (a) Find the mean and the median birth weight.
- (b) Describe the shape of the distribution.
- (c) Which measure of central tendency better describes the typical birth weight?

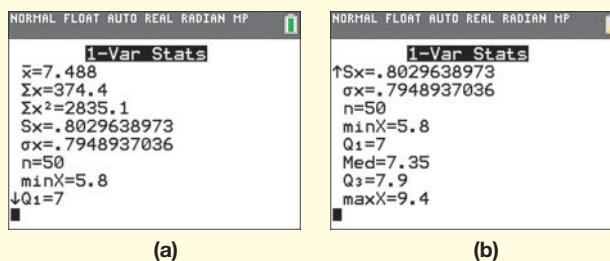
Approach

- (a) This can be done either by hand or technology. We will use a TI-84 Plus CE.
- (b) Draw a histogram to identify the shape of the distribution.
- (c) If the data are roughly symmetric, the mean is the better measure of central tendency. If the data are skewed, the median is the better measure.

Solution

- (a) Using a TI-84 Plus CE, we find $\bar{x} = 7.49$ and $M = 7.35$. See Figure 7.

Figure 7

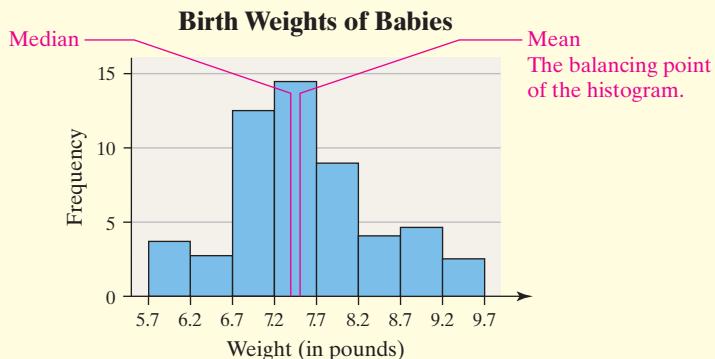


- (b) Figure 8 on the following page shows the frequency histogram with the mean and median labeled. The distribution is bell shaped. We have further evidence of the shape because the mean and median are close in value to each other.

(continued)

*This idea is discussed in “Mean, Median, and Skew: Correcting a Textbook Rule” by Paul T. von Hippel. *Journal of Statistics Education*, Volume 13, Number 2 (2005).

Figure 8 Birth weights of 50 randomly selected babies



- (c) Because the mean and median are close in value, we use the mean as the measure of central tendency.

NW Now Work Problem 23



4 Determine the Mode of a Variable from Raw Data

A third measure of central tendency is the mode, which can be computed for either quantitative or qualitative data.

Definition

The **mode** of a variable is the most frequent observation of the variable that occurs in the data set.

To compute the mode, tally the number of observations that occur for each data value. The data value that occurs most often is the mode. A set of data can have no mode, one mode, or more than one mode. If no observation occurs more than once, we say the data have **no mode**.

EXAMPLE 7

Finding the Mode of Quantitative Data

Problem The following data represent the number of O-ring failures on all space shuttle flights prior to the fatal flight of the space shuttle *Columbia* that occurred on January 28, 1986. *Source: Report of the Presidential Commission on the Space Shuttle Challenger Accident*, 6 June 1986, Volume 1, Page 145.

Find the mode number of O-ring failures.

Approach Tally the number of times we observe each data value. The data value with the highest frequency is the mode.

Solution The mode is 0 because it occurs most frequently (17 times).



EXAMPLE 8

Finding the Mode of Quantitative Data

Problem Find the mode of the exam score data listed in Table 1, which is repeated here:

82, 77, 90, 71, 62, 68, 74, 84, 94, 88

Approach Tally the number of times we observe each data value. The data value with the highest frequency is the mode.

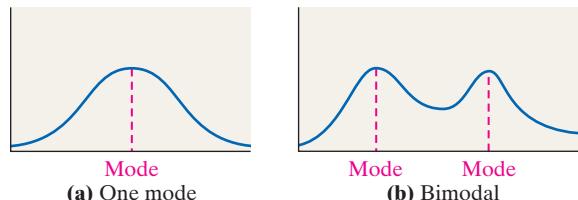
Solution Since each data value occurs only once, there is no mode.

1

NW Now compute the mode of the data in Problem 15.

A data set can have more than one mode. For example, if the data set in Table 1 had two scores of 77 and 88, the data set would have two modes: 77 and 88. In this case, we say the data are **bimodal**. If a data set has three or more data values that occur with the highest frequency, the data set is **multimodal**. The mode is usually not reported for multimodal data because it is not representative of a typical value. Figure 9(a) shows a distribution with one mode. Figure 9(b) shows a distribution that is bimodal.

Figure 9

**IN OTHER WORDS**

Remember, nominal data are qualitative data that cannot be written in any meaningful order.

We cannot determine the value of the mean or median of data that are nominal. The only measure of central tendency that can be determined for nominal data is the mode.

EXAMPLE 9**Determining the Mode of Qualitative Data**

Problem The data in Table 6 represent the location of injuries that required rehabilitation by a physical therapist. Determine the mode location of injury.

Table 6

Hip	Back	Back	Back	Hand	Neck	Knee	Knee	Knee	Hand
Knee	Shoulder	Wrist	Back	Groin	Shoulder	Shoulder	Back	Knee	Back
Hip	Shoulder	Elbow	Back	Back	Back	Back	Back	Back	Wrist

Source: Krystal Catton, student at Joliet Junior College

Approach Determine the location of injury that occurs with the highest frequency.

NW Now Work Problem 29

Solution The mode location of injury is the back, with 12 instances.

Summary: Measures of Central Tendency

Measure of Central Tendency	Computation	Interpretation	When to Use
Mean	Population mean: $\mu = \frac{\sum x_i}{N}$ Sample mean: $\bar{x} = \frac{\sum x_i}{n}$	Center of gravity	When data are quantitative and the frequency distribution is roughly symmetric
Median	Arrange data in ascending order and determine the value of the observation that lies in the middle.	Divides the bottom 50% of the data from the top 50%	When the data are quantitative and the frequency distribution is skewed left or right
Mode	Tally data to determine most frequent observation	Most frequent observation	When the most frequent observation is the desired measure of central tendency or the data are qualitative

Technology Step-by-Step

Determining the Mean and Median

TI-83/84 Plus

- Enter raw data in L1 by pressing STAT and then selecting 1:Edit.
- Press STAT, highlight the CALC menu, and select 1:1-Var Stats. In the menu, select L1 for List. Clear the entry in FreqList. Highlight Calculate and press ENTER.

Minitab

- Enter the data in C1.
- Select the Stat menu, highlight Basic Statistics, and then select Display Descriptive Statistics
- In the Variables window, enter C1. Click OK.

Excel

- Enter the data in column A.
- Be sure you have the “Data Analysis” ToolPak loaded. Select the Data menu and click Data Analysis.

- In the Data Analysis window, highlight Descriptive Statistics and click OK.

- With the cursor in the Input Range window, use the mouse to highlight the data in column A.

- Select the Summary statistics option and click OK.

StatCrunch

- Enter the raw data into the spreadsheet. Name the column variable.
- Select Stat, highlight Summary Stats, and select Columns.
- Click on the variable you want to summarize. If you wish to compute certain statistics, hold down the Control (Ctrl) key when selecting the statistic (or Command on an Apple). Click Compute!.



3.1 Assess Your Understanding

Vocabulary and Skill Building

- What does it mean if a statistic is resistant?
- The Federal Reserve Bank of St. Louis reported two average personal incomes for 2018: \$33,706 and \$50,731. One of these averages is the mean and the other is the median. Which is the mean? Support your answer.
- The U.S. Department of Housing and Urban Development (HUD) uses the median to report the average price of a home in the United States. Why do you think HUD uses the median?
- A histogram of a set of data indicates that the distribution of the data is skewed right. Which measure of central tendency will likely be larger, the mean or the median? Why?
- If a data set contains 10,000 values arranged in increasing order, where is the median located?
- True or False:* A data set will always have exactly one mode.

In Problems 7–10, find the population mean or sample mean as indicated.

- Sample: 20, 13, 4, 8, 10
- Sample: 83, 65, 91, 87, 84
- Population: 3, 6, 10, 12, 14
- Population: 1, 19, 25, 15, 12, 16, 28, 13, 6
- For Super Bowl LIII, there were 62,612 tickets sold for a total value of \$108,005,700. What was the mean price per ticket?

- The median for the given set of six ordered data values is 26.5. What is the missing value? 7 12 21 ____ 41 50

- Miles per Gallon** The following data represent the miles per gallon for a 2013 Ford Fusion for six randomly selected vehicles. Compute the mean, median, and mode miles per gallon.

Source: www.fueleconomy.gov

34.0, 33.2, 37.0, 29.4, 23.6, 25.9

- Exam Time** The following data represent the amount of time (in minutes) a random sample of eight students took to complete the online portion of an exam in Sullivan’s Statistics course. Compute the mean, median, and mode time.

60.5, 128.0, 84.6, 122.3, 78.9, 94.7, 85.9, 89.9

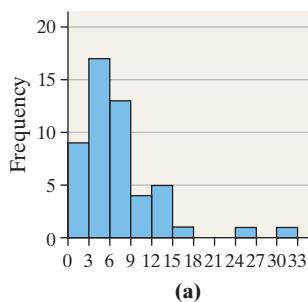
- NW 15. Concrete Mix** A certain type of concrete mix is designed to withstand 3000 pounds per square inch (psi) of pressure. The concrete is poured into casting cylinders and allowed to set for 28 days. The concrete’s strength is then measured. The following data represent the strength of nine randomly selected casts (in psi). Compute the mean, median, and mode strength of the concrete (in psi).

3960, 4090, 3200, 3100, 2940, 3830, 4090, 4040, 3780

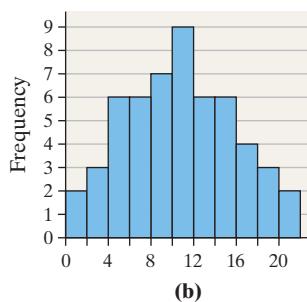
- Flight Time** The following data represent the flight time (in minutes) of a random sample of seven flights from Las Vegas, Nevada, to Newark, New Jersey, on United Airlines. Compute the mean, median, and mode flight time.

282, 270, 260, 266, 257, 260, 267

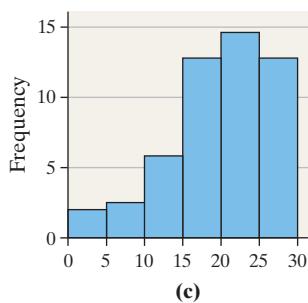
- 17.** For each of the three histograms shown, determine whether the mean is greater than, less than, or approximately equal to the median. Justify your answer.



(a)



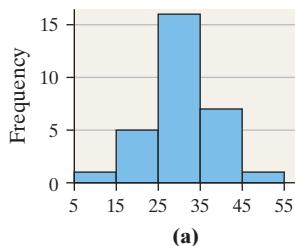
(b)



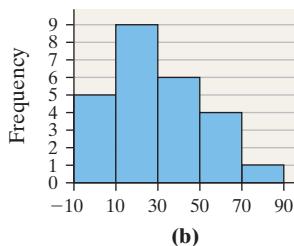
(c)

- 18.** Match the histograms shown to the summary statistics:

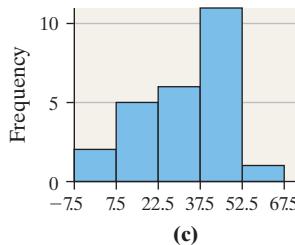
	Mean	Median
I	42	42
II	31	36
III	31	26
IV	31	32



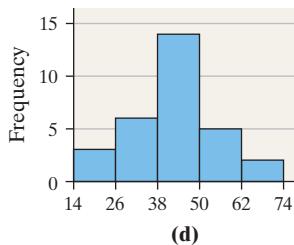
(a)



(b)



(c)



(d)

Applying the Concepts

- 19. Exam Scores** The data in the next column represent exam scores in a statistics class taught using traditional lecture and a class taught using a “flipped” classroom. The “flipped” classroom is one where the content is delivered via video and watched at home, while class time is used for homework and activities.

Traditional	70.8	69.1	79.4	67.6	85.3	78.2	56.2
	81.3	80.9	71.5	63.7	69.8	59.8	
Flipped	76.4	71.6	63.4	72.4	77.9	91.8	78.9
	76.8	82.1	70.2	91.5	77.8	76.5	

Source: Michael Sullivan.

- (a)** Determine the mean and median score for each class.

Comment on any differences.

- (b)** Suppose the score of 59.8 in the traditional course was incorrectly recorded as 598. How does this affect the mean? the median? What property does this illustrate?

- DATA 20. pH in Water** The acidity or alkalinity of a solution is measured using pH. A pH less than 7 is acidic; a pH greater than 7 is alkaline. The following data represent the pH in samples of bottled water and tap water.

Tap	7.64	7.45	7.47	7.50	7.68	7.69
	7.45	7.10	7.56	7.47	7.52	7.47
Bottled	5.15	5.09	5.26	5.20	5.02	5.23
	5.28	5.26	5.13	5.26	5.21	5.24

Source: Emily McCarney, student at Joliet Junior College.

- (a)** Determine the mean, median, and mode pH for each type of water. Comment on the differences between the two water types.

- (b)** Suppose the pH of 7.10 in tap water was incorrectly recorded as 1.70. How does this affect the mean? the median? What property of the median does this illustrate?

- NW 21. Pulse Rates** The following data represent the pulse rates (beats per minute) of nine students enrolled in a section of Sullivan’s Introductory Statistics course. Treat the nine students as a population.

Student	Pulse
Perpetual Bempah	76
Megan Brooks	60
Jeff Honeycutt	60
Clarice Jefferson	81
Crystal Kurtenbach	72
Janette Lantka	80
Kevin McCarthy	80
Tammy Ohm	68
Kathy Wojdyla	73

- (a)** Determine the population mean pulse.

- (b)** Find three simple random samples of size 3 and determine the sample mean pulse of each sample.

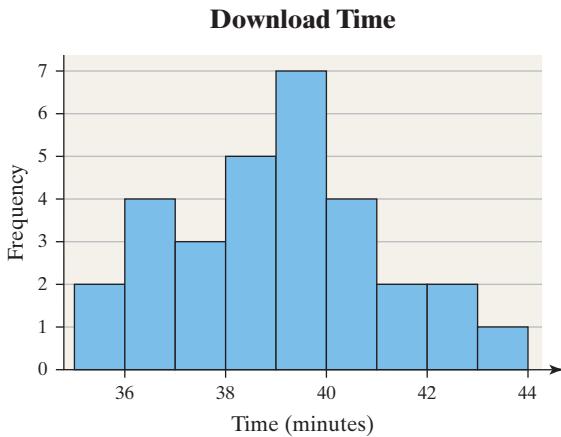
- (c)** Which samples result in a sample mean that overestimates the population mean? Which samples result in a sample mean that underestimates the population mean? Do any samples lead to a sample mean that equals the population mean?

- DATA 22. Travel Time** The data on the next page represent the travel time (in minutes) to school for nine students enrolled in Sullivan’s College Algebra course. Treat the nine students as a population.

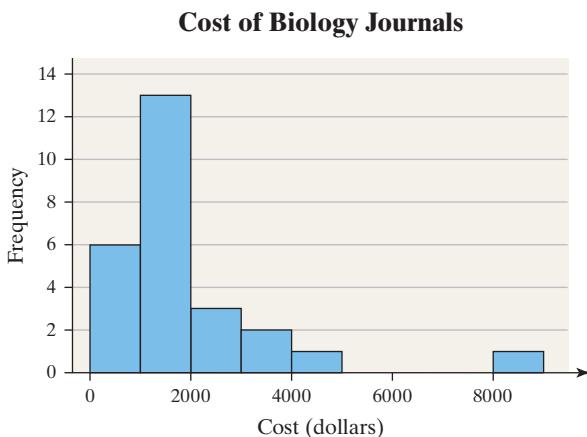
Student	Travel Time	Student	Travel Time
Amanda	39	Scot	45
Amber	21	Erica	11
Tim	9	Tiffany	12
Mike	32	Glenn	39
Nicole	30		

- (a) Determine the population mean travel time.
 (b) Find three simple random samples of size 4 and determine the sample mean travel time of each sample.
 (c) Which samples result in a sample mean that overestimates the population mean? Which samples result in a sample mean that underestimates the population mean? Do any samples lead to a sample mean that equals the population mean?

NW 23. Download Time A histogram of the download time of a movie, in minutes, for 30 randomly selected movies is shown. The mean download time is 39.007 minutes and the median download time is 39.065 minutes. Identify the shape of the distribution. Which measure of central tendency better describes the “center” of the distribution?



24. Journal Costs A histogram of the annual subscription cost (in dollars) for 26 biology journals is shown. The mean subscription cost is \$1846 and the median subscription cost is \$1142. Identify the shape of the distribution. Which measure of central tendency better describes the “center” of the distribution?



Source: Carol Wesolowski, student at Joliet Junior College.

DATA 25. M&Ms The following data represent the weights (in grams) of a simple random sample of 50 M&M plain candies.

0.87	0.88	0.82	0.90	0.90	0.84	0.84
0.91	0.94	0.86	0.86	0.86	0.88	0.87
0.89	0.91	0.86	0.87	0.93	0.88	
0.83	0.95	0.87	0.93	0.91	0.85	
0.91	0.91	0.86	0.89	0.87	0.84	
0.88	0.88	0.89	0.79	0.82	0.83	
0.90	0.88	0.84	0.93	0.81	0.90	
0.88	0.92	0.85	0.84	0.84	0.86	

Source: Michael Sullivan.

Determine the shape of the distribution of weights of M&Ms by drawing a frequency histogram. Find the mean and median. Which measure of central tendency better describes the weight of a plain M&M?

DATA 26. Old Faithful We have all heard of the Old Faithful geyser in Yellowstone National Park. However, there is another, less famous, Old Faithful geyser in Calistoga, California. The following data represent the length of eruption (in seconds) for a random sample of eruptions of the California Old Faithful.

108	108	99	105	103	103	94
102	99	106	90	104	110	110
103	109	109	111	101	101	
110	102	105	110	106	104	
104	100	103	102	120	90	
113	116	95	105	103	101	
100	101	107	110	92	108	

Source: Ladonna Hansen, Park Curator.

Determine the shape of the distribution of length of eruption by drawing a frequency histogram. Find the mean and median. Which measure of central tendency better describes the length of eruption?

DATA 27. Wait Time The following data represent the wait time (in minutes) for a random sample of 40 visitors to Disney's Dinosaur Ride in Animal Kingdom. Determine the shape of the distribution of wait time by drawing a frequency histogram. Find the mean and median wait time. Which measure of central tendency better describes wait time?

6	31	8	0	21	16	0	7
15	6	44	27	7	52	3	7
4	5	10	5	21	3	6	14
5	24	10	9	9	10	12	8
4	8	39	5	28	30	4	15

Source: touringplans.com

DATA 28. Online Shopping The following data represent the number of days between grocery orders at the online delivery company Instacart. Determine the shape of the distribution of days between orders by drawing a frequency histogram. Find the mean and median days between orders.

Which measure of central tendency better describes days between orders?

14	5	7	30	14	30	15	3
1	30	11	28	9	6	6	6
8	5	6	6	14	15	10	12
5	5	8	30	7	10	7	5
30	30	29	5	4	14	6	30

Source: Instacart.

- NW DATA 29. Political Views** A sample of 30 registered voters was surveyed in which the respondents were asked, "Do you consider your political views to be conservative, moderate, or liberal?" The results of the survey are shown in the table.

Liberal	Conservative	Moderate
Moderate	Liberal	Moderate
Liberal	Moderate	Conservative
Moderate	Conservative	Moderate
Moderate	Moderate	Liberal
Liberal	Moderate	Liberal
Conservative	Moderate	Moderate
Liberal	Conservative	Liberal
Liberal	Conservative	Liberal
Conservative	Moderate	Conservative

Source: Based on data from the General Social Survey.

- (a) Determine the mode political view.
 (b) Do you think it would be a good idea to rotate the choices conservative, moderate, or liberal in the question? Why?

- DATA 30. Hospital Admissions** The following data represent the diagnosis of a random sample of 20 patients admitted to a hospital. Determine the mode diagnosis.

Cancer	Motor vehicle accident	Congestive heart failure
Gunshot wound	Fall	Gunshot wound
Gunshot wound	Motor vehicle accident	Gunshot wound
Assault	Motor vehicle accident	Gunshot wound
Motor vehicle accident	Motor vehicle accident	Gunshot wound
Motor vehicle accident	Gunshot wound	Motor vehicle accident
Fall	Gunshot wound	

Source: Tamela Ohm, student at Joliet Junior College.

- DATA 31. Resistance and Sample Size** Each of the following three data sets in the next column represents the IQ scores of a random sample of adults. IQ scores are known to have a mean and median of 100. For each data set, determine the mean and median. For each data set recalculate the mean and median, assuming that the individual whose IQ is 106 is accidentally recorded as 160. For each sample size, state what happens to the mean and the median. Comment on the role the number of observations plays in resistance.

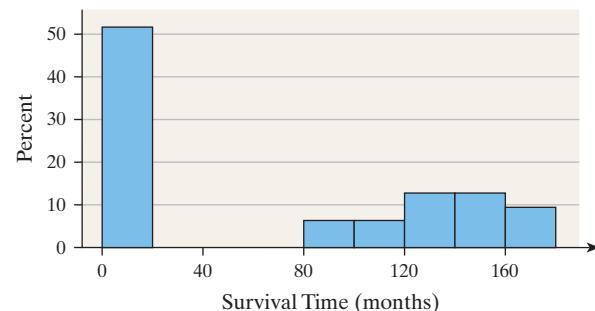
Sample of Size 5				
106	92	98	103	100
Sample of Size 12				
106	92	98	103	100
98	124	83	70	108
Sample of Size 30				
106	92	98	103	100
98	124	83	70	108
102	87	121	107	97
140	93	130	72	81
103	97	89	98	88

- 32.** Mr. Zuro finds the mean height of all 14 students in his statistics class to be 68.0 inches. Just as Mr. Zuro finishes explaining how to get the mean, Danielle walks in late. Danielle is 65 inches tall. What is the mean height of the 15 students in the class including Danielle?

- 33. Missing Exam Grade** A professor has recorded exam grades for 20 students in his class, but one of the grades is no longer readable. If the mean score on the exam was 82 and the mean of the 19 readable scores is 84, what is the value of the unreadable score?

- 34. Survival Rates** Unfortunately, a friend of yours has been diagnosed with cancer. A histogram of the survival time (in months) of patients diagnosed with this form of cancer is shown in the figure; the median survival time is 11 months, while the mean survival time is 69 months. What words of encouragement should you share with your friend from a statistical point of view?

Survival Time of Patients



- DATA 35. Blood Alcohol Concentration** Go to <http://www.pearsonhighered.com/sullivanstats> to obtain the data 3_1_35. The data represent the blood alcohol concentration (BAC), in percent, of a random sample of drivers involved in fatal car accidents. A BAC of 0 indicates that no alcohol was present. Draw a histogram of the data, describe the shape, and determine the mean and median BAC of drivers in fatal accidents. Which measure of central tendency better describes the typical BAC of drivers in fatal accidents? Explain.

- DATA 36. Sullivan Survey** Go to <http://www.pearsonhighered.com/sullivanstats> and download the SullivanStatsSurvey I data. The data represent the results of a survey conducted by the author. The column "Texts" represents the number of texts the individual sent for the month prior to the survey. Draw a relative

frequency histogram and determine the mean and median number of texts for each gender. Make some general comments about the results.

- 37. Linear Transformations** Benjamin owns a small Internet business. Besides himself, he employs nine other people. The salaries earned by the employees are given next in thousands of dollars (Benjamin's salary is the largest, of course):

$$30, 30, 45, 50, 50, 50, 55, 55, 60, 75$$

- (a) Determine the mean, median, and mode for salary.
- (b) Business has been good! As a result, Benjamin has a total of \$25,000 in bonus pay to distribute to his employees. One option for distributing bonuses is to give each employee (including himself) \$2500. Add the bonuses under this plan to the original salaries to create a new data set. Recalculate the mean, median, and mode. How do they compare to the originals?
- (c) As a second option, Benjamin can give each employee a bonus of 5% of his or her original salary. Add the bonuses under this second plan to the original salaries to create a new data set. Recalculate the mean, median, and mode. How do they compare to the originals?
- (d) As a third option, Benjamin decides not to give his employees a bonus at all. Instead, he keeps the \$25,000 for himself. Use this plan to create a new data set. Recalculate the mean, median, and mode. How do they compare to the originals?

- 38. Linear Transformations** Use the five test scores of 65, 70, 71, 75, and 95 to answer the following questions:

- (a) Find the sample mean.
- (b) Find the median.
- (c) Which measure of central tendency better describes the typical test score?
- (d) Suppose the professor decides to curve the exam by adding 4 points to each test score. Compute the sample mean based on the adjusted scores.
- (e) Compare the unadjusted test score mean with the curved test score mean. What effect did adding 4 to each score have on the mean?

- 39. Trimmed Mean** Another measure of central tendency is the trimmed mean. It is computed by determining the mean of a data set after deleting the smallest and largest observed values. Compute the trimmed mean for the data in Problem 25. Is the trimmed mean resistant? Explain.

- 40. Midrange** The midrange is also a measure of central tendency. It is computed by adding the smallest and largest observed values of a data set and dividing the result by 2; that is,

$$\text{Midrange} = \frac{\text{largest data value} + \text{smallest data value}}{2}$$

Compute the midrange for the data in Problem 25. Is the midrange resistant? Explain.

- 41. Threaded Problem: Tornado** The data set "Tornadoes_2017" located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a) Determine the mean and median length for all tornadoes. What shape would you expect a histogram of tornado length to have? Why? Confirm your answer by consulting the histogram drawn in Problem 41(a) from Section 2.2.

- (b) Determine the mean and median length for tornadoes in Texas (TX). **Hint:** If you are using StatCrunch, enter "State=TX" in the Where: box of the Summary Stats dialogue window. Determine the mean and median length for tornadoes in Georgia (GA). Which state has longer tornadoes?

- (c) Determine the mean and median property loss for all tornadoes. What does this result suggest? Draw a graph of the data and explore the cause of this result.
- (d) Does it make sense to determine the mean F scale? Explain.

- 42. Putting It Together: Shape, Mean and Median** As part of a semester project in a statistics course, Carlos surveyed a sample of 50 high school students and asked, "How many days in the past week have you consumed an alcoholic beverage?" The results of the survey are shown next.

0	0	1	4	1	1	1	5	1	3
0	1	0	1	0	4	0	1	0	1
0	0	0	0	2	0	0	0	0	0
1	0	2	0	0	0	1	2	1	1
2	0	1	0	1	3	1	1	0	3

- (a) Is this data discrete or continuous?
- (b) Draw a histogram of the data and describe its shape.
- (c) Based on the shape of the histogram, do you expect the mean to be more than, equal to, or less than the median? –
- (d) Determine the mean and the median. What does this tell you?
- (e) Determine the mode.
- (f) Do you believe that Carlos's survey suffers from sampling bias? Why?

Explaining the Concepts

- 43. FICO Scores** The Fair Isaacs Corporation has devised a model that is used to determine creditworthiness of individuals, called a FICO score. FICO scores range in value from 300 to 850, with a higher score indicating a more creditworthy individual. The distribution of FICO scores is skewed left with a median score of 723.

- (a) Do you think the mean FICO score is greater than, less than, or equal to 723? Justify your response.
- (b) What proportion of individuals have a FICO score above 723?

- 44. Why is the median resistant, but the mean is not?**

- 45. A researcher with the Department of Energy wants to determine the mean natural gas bill of households throughout the United States. He knows the mean natural gas bill of households for each state, so he adds together these 50 values and divides by 50 to arrive at his estimate. Is this a valid approach? Why or why not?**

- 46. Net Worth** According to MarketWatch, the mean net worth of all individuals in the United States in 2016 was \$692,100, while the median net worth was \$97,300.

- (a) Which measure do you believe better describes the typical individual's net worth? Support your opinion.
- (b) What shape would you expect the distribution of net worth to have? Why?
- (c) What do you think causes the disparity in the two measures of central tendency?

- 47.** You are negotiating a contract for the Players Association of the NBA. Which measure of central tendency will you use to support your claim that the average player's salary needs to be increased? Why? As the chief negotiator for the owners, which measure would you use to refute the claim made by the Players Association?
- 48.** In January 2020, the mean amount of money lost per visitor to a local riverboat casino was \$135. Do you think the median was more than, less than, or equal to this amount? Why?
- 49.** For each of the following situations, determine which measure of central tendency is most appropriate and justify your reasoning.

- (a) Average price of a home sold in Pittsburgh, Pennsylvania, in 2019
- (b) Most popular major for students enrolled in a statistics course
- (c) Average test score when the scores are distributed symmetrically
- (d) Average test score when the scores are skewed right
- (e) Average income of a player in the National Football League
- (f) Most requested song at a radio station
- (g) Typical number on the jersey of a player in the National Hockey League.

3.2 Measures of Dispersion



Objectives

- 1 Determine the range of a variable from raw data
- 2 Determine the standard deviation of a variable from raw data
- 3 Determine the variance of a variable from raw data
- 4 Use the Empirical Rule to describe data that are bell shaped
- 5 Use Chebyshev's Inequality to describe any set of data

In Section 3.1, we discussed measures of central tendency. These measures describe the typical *value* of a variable. We would also like to know the amount of *dispersion* in the variable. **Dispersion** is the degree to which the data are spread out. Example 1 demonstrates why measures of central tendency alone are not sufficient in describing a distribution.

EXAMPLE 1 Comparing Two Sets of Data

Problem The data in Table 7 represent the IQ scores of a random sample of 100 students from two different universities. For each university, compute the mean IQ score and draw a histogram using a lower class limit of 55 for the first class and a class width of 15. Comment on the results.

Table 7

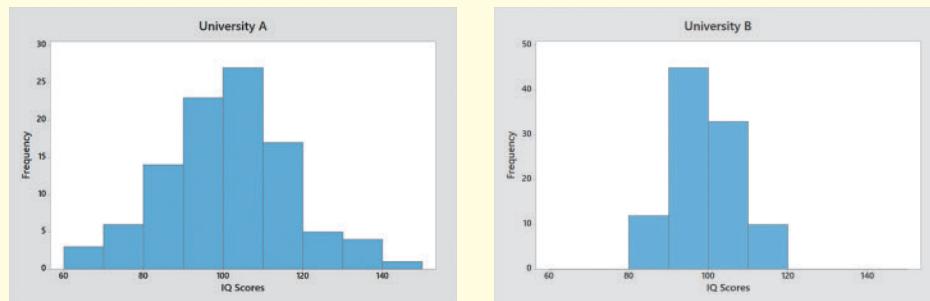
University A										University B									
73	103	91	93	136	108	92	104	90	78	86	91	107	94	105	107	89	96	102	96
108	93	91	78	81	130	82	86	111	93	92	109	103	106	98	95	97	95	109	109
102	111	125	107	80	90	122	101	82	115	93	91	92	91	117	108	89	95	103	109
103	110	84	115	85	83	131	90	103	106	110	88	97	119	90	99	96	104	98	95
71	69	97	130	91	62	85	94	110	85	87	105	111	87	103	92	103	107	106	97
102	109	105	97	104	94	92	83	94	114	107	108	89	96	107	107	96	95	117	97
107	94	112	113	115	106	97	106	85	99	98	89	104	99	99	87	91	105	109	108
102	109	76	94	103	112	107	101	91	107	116	107	90	98	98	92	119	96	118	98
107	110	106	103	93	110	125	101	91	119	97	106	114	87	107	96	93	99	89	94
118	85	127	141	129	60	115	80	111	79	104	88	99	97	106	107	112	97	94	107

Approach We will use Minitab to compute the mean and draw a histogram for each university.

(continued)

Solution Enter the data into Minitab and determine that the mean IQ score of both universities is 100.0. Figure 10 shows the histograms.

Figure 10



Both universities have the same mean IQ, but the histograms suggest the IQs from University A are more spread out, that is, more dispersed. While an IQ of 100.0 is typical for both universities, it appears to be a more reliable description of the typical student from University B than from University A. That is, a higher proportion of students from University B have IQ scores within, say, 10 points of the mean of 100.0 than students from University A.



The goal of this section is to discuss numerical measures of dispersion so that the spread of data may be quantified. Three numerical measures for describing the dispersion, or spread, of data will be discussed: the *range*, *standard deviation*, and *variance*. In Section 3.4, we will discuss another measure of dispersion, the *interquartile range* (IQR).

1 Determine the Range of a Variable from Raw Data

The simplest measure of dispersion is the range. To compute the range, the data must be quantitative.

Definition

The **range**, R , of a variable is the difference between the largest and the smallest data value. That is,

$$\text{Range} = R = \text{largest data value} - \text{smallest data value}$$

EXAMPLE 2

Computing the Range of a Set of Data

Table 8

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

NW Now compute the range of the data in Problem 13

Problem The data in Table 8 represent the scores on the first exam of 10 students enrolled in Introductory Statistics. Compute the range.

Approach The range is the difference between the largest and smallest data values.

Solution The highest test score is 94 and the lowest test score is 62. The range is

$$R = 94 - 62 = 32$$

All the students in the class scored between 62 and 94 on the exam. The difference between the best score and the worst score is 32 points.



Notice that the range is affected by extreme values in the data set, so the range is not resistant. If Jennifer scored 28, the range becomes $R = 94 - 28 = 66$. Also, the range is computed using only two values in the data set (the largest and smallest). The *standard deviation*, on the other hand, uses all the data values in the computations.

② Determine the Standard Deviation of a Variable from Raw Data

NOTE

Recall, $|a| = a$ if $a \geq 0$, and $|a| = -a$ if $a < 0$ so $|3| = 3$ and $|-3| = 3$.

NOTE

$(-3)^2 = 9$, so squaring a negative number results in a positive number.

Measures of dispersion are meant to describe how spread out data are. In other words, they describe how far, on average, each observation is from the typical data value. Standard deviation is based on the **deviation about the mean**. For a population, the deviation about the mean for the i th observation is $x_i - \mu$. For a sample, the deviation about the mean for the i th observation is $x_i - \bar{x}$. The further an observation is from the mean, the larger the absolute value of the deviation.

The sum of all deviations about the mean must equal zero. That is,

$$\Sigma(x_i - \mu) = 0 \quad \text{and} \quad \Sigma(x_i - \bar{x}) = 0$$

This result follows from the fact that observations greater than the mean are offset by observations less than the mean. Because this sum is zero, we cannot use the average deviation about the mean as a measure of spread. There are two possible solutions to this “problem.” We could either find the mean of the absolute values of the deviations about the mean, or we could find the mean of the squared deviations because squaring a nonzero number always results in a positive number. The first approach yields a measure of dispersion called the mean absolute deviation (MAD) (see Problem 45). The second approach leads to *variance*. The problem with variance is that squaring the deviations about the mean leads to squared units of measure, such as dollars squared. It is difficult to have a reasonable interpretation of dollars squared, so we “undo” the squaring process by taking the square root of the sum of squared deviations. We have the following definition for the *population standard deviation*.

Definition

The **population standard deviation** of a variable is the square root of the sum of squared deviations about the population mean divided by the number of observations in the population, N . That is, it is the square root of the mean of the squared deviations about the population mean.

NOTE

The symbol $\sqrt{}$ is called a radical. It tells us to find the number whose square is the value under the radical, called the radicand. So, in $\sqrt{25}$, 25 is the radicand. And,

$$\sqrt{25} = 5$$

because

$$5^2 = 25$$

The population standard deviation is symbolically represented by σ (lowercase Greek sigma).

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}} = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}} \quad (1)$$

where x_1, x_2, \dots, x_N are the N observations in the population and μ is the population mean.

Formula (1) is sometimes referred to as the **conceptual formula** because it allows us to see how standard deviation measures spread.

A formula that is equivalent to Formula (1), called the **computational formula**, for determining the population standard deviation is

$$\sigma = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N}} \quad (2)$$

where $\sum x_i^2$ means to square each observation and then sum these squared values, and $(\sum x_i)^2$ means to add up all the observations and then square the sum.

Example 3 illustrates how to use both formulas.

EXAMPLE 3 Computing a Population Standard Deviation

Problem Compute the population standard deviation of the test scores in Table 8.

Approach Using Conceptual Formula (1)

Step 1 Create a table with four columns. Enter the population data in Column 1. In Column 2, enter the population mean.

Step 2 Compute the deviation about the mean for each data value, $x_i - \mu$. Enter the result in Column 3.

Step 3 In Column 4, enter the squares of the values in Column 3.

Step 4 Sum the entries in Column 4, and divide this result by the size of the population, N .

Step 5 Determine the square root of the value found in Step 4.

Solution Using Formula (1)

Step 1 See Table 9. Column 1 lists the observations in the data set, and Column 2 contains the population mean.

Table 9

Score, x_i	Population Mean, μ	Deviation about the Mean, $x_i - \mu$	Squared Deviations about the Mean, $(x_i - \mu)^2$
82	79	$82 - 79 = 3$	$3^2 = 9$
77	79	$77 - 79 = -2$	$(-2)^2 = 4$
90	79	11	121
71	79	-8	64
62	79	-17	289
68	79	-11	121
74	79	-5	25
84	79	5	25
94	79	15	225
88	79	9	81
$\Sigma(x_i - \mu) = 0$		$\Sigma(x_i - \mu)^2 = 964$	

Step 2 Column 3 contains the deviations about the mean for each observation. For example, the deviation about the mean for Michelle is $82 - 79 = 3$. It is a good idea to add the entries in this column to make sure they sum to 0.

Step 3 Column 4 shows the squared deviations about the mean. Notice the farther an observation is from the mean, the larger the squared deviation.

Step 4 Sum the entries in Column 4 to obtain $\Sigma(x_i - \mu)^2$. Divide this sum by the number of students, 10:

$$\frac{\Sigma(x_i - \mu)^2}{N} = \frac{964}{10} = 96.4 \text{ points}^2$$

Step 5 The square root of the result in Step 4 is the population standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}} = \sqrt{96.4 \text{ points}^2} \approx 9.8 \text{ points}$$

Approach Using Computational Formula (2)

Step 1 Create a table with two columns. Enter the population data in Column 1. Square each value in Column 1 and enter the result in Column 2.

Step 2 Sum the entries in Column 1. That is, find Σx_i . Sum the entries in Column 2. That is, find Σx_i^2 .

Step 3 Substitute the values found in Step 2 and the value for N into the computational formula and simplify.

Solution Using Formula (2)

Step 1 See Table 10. Column 1 lists the observations in the data set, and Column 2 contains the values in Column 1 squared.

Table 10

Score, x_i	Score Squared, x_i^2
82	$82^2 = 6724$
77	$77^2 = 5929$
90	8100
71	5041
62	3844
68	4624
74	5476
84	7056
94	8836
88	7744
$\Sigma x_i = 790$	$\Sigma x_i^2 = 63,374$

Step 2 The last rows of Columns 1 and 2 show that $\Sigma x_i = 790$ and $\Sigma x_i^2 = 63,374$.

Step 3 Substitute 790 for Σx_i , 63,374 for Σx_i^2 , and 10 for N into Formula (2):

$$\begin{aligned}\sigma &= \sqrt{\frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{N}}{N}} = \sqrt{\frac{63,374 - \frac{(790)^2}{10}}{10}} \\ &= \sqrt{\frac{964}{10}} \\ &= \sqrt{96.4 \text{ points}^2} \\ &\approx 9.8 \text{ points}\end{aligned}$$



NOTE

The symbol \approx means “approximately.”

Look at Table 9 in Example 3. The further an observation is from the mean, 79, the larger the squared deviation. For example, because the second observation, 77, is not “far” from 79, the squared deviation, 4, is not large. However, the fifth observation, 62, is further from 79, so the squared deviation, 289, is much larger.

So, if a data set has many observations that are “far” from the mean, the sum of the squared deviations will be large, and therefore the standard deviation will be large. This is how standard deviation measures dispersion, or spread.

Now let’s look at the definition of the *sample standard deviation*.

Definition**CAUTION!**

When using Formula (3), be sure to use \bar{x} with as many decimal places as possible to avoid round-off error.

The **sample standard deviation**, s , of a variable is the square root of the sum of squared deviations about the sample mean divided by $n - 1$, where n is the sample size.

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} \quad (3)$$

where x_1, x_2, \dots, x_n are the n observations in the sample and \bar{x} is the sample mean.

CAUTION!

When computing the sample standard deviation, be sure to divide by $n - 1$, not n .

Formula (3) is often referred to as the conceptual formula for determining the sample standard deviation.

A computational formula that is equivalent to Formula (3) for computing the sample standard deviation is

$$s = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n - 1}} \quad (4)$$

Notice that the sample standard deviation is obtained by dividing by $n - 1$. Although showing why we divide by $n - 1$ is beyond the scope of the text, the following explanation has intuitive appeal. We already know that the sum of the deviations about the mean, $\sum (x_i - \bar{x})$, must equal zero. Therefore, if the sample mean is known and the first $n - 1$ observations are known, then the n th observation must be the value that causes the sum of the deviations to equal zero. For example, suppose $\bar{x} = 4$ based on a sample of size 3. In addition, if $x_1 = 2$ and $x_2 = 3$, then we can determine x_3 .

$$\begin{aligned} \frac{x_1 + x_2 + x_3}{3} &= \bar{x} \\ \frac{2 + 3 + x_3}{3} &= 4 \quad x_1 = 2, x_2 = 3, \bar{x} = 4 \\ 5 + x_3 &= 12 \\ x_3 &= 7 \end{aligned}$$

IN OTHER WORDS

We have $n - 1$ degrees of freedom in the computation of s because an unknown parameter, μ , is estimated with \bar{x} . For each parameter estimated, we lose 1 degree of freedom.

We call $n - 1$ the **degrees of freedom** because the first $n - 1$ observations have freedom to be whatever value they wish, but the n th value has no freedom. It must be whatever value forces the sum of the deviations about the mean to equal zero.

Again, you should notice that Greek letters are typically used for parameters, while Roman letters are used for statistics. Do not use rounded values of the sample mean in Formula (3).

EXAMPLE 4**Computing a Sample Standard Deviation**

Problem Compute the sample standard deviation of the sample obtained in Example 1(b) on page 109 from Section 3.1.

Approach We follow the same approach that we used to compute the population standard deviation, but this time using the sample data. In looking back at Example 1(b) from Section 3.1, we see that Jennifer (62), Juan (88), Ryanne (77), and Dave (68) are in the sample.

(continued)

Solution Using Conceptual Formula (3)

Step 1 Create a table with four columns. Enter the sample data in Column 1. In Column 2, enter the unrounded sample mean. See Table 11.

Table 11

Score, x_i	Sample Mean, \bar{x}	Deviation about the Mean, $x_i - \bar{x}$	Squared Deviations about the Mean, $(x_i - \bar{x})^2$
62	73.75	$62 - 73.75 = -11.75$	$(-11.75)^2 = 138.0625$
88	73.75	14.25	203.0625
77	73.75	3.25	10.5625
68	73.75	-5.75	33.0625
$\Sigma(x_i - \bar{x}) = 0$		$\Sigma(x_i - \bar{x})^2 = 384.75$	

Step 2 Column 3 contains the deviations about the mean for each observation. For example, the deviation about the mean for Jennifer is $62 - 73.75 = -11.75$. It is a good idea to add the entries in this column to make sure they sum to 0.

Step 3 Column 4 shows the squared deviations about the mean.

Step 4 Sum the entries in Column 4 to obtain $\Sigma(x_i - \bar{x})^2$. Divide the sum of the entries in Column 4 by one fewer than the number of students, $4 - 1$:

$$\frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{384.75}{4 - 1} = 128.25 \text{ points}^2$$

Step 5 The square root of the result in Step 4 is the sample standard deviation.

$$\begin{aligned}s &= \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} \\&= \sqrt{128.25 \text{ points}^2} \\&\approx 11.3 \text{ points}\end{aligned}$$

Solution Using Computational Formula (4)

Step 1 See Table 12. Column 1 lists the observations in the data set, and Column 2 contains the values in Column 1 squared.

Table 12

Score, x_i	Score squared, x_i^2
62	$62^2 = 3844$
88	$88^2 = 7744$
77	5929
68	4624
$\Sigma x_i = 295$	$\Sigma x_i^2 = 22,141$

Step 2 The last rows of Columns 1 and 2 show that $\Sigma x_i = 295$ and $\Sigma x_i^2 = 22,141$.

Step 3 Substitute 295 for Σx_i , 22,141 for Σx_i^2 , and 4 for n into the computational formula (4).

$$\begin{aligned}s &= \sqrt{\frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}{n - 1}} = \sqrt{\frac{22,141 - \frac{(295)^2}{4}}{4 - 1}} \\&= \sqrt{\frac{384.75}{3}} \\&= \sqrt{128.25 \text{ points}^2} \\&\approx 11.3 \text{ points}\end{aligned}$$

EXAMPLE 5

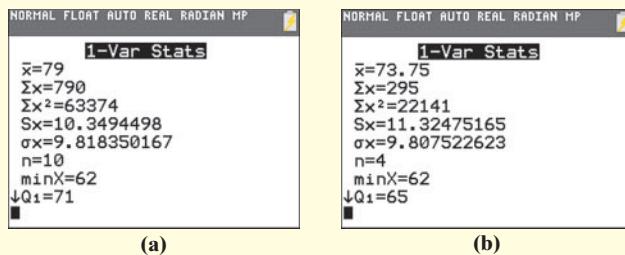
Determining the Standard Deviation Using Technology

Problem Use statistical software or a calculator to determine the population standard deviation of the data listed in Table 8 and the sample standard deviation of the sample data from Example 4.

Approach We will use a TI-84 Plus CE graphing calculator. The steps for determining the standard deviation using the TI-83 or TI-84 Plus graphing calculator, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on page 132.

Solution Figure 11(a) shows the population standard deviation, and Figure 11(b) shows the sample standard deviation. Notice that the TI graphing calculators provide both a population and sample standard deviation as output. This is because the calculator does not know whether the data entered are population data or sample data. It is up to the user of the calculator to choose the correct standard deviation. The results agree with those found in Examples 3 and 4.

Figure 11

**NW Now Work Problem 21**

Is the standard deviation resistant? To help determine the answer, suppose Jennifer's exam score (Table 8 on page 122) is 26, not 62. The population standard deviation increases from 9.8 to 18.3. Clearly, the standard deviation is *not* resistant. Therefore, the standard deviation should only be used for distributions that do not have extreme observations. We will introduce a different measure of spread in Section 3.4 that may be used when extreme observations are present in the data.

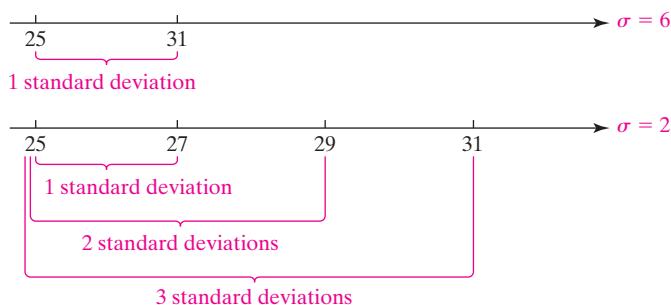
Notice the sample standard deviation is larger than the population standard deviation. Why? Our sample happened to include scores that are more dispersed than the population. A different random sample of, say, Ryanne (77), Pam (71), Joel (74), and Juan (88), would result in a sample standard deviation of 7.4 points. The different sample standard deviation occurs because we have different individuals in the sample.

Interpretations of the Standard Deviation

How does the value of the standard deviation relate to the spread of the distribution?

Standard deviation represents the “typical” deviation from the mean. As such, the standard deviation may be used to judge whether a particular observation is “far away” from the mean of a data set. For example, is a measure of 31 cm far from 25 cm? It depends. If the standard deviation of the data is 6 cm, then the answer is no because 31 cm would be only 1 standard deviation from 25 cm. However, if the standard deviation of the data is 2 cm, then the answer is yes because 31 cm would be 3 standard deviations from 25 cm. See Figure 12.

Figure 12



A good rule of thumb is to consider an observation “far away” if it is more than 2 standard deviations from the other observation (such as the mean). So, when judging the unusualness of an observation, it is vital that you consider the underlying variation in the data as measured by the standard deviation.

When comparing two populations, **the larger the standard deviation, the greater the dispersion, or spread, of the distribution** provided the variable of interest from the two populations has the same unit of measure. The units of measure must be the same so that we are comparing “apples with apples.” For example, \$100 is not the same as 100 Japanese yen (because recently, \$1 was equivalent to about 114 yen). So, a standard deviation of \$100 is substantially higher than a standard deviation of 100 yen.

EXAMPLE 6**Comparing the Standard Deviation of Two Data Sets**

Problem Refer to the data in Example 1. Use the standard deviation to determine whether University A or University B has more dispersion in its students' IQ scores.

Approach We will use Minitab to determine the standard deviation of IQ for each university. The university with the higher standard deviation will be the university with more dispersion in IQ scores. The histograms shown in Figure 10 on page 122 imply that University A has more dispersion. Therefore, we expect University A to have a higher sample standard deviation.

Solution Enter the data into Minitab and compute the descriptive statistics. See Figure 13.

Figure 13

Descriptive Statistics: A, B

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
A	100	0	100.00	1.61	16.08	60.00	90.00	102.00	110.00	141.00
B	100	0	100.01	0.836	8.36	86.00	94.00	98.00	107.00	119.00

The sample standard deviation is larger for University A (16.1) than for University B (8.4). Therefore, University A has IQ scores that are more dispersed. 

③ Determine the Variance of a Variable from Raw Data

A third measure of dispersion is called the *variance*.

Definitions

The **variance** of a variable is the square of the standard deviation. The **population variance** is σ^2 and the **sample variance** is s^2 .

The units of measure in variance are squared values. So, if the variable is measured in dollars, the variance is measured in dollars squared. This makes interpreting the variance difficult. However, the variance is important for conducting certain types of statistical inference, which we discuss later in the text.

EXAMPLE 7**Determining the Variance of a Variable for a Population and a Sample**

Problem Use the results of Examples 3 and 4 to determine the population and sample variance of test scores on the statistics exam.

Approach The population variance is found by squaring the population standard deviation. The sample variance is found by squaring the sample standard deviation.

Solution The population standard deviation in Example 3 was found to be $\sigma = 9.8$ points, so the population variance is $\sigma^2 = (9.8 \text{ points})^2 = 96.04 \text{ points}^2$. The sample standard deviation in Example 4 was found to be $s = 11.3$ points, so the sample variance is $s^2 = (11.3 \text{ points})^2 = 127.7 \text{ points}^2$. 

If you look carefully at Formulas (1) and (3), you will notice that the value under the radical ($\sqrt{}$) represents the variance. So we could also find the population or sample variance while in the process of finding the population or sample standard deviation. Using this approach, we obtain a population variance of 96.4 points² and a sample variance of 128.3 points². Using a rounded value of the standard deviation to obtain the variance results in round-off error. Therefore, it is recommended that you use the value under the radical while finding standard deviation, or as many decimal places as possible when using the standard deviation, to obtain the variance.

A Final Thought on Variance and Standard Deviation

The sample variance is obtained using the formula $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$. What if we divided by n instead of $n - 1$ to obtain the sample variance, as one might expect? Then the sample variance would consistently underestimate the population variance. Whenever a statistic consistently underestimates (or overestimates) a parameter, it is said to be **biased**. To obtain an unbiased estimate of the population variance, divide the sum of squared deviations about the sample mean by $n - 1$.

To help understand the concept of a biased estimator, consider the following scenario. Suppose you work for a carnival in which you must guess a person's age. After 20 people come to your booth, you notice that you have a tendency to underestimate people's age. (You guess too low.) What would you do to correct this? You could adjust your guess higher to avoid underestimating. In other words, originally your guesses were biased. To remove the bias, you increase your guess. This is what dividing by $n - 1$ in the sample variance formula accomplishes.

Unfortunately, the sample standard deviation given in Formulas (3) and (4) is not an unbiased estimate of the population standard deviation. In fact, it is not possible to provide an unbiased estimator of the population standard deviation for all distributions. The explanation for this concept is beyond the scope of this class (it has to do with the shape of the graph of the square root function). However, for the applications presented in this text, the bias is minor and does not impact results.

④ Use the Empirical Rule to Describe Data That Are Bell Shaped

If data have a distribution that is bell shaped, the *Empirical Rule* can be used to determine the percentage of data that will lie within k standard deviations of the mean.

The Empirical Rule

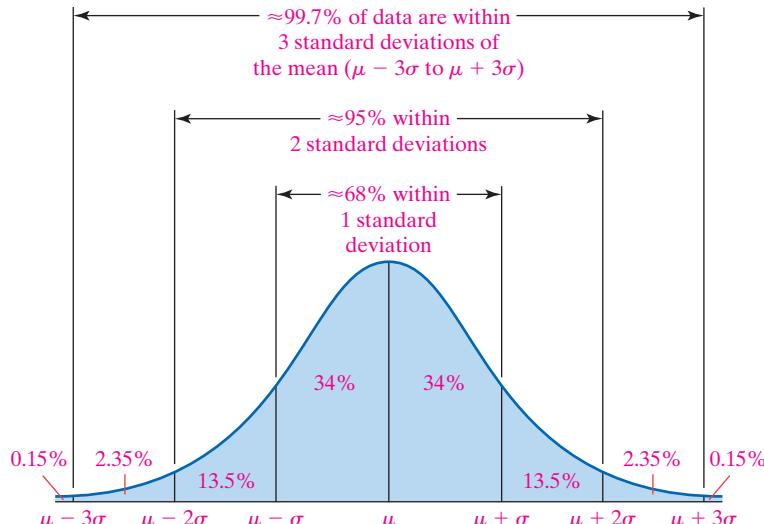
If a distribution is roughly bell shaped, then

- Approximately 68% of the data will lie within 1 standard deviation of the mean. That is, approximately 68% of the data lie between $\mu - 1\sigma$ and $\mu + 1\sigma$.
- Approximately 95% of the data will lie within 2 standard deviations of the mean. That is, approximately 95% of the data lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- Approximately 99.7% of the data will lie within 3 standard deviations of the mean. That is, approximately 99.7% of the data lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Note: We can also use the Empirical Rule based on sample data with \bar{x} used in place of μ and s used in place of σ .

Figure 14 illustrates the Empirical Rule.

Figure 14



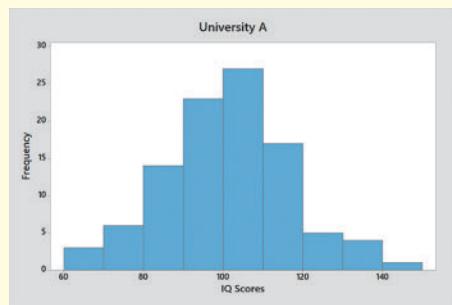
EXAMPLE 8 Using the Empirical Rule

Problem Use the data from University A in Table 7.

- Determine the percentage of students who have IQ scores within 3 standard deviations of the mean according to the Empirical Rule.
- Determine the percentage of students who have IQ scores between 67.8 and 132.2 according to the Empirical Rule.
- Determine the actual percentage of students who have IQ scores between 67.8 and 132.2.
- According to the Empirical Rule, what percentage of students have IQ scores above 132.2?

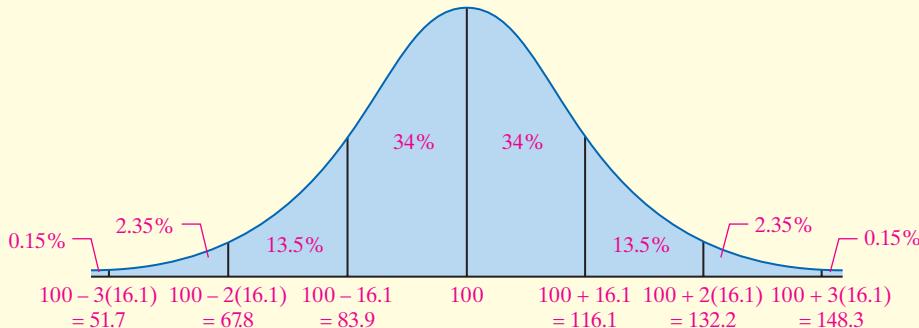
Approach To use the Empirical Rule, a histogram of the data must be roughly bell shaped. Figure 15 shows the histogram drawn in Minitab of the data from University A.

Figure 15



Solution The histogram is roughly bell shaped. From Examples 1 and 6, the mean IQ score of the students enrolled in University A is 100 and the standard deviation is 16.1. To make the analysis easier, draw a bell-shaped curve like the one in Figure 14, with $\bar{x} = 100$ and $s = 16.1$. See Figure 16.

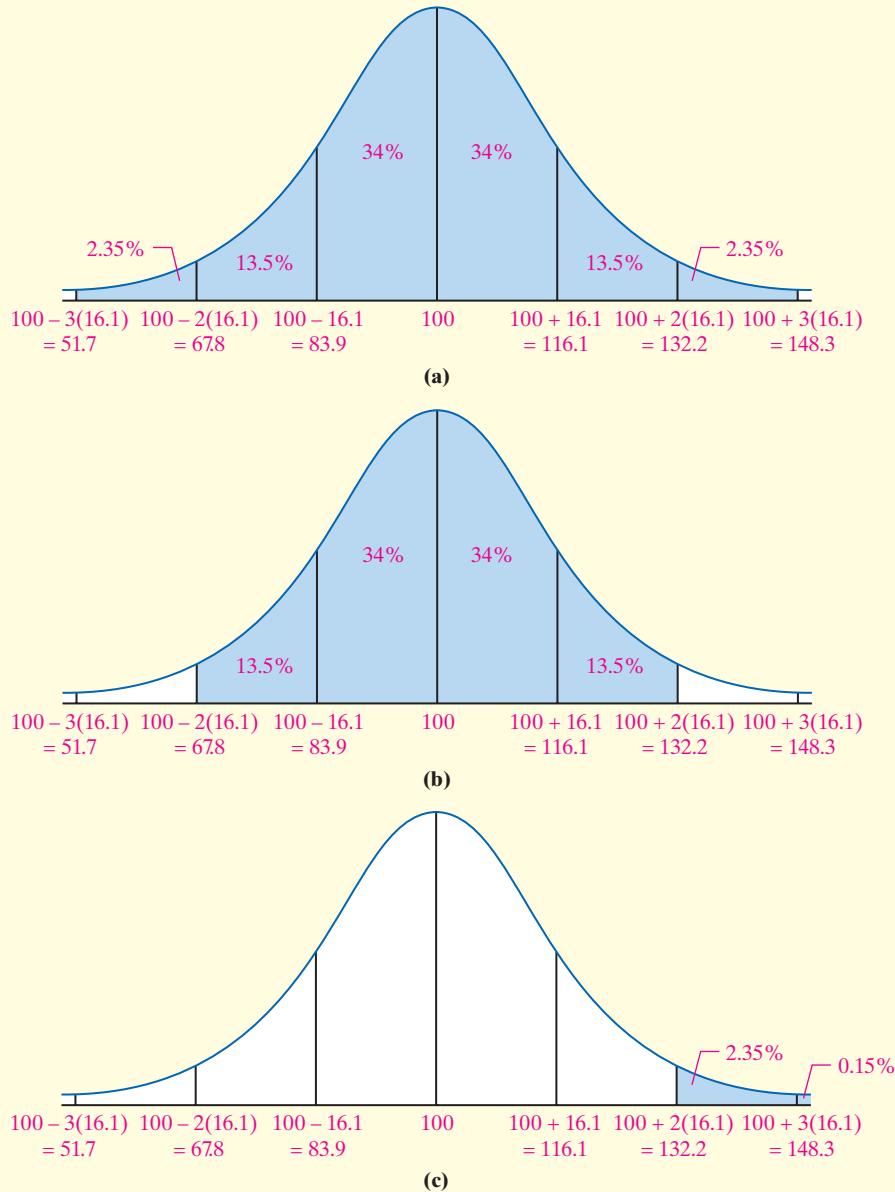
Figure 16



- According to the Empirical Rule, approximately 99.7% of the IQ scores are within 3 standard deviations of the mean [that is, greater than or equal to $100 - 3(16.1) = 51.7$ and less than or equal to $100 + 3(16.1) = 148.3$]. Figure 17(a) shows a bell-shaped curve similar to Figure 16 with the regions between 51.7 and 148.3 shaded.
- Since 67.8 is exactly 2 standard deviations below the mean [$100 - 2(16.1) = 67.8$] and 132.2 is exactly 2 standard deviations above the mean [$100 + 2(16.1) = 132.2$], the Empirical Rule tells us that approximately 95% of the IQ scores lie between 67.8 and 132.2. Figure 17(b) shows a bell-shaped curve similar to Figure 16 with the regions between 67.8 and 132.2 shaded.
- Of the 100 IQ scores listed in Table 7, 96, or 96%, are between 67.8 and 132.2. This is very close to the Empirical Rule's approximation.

- (d) Figure 17(c) shows a bell-shaped curve similar to Figure 16 with the region to the right of 132.2 shaded. Based on Figure 17(c) approximately $2.35\% + 0.15\% = 2.5\%$ of students at University A will have IQ scores above 132.2.

Figure 17



NW Now Work Problem 31



5 Use Chebyshev's Inequality to Describe Any Set of Data

The Russian mathematician Pafnuty Chebyshev (1821–1894) developed an inequality that determines a minimum percentage of observations that lie within k standard deviations of the mean, where $k > 1$. What's amazing is that the result is obtained regardless of the basic shape of the distribution (skewed left, skewed right, or symmetric).

Chebyshev's Inequality

For any data set or distribution, at least $\left(1 - \frac{1}{k^2}\right) \cdot 100\%$ of the observations lie within k standard deviations of the mean, where k is any number greater than 1. That is, at least $\left(1 - \frac{1}{k^2}\right) \cdot 100\%$ of the data lie between $\mu - k\sigma$ and $\mu + k\sigma$ for $k > 1$.

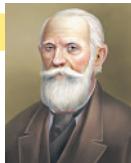
Note: We can also use Chebyshev's Inequality based on sample data.

CAUTION!

The Empirical Rule holds only if the distribution is bell shaped. Chebyshev's Inequality holds regardless of the shape of the distribution.

For example, at least $(1 - \frac{1}{2^2}) \cdot 100\% = 75\%$ of all observations lie within $k = 2$ standard deviations of the mean and at least $(1 - \frac{1}{3^2}) \cdot 100\% = 88.9\%$ of all observations lie within $k = 3$ standard deviations of the mean.

Notice the result does not state that exactly 75% of all observations lie within 2 standard deviations of the mean, but instead states that 75% or more of the observations will lie within 2 standard deviations of the mean.

EXAMPLE 9**Using Chebyshev's Inequality****Historical Note**

Pafnuty Chebyshev was born on May 16, 1821, in Okatovo, Russia. In 1847, he began teaching mathematics at the University of St. Petersburg. Some of his more famous work was done on prime numbers. In particular, he discovered a way to determine the number of prime numbers less than or equal to a given number. Chebyshev also studied mechanics, including rotary motion. Chebyshev was elected a Fellow of the Royal Society in 1877. He died on November 26, 1894, in St. Petersburg.

Problem Use the data from University A in Table 7.

- Determine the minimum percentage of students who have IQ scores within 3 standard deviations of the mean according to Chebyshev's Inequality.
- Determine the minimum percentage of students who have IQ scores between 67.8 and 132.2, according to Chebyshev's Inequality.
- Determine the actual percentage of students who have IQ scores between 67.8 and 132.2.

Approach

- Use Chebyshev's Inequality with $k = 3$.
- Determine the number of standard deviations 67.8 and 132.2 are from the mean of 100.0. Then substitute this value for k into Chebyshev's Inequality.
- Refer to Table 7 and count the number of observations between 67.8 and 132.2. Divide this result by 100, the number of observations in the data set.

Solution

- Using Chebyshev's Inequality with $k = 3$, at least $(1 - \frac{1}{3^2}) \cdot 100\% = 88.9\%$ of all students have IQ scores within 3 standard deviations of the mean. Since the mean of the data set is 100.0 and the standard deviation is 16.1, at least 88.9% of the students have IQ scores between $\bar{x} - ks = 100.0 - 3(16.1) = 51.7$ and $\bar{x} + ks = 100 + 3(16.1) = 148.3$.
- Since 67.8 is exactly 2 standard deviations below the mean [$100 - 2(16.1) = 67.8$] and 132.2 is exactly 2 standard deviations above the mean [$100 + 2(16.1) = 132.2$], Chebyshev's Inequality (with $k = 2$) says that at least $(1 - \frac{1}{2^2}) \cdot 100\% = 75\%$ of all IQ scores lie between 67.8 and 132.2.
- Of the 100 IQ scores listed, 96 or 96% are between 67.8 and 132.2. Notice that Chebyshev's Inequality provides a conservative result.

NW Now Work Problem 37

Because the Empirical Rule requires that the distribution be bell shaped, while Chebyshev's Inequality applies to all distributions, the Empirical Rule gives more precise results.

Technology Step-by-Step**Determining the Range, Variance, and Standard Deviation**

The same steps followed to obtain the measures of central tendency from raw data can be used to obtain the measures of dispersion.

Refer to the Technology Step-by-Step on page 116.

**3.2 Assess Your Understanding****Vocabulary and Skill Building**

- The sum of the deviations about the mean always equals _____.
- The standard deviation is used in conjunction with the _____ to numerically describe distributions that are bell shaped.

The _____ measures the center of the distribution, while the standard deviation measures the _____ of the distribution.

- True or False:** When comparing two populations, the larger the standard deviation, the more dispersion the distribution has, provided that the variable of interest from the two populations has the same unit of measure.

- 4. True or False:** Chebyshev's Inequality applies to all distributions regardless of shape, but the Empirical Rule holds only for distributions that are bell shaped.

In Problems 5–10, by hand, find the population variance and standard deviation or the sample variance and standard deviation as indicated.

5. Sample: 20, 13, 4, 8, 10
6. Sample: 83, 65, 91, 87, 84
7. Population: 3, 6, 10, 12, 14
8. Population: 1, 19, 25, 15, 12, 16, 28, 13, 6
9. Sample: 6, 52, 13, 49, 35, 25, 31, 29, 31, 29
10. Population: 4, 10, 12, 12, 13, 21

- 11. Miles per Gallon** The following data represent the miles per gallon for a 2013 Ford Fusion for six randomly selected vehicles. Compute the range, sample variance, and sample standard deviation miles per gallon.

Source: www.fueleconomy.gov

34.0, 33.2, 37.0, 29.4, 23.6, 25.9

- 12. Exam Time** The following data represent the amount of time (in minutes) a random sample of eight students took to complete the online portion of an exam in Sullivan's Statistics course. Compute the range, sample variance, and sample standard deviation time.

60.5, 128.0, 84.6, 122.3, 78.9, 94.7, 85.9, 89.9

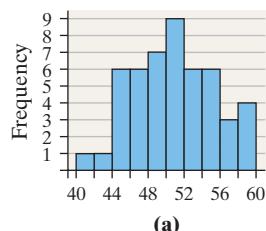
- NW 13. Concrete Mix** A certain type of concrete mix is designed to withstand 3000 pounds per square inch (psi) of pressure. The strength of concrete is measured by pouring the mix into casting cylinders after it is allowed to set up for 28 days. The following data represent the strength of nine randomly selected casts. Compute the range and sample standard deviation for the strength of the concrete (in psi).

3960, 4090, 3200, 3100, 2940, 3830, 4090, 4040, 3780

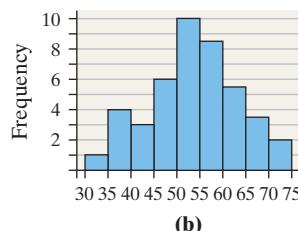
- 14. Flight Time** The following data represent the flight time (in minutes) of a random sample of seven flights from Las Vegas, Nevada, to Newark, New Jersey, on United Airlines. Compute the range and sample standard deviation of flight time.

282, 270, 260, 266, 257, 260, 267

- 15.** Which histogram depicts a higher standard deviation? Justify your answer.



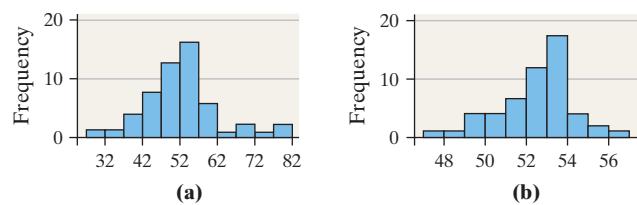
(a)



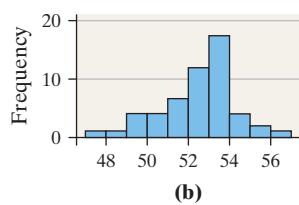
(b)

- 16.** Match the histograms in the next column to the summary statistics given.

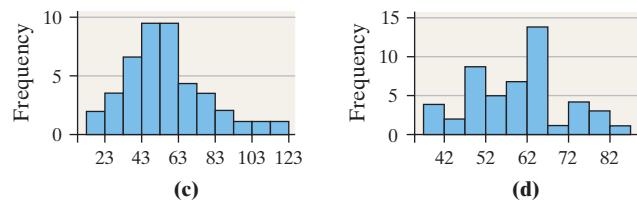
	Mean	Median	Standard Deviation
I	53	53	1.8
II	60	60	11
III	53	53	10
IV	53	53	22



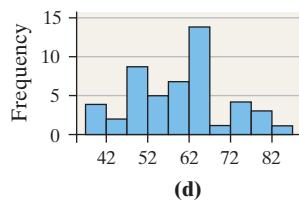
(a)



(b)



(c)



(d)

- 17.** For each of the following data sets, decide which has the higher standard deviation (set 1 or set 2), if any, *without doing any computation*. Explain the rationale behind your choice. Then, verify your choice by computing the standard deviation by hand.

	Set 1	Set 2
(a)	4, 6, 7, 8, 10	4, 7, 7, 7, 10
(b)	4, 8, 9, 10, 15	40, 80, 90, 100, 150
(c)	3, 6, 8, 10, 12, 15	93, 96, 98, 100, 102, 105

- 18. (a)** Suppose in a certain set of data, there are two observations of 90 and 100. Is 90 far away from 100 if the standard deviation of the data is 10?
- (b)** Suppose in a certain set of data, there are two observations of 90 and 100. Is 90 far away from 100 if the standard deviation of the data is 3?

Applying the Concepts

- 19. Exam Scores** The following data represent exam scores in a statistics class taught using traditional lecture and a class taught using a "flipped" classroom. The "flipped" classroom is one where the content is delivered via video and watched at home, while class time is used for homework and activities.

Traditional	70.8	69.1	79.4	67.6	85.3	78.2	56.2
	81.3	80.9	71.5	63.7	69.8	59.8	
Flipped	76.4	71.6	63.4	72.4	77.9	91.8	78.9
	76.8	82.1	70.2	91.5	77.8	76.5	

Source: Michael Sullivan.

- (a)** Which course has more dispersion in exam scores using the range as the measure of dispersion?
- (b)** Which course has more dispersion in exam scores using the standard deviation as the measure of dispersion?
- (c)** Suppose the score of 59.8 in the traditional course was incorrectly recorded as 598. How does this affect the range? the standard deviation? What property does this illustrate?

- DATA 20. pH in Water** The acidity or alkalinity of a solution is measured using pH. A pH less than 7 is acidic, while a pH greater than 7 is alkaline. The following data represent the pH in samples of bottled water and tap water.

Tap	7.64	7.45	7.47	7.50	7.68	7.69
	7.45	7.10	7.56	7.47	7.52	7.47
Bottled	5.15	5.09	5.26	5.20	5.02	5.23
	5.28	5.26	5.13	5.26	5.21	5.24

Source: Emily McCarney, student at Joliet Junior College.

- (a) Which type of water has more dispersion in pH using the range as the measure of dispersion?
 (b) Which type of water has more dispersion in pH using the standard deviation as the measure of dispersion?

- NW 21. Pulse Rates** The data represent the pulse rates (beats per minute) of nine students enrolled in a section of Sullivan's course in Introductory Statistics. Treat the nine students as a population.

Student	Pulse
Perpetual Bempah	76
Megan Brooks	60
Jeff Honeycutt	60
Clarice Jefferson	81
Crystal Kurtenbach	72
Janette Lantka	80
Kevin McCarthy	80
Tammy Ohm	68
Kathy Wojdyla	73

- (a) Determine the population standard deviation.
 (b) Find three simple random samples of size 3, and determine the sample standard deviation of each sample.
 (c) Which samples underestimate the population standard deviation? Which overestimate the population standard deviation?

- DATA 22. Travel Time** The following data represent the travel time (in minutes) to school for nine students enrolled in Sullivan's College Algebra course. Treat the nine students as a population.

Student	Travel Time	Student	Travel Time
Amanda	39	Scot	45
Amber	21	Erica	11
Tim	9	Tiffany	12
Mike	32	Glenn	39
Nicole	30		

- (a) Determine the population standard deviation.
 (b) Find three simple random samples of size 4, and determine the sample standard deviation of each sample.
 (c) Which samples underestimate the population standard deviation? Which overestimate the population standard deviation?

- DATA 23. A Fish Story** Ethan and Drew went on a 10-day fishing trip. The number of smallmouth bass caught and released by the two boys each day was as follows:

Ethan	9	24	8	9	5	8	9	10	8	10
Drew	15	2	3	18	20	1	17	2	19	3

- (a) Find the population mean and the range for the number of smallmouth bass caught per day by each fisherman. Do these values indicate any differences between the two fishermen's catches per day? Explain.
 (b) Draw a dot plot for Ethan. Draw a dot plot for Drew. Which fisherman seems more consistent?
 (c) Find the population standard deviation for the number of smallmouth bass caught per day by each fisherman. Do these values present a different story about the two fishermen's catches per day? Which fisherman has the more consistent record? Explain.
 (d) Discuss limitations of the range as a measure of dispersion.

- DATA 24. Soybean Yield** The data below represent the number of pods on a sample of soybean plants for two different plot types. Which plot type do you think is superior? Why?

Plot Type	Pods								
Liberty	32	31	36	35	44	31	39	37	38
No Till	35	31	32	30	43	33	37	42	40

Source: Andrew Dieter and Brad Schmidgall, students at Joliet Junior College.

- DATA 25. The Empirical Rule** The following data represent the weights (in grams) of a random sample of 50 M&M plain candies.

0.87	0.88	0.82	0.90	0.90	0.84	0.84
0.91	0.94	0.86	0.86	0.86	0.88	0.87
0.89	0.91	0.86	0.87	0.93	0.88	
0.83	0.95	0.87	0.93	0.91	0.85	
0.91	0.91	0.86	0.89	0.87	0.84	
0.88	0.88	0.89	0.79	0.82	0.83	
0.90	0.88	0.84	0.93	0.81	0.90	
0.88	0.92	0.85	0.84	0.84	0.86	

Source: Michael Sullivan.

- (a) Determine the sample standard deviation weight. Express your answer rounded to three decimal places.
 (b) On the basis of the histogram drawn in Section 3.1, Problem 25, comment on the appropriateness of using the Empirical Rule to make any general statements about the weights of M&Ms.
 (c) Use the Empirical Rule to determine the percentage of M&Ms with weights between 0.803 and 0.947 gram. Hint: $\bar{x} = 0.875$.
 (d) Determine the actual percentage of M&Ms that weigh between 0.803 and 0.947 gram, inclusive.
 (e) Use the Empirical Rule to determine the percentage of M&Ms with weights more than 0.911 gram.
 (f) Determine the actual percentage of M&Ms that weigh more than 0.911 gram.

- DATA 26. The Empirical Rule** The following data represent the length of eruption for a random sample of eruptions at the Old Faithful geyser in Calistoga, California.

108	108	99	105	103	103	94
102	99	106	90	104	110	110
103	109	109	111	101	101	
110	102	105	110	106	104	
104	100	103	102	120	90	
113	116	95	105	103	101	
100	101	107	110	92	108	

Source: Ladonna Hansen, Park Curator.

- (a) Determine the sample standard deviation length of eruption. Express your answer rounded to the nearest whole number.
- (b) On the basis of the histogram drawn in Section 3.1, Problem 26, comment on the appropriateness of using the Empirical Rule to make any general statements about the length of eruptions.
- (c) Use the Empirical Rule to determine the percentage of eruptions that last between 92 and 116 seconds.
Hint: $\bar{x} = 104$.
- (d) Determine the actual percentage of eruptions that last between 92 and 116 seconds, inclusive.
- (e) Use the Empirical Rule to determine the percentage of eruptions that last less than 98 seconds.
- (f) Determine the actual percentage of eruptions that last less than 98 seconds.

- DATA 27. Which Car Would You Buy?** Suppose that you are in the market to purchase a car. You have narrowed it down to two choices and will let gas mileage be the deciding factor. You decide to conduct a little experiment in which you put 10 gallons of gas in the car and drive it on a closed track until it runs out of gas. You conduct this experiment 15 times on each car and record the number of miles driven. Describe each data set. That is, determine the shape, center, and spread. Which car would you buy and why?

Car 1				
228	223	178	220	220
233	233	271	219	223
217	214	189	236	248

Car 2				
277	164	326	215	259
217	321	263	160	257
239	230	183	217	230

- DATA 28. Which Investment Is Better?** You have received a year-end bonus of \$5000. You decide to invest the money in the stock market and have narrowed your investment options down to two mutual funds. The following data represent the historical quarterly rates of return of each mutual fund for the past 20 quarters (5 years). Describe each data set. That is, determine the shape, center, and spread. Which mutual fund would you invest in and why?

Mutual Fund A			
1.3	-0.3	0.6	6.8
5.2	4.8	2.4	3.0
7.3	8.6	3.4	3.8
6.4	1.9	-0.5	-2.3
5.0	1.8	-1.3	3.1

Mutual Fund B			
-5.4	6.7	11.9	4.3
3.5	10.5	2.9	3.8
-6.7	1.4	8.9	0.3
-4.7	-1.1	3.4	7.7
4.3	5.9	-2.4	12.9

- DATA 29. Rates of Returns of Stocks** Stocks may be categorized by sectors. Go to www.pearsonhighered.com/sullivanstats and download 3_2_29 using the file format of your choice. The data represent the one-year rate of return (in percent) for a sample of consumer cyclical stocks and industrial stocks.

Note: Consumer cyclical stocks include names such as Starbucks and Home Depot. Industrial stocks include names such as 3M and FedEx.

- (a) Draw a relative frequency histogram for each sector using a lower class limit for the first class of -50 and a class width of 10. Which sector appears to have more dispersion?
- (b) Determine the mean and median rate of return for each sector. Which sector has the higher mean rate of return? Which sector has the higher median rate of return?
- (c) Determine the standard deviation rate of return for each sector. In finance, the standard deviation rate of return is called **risk**. Typically, an investor "pays" for a higher return by accepting more risk. Is the investor paying for higher returns for these sectors? Do you think the higher returns are worth the cost? Explain.

- DATA 30. Temperatures** It is well known that San Diego has milder weather than Chicago, but which city has more deviation from normal temperatures? Use the following data, which represent the deviation from normal high temperatures for a random sample of days. In which city would you rather be a meteorologist? Why?

Deviation from Normal High

Temperature, Chicago (°F)

8	2	22	-2	0	-9	7	-15
-7	13	17	1	-5	0	7	-2
-9	11	19	-3	-4	2	1	-13
-5	15	11	6	6	8	-17	

Deviation from Normal High

Temperature, San Diego (°F)

4	-5	3	-1	-5	-3	-4	4
5	-6	7	-4	-5	-4	-3	-6
1	-1	-2	-5	-4	-5	10	-6
-4	-2	-1	-3	-6	-3	8	

Source: National Climatic Data Center.

- NW 31. The Empirical Rule** The Stanford—Binet Intelligence Quotient (IQ) measures intelligence. IQ scores have a bell-shaped distribution with a mean of 100 and a standard deviation of 15.

- (a) What percentage of people has an IQ score between 70 and 130?
- (b) What percentage of people has an IQ score less than 70 or greater than 130?
- (c) What percentage of people has an IQ score greater than 130?

- 32. The Empirical Rule** SAT Math scores have a bell-shaped distribution with a mean of 515 and a standard deviation of 114.

Source: College Board.

- (a) What percentage of SAT scores is between 401 and 629?
 (b) What percentage of SAT scores is less than 401 or greater than 629?
 (c) What percentage of SAT scores is greater than 743?

33. The Empirical Rule The weight, in grams, of the pair of kidneys in adult males between the ages of 40 and 49 has a bell-shaped distribution with a mean of 325 grams and a standard deviation of 30 grams.

- (a) About 95% of kidney pairs will be between what weights?
 (b) What percentage of kidney pairs weighs between 235 grams and 415 grams?
 (c) What percentage of kidney pairs weighs less than 235 grams or more than 415 grams?
 (d) What percentage of kidney pairs weighs between 295 grams and 385 grams?

34. The Empirical Rule The distribution of the length of bolts has a bell shape with a mean of 4 inches and a standard deviation of 0.007 inch.

- (a) About 68% of bolts manufactured will be between what lengths?
 (b) What percentage of bolts will be between 3.986 inches and 4.014 inches?
 (c) If the company discards any bolts less than 3.986 inches or greater than 4.014 inches, what percentage of bolts manufactured will be discarded?
 (d) What percentage of bolts manufactured will be between 4.007 inches and 4.021 inches?

35. Which Professor? Suppose Professor Alpha and Professor Omega each teach Introductory Biology. You need to decide which professor to take the class from and have just completed your Introductory Statistics course. Records obtained from past students indicate that students in Professor Alpha's class have a mean score of 80% with a standard deviation of 5%, while past students in Professor Omega's class have a mean score of 80% with a standard deviation of 10%. Decide which instructor to take for Introductory Biology using a statistical argument.

36. Larry Summers Lawrence Summers (former Secretary of the Treasury and former president of Harvard) infamously claimed that women have a lower standard deviation IQ than men. He went on to suggest that this was a potential explanation as to why there are fewer women in top math and science positions. Suppose an IQ of 145 or higher is required to be a researcher at a top-notch research institution. Use the idea of standard deviation, the Empirical Rule, and the fact that the mean and standard deviation IQ of humans is 100 and 15, respectively, to explain Summers' argument.

NW 37. Chebyshev's Inequality In December 2018, the average price of regular unleaded gasoline excluding taxes in the United States was \$3.06 per gallon, according to the Energy Information Administration. Assume that the standard deviation price per gallon is \$0.06 per gallon to answer the following.

- (a) What minimum percentage of gasoline stations had prices within 3 standard deviations of the mean?
 (b) What minimum percentage of gasoline stations had prices within 2.5 standard deviations of the mean? What are the gasoline prices that are within 2.5 standard deviations of the mean?
 (c) What is the minimum percentage of gasoline stations that had prices between \$2.94 and \$3.18?

38. Chebyshev's Inequality According to the U.S. Census Bureau, the mean of the commute time to work for a resident of

Boston, Massachusetts, is 27.3 minutes. Assume that the standard deviation of the commute time is 8.1 minutes to answer the following:

- (a) What minimum percentage of commuters in Boston has a commute time within 2 standard deviations of the mean?
 (b) What minimum percentage of commuters in Boston has a commute time within 1.5 standard deviations of the mean? What are the commute times within 1.5 standard deviations of the mean?
 (c) What is the minimum percentage of commuters who have commute times between 3 minutes and 51.6 minutes?

39. Comparing Standard Deviations The standard deviation of batting averages of all teams in the American League is 0.008. The standard deviation of all players in the American League is 0.02154. Why is there less variability in team batting averages?

40. Linear Transformations Benjamin owns a small Internet business. Besides himself, he employs nine other people. The salaries earned by the employees are given next in thousands of dollars (Benjamin's salary is the largest, of course):

30, 30, 45, 50, 50, 50, 55, 55, 60, 75

- (a) Determine the range, population variance, and population standard deviation for the data.
 (b) Business has been good! As a result, Benjamin has a total of \$25,000 in bonus pay to distribute to his employees. One option for distributing bonuses is to give each employee (including himself) \$2500. Add the bonuses under this plan to the original salaries to create a new data set. Recalculate the range, population variance, and population standard deviation. How do they compare to the originals?
 (c) As a second option, Benjamin can give each employee a bonus of 5% of his or her original salary. Add the bonuses under this second plan to the original salaries to create a new data set. Recalculate the range, population variance, and population standard deviation. How do they compare to the originals?
 (d) As a third option, Benjamin decides not to give his employees a bonus at all. Instead, he keeps the \$25,000 for himself. Use this plan to create a new data set. Recalculate the range, population variance, and population standard deviation. How do they compare to the originals?

DATA 41. Resistance and Sample Size Each of the following three data sets represents the IQ scores of a random sample of adults. IQ scores are known to have a mean and median of 100. For each data set, determine the sample standard deviation. Then recompute the sample standard deviation assuming that the individual whose IQ is 106 is accidentally recorded as 160. For each sample size, state what happens to the standard deviation. Comment on the role that the number of observations plays in resistance.

Sample of Size 5

106	92	98	103	100
-----	----	----	-----	-----

Sample of Size 12

106	92	98	103	100	102
98	124	83	70	108	121

Sample of Size 30					
106	92	98	103	100	102
98	124	83	70	108	121
102	87	121	107	97	114
140	93	130	72	81	90
103	97	89	98	88	103

- 42. Identical Values** Compute the sample standard deviation of the following test scores: 78, 78, 78, 78. What can be said about a data set in which all the values are identical?

- DATA 43. Buying a Car** The following data represent the asking price, in dollars, for a random sample of 2014 coupes (a two-door car) and a random sample of 2014 Chevy Camaros.

Coupes			Camaros		
25,991	15,900	16,900	24,949	24,948	23,061
24,948	23,791	20,990	22,150	21,855	20,990
19,900	17,900	16,888	19,950	19,593	18,995
15,995	15,891	13,991	17,849	16,900	16,440
12,900	11,995	9900	15,987	15,900	15,891

Source: autotrader.com

- (a) Find the mean and standard deviation price for each sample.
(b) Explain why the mean is higher for Camaros yet the standard deviation is less.

- 44. Blocking and Variability** Blocking refers to the idea that we can reduce the variability in a variable by segmenting the data by some other variable. The given data represent the recumbent length (in centimeters) of a sample of 10 males and 10 females who are 40 months of age.

Males		Females	
104.0	94.4	102.5	100.8
93.7	97.6	100.4	96.3
98.3	100.6	102.7	105.0
86.2	103.0	98.1	106.5
90.7	100.9	95.4	114.5

Source: National Center for Health Statistics.

- (a) Determine the standard deviation of recumbent length for all 20 observations.
(b) Determine the standard deviation of recumbent length for the males.
(c) Determine the standard deviation of recumbent length for the females.
(d) What effect does blocking by gender have on the standard deviation of recumbent length for each gender?

- 45. Mean Absolute Deviation** Another measure of variation is the mean absolute deviation. It is computed using the formula

$$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

Compute the mean absolute deviation of the data in Problem 11 and compare the results with the sample standard deviation.

- 46. Coefficient of Skewness** Karl Pearson developed a measure that describes the skewness of a distribution, called the **coefficient of skewness**. The formula is

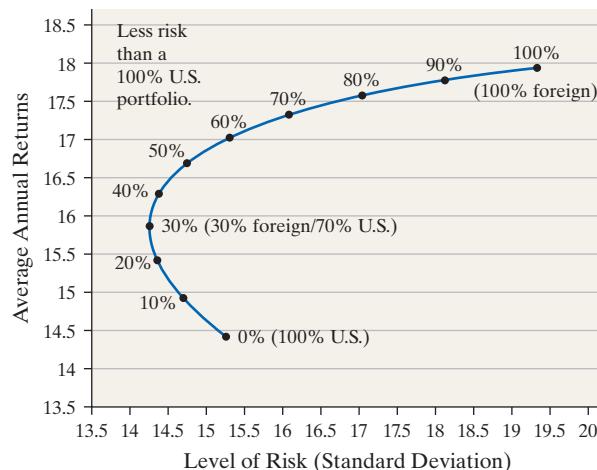
$$\text{Skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

The value of this measure generally lies between -3 and $+3$. The closer the value lies to -3 , the more the distribution is skewed left. The closer the value lies to $+3$, the more the distribution is skewed right. A value close to 0 indicates a symmetric distribution. Find the coefficient of skewness of the following distributions and comment on the skewness.

- (a) Mean = 50, median = 40, standard deviation = 10
(b) Mean = 100, median = 100, standard deviation = 15
(c) Mean = 400, median = 500, standard deviation = 120
(d) Compute the coefficient of skewness for the data in Problem 25.
(e) Compute the coefficient of skewness for the data in Problem 26.

- 47. Diversification** A popular theory in investment states that you should invest a certain amount of money in foreign investments to reduce your risk. The risk of a portfolio is defined as the standard deviation of the rate of return. Refer to the graph, which depicts the relation between risk (standard deviation of rate of return) and reward (mean rate of return).

How Foreign Stocks Benefit a Domestic Portfolio



Source: T. Rowe Price.

- (a) Determine the average annual return and level of risk in a portfolio that is 10% foreign.
(b) Determine the percentage that should be invested in foreign stocks to best minimize risk.
(c) Why do you think risk initially decreases as the percent of foreign investments increases?
(d) A portfolio that is 30% foreign and 70% American has a mean rate of return of about 15.8%, with a standard deviation of 14.3%. According to Chebyshev's Inequality, at least 75% of returns will be between what values? According to Chebyshev's Inequality, at least 88.9% of returns will be between what two values? Should an investor be surprised if she has a negative rate of return? Why?

- 48. More Spread?** The data set on the left represents the annual rate of return (in percent) of eight randomly sampled bond mutual funds, and the data set on the right represents the annual rate of return (in percent) of eight randomly sampled stock mutual funds.

2.0	1.9	1.8
3.2	2.4	3.4
1.6	2.7	

8.4	7.2	7.6
7.4	6.9	9.4
9.1	8.1	

- DATA** (a) Determine the mean and standard deviation of each data set.
 (b) Based only on the standard deviation, which data set has more spread?
 (c) What proportion of the observations is within one standard deviation of the mean for each data set?
 (d) The **coefficient of variation**, CV, is defined as the ratio of the standard deviation to the mean of a data set, so

$$CV = \frac{\text{standard deviation}}{\text{mean}}$$

The coefficient of variation is unitless and allows for comparison in spread between two data sets by describing the amount of spread per unit mean. After all, larger numbers will likely have a larger standard deviation simply due to the size of the numbers. Compute the coefficient of variation for both data sets. Which data set do you believe has more “spread”?

- DATA** (e) Let’s take this idea one step further. The following data represent the height of a random sample of 8 male college students. The data set on the left has their height measured in inches, and the data set on the right has their height measured in centimeters.

74	68	71
66	72	69
69	71	

187.96	172.72	180.34
167.64	182.88	175.26
175.26	180.34	

For each data set, determine the mean and the standard deviation. Would you say that the height of the males is more dispersed based on the standard deviation of height measured in centimeters? Why? Now, determine the coefficient of variation for each data set. What did you find?

- DATA** 49. **Sullivan Survey** Choose two quantitative variables from the SullivanStatsSurveyI at www.pearsonhighered.com/sullivanstats for which a comparison is reasonable, such as number of hours of television versus number of hours of Internet. Draw a histogram for each variable. Which variable appears to have more dispersion? Determine the range and standard deviation of each variable. Based on the numerical measure, which variable has more dispersion?

- DATA** 50. **Sullivan Survey** Choose any quantitative variable from the SullivanStatsSurveyI at www.pearsonhighered.com/sullivanstats. Now choose a qualitative variable, such as gender or political philosophy. Determine the range and standard deviation by the qualitative variable chosen. For example, if you chose gender as the qualitative variable, determine the range and standard deviation by gender. Does there appear to be any difference in the measure of dispersion for each level of the qualitative variable?

- 51. Threaded Problem: Tornado** The data set “Tornadoes_2017” located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a) Determine the range and standard deviation of the length for all tornadoes.
 (b) Draw a histogram of the length of tornadoes in Texas (TX).
Hint: If you are using StatCrunch, enter “State=TX” in the Where: box of the Histogram dialogue window. Draw a histogram of the length of tornadoes in Oklahoma (OK). Which state has lengths of tornadoes that are more dispersed? Explain.
 (c) Determine the range and standard deviation of the length for tornadoes in Texas (TX). Determine the range and the standard deviation of the length for tornadoes in Oklahoma (OK). Which state has more dispersion in the length of tornadoes?
 (d) Determine the standard deviation of the length for tornadoes in Idaho (ID). What does this result suggest?

Explaining the Concepts

52. Would it be appropriate to say that a distribution with a standard deviation of 10 centimeters is more dispersed than a distribution with a standard deviation of 5 inches? Support your position.
 53. What is meant by the phrase *degrees of freedom* as it pertains to the computation of the sample standard deviation?
 54. Are any of the measures of dispersion mentioned in this section resistant? Explain.
 55. What does it mean when a statistic is biased?
 56. What makes the range less desirable than the standard deviation as a measure of dispersion?
 57. In one of Sullivan’s statistics sections, the standard deviation of the heights of all students was 3.9 inches. The standard deviation of the heights of males was 3.4 inches and the standard deviation of females was 3.3 inches. Why is the standard deviation of the entire class more than the standard deviation of the males and females considered separately?
 58. Explain how standard deviation measures spread. In your explanation include the computation of the same standard deviation for two data sets: Data set I: 3, 4, 5; Data set II: 0, 4, 8.
 59. Which of the following would have a higher standard deviation? (a) IQ of students on your campus or (b) IQ of residents in your home town? Why?
 60. Develop a sample of size $n = 8$ such that $\bar{x} = 15$ and $s = 0$.
 61. Draw two histograms with different standard deviations and label them I and II. Which histogram has the larger standard deviation?
 62. **Fast Pass** In 2000, the Walt Disney Company created the “fast pass”—a ticket issued to a rider for a popular ride and the rider is told to return at a specific time during the day. At that time, the rider is allowed to bypass the regular line, thereby reducing the wait time for that particular rider. When compared to wait times prior to creating the fast pass, overall wait times for rides at the park increased, on average, yet patrons to the park indicated they were happier with the fast pass system. Use the concepts of central tendency and dispersion to explain why.

3.3 Measures of Central Tendency and Dispersion from Grouped Data



Preparing for This Section Before getting started, review the following:

- Organizing discrete data in tables (Section 2.2, pp. 76–77)
- Organizing continuous data in tables (Section 2.2, pp. 78–80)

Objectives

- Approximate the mean of a variable from grouped data
- Compute the weighted mean
- Approximate the standard deviation of a variable from grouped data

We have discussed how to compute descriptive statistics from raw data, but often the only data available have already been summarized in frequency distributions (**grouped data**). Although we cannot find exact values of the mean or standard deviation without raw data, we can approximate these measures.

1 Approximate the Mean of a Variable from Grouped Data

Before providing a method for computing the mean from grouped data, we must learn how to determine the *class midpoint* of a class.

Definition

A **class midpoint** is the sum of consecutive lower class limits divided by 2.

Because raw data cannot be retrieved from a frequency table, we assume that within each class the mean of the data is equal to the class midpoint. We then multiply the class midpoint by the frequency. This product is expected to be close to the sum of the data that lie within the class. We repeat the process for each class and add the results. This sum approximates the sum of all the data.

Definition

Approximate Mean of a Variable from a Frequency Distribution

Population Mean

$$\begin{aligned}\mu &= \frac{\sum x_i f_i}{\sum f_i} \\ &= \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n}\end{aligned}$$

Sample Mean

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} \quad (1)$$

where x_i is the midpoint or value of the i th class

f_i is the frequency of the i th class

n is the number of classes

In Formula (1), $x_1 f_1$ approximates the sum of all the data in the first class, $x_2 f_2$ approximates the sum of all the data in the second class, and so on. Notice that the formulas for the population mean and sample mean are essentially identical, just as they were for computing the mean from raw data.

EXAMPLE 1

Approximating the Mean for Continuous Quantitative Data from a Frequency Distribution

Problem The frequency distribution in Table 13 on the next page represents the total fine including late penalties, in dollars, for a simple random sample of 50 parking and camera violations in the City of New York. Approximate the mean fine.

(continued)

Table 13

Class (Amount of Fine)	Frequency
50–74.99	1
75–99.99	0
100–124.99	7
125–149.99	10
150–174.99	5
175–199.99	4
200–224.99	13
225–249.99	4
250–274.99	5
275–299.99	0
300–324.99	1

Approach

Step 1 Determine the class midpoint of each class by adding consecutive lower class limits and dividing the result by 2.

Step 2 Compute the sum of the frequencies, $\sum f_i$.

Step 3 Multiply the class midpoint by the frequency to obtain $x_i f_i$ for each class.

Step 4 Compute $\sum x_i f_i$.

Step 5 Substitute into Sample Mean Formula (1) to obtain the mean from grouped data.

Solution

Step 1 The first two lower class limits are 50 and 75. Therefore, the class midpoint of the first class is $\frac{50 + 75}{2} = 62.5$, so $x_1 = 62.5$. The remaining class midpoints are listed in column 2 of Table 14.

Table 14

Class (Amount of Fine)	Class Midpoint, x_i	Frequency, f_i	$x_i f_i$
50–74.99	62.5	1	62.5(1) = 62.5
75–99.99	87.5	0	87.5(0) = 0
100–124.99	112.5	7	112.5(7) = 787.5
125–149.99	137.5	10	1375
150–174.99	162.5	5	812.5
175–199.99	187.5	4	750
200–224.99	212.5	13	2762.5
225–249.99	237.5	4	950
250–274.99	262.5	5	1312.5
275–299.99	287.5	0	0
300–324.99	312.5	1	312.5
$\sum f_i = 50$		$\sum x_i f_i = 9125$	

Step 2 Add the frequencies in column 3 to obtain $\sum f_i = 1 + 0 + \dots + 1 = 50$.

Step 3 Compute the values of $x_i f_i$ by multiplying each class midpoint by the corresponding frequency and obtain the results shown in column 4 of Table 14.

Step 4 Add the values in column 4 of Table 14 to obtain $\sum x_i f_i = 9125$.

Step 5 Substituting into Sample Mean Formula (1), we obtain

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{9125}{50} = 182.5$$

The approximate mean fine is \$182.50.

**CAUTION!**

We computed the mean from grouped data in Example 1 even though the raw data is available. The reason for doing this was to illustrate how close the two values can be. In practice, use raw data whenever possible.

NW Now compute the approximate mean of the frequency distribution in Problem 1

The mean fine from the raw data listed in Example 3 on page 79 from Section 2.2 is \$181.91.

2 Compute the Weighted Mean

When observations have different importance, or *weight*, associated with them, we compute the *weighted mean*. For example, your grade-point average is a weighted mean, with the weights equal to the number of credit hours in each course. The value of the variable is equal to the grade converted to a point value.

Definition

The **weighted mean**, \bar{x}_w , of a variable is found by multiplying each value of the variable by its corresponding weight, adding these products, and dividing this sum by the sum of the weights. It can be expressed using the formula

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} \quad (2)$$

where w_i is the weight of the i th observation
 x_i is the value of the i th observation

EXAMPLE 2 Computing the Weighted Mean

Problem Marissa just completed her first semester in college. She earned an A in her four-hour statistics course, a B in her three-hour sociology course, an A in her three-hour psychology course, a C in her five-hour computer programming course, and an A in her one-hour drama course. Determine Marissa's grade-point average.

Approach Assign point values to each grade. Let an A equal 4 points, a B equal 3 points, and a C equal 2 points. The number of credit hours for each course determines its weight. So a five-hour course gets a weight of 5, a four-hour course gets a weight of 4, and so on. Multiply the weight of each course by the points earned in the course, add these products, and divide this sum by the sum of the weights, number of credit hours.

Solution

$$\text{GPA} = \bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{4(4) + 3(3) + 3(4) + 5(2) + 1(4)}{4 + 3 + 3 + 5 + 1} = \frac{51}{16} = 3.19$$

NW Now Work Problem 9

Marissa's grade-point average for her first semester is 3.19.



3 Approximate the Standard Deviation of a Variable from Grouped Data

The procedure for approximating the standard deviation from grouped data is similar to that of finding the mean from grouped data. Again, because we do not have access to the original data, the standard deviation is approximate.

Definition**Approximate Standard Deviation of a Variable from a Frequency Distribution****Population Standard Deviation**

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2 f_i}{\sum f_i}}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{(\sum f_i) - 1}} \quad (3)$$

where x_i is the midpoint or value of the i th class

f_i is the frequency of the i th class

An algebraically equivalent formula for the population standard deviation is

$$\sigma = \sqrt{\frac{\sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{\sum f_i}}{\sum f_i}}$$

EXAMPLE 3 Approximating the Standard Deviation from a Frequency Distribution

Problem The data in Table 13 on page 140 represent the total fine including late penalties, in dollars, for a simple random sample of 50 parking and camera violations in the City of New York. Approximate the standard deviation of the fine.

Approach Use the sample standard deviation Formula (3).

Step 1 Create a table with the class in column 1, the class midpoint in column 2, the frequency in column 3, and the unrounded mean in column 4.

Step 2 Compute the deviation about the mean, $x_i - \bar{x}$, for each class, where x_i is the class midpoint of the i th class and \bar{x} is the sample mean. Enter the results in column 5.

Step 3 Square the deviation about the mean and multiply this result by the frequency to obtain $(x_i - \bar{x})^2 f_i$. Enter the results in column 6.

Step 4 Add the entries in columns 3 and 6 to obtain $\sum f_i$ and $\sum (x_i - \bar{x})^2 f_i$.

Step 5 Substitute the values found in Step 4 into Formula (3) to obtain an approximate value for the sample standard deviation.

Solution

Step 1 Create Table 15. Column 1 contains the classes. Column 2 contains the class midpoint of each class. Column 3 contains the frequency of each class. Column 4 contains the unrounded sample mean obtained in Example 1.

Table 15

Class (Amount of Fine)	Class Midpoint, x_i	Frequency, f_i	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2 f_i$
50–74.99	62.5	1	182.5	-120	14,400
75–99.99	87.5	0	182.5	-95	0
100–124.99	112.5	7	182.5	-70	34,300
125–149.99	137.5	10	182.5	-45	20,250
150–174.99	162.5	5	182.5	-20	2000
175–199.99	187.5	4	182.5	5	100
200–224.99	212.5	13	182.5	30	11,700
225–249.99	237.5	4	182.5	55	12,100
250–274.99	262.5	5	182.5	80	32,000
275–299.99	287.5	0	182.5	105	0
300–324.99	312.5	1	182.5	130	16,900
$\sum f_i = 50$			$\sum (x_i - \bar{x})^2 f_i = 143,750$		

Step 2 Column 5 contains the deviation about the mean, $x_i - \bar{x}$, for each class.

Step 3 Column 6 contains the values of the squared deviation about the mean multiplied by the frequency, $(x_i - \bar{x})^2 f_i$.

Step 4 Add the entries in columns 3 and 6 to obtain $\sum f_i = 50$ and $\sum (x_i - \bar{x})^2 f_i = 143,750$.

Step 5 Substitute these values into Formula (3) to obtain an approximate value for the sample standard deviation.

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{(\sum f_i) - 1}} = \sqrt{\frac{143,750}{50 - 1}} \approx \$54.163$$

The approximate standard deviation of the fine is \$54.163. The standard deviation from the raw data listed in Example 3 from Section 2.2 is \$54.151.

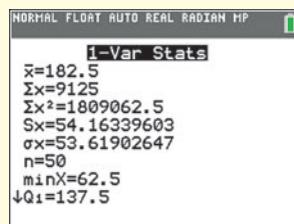
EXAMPLE 4**Approximating the Mean and Standard Deviation of Grouped Data Using Technology**

Problem Approximate the mean and standard deviation of the fine data in Table 13 using a TI-84 Plus CE graphing calculator.

Approach The steps for approximating the mean and standard deviation of grouped data using the TI-83/84 Plus graphing calculator and StatCrunch are given in the Technology Step-by-Step below.

Solution Figure 18 shows the result from the TI-84 Plus CE. From the output, we can see that the approximate mean is \$182.5 and the approximate standard deviation is \$54.163. The results agree with our by-hand solutions.

Figure 18



NW Now compute the approximate standard deviation from the frequency distribution from the frequency distribution in Problem 1

Technology Step-by-Step**Determining the Mean and Standard Deviation from Grouped Data****TI-83/84 Plus**

- Enter the class midpoint in L1 and the frequency or relative frequency in L2 by pressing STAT and then selecting 1 : Edit.
- Press STAT, highlight the CALC menu, and select 1 : 1-Var Stats. In the menu, select L1 for List. Select L2 for FreqList:. Highlight Calculate and press ENTER.

StatCrunch

- If necessary, enter the summarized data into the spreadsheet. Name the columns.
- Select Stat, highlight Summary Stats, and select Grouped/Binned data.
- Choose the column that contains the class under the “Bins in:” drop-down menu. Choose the column that contains the frequencies in the “Counts in:” drop-down menu. Select the “Consecutive lower limits” radio button for defining the midpoints. Click Compute!.

**3.3 Assess Your Understanding****Applying the Concepts**

- NW 1. Savings** Recently, a random sample of 25–34 year olds was asked, “How much do you currently have in savings, not including retirement savings?” The following data represent the responses to the survey. Approximate the mean and standard deviation amount of savings.

Savings	Frequency
\$0–\$19,999	344
\$20,000–\$39,999	98
\$40,000–\$59,999	52
\$60,000–\$79,999	19
\$80,000–\$99,999	13
\$100,000–\$119,999	6
\$120,000–\$139,999	2

Source: Based on a poll by LearnVest.



- 2. Square Footage of Housing** The frequency distribution below represents the square footage of a random sample of 500 houses that are owner occupied year round. Approximate the mean and standard deviation square footage.

Square Footage	Frequency
0–499	5
500–999	17
1000–1499	36
1500–1999	121
2000–2499	119
2500–2999	81
3000–3499	47
3500–3999	45
4000–4499	22
4500–4999	7

Source: Based on data from the U.S. Census Bureau.

- DATA 3. Household Winter Temperature** Often, frequency distributions are reported using unequal class widths because the frequencies of some groups would otherwise be small or very large. Consider the following data, which represent the daytime household temperature the thermostat is set to when someone is home for a random sample of 750 households. Determine the class midpoint, if necessary, for each class and approximate the mean and standard deviation temperature.

Temperature (°F)	Frequency
61–64	31
65–67	67
68–69	198
70	195
71–72	120
73–76	89
77–80	50

Source: Based on data from the U.S. Department of Energy.

- DATA 4. Living in Poverty** (See Problem 3.) The following frequency distribution represents the age of people living in poverty in 2018 based on a random sample of residents of the United States. In this frequency distribution, the class widths are not the same for each class. Approximate the mean and standard deviation age of a person living in poverty. For the open-ended class 65 and older, use 70 as the class midpoint.

Age	Frequency
0–17	14,659
18–24	5819
25–34	6694
35–44	4871
45–54	4533
55–59	2476
60–64	2036
65 and older	4231

Source: U.S. Census Bureau.

- DATA 5. Multiple Births** The following data represent the number of live multiple-delivery births (three or more babies) in 2017 for women 15 to 54 years old.

Age	Number of Multiple Births
15–19	43
20–24	365
25–29	964
30–34	1442
35–39	837
40–44	197
45–49	48
50–54	21

Source: National Vital Statistics Reports.

- (a) Approximate the mean and standard deviation for age.
 (b) Draw a frequency histogram of the data to verify that the distribution is bell shaped.

- (c) According to the Empirical Rule, 95% of mothers of multiple births will be between what two ages?
DATA 6. Birth Weight The following frequency distribution represents the birth weight of all babies born in the United States in 2017.

Weight (grams)	Number of Babies
0–999	25,446
1000–1999	91,375
2000–2999	924,163
3000–3999	2,513,786
4000–4999	296,874
5000–5999	4241

Source: National Vital Statistics Report.

- (a) Approximate the mean and standard deviation birth weight.
 (b) Draw a frequency histogram of the data to verify that the distribution is bell shaped.
 (c) According to the Empirical Rule, 68% of all babies will weigh between what two values?
7. Exit Velocity Use the frequency distribution whose class width is 4 obtained in Problem 29 in Section 2.2 to approximate the mean and standard deviation exit velocity. Compare these results to the actual mean and standard deviation exit velocity.
8. Cigarette Tax Rates Use the frequency distribution whose class width is 0.5 obtained in Problem 30 in Section 2.2 to approximate the mean and standard deviation for cigarette tax rates. Compare these results to the actual mean and standard deviation.
NW 9. Grade-Point Average Marissa has just completed her second semester in college. She earned a B in her five-hour calculus course, an A in her three-hour social work course, an A in her four-hour biology course, and a C in her three-hour American literature course. Assuming that an A equals 4 points, a B equals 3 points, and a C equals 2 points, determine Marissa's grade-point average for the semester.

- 10. Computing Class Average** In Marissa's calculus course, attendance counts for 5% of the grade, quizzes count for 10% of the grade, exams count for 60% of the grade, and the final exam counts for 25% of the grade. Marissa had a 100% average for attendance, 93% for quizzes, 86% for exams, and 85% on the final. Determine Marissa's course average.

- 11. Mixed Chocolates** Michael and Kevin want to buy chocolates. They can't agree on whether they want chocolate-covered almonds, chocolate-covered peanuts, or chocolate-covered raisins. They agree to create a mix. They bought 4 pounds of chocolate-covered almonds at \$3.50 per pound, 3 pounds of chocolate-covered peanuts for \$2.75 per pound, and 2 pounds of chocolate-covered raisins for \$2.25 per pound. Determine the cost per pound of the mix.

- 12. Nut Mix** Michael and Kevin return to the candy store, but this time they want to purchase nuts. They can't decide among peanuts, cashews, or almonds. They again agree to create a mix. They bought 2.5 pounds of peanuts for \$1.30 per pound, 4 pounds of cashews for \$4.50 per pound, and 2 pounds of almonds for \$3.75 per pound. Determine the price per pound of the mix.

- DATA 13. Population** The data represent the male and female population, by age, of the United States in 2017.

Note: Use 85 for the class midpoint of ≥ 80 .

Age	Male	Female
0–9	20,361,845	19,529,261
10–19	21,858,713	20,769,221
20–29	23,000,947	22,034,897
30–39	20,444,193	21,565,505
40–49	21,687,685	20,516,540
50–59	20,177,011	22,064,308
60–69	17,691,301	10,541,679
70–79	9,847,385	11,812,176
≥ 80	4,694,676	7,534,539

Source: U.S. Census Bureau.

- (a) Approximate the population mean and standard deviation of age for males.
- (b) Approximate the population mean and standard deviation of age for females.
- (c) Which gender has the higher mean age?
- (d) Which gender has more dispersion in age?

- DATA 14. Age of Mother** The following data represent the age of the mother at childbirth for 1980 and 2017.

Age	1980 Births (thousands)	2017 Births (thousands)
10–14	9.8	2.0
15–19	551.9	194.4
20–24	1226.4	764.8
25–29	1108.2	1123.6
30–34	549.9	1091.9
35–39	140.7	554.8
40–44	23.2	114.8
45–49	1.1	8.5
50–54	0	0.8

Source: National Vital Statistics Reports.

- (a) Approximate the population mean and standard deviation of age for mothers in 1980.
- (b) Approximate the population mean and standard deviation of age for mothers in 2017.
- (c) Which year has the higher mean age?
- (d) Which year has more dispersion in age?

Problems 15 and 16 use the following steps to approximate the median from grouped data.

Approximating the Median from Grouped Data

- Step 1** Construct a cumulative frequency distribution.
- Step 2** Identify the class in which the median lies. Remember, the median can be obtained by determining the observation that lies in the middle.
- Step 3** Interpolate the median using the formula

$$\text{Median} = L + \frac{\frac{n}{2} - CF}{f}(i)$$

where L is the lower class limit of the class containing the median

n is the number of data values in the frequency distribution

CF is the cumulative frequency of the class immediately preceding the class containing the median

f is the frequency of the median class

i is the class width of the class containing the median

- 15.** Approximate the median of the frequency distribution in Problem 2.

- 16.** Approximate the median of the frequency distribution in Problem 4.

Problems 17 and 18 use the following definition of the modal class. The **modal class** of a variable can be obtained from data in a frequency distribution by determining the class that has the largest frequency.

- 17.** Determine the modal class of the frequency distribution in Problem 1.

- 18.** Determine the modal class of the frequency distribution in Problem 2.

3.4 Measures of Position and Outliers



Objectives

- ① Determine and interpret z-scores
- ② Interpret percentiles
- ③ Determine and interpret quartiles
- ④ Determine and interpret the interquartile range
- ⑤ Check a set of data for outliers

In Section 3.1, we determined measures of central tendency, which describe the *typical* data value. Section 3.2 discussed measures of dispersion, which describe the amount of

spread in a set of data. In this section, we discuss measures of position, which describe the *relative position* of a certain data value within the entire set of data.

1 Determine and Interpret z-Scores

At the end of the 2018 season, the Boston Red Sox led the American League with 876 runs scored, while the Los Angeles Dodgers led the National League with 804 runs scored. It appears that the Red Sox are the better run-producing team. However, this comparison is unfair because the two teams play in different leagues. The Red Sox play in the American League, where the designated hitter bats for the pitcher, whereas the Dodgers play in the National League, where the pitcher must bat (pitchers are typically poor hitters). To compare the two teams' scoring of runs, we need to determine their relative standings in their respective leagues. We can do this using a *z-score*.

Definition

The **z-score** represents the distance that a data value is from the mean in terms of the number of standard deviations. We find it by subtracting the mean from the data value and dividing this result by the standard deviation. There is both a population z-score and a sample z-score:

Population z-Score

$$z = \frac{x - \mu}{\sigma}$$

Sample z-Score

$$z = \frac{x - \bar{x}}{s} \quad (1)$$

The z-score is unitless. It has mean 0 and standard deviation 1.

IN OTHER WORDS

The z-score provides a way to compare apples to oranges by converting variables with different centers or spreads to variables with the same center (0) and spread (1).

If a data value is larger than the mean, the z-score is positive. If a data value is smaller than the mean, the z-score is negative. If the data value equals the mean, the z-score is zero. A z-score measures the number of standard deviations an observation is above or below the mean. For example, a z-score of 1.24 means the data value is 1.24 standard deviations above the mean. A z-score of -2.31 means the data value is 2.31 standard deviations below the mean.

EXAMPLE 1

Comparing z-Scores

Problem Determine whether the Boston Red Sox or the Los Angeles Dodgers had a relatively better run-producing season. The Red Sox scored 876 runs and play in the American League, where the mean number of runs scored was $\mu = 731.2$ and the standard deviation was $\sigma = 72.6$ runs. The Dodgers scored 804 runs and play in the National League, where the mean number of runs scored was $\mu = 710.8$ and the standard deviation was $\sigma = 60.2$ runs.

Approach To determine which team had the relatively better run-producing season, compute each team's z-score. The team with the higher z-score had the better season. Because we know the values of the population parameters, compute the population z-score.

Solution We compute each team's z-score, rounded to two decimal places.

$$\text{Red Sox: } z\text{-score} = \frac{x - \mu}{\sigma} = \frac{876 - 731.2}{72.6} = 1.99$$

$$\text{Dodgers: } z\text{-score} = \frac{x - \mu}{\sigma} = \frac{804 - 710.8}{60.2} = 1.55$$

So the Red Sox had run production 1.99 standard deviations above the mean, while the Dodgers had run production 1.55 standard deviations above the mean. Therefore, the Red Sox had a relatively better year at scoring runs than the Dodgers.



In Example 1, the team with the higher z -score was said to have a relatively better season in producing runs. With negative z -scores, we need to be careful when deciding the better outcome. For example, suppose Bob and Mary run a marathon. If Bob finished the marathon in 213 minutes, where the mean finishing time among all men was 242 minutes with a standard deviation of 57 minutes, and Mary finished the marathon in 241 minutes, where the mean finishing time among all women was 273 minutes with a standard deviation of 52 minutes, who did better in the race? Since Bob's z -score is $z_{\text{Bob}} = \frac{213 - 242}{57} = -0.51$ and Mary's z -score is $z_{\text{Mary}} = \frac{241 - 273}{52} = -0.62$, Mary did better. Even though Bob's z -score is larger, Mary did better because her finishing time is more standard deviations below the mean.

2 Interpret Percentiles

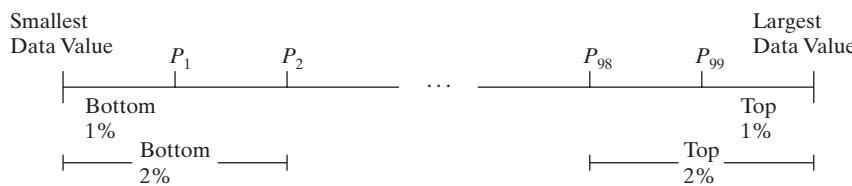
Recall that the median divides the lower 50% of a set of data from the upper 50%. The median is a special case of a general concept called the *percentile*.

Definition

The **k th percentile**, denoted P_k , of a set of data is a value such that k percent of the observations are less than or equal to the value.

So percentiles divide a set of data that is written in ascending order into 100 parts; thus 99 percentiles can be determined. For example, P_1 divides the bottom 1% of the observations from the top 99%, P_2 divides the bottom 2% of the observations from the top 98%, and so on. Figure 19 displays the 99 possible percentiles.

Figure 19



Percentiles are used to give the relative standing of an observation. Many standardized exams, such as the SAT college entrance exam, use percentiles to let students know how they scored on the exam in relation to all other students who took the exam.

EXAMPLE 2 Interpret a Percentile

Problem Jennifer just received the results of her SAT exam. Her SAT Mathematics score of 600 is in the 74th percentile. What does this mean?

Approach The k th percentile of an observation means that k percent of the observations are less than or equal to the observation.

Interpretation Being at the 74th percentile means that 74% of SAT Mathematics scores are less than or equal to 600 and 26% of the scores are greater. So 26% of the students who took the exam scored better than Jennifer.

NW Now Work Problem 15



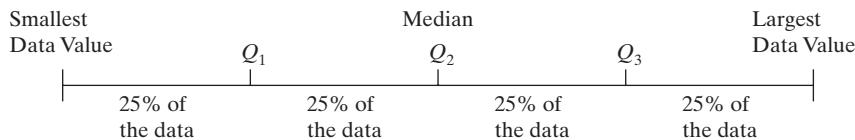
3 Determine and Interpret Quartiles

The most common percentiles are quartiles. **Quartiles** divide data sets into fourths, or four equal parts.

IN OTHER WORDS

The first quartile, Q_1 , is equivalent to the 25th percentile, P_{25} . The 2nd quartile, Q_2 , is equivalent to the 50th percentile, P_{50} , which is equivalent to the median, M . Finally, the third quartile, Q_3 , is equivalent to the 75th percentile, P_{75} .

Figure 20

**Finding Quartiles**

Step 1 Arrange the data in ascending order.

Step 2 Determine the median, M , or second quartile, Q_2 .

Step 3 Divide the data set into halves: the observations below (to the left of) M and the observations above M . The first quartile, Q_1 , is the median of the bottom half of the data and the third quartile, Q_3 , is the median of the top half of the data.

EXAMPLE 3**Finding and Interpreting Quartiles**

Problem The Highway Loss Data Institute routinely collects data on collision coverage claims. Collision coverage insures against physical damage to an insured individual's vehicle. The data in Table 16 represent a random sample of 18 collision coverage claims based on data obtained from the Highway Loss Data Institute. Find and interpret the first, second, and third quartiles for collision coverage claims.

Table 16

\$6751	\$9908	\$3461	\$2336	\$21,147	\$2332
\$189	\$1185	\$370	\$1414	\$4668	\$1953
\$10,034	\$735	\$802	\$618	\$180	\$1657

Approach Follow the steps given above.

Solution

Step 1 The data written in ascending order are given as follows:

\$180	\$189	\$370	\$618	\$735	\$802	\$1185	\$1414	\$1657
\$1953	\$2332	\$2336	\$3461	\$4668	\$6751	\$9908	\$10,034	\$21,147

Step 2 There are $n = 18$ observations, so the median, or second quartile, Q_2 , is the mean of the 9th and 10th observations. Therefore, $M = Q_2 = \frac{\$1657 + \$1953}{2} = \$1805$.

Step 3 The median of the bottom half of the data is the first quartile, Q_1 . As shown next, the median of these data is the 5th observation, so $Q_1 = \$735$.

$$\begin{array}{ccccccccc} \$180 & \$189 & \$370 & \$618 & \$\textcolor{blue}{735} & \$802 & \$1185 & \$1414 & \$1657 \\ & & & & \uparrow & & & & \\ & & & & Q_1 & & & & \end{array}$$

NOTE

If the number of observations is odd, do not include the median when determining Q_1 and Q_3 by hand.

The median of the top half of the data is the third quartile, Q_3 . As shown next, the median of these data is the 5th observation, so $Q_3 = \$4668$.

\$1953	\$2332	\$2336	\$3461	\$4668	\$6751	\$9908	\$10,034	\$21,147
↑				Q_3				

Interpretation Interpret the quartiles as percentiles. For example, 25% of the collision claims are less than or equal to the first quartile, \$735, and 75% of the collision claims are greater than \$735. Also, 50% of the collision claims are less than or equal to \$1805, the second quartile, and 50% of the collision claims are greater than \$1805. Finally, 75% of the collision claims are less than or equal to \$4668, the third quartile, and 25% of the collision claims are greater than \$4668.

EXAMPLE 4

Finding Quartiles Using Technology

Problem Find the quartiles of the collision coverage claims data in Table 16.

Approach Use both StatCrunch and Minitab to obtain the quartiles. The steps for obtaining quartiles using a TI-83/84 Plus graphing calculator, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on pages 151–152.

Solution The results obtained from StatCrunch [Figure 21(a)] agree with our “by hand” solution. In Figure 21(b), notice that the first quartile, 706, and the third quartile, 5189, reported by Minitab disagree with our “by hand” and StatCrunch result. This difference is due to the fact that StatCrunch and Minitab use different algorithms for obtaining quartiles.

Figure 21

Summary statistics:

Column	n	Median	Min	Max	Q1	Q3
Claim	18	1805	180	21147	735	4668

(a)

Descriptive Statistics: Claim

Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Claim	18	0	3874	1250	5302	180	706	1805	5189	21147

(b)

NW Now Work Problem 21(b)

④

Determine and Interpret the Interquartile Range

So far we have discussed three measures of dispersion: range, standard deviation, and variance. None of these measures of dispersion are resistant. Quartiles, however, are resistant. For this reason, quartiles are used to define a fourth measure of dispersion.

Definition

The **interquartile range, IQR**, is the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the third and first quartiles and is found using the formula

$$\text{IQR} = Q_3 - Q_1$$

The interpretation of the interquartile range is similar to that of the range and standard deviation. That is, the more spread a set of data has, the higher the interquartile range will be.

EXAMPLE 5 Determining and Interpreting the Interquartile Range

Problem Determine and interpret the interquartile range of the collision claim data from Example 3.

Approach Use the quartiles found by hand in Example 3. The interquartile range, IQR, is found by computing the difference between the third and first quartiles. It represents the range of the middle 50% of the observations.

Solution The interquartile range is

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ &= \$4668 - \$735 \\ &= \$3933 \end{aligned}$$

Interpretation The IQR, that is, the range of the middle 50% of the observations, in the collision claim data is \$3933. 

NW Now Work Problem 21(c)

Let's compare the measures of central tendency and dispersion discussed thus far for the collision claim data. The mean collision claim is \$3874.4 and the median is \$1805. The median is more representative of the "center" because the data are skewed to the right (only 5 of the 18 observations are greater than the mean). The range is \$21,147 – \$180 = \$20,967. The standard deviation is \$5301.6 and the interquartile range is \$3933. The values of the range and standard deviation are affected by the extreme claim of \$21,147. In fact, if this claim had been \$120,000 (let's say the claim was for a totaled Mercedes S-class AMG), then the range and standard deviation would increase to \$119,820 and \$27,782.5, respectively. The interquartile range would not be affected. Therefore, when the distribution of data is highly skewed or contains extreme observations, it is best to use the interquartile range as the measure of dispersion because it is resistant.

Summary: Which Measures to Report

Shape of Distribution	Measure of Central Tendency	Measure of Dispersion
Symmetric	Mean	Standard deviation
Skewed left or skewed right	Median	Interquartile range

For the remainder of this text, the direction **describe the distribution** will mean to describe its shape (skewed left, skewed right, symmetric), its center (mean or median), and its spread (standard deviation or interquartile range).

5 Check a Set of Data for Outliers

CAUTION!

Outliers distort both the mean and the standard deviation, because neither is resistant. Because these measures often form the basis for most statistical inference, any conclusions drawn from a set of data that contains outliers can be flawed.

When performing any type of data analysis, we should always check for extreme observations in the data set. Extreme observations are referred to as **outliers**. Outliers can occur by chance, because of error in the measurement of a variable, during data entry, or from errors in sampling.

Outliers do not always occur because of error. Sometimes extreme observations are common within a population. For example, suppose we wanted to estimate the mean price of a European car. We might take a random sample of size 5 from the population of all European automobiles. If our sample included a Ferrari 488 Spider (approximately \$280,900), it probably would be an outlier, because this car costs much more than the typical European automobile. The value of this car would be considered *unusual* because it is not a typical value from the data set.

Use the following steps to check for outliers using quartiles.

Checking for Outliers by Using Quartiles

Step 1 Determine the first and third quartiles of the data.

Step 2 Compute the interquartile range.

Step 3 Determine the fences. **Fences** serve as cutoff points for determining outliers.

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$

$$\text{Upper fence} = Q_3 + 1.5(\text{IQR})$$

Step 4 If a data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

EXAMPLE 6 Checking for Outliers

Problem Check the collision coverage claims data in Table 16 for outliers.

Approach Follow the preceding steps. Any data value that is less than the lower fence or greater than the upper fence will be considered an outlier.

Solution

Step 1 The quartiles found in Example 3 are $Q_1 = \$735$ and $Q_3 = \$4668$.

Step 2 The interquartile range, IQR, is

$$\begin{aligned}\text{IQR} &= Q_3 - Q_1 \\ &= \$4668 - \$735 \\ &= \$3933\end{aligned}$$

Step 3 The lower fence, LF, is

$$\begin{aligned}\text{LF} &= Q_1 - 1.5(\text{IQR}) \\ &= \$735 - 1.5(\$3933) \\ &= -\$5164.5\end{aligned}$$

The upper fence, UF, is

$$\begin{aligned}\text{UF} &= Q_3 + 1.5(\text{IQR}) \\ &= \$4668 + 1.5(\$3933) \\ &= \$10,567.5\end{aligned}$$

Step 4 There are no observations below the lower fence. However, there is an observation above the upper fence. The claim of \$21,147 is an outlier.

NW Now Work Problem 21(d)



Technology Step-by-Step

Determining Quartiles

TI-83/84 Plus

Follow the same steps given to compute the mean and median from raw data. (Section 3.1)

Minitab

Follow the same steps given to compute the mean and median from raw data. (Section 3.1)

Excel

- Enter the raw data into column A.
- With the data analysis Tool Pak enabled, select the Data tab and click on **Data Analysis**.
- Select **Rank and Percentile** from the Data Analysis window. Press OK.
- With the cursor in the **Input Range** cell, highlight the data. Press OK.

StatCrunch

Follow the same steps given to compute the mean and median from raw data. (Section 3.1)



3.4 Assess Your Understanding

Vocabulary

- The _____ represents the number of standard deviations an observation is from the mean.
- The _____ of a data set is a value such that k percent of the observations are less than or equal to the value.
- _____ divide data sets into fourths.
- The _____ is the range of the middle 50% of the observations in a data set.

Applying the Concepts

NW 5. Birth Weights Babies born after a gestation period of 32–35 weeks have a mean weight of 2600 grams and a standard deviation of 660 grams. Babies born after a gestation period of 40 weeks have a mean weight of 3500 grams and a standard deviation of 470 grams. Suppose a 34-week gestation period baby weighs 2400 grams and a 40-week gestation period baby weighs 3300 grams. What is the z -score for the 34-week gestation period baby? What is the z -score for the 40-week gestation period baby? Which baby weighs less relative to the gestation period?

6. Birth Weights Babies born after a gestation period of 32–35 weeks have a mean weight of 2600 grams and a standard deviation of 660 grams. Babies born after a gestation period of 40 weeks have a mean weight of 3500 grams and a standard deviation of 470 grams. Suppose a 34-week gestation period baby weighs 3000 grams and a 40-week gestation period baby weighs 3900 grams. What is the z -score for the 34-week gestation period baby? What is the z -score for the 40-week gestation period baby? Which baby weighs less relative to the gestation period?

7. Men versus Women The average 20- to 29-year-old man is 69.6 inches tall, with a standard deviation of 3.0 inches, while the average 20- to 29-year-old woman is 64.1 inches tall, with a standard deviation of 3.8 inches. Who is relatively taller, a 75-inch man or a 70-inch woman?

Source: CDC Vital and Health Statistics, Advance Data, Number 361, July 5, 2005.

8. Men versus Women The average 20- to 29-year-old man is 69.6 inches tall, with a standard deviation of 3.0 inches, while the average 20- to 29-year-old woman is 64.1 inches tall, with a standard deviation of 3.8 inches. Who is relatively taller, a 67-inch man or a 62-inch woman?

Source: CDC Vital and Health Statistics, Advance Data, Number 361, July 5, 2005.

9. ERA Champions In 2018, Jacob deGrom of the New York Mets had the lowest earned-run average (ERA is the mean number of runs yielded per nine innings pitched) of any starting pitcher in the National League, with an ERA of 1.70. Also in 2018, Blake Snell of the Tampa Bay Rays had the lowest ERA of any starting pitcher in the American League with an ERA of 1.89. In the National League, the mean ERA in 2018 was 3.611 and the standard deviation was 0.772. In the American League, the mean ERA in 2018 was 3.744 and the standard deviation was 0.893. Which player had the better year relative to his peers? Why?

10. Batting Champions The highest batting average ever recorded in Major League Baseball was by Ted Williams in 1941 when he hit 0.406. That year, the mean and standard deviation for batting average were 0.2806 and 0.0328. In 2018, Mookie Betts was the American League batting champion, with a batting average of 0.346. In 2018, the mean and standard deviation for batting average were 0.2621 and 0.0313. Who had the better year relative to his peers, Williams or Betts? Why?

11. Swim Ryan Murphy, nephew of the author, swims for U.S.A. Swimming. While he was in college at the University of California at Berkeley, Ryan's best time in the 100-meter backstroke was 45.3 seconds. The mean of all NCAA swimmers in this event is 48.62 seconds with a standard deviation of 0.98 second. Ryan's best time in the 200-meter backstroke was 99.32 seconds. The mean of all NCAA swimmers in this event is 106.58 seconds with a standard deviation of 2.38 seconds. In which race was Ryan better in college?

12. Triathlon Roberto finishes a triathlon (750-meter swim, 5-kilometer run, and 20-kilometer bicycle) in 63.2 minutes. Among all men in the race, the mean finishing time was 69.4 minutes with a standard deviation of 8.9 minutes. Zandra finishes the same triathlon in 79.3 minutes. Among all women in the race, the mean finishing time was 84.7 minutes with a standard deviation of 7.4 minutes. Who did better in relation to their gender?

13. School Admissions A highly selective boarding school will only admit students who place at least 1.5 standard deviations above the mean on a standardized test that has a mean of 200 and a standard deviation of 26. What is the minimum score that an applicant must make on the test to be accepted?

14. Quality Control A manufacturer of bolts has a quality-control policy that requires it to destroy any bolts that are more than 2 standard deviations from the mean. The quality-control

engineer knows that the bolts coming off the assembly line have a mean length of 8 cm with a standard deviation of 0.05 cm. For what lengths will a bolt be destroyed?

- NW 15. You Explain It! Percentiles** Explain the meaning of the following percentiles.

Source: Advance Data from Vital and Health Statistics.

- (a) The 15th percentile of the head circumference of males 3 to 5 months of age is 41.0 cm.
- (b) The 90th percentile of the waist circumference of females 2 years of age is 52.7 cm.
- (c) Anthropometry involves the measurement of the human body. One goal of these measurements is to assess how body measurements may be changing over time. The following table represents the standing height of males aged 20 years or older for various age groups. Based on the percentile measurements of the different age groups, what might you conclude?

Age	Percentile				
	10th	25th	50th	75th	90th
20–29	166.8	171.5	176.7	181.4	186.8
30–39	166.9	171.3	176.0	181.9	186.2
40–49	167.9	172.1	176.9	182.1	186.0
50–59	166.0	170.8	176.0	181.2	185.4
60–69	165.3	170.1	175.1	179.5	183.7
70–79	163.2	167.5	172.9	178.1	181.7
80 or older	161.7	166.1	170.5	175.3	179.4

- 16. You Explain It! Percentiles** Explain the meaning of the following percentiles.

Source: National Center for Health Statistics.

- (a) The 5th percentile of the weight of males 36 months of age is 12.0 kg.
- (b) The 95th percentile of the length of newborn females is 53.8 cm.

17. You Explain It! Quartiles Violent crimes include rape, robbery, assault, and homicide. The following is a summary of the violent-crime rate (violent crimes per 100,000 population) for all 50 states in the United States plus Washington, D.C., in 2017.

$$Q_1 = 244.8 \quad Q_2 = 357.6 \quad Q_3 = 454.8$$

- (a) Provide an interpretation of these results.
- (b) Determine and interpret the interquartile range.
- (c) The violent-crime rate in Washington, D.C., in 2017 was 1004.9. Would this be an outlier?
- (d) Do you believe that the distribution of violent-crime rates is skewed or symmetric? Why?

18. You Explain It! Quartiles One variable that is measured by online homework systems is the amount of time a student spends on homework for each section of the text. The following is a summary of the number of minutes a student spends for each section of the text for the fall 2018 semester in a College Algebra class at Joliet Junior College.

$$Q_1 = 42 \quad Q_2 = 51.5 \quad Q_3 = 72.5$$

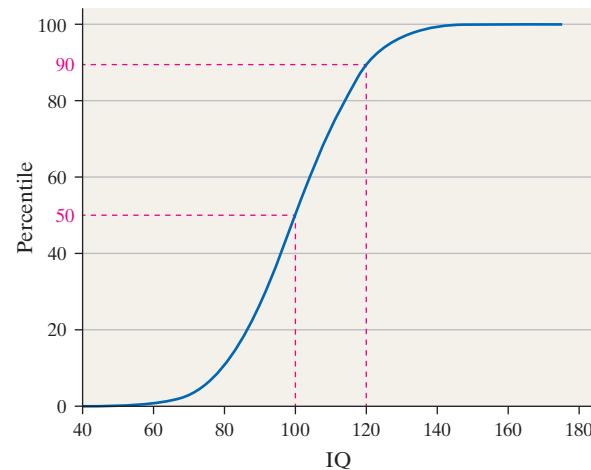
- (a) Provide an interpretation of these results.
- (b) Determine and interpret the interquartile range.

- (c) Suppose a student spent 2 hours doing homework for a section. Is this an outlier?

- (d) Do you believe that the distribution of time spent doing homework is skewed or symmetric? Why?

19. Ogives and Percentiles The following graph is an ogive of IQ scores. The vertical axis in an ogive is the cumulative relative frequency and can also be interpreted as a percentile.

Percentile Ranks of IQ Scores



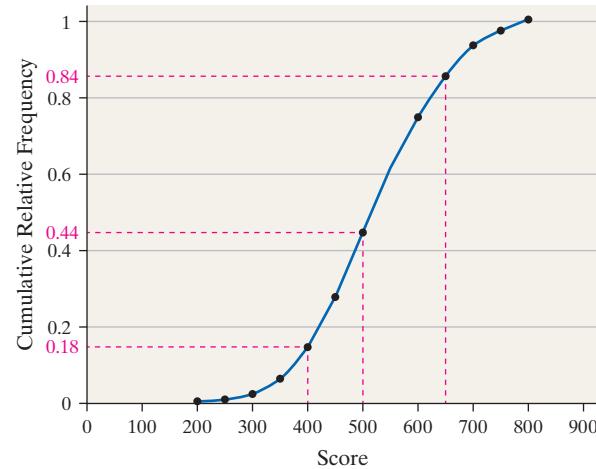
- (a) Find and interpret the percentile rank of an individual whose IQ is 100.

- (b) Find and interpret the percentile rank of an individual whose IQ is 120.

- (c) What score corresponds to the 60th percentile for IQ?

20. Ogives and Percentiles The following graph is an ogive of the mathematics scores on the SAT. The vertical axis in an ogive is the cumulative relative frequency and can also be interpreted as a percentile.

SAT Mathematics Scores



- (a) Find and interpret the percentile rank of a student who scored 400 on the SAT mathematics exam.

- (b) Find and interpret the percentile rank of a student who scored 700 on the SAT mathematics exam.

- (c) If Jane scored at the 44th percentile, what was her score?

- NW 21. SMART Car** The following data represent the miles per gallon of a random sample of SMART cars with a three-cylinder, 1.0-liter engine.

31.5	36.0	37.8	38.4	40.1	42.3
34.3	36.3	37.9	38.8	40.6	42.7
34.5	37.4	38.0	39.3	41.4	43.5
35.5	37.5	38.3	39.5	41.5	47.5

Source: www.fueleconomy.gov

- (a) Compute the z -score corresponding to the individual who obtained 36.3 miles per gallon. Interpret this result.
- (b) Determine the quartiles.
- (c) Compute and interpret the interquartile range, IQR.
- (d) Determine the lower and upper fences. Are there any outliers?

- DATA 22. Hemoglobin in Cats** The following data represent the hemoglobin (in g/dL) for 20 randomly selected cats.

5.7	8.9	9.6	10.6	11.7
7.7	9.4	9.9	10.7	12.9
7.8	9.5	10.0	11.0	13.0
8.7	9.6	10.3	11.2	13.4

Source: Joliet Junior College Veterinarian Technology Program.

- (a) Compute the z -score corresponding to the hemoglobin of Blackie, 7.8 g/dL. Interpret this result.
- (b) Determine the quartiles.
- (c) Compute and interpret the interquartile range, IQR.
- (d) Determine the lower and upper fences. Are there any outliers?

- DATA 23. Wait Time** The following data represent the wait time (in minutes) for a random sample of 40 visitors to Disney's Dinosaur Ride in Animal Kingdom.

6	31	8	0	21	16	0	7
15	6	44	27	7	52	3	7
4	5	10	5	21	3	6	14
5	24	10	9	9	10	12	8
4	8	39	5	28	30	4	15

Source: touringplans.com

- (a) Determine and interpret the quartiles.
- (b) Check the data set for outliers.

- DATA 24. Online Shopping** The following data represent the number of days between grocery orders at the online delivery company Instacart.

14	5	7	30	14	30	15	3
1	30	11	28	9	6	6	6
8	5	6	6	14	15	10	12
5	5	8	30	7	10	7	5
30	30	29	5	4	14	6	30

Source: Instacart.

- (a) Determine and interpret the quartiles.
 - (b) Check the data set for outliers.
- DATA 25. Fraud Detection** As part of its "Customers First" program, a cellular phone company monitors monthly phone usage. The program identifies unusual use and alerts the customer that their phone may have been used by another person. The data

represent the monthly phone use in minutes of a customer enrolled in this program for the past 20 months. The phone company decides to use the upper fence as the cutoff point for the number of minutes at which the customer should be contacted. What is the cutoff point?

346	345	489	358	471
442	466	505	466	372
442	461	515	549	437
480	490	429	470	516

- DATA 26. Stolen Credit Card** A credit card company has a fraud-detection service that determines if a card has any unusual activity. The company maintains a database of daily charges on a customer's credit card. Days when the card was inactive are excluded from the database. If a day's worth of charges appears unusual, the customer is contacted to make sure that the credit card has not been compromised. Use the following daily charges (rounded to the nearest dollar) to determine the amount the daily charges must exceed before the customer is contacted.

143	166	113	188	133
90	89	98	95	112
111	79	46	20	112
70	174	68	101	212

- DATA 27. Student Survey of Income** A survey of 50 randomly selected full-time Joliet Junior College students was conducted during the Fall 2019 semester. In the survey, the students were asked to disclose their weekly income from employment. If the student did not work, \$0 was entered.

0	262	0	635	0	0	671
244	521	476	100	650	454	95
12,777	567	310	527	0	67	736
83	159	0	547	188	389	300
719	0	367	316	0	0	181
479	0	82	579	289		
375	347	331	281	628		
0	203	149	0	403		

- (a) Check the data set for outliers.
- (b) Draw a histogram of the data and label the outliers on the histogram.
- (c) Provide a possible explanation for the outliers.

- DATA 28. Student Survey of Entertainment Spending** A survey of 40 randomly selected full-time Joliet Junior College students was conducted in the Fall 2019 semester. In the survey, the students were asked to disclose their weekly spending on entertainment. The results of the survey are as follows:

21	54	64	33	65	32	21	16
22	39	67	54	22	51	26	14
115	7	80	59	20	33	13	36
36	10	12	101	1000	26	38	8
28	28	75	50	27	35	9	48

- (a) Check the data set for outliers.
- (b) Draw a histogram of the data and label the outliers on the histogram.
- (c) Provide a possible explanation for the outliers.

 **29. Threaded Problem: Tornado** The data set “Tornadoes_2017” located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a) Determine and interpret the quartiles of the length for all tornadoes.
- (b) Determine the interquartile range of the length of tornadoes in Iowa (IA). **Hint:** If you are using StatCrunch, enter “State=IA” in the Where: box of the Summary Stats dialogue window. Determine the interquartile range of the length of tornadoes in Kansas (KS). Which state has lengths of tornadoes that are more dispersed? Explain.

30. Travel Time Use the results of Problem 22 in Section 3.1 and Problem 22 in Section 3.2 to compute the z -scores for all the students. Compute the mean and standard deviation of these z -scores.

31. Fraud Detection Revisited Use the fraud-detection data from Problem 25 to do the following.

- (a) Determine the standard deviation and interquartile range of the data.
- (b) Suppose the month in which the customer used 346 minutes was not actually that customer’s phone. That particular month, the customer did not use her phone at all, so 0 minutes were used. How does changing the observation from 346 to 0 affect the standard deviation and interquartile range? What property does this illustrate?

Explaining the Concepts

- 32. Write a paragraph that explains the meaning of percentiles.
- 33. Suppose you received the highest score on an exam. Your friend scored the second-highest score, yet you both were in the 99th percentile. How can this be?
- 34. Morningstar is a mutual fund rating agency. It ranks a fund’s performance by using one to five stars. A one-star mutual fund is in the bottom 10% of its investment class; a five-star mutual fund is at the 90th percentile of its investment class. Interpret the meaning of a five-star mutual fund.
- 35. When outliers are discovered, should they always be removed from the data set before further analysis?
- 36. Mensa is an organization designed for people of high intelligence. One qualifies for Mensa if one’s intelligence is measured at or above the 98th percentile. Explain what this means.
- 37. Explain the advantage of using z -scores to compare observations from two different data sets.
- 38. Explain the circumstances for which the interquartile range is the preferred measure of dispersion. What is an advantage that the standard deviation has over the interquartile range?
- 39. Explain what each quartile represents.

3.5 The Five-Number Summary and Boxplots



Objectives

- ① Compute the five-number summary
- ② Draw and interpret boxplots

Historical Note



John Tukey was born on July 16, 1915, in New Bedford, Massachusetts. His parents graduated numbers 1 and 2 from Bates College and were voted “the couple most likely to give birth to a genius.” Tukey earned his undergraduate and master’s degrees in chemistry from Brown University. In 1939, he earned his doctorate in mathematics from Princeton University. He remained at Princeton and, in 1965, became the founding chair of the Department of Statistics. Among his many accomplishments, Tukey is credited with coining the terms *software* and *bit*. In the early 1970s, he discussed the negative effects of aerosol cans on the ozone layer. In December 1976, he published *Exploratory Data Analysis*, from which the following quote appears: “Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone—as the first step” (p. 3). Tukey also recommended that the 1990 Census be adjusted by means of statistical formulas. John Tukey died in New Brunswick, New Jersey, on July 26, 2000.

Let’s consider what we have learned so far. In Chapter 2, we discussed techniques for graphically representing data. These summaries included bar graphs, pie charts, histograms, stem-and-leaf plots, and time-series graphs. In Sections 3.1 to 3.4, we presented techniques for measuring the center of a distribution, spread in a distribution, and relative position of observations in a distribution of data. Why do we want these summaries? What purpose do they serve?

Well, we want these summaries to see what the data can tell us. We *explore* the data to see if they contain interesting information that may be useful in our research. The summaries make this exploration much easier. In fact, because these summaries represent an exploration, a famous statistician named John Tukey called this material **exploratory data analysis**.

Tukey defined exploratory data analysis as “detective work—numerical detective work—or graphical detective work.” He believed exploration of data is best carried out the way a detective searches for evidence when investigating a crime. Our goal is only to collect and present evidence. Drawing conclusions (or inference) is like the deliberations of the jury. What we have done so far falls under the category of exploratory data analysis. We have only collected information and presented summaries, not reached any conclusions.

We have already seen one of Tukey’s graphical summaries, the stem-and-leaf plot. In this section, we look at two more summaries: the five-number summary and the boxplot.

1 Compute the Five-Number Summary

Remember that the median is a measure of central tendency that divides the lower 50% of the data from the upper 50%. It is resistant to extreme values and is the preferred measure of central tendency when data are skewed right or left.

The three measures of dispersion presented in Section 3.2 (range, standard deviation, and variance) are not resistant to extreme values. However, the interquartile range, $Q_3 - Q_1$, the difference between the 75th and 25th percentiles, is resistant. It is interpreted as the range of the middle 50% of the data. However, the median, Q_1 , and Q_3 do not provide information about the extremes of the data, the smallest and largest values in the data set.

The **five-number summary** of a set of data consists of the smallest data value, Q_1 , the median, Q_3 , and the largest data value. We organize the five-number summary as follows:

Five-Number Summary

MINIMUM	Q_1	M	Q_3	MAXIMUM
---------	-------	-----	-------	---------

EXAMPLE 1 Obtaining the Five-Number Summary

Problem The data shown in Table 17 show the finishing times (in minutes) of the men in the 60- to 64-year-old age group in a 5-kilometer race. Determine the five-number summary of the data.

Table 17

19.95	23.25	23.32	25.55	25.83	26.28	42.47
28.58	28.72	30.18	30.35	30.95	32.13	49.17
33.23	33.53	36.68	37.05	37.43	41.42	54.63

Source: Laura Gillogly, student at Joliet Junior College.

Approach The five-number summary requires the minimum data value, Q_1 , M (the median), Q_3 , and the maximum data value. Arrange the data in ascending order and then use the procedures introduced in Section 3.4 to obtain Q_1 , M , and Q_3 .

Solution The data in ascending order are as follows:

19.95, 23.25, 23.32, 25.55, 25.83, 26.28, 28.58, 28.72, 30.18, 30.35, 30.95,
32.13, 33.23, 33.53, 36.68, 37.05, 37.43, 41.42, 42.47, 49.17, 54.63

The smallest number (the fastest time) in the data set is 19.95. The largest number in the data set is 54.63. The first quartile, Q_1 , is 26.06. The median, M , is 30.95. The third quartile, Q_3 , is 37.24. The five-number summary is

19.95 26.06 30.95 37.24 54.63

EXAMPLE 2 Obtaining the Five-Number Summary Using Technology

Problem Using statistical software or a graphing calculator, determine the five-number summary of the data presented in Table 17.

Approach We will use Minitab to obtain the five-number summary.

Solution Figure 22 shows the output supplied by Minitab. The five-number summary is highlighted.

Figure 22

Descriptive Statistics: Time**Statistics**

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Time	21	0	32.89	1.90	8.68	19.95	26.06	30.95	37.24	54.63

② Draw and Interpret Boxplots

The five-number summary can be used to create another graph, called the **boxplot**.

Drawing a Boxplot

Step 1 Determine the lower and upper fences:

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$

where $\text{IQR} = Q_3 - Q_1$

$$\text{Upper fence} = Q_3 + 1.5(\text{IQR})$$

Step 2 Draw a number line long enough to include the maximum and minimum values. Insert vertical lines at Q_1 , M , and Q_3 . Enclose these vertical lines in a box.

Step 3 Label the lower and upper fences.

Step 4 Draw a line from Q_1 to the smallest data value that is larger than the lower fence. Draw a line from Q_3 to the largest data value that is smaller than the upper fence. These lines are called **whiskers**.

Step 5 Any data values less than the lower fence or greater than the upper fence are outliers and are marked with an asterisk (*).

EXAMPLE 3 Constructing a Boxplot

Problem Use the results from Example 1 to construct a boxplot of the finishing times of the men in the 60- to 64-year-old age group.

Approach Follow the steps presented above.

Solution In Example 1, we found that $Q_1 = 26.06$, $M = 30.95$, and $Q_3 = 37.24$. Therefore, the interquartile range = IQR = $Q_3 - Q_1 = 37.24 - 26.06 = 11.18$. The difference between the 75th percentile and 25th percentile is a time of 11.18 minutes.

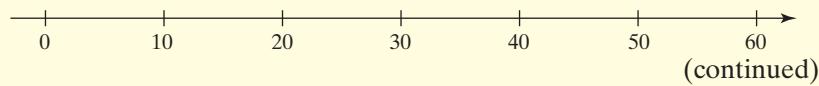
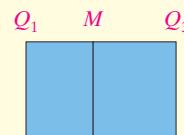
Step 1 Compute the lower and upper fences:

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR}) = 26.06 - 1.5(11.18) = 9.29$$

$$\text{Upper fence} = Q_3 + 1.5(\text{IQR}) = 37.24 + 1.5(11.18) = 54.01$$

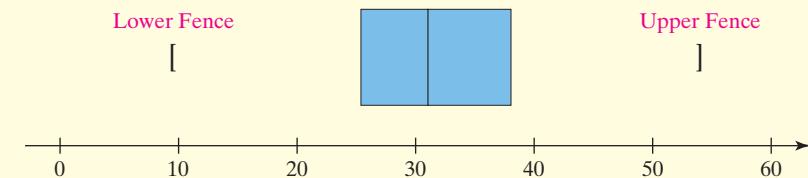
Step 2 Draw a horizontal number line with a scale that will accommodate our graph. Draw vertical lines at $Q_1 = 26.06$, $M = 30.95$, and $Q_3 = 37.24$. Enclose these lines in a box. See Figure 23(a).

Figure 23 (a)



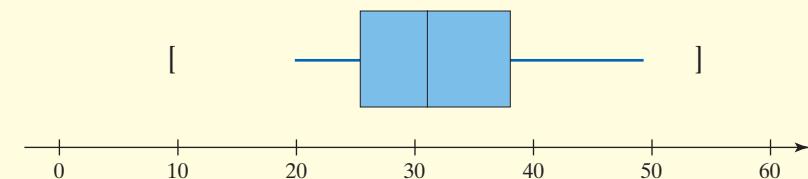
Step 3 Temporarily mark the location of the lower and upper fence with brackets ([and]). See Figure 23(b).

Figure 23 (b)



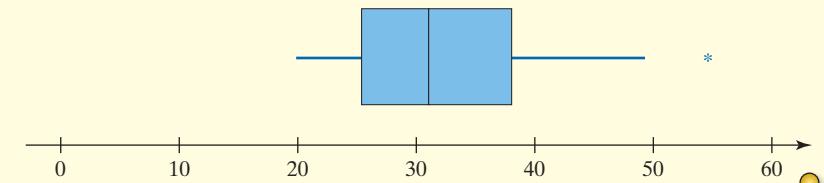
Step 4 Draw a horizontal line from Q_1 to 19.95, the smallest data value that is larger than 9.29 (the lower fence). Draw a horizontal line from Q_3 to 49.17, the largest data value that is smaller than 54.01 (the upper fence). See Figure 23(c).

Figure 23 (c)



Step 5 Plot any outliers, which are values less than 9.29 (the lower fence) or greater than 54.01 (the upper fence) using an asterisk (*). So 54.63 is an outlier. Remove the temporary brackets from the graph. See Figure 23(d).

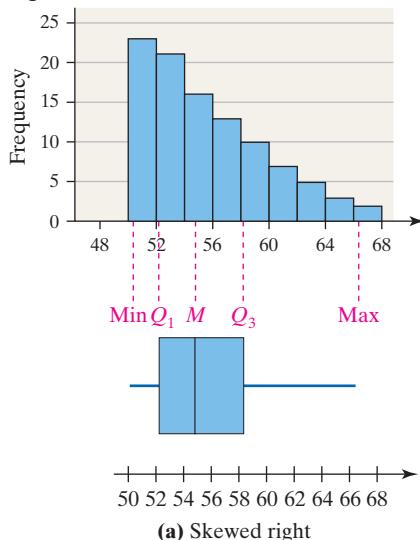
Figure 23 (d)



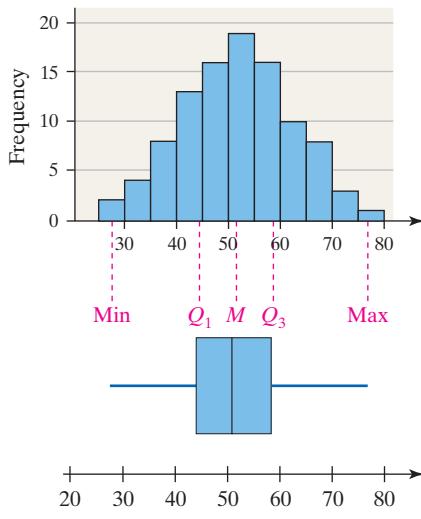
Using a Boxplot and Quartiles to Describe the Shape of a Distribution

Figure 24 shows three histograms and their corresponding boxplots with the five-number summary labeled.

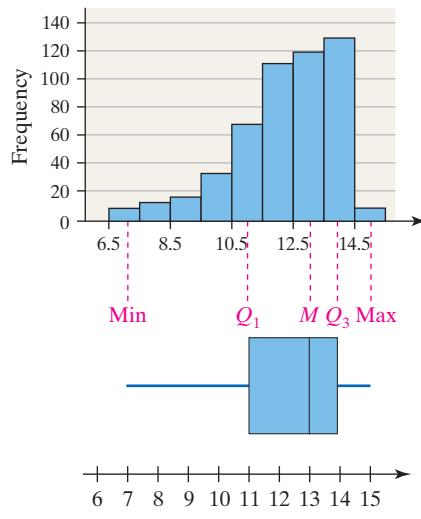
Figure 24



(a) Skewed right



(b) Symmetric



(c) Skewed left

Notice the following from the figure.

CAUTION!

Identifying the shape of a distribution from a boxplot (or from a histogram, for that matter) is subjective. When identifying the shape of a distribution from a graph, be sure to support your opinion.

- In Figure 24(a), the histogram shows the distribution is skewed right. Notice that the median is left of center in the box, which means the distance from M to Q_1 is less than the distance from M to Q_3 . In addition, the right whisker is longer than the left whisker. Finally, the distance from the median to the minimum value in the data set is less than the distance from the median to the maximum value in the data set.
- In Figure 24(b), the histogram shows the distribution is symmetric. Notice that the median is in the center of the box, so the distance from M to Q_1 is the same as the distance from M to Q_3 . In addition, the left and right whiskers are roughly the same length. Finally, the distance from the median to the minimum value in the data set is the same as the distance from the median to the maximum value in the data set.
- In Figure 24(c), the histogram shows the distribution is skewed left. Notice that the median is right of center in the box, so the distance from M to Q_1 is more than the distance from M to Q_3 . In addition, the left whisker is longer than the right whisker. Finally, the distance from the median to the minimum value in the data set is more than the distance from the median to the maximum value in the data set.

The guidelines given above are just that—guidelines. Judging the shape of a distribution is a subjective practice.

The boxplot in Figure 23(d) suggests that the distribution is skewed right, since the right whisker is longer than the left whisker and the median is left of center in the box. We can also assess the shape using the quartiles. The distance from M to Q_1 is 4.89 ($= 30.95 - 26.06$), while the distance from M to Q_3 is 6.29 ($= 37.24 - 30.95$). Also, the distance from M to the minimum value is 11 ($= 30.95 - 19.95$), while the distance from M to the maximum value is 23.68 ($= 54.63 - 30.95$).

NW Now Work Problem 11
EXAMPLE 4
Comparing Two Distributions by Using Boxplots

Problem In the Spacelab Life Sciences 2, led by Paul X. Callahan, 14 male rats were sent to space. The red blood cell mass (in milliliters) of the rats was determined when they returned. A control group of 14 male rats was held under the same conditions (except for space flight) as the space rats, and their red blood cell mass was also measured when the space rats returned. The data are in Table 18. Construct side-by-side boxplots for red blood cell mass for the flight group and control group. Does it appear that space flight affected the rats' red blood cell mass?

Table 18

Flight				Control			
7.43	7.21	8.59	8.64	8.65	6.99	8.40	9.66
9.79	6.85	6.87	7.89	7.62	7.44	8.55	8.70
9.30	8.03	7.00	8.80	7.33	8.58	9.88	9.94
6.39	7.54			7.14	9.14		

Source: NASA Life Sciences Data Archive.

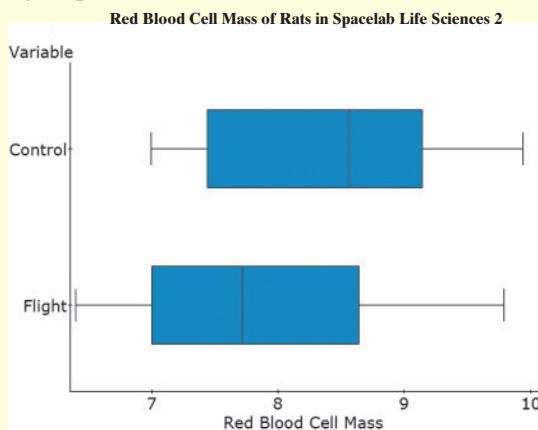
Approach Comparing two data sets is easy if we draw side-by-side boxplots on the same horizontal number line. Graphing calculators with advanced statistical features, as well as statistical spreadsheets such as Minitab, Excel, and StatCrunch, can draw boxplots. We will use StatCrunch to draw the boxplots. The steps for drawing boxplots using a TI-83/84 Plus graphing calculator, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on page 160.

Solution Figure 25 on the following page shows the side-by-side boxplots drawn in StatCrunch. It appears that the space flight has reduced the red blood cell mass of the rats since the median for the flight group ($M \approx 7.7$) is less than the median for the

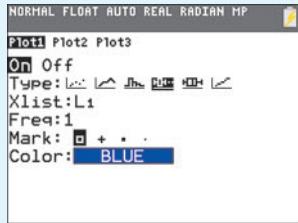
(continued)

control group ($M \approx 8.6$). The spread, as measured by the interquartile range, appears to be similar for both groups.

Figure 25

**NW Now Work Problem 15****Technology Step-by-Step****Drawing Boxplots Using Technology****TI-83/84 Plus**

- Enter the raw data into L1.
- Press 2nd Y= and select 1:Plot 1.
- Turn the plots ON. Use the cursor to highlight the modified boxplot icon. Your screen should appear as follows:



- Press ZOOM and select 9:ZoomStat.

Minitab

- Enter the raw data into column C1.
- Select the **Graph** menu and then **Boxplot**
- For a single boxplot, select One Y, simple. For two or more boxplots, select Multiple Y's, simple.
- Select the data to be graphed. If you want the boxplot to be horizontal rather than vertical, select the Scale button, then transpose value and category scales. Click OK.

Excel

- Load the XLSTAT Add-in.
- Enter the raw data into column A. If you are drawing side-by-side boxplots, enter each category of data in a separate column.
- Select the **XLSTAT** menu and highlight **Describing data**. Then select **Descriptive statistics**.
- In the Descriptive statistics dialogue box, place the cursor in the Quantitative data cell. Then highlight the data in column A. If you are drawing side-by-side boxplots, highlight all the data.
- Select the Charts tab. Check Box plots under the Chart types tab. Click the Options tab. If you want the graph drawn horizontal, select the Horizontal radio button. Check the Outliers box. If you are drawing side-by-side boxplots, check the Group plots box. Click OK.

StatCrunch

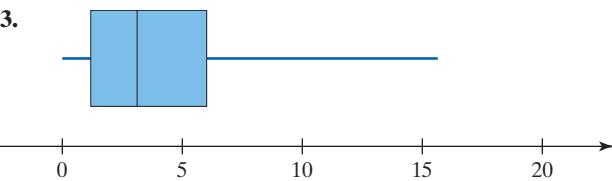
- If necessary, enter the raw data into the spreadsheet. Name the column variable.
- Select Graph and highlight Boxplot.
- Click on the variable whose boxplot you want to draw. Check the boxes "Use fences to identify outliers" and "Draw boxes horizontally." Enter label for the X-axis. Enter a title for the graph. Click Compute!.

**3.5 Assess Your Understanding****Vocabulary and Skill Building**

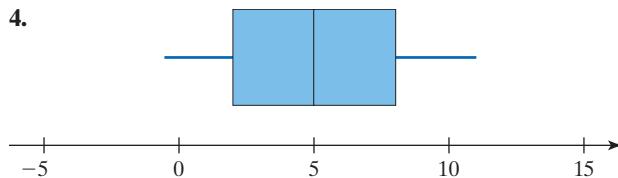
- What does the five-number summary consist of?
- In a boxplot, if the median is to the left of the center of the box and the right whisker is substantially longer than the left whisker, the distribution is skewed _____.

In Problems 3 and 4, (a) identify the shape of the distribution and (b) determine the five-number summary. Assume that each number in the five-number summary is an integer.

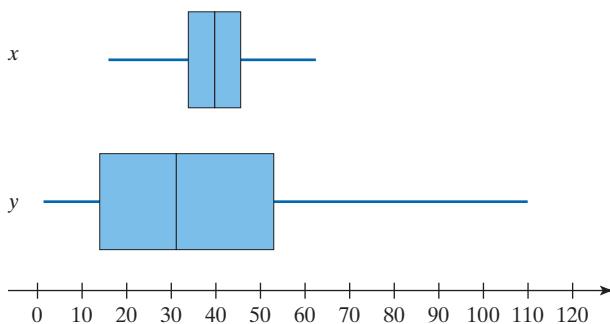
3.



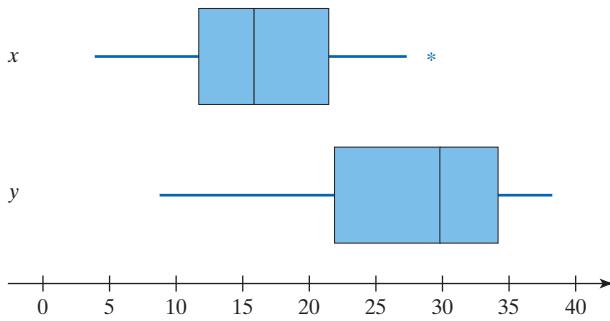
4.



5. Use the side-by-side boxplots shown to answer the questions that follow.



- (a) To the nearest integer, what is the median of variable x ?
 (b) To the nearest integer, what is the third quartile of variable y ?
 (c) Which variable has more dispersion? Why?
 (d) Describe the shape of the variable x . Support your position.
 (e) Describe the shape of the variable y . Support your position.
6. Use the side-by-side boxplots shown to answer the questions that follow.



- (a) To the nearest integer, what is the median of variable x ?
 (b) To the nearest integer, what is the first quartile of variable y ?
 (c) Which variable has more dispersion? Why?
 (d) Does the variable x have any outliers? If so, what is the value of the outlier(s)?
 (e) Describe the shape of the variable y . Support your position.

7. **Exam Scores** After giving a statistics exam, Professor Dang determined the following five-number summary for her class results: 60 68 77 89 98. Use this information to draw a boxplot of the exam scores.

8. **Speed Reading** Jessica enrolled in a course that promised to increase her reading speed. To help judge the effectiveness of the course, Jessica measured the number of words per minute she could read prior to enrolling in the course. She obtained the following five-number summary: 110 140 157 173 205. Use this information to draw a boxplot of the reading speed.

Applying the Concepts

- DATA** 9. **Age at Inauguration** The following data represent the age of U.S. presidents on their respective inauguration days (through Donald Trump).

42	47	50	52	54	55	57	61	64
43	48	51	52	54	56	57	61	65
46	49	51	54	55	56	57	61	68
46	49	51	54	55	56	58	62	69
47	50	51	54	55	57	60	64	70

Source: factmonster.com

- (a) Find the five-number summary.

- (b) Construct a boxplot.

- (c) Comment on the shape of the distribution.

- DATA** 10. **Carpoolers** The following data represent the percentage of workers who carpool to work for the 50 states plus Washington, D.C. **Note:** The minimum observation of 7.2% corresponds to Maine and the maximum observation of 16.4% corresponds to Hawaii.

7.2	8.5	9.0	9.4	10.0	10.3	11.2	11.5	13.8
7.8	8.6	9.1	9.6	10.0	10.3	11.2	11.5	14.4
7.8	8.6	9.2	9.7	10.0	10.3	11.2	11.7	16.4
7.9	8.6	9.2	9.7	10.1	10.7	11.3	12.4	
8.1	8.7	9.2	9.9	10.2	10.7	11.3	12.5	
8.3	8.8	9.4	9.9	10.3	10.9	11.3	13.6	

Source: American Community Survey by the U.S. Census Bureau.

- (a) Find the five-number summary.

- (b) Construct a boxplot.

- (c) Comment on the shape of the distribution.

- NW DATA** 11. **Age of Mother at Birth** The data below represent the age of the mother at the time of her first birth for a random sample of 30 mothers.

21	35	33	25	22	26
21	24	16	32	25	20
30	20	20	29	21	19
18	24	33	22	23	25
17	23	25	29	25	19

Source: General Social Survey.

- (a) Construct a boxplot of the data.

- (b) Use the boxplot and quartiles to describe the shape of the distribution.

- DATA** 12. **Got a Headache?** The following data represent the weight (in grams) of a random sample of 25 Tylenol tablets.

0.608	0.601	0.606	0.602	0.611
0.608	0.610	0.610	0.607	0.600
0.608	0.608	0.605	0.609	0.605
0.610	0.607	0.611	0.608	0.610
0.612	0.598	0.600	0.605	0.603

Source: Kelly Roe, student at Joliet Junior College.

- (a) Construct a boxplot.

- (b) Use the boxplot and quartiles to describe the shape of the distribution.

- DATA** **13. M&Ms** In Problem 25 from Section 3.1, we drew a histogram of the weights of M&Ms and found that the distribution is symmetric. Draw a boxplot of these data. Use the boxplot and quartiles to confirm the distribution is symmetric. For convenience, the data are displayed again.

0.87	0.88	0.82	0.90	0.90	0.84	0.84
0.91	0.94	0.86	0.86	0.86	0.88	0.87
0.89	0.91	0.86	0.87	0.93	0.88	
0.83	0.95	0.87	0.93	0.91	0.85	
0.91	0.91	0.86	0.89	0.87	0.84	
0.88	0.88	0.89	0.79	0.82	0.83	
0.90	0.88	0.84	0.93	0.81	0.90	
0.88	0.92	0.85	0.84	0.84	0.86	

Source: Michael Sullivan.

- DATA** **14. Old Faithful** In Problem 26 from Section 3.1, we drew a histogram of the length of eruption of California's Old Faithful geyser and found that the distribution is symmetric. Draw a boxplot of these data. Use the boxplot and quartiles to confirm the distribution is symmetric. For convenience, the data are displayed again.

108	108	99	105	103	103	94
102	99	106	90	104	110	110
103	109	109	111	101	101	
110	102	105	110	106	104	
104	100	103	102	120	90	
113	116	95	105	103	101	
100	101	107	110	92	108	

Source: Ladonna Hansen, Park Curator.

- NW 15. Dissolving Rates of Vitamins** A student wanted to know whether Centrum vitamins dissolve faster than the corresponding generic brand. The student used vinegar as a proxy for stomach acid and measured the time (in minutes) it took for a vitamin to completely dissolve. The results are shown next.

Centrum					Generic Brand				
2.73	3.07	3.30	3.35	3.12	6.57	6.23	6.17	7.17	5.77
2.95	2.15	2.67	2.80	2.25	6.73	5.78	5.38	5.25	5.55
2.60	2.57	4.02	3.02	2.15	5.50	6.50	7.42	6.47	6.30
3.03	3.53	2.63	2.30	2.73	6.33	7.42	5.57	6.35	5.92
3.92	2.38	3.25	4.00	3.63	5.35	7.25	7.58	6.50	4.97
3.02	4.17	4.33	3.85	2.23	7.13	5.98	6.60	5.03	7.18

Source: Amanda A. Sindewald, student at Joliet Junior College.

- (a) Draw side-by-side boxplots for each vitamin type.
(b) Which vitamin type has more dispersion?
(c) Which vitamin type appears to dissolve faster?
- DATA** **16. Chips per Cookie** Do store-brand chocolate chip cookies have fewer chips per cookie than Keebler's Chips Deluxe Chocolate Chip Cookies? To find out, a student randomly selected 21 cookies of each brand and counted the number of chips in the cookies. The results are shown in the next column.

Keebler			Store Brand		
32	23	28	21	23	24
28	28	29	24	25	27
25	20	25	26	26	21
22	21	24	18	16	24
21	24	21	21	30	17
26	28	24	23	28	31
33	20	31	27	33	29

Source: Trina McNamara, student at Joliet Junior College.

- (a) Draw side-by-side boxplots for each brand of cookie.
(b) Does there appear to be a difference in the number of chips per cookie?
(c) Does one brand have a more consistent number of chips per cookie?
- DATA** **17. Flint Water Crisis** In April, 2014, the city of Flint, Michigan stopped buying water that came from Lake Huron and started using water from the Flint River to save money. After the move, residents of Flint started experiencing symptoms that could be attributed to the consumption of lead. Open the data file 3_5_17 at www.pearsonhighered.com/sullivanstats.
- (a) Draw a boxplot of Lead (ppb), which represents the lead content in parts per billion of a sample of 71 properties in Flint, Michigan between February and June 2015.
(b) The Lead and Copper Rule of 1991 states that if more than 10% of homes have lead readings greater than 15 ppb, then action must be taken to reduce lead readings in the water. What percent of the properties had a lead reading greater than 15 ppb?
(c) After considering the data, Flint officials excluded the home at 212 Browning Avenue because it had a water filtration system and excluded the property at 625 S. Grand Traverse because it was a business (not a residential property). Draw a boxplot of the data with these properties removed.
(d) In the data set with these properties removed, what percent of the homes had a lead reading greater than 15 ppb?
(e) Discuss the impact of removing outliers from data sets as it pertains to the Flint water crisis.

- DATA** **18. Putting It Together: Paternal Smoking** It is well-documented that active maternal smoking during pregnancy is associated with lower-birth-weight babies. Researchers wanted to determine if there is a relationship between paternal smoking habits and birth weight. The researchers administered a questionnaire to each parent of newborn infants. One question asked whether the individual smoked regularly. Because the survey was administered within 15 days of birth, it was assumed that any regular smokers were also regular smokers during pregnancy. Birth weights for the babies (in grams) of nonsmoking mothers were obtained and divided into two groups, nonsmoking fathers and smoking fathers. The given data are representative of the data collected by the researchers. The researchers concluded that the birth weight of babies whose father smoked was less than the birth weight of babies whose father did not smoke.

Source: "The Effect of Paternal Smoking on the Birthweight of Newborns Whose Mothers Did Not Smoke," Fernando D. Martinez, Anne L. Wright, Lynn M. Taussig, *American Journal of Public Health* Vol. 84, No. 9.

Nonsmokers			Smokers		
4194	3522	3454	3998	3455	3066
3062	3771	3783	3150	2986	2918
3544	3746	4019	4216	3502	3457
4054	3518	3884	3493	3255	3234
4248	3719	3668	2860	3282	2746
3128	3290	3423	3686	2851	3145
3471	4354	3544	3807	3548	4104
3994	2976	4067	3963	3892	2768
3732	3823	3302	3769	3509	3629
3436	3976	3263	4131	3129	4263

- (a) Is this an observational study or a designed experiment? Why?
- (b) What is the explanatory variable? What is the response variable?
- (c) Can you think of any lurking variables that may affect the results of the study?
- (d) In the article, the researchers stated that “birthweights were adjusted for possible confounders.” What does this mean?
- (e) Determine summary statistics (mean, median, standard deviation, quartiles) for each group.
- (f) Interpret the first quartile for both the nonsmoker and smoker group.
- (g) Draw a side-by-side boxplot of the data. Does the side-by-side boxplot confirm the conclusions of the study?



19. Putting It Together: Taxi Ride Open the data 3_5_19 at www.pearsonhighered.com/sullivanstats. This data represents the amount of time of a cab ride (in seconds), the fare collected, and the payment method for a random sample of 100 rides in the City of Chicago. *Source:* Chicago Data Portal.

- (a) Create a graphical summary that would answer the question, “Which payment method is more popular, cash or credit card?”
- (b) Create a graphical summary that would answer the question, “What proportion of fares are less than \$10?”
- (c) Create a graphical summary that would answer the question, “How many of the 100 rides lasted at least 2000 seconds?”
- (d) Create a graphical summary that would answer the question, “Which payment method has more outliers?”
- (e) Determine a numerical summary that answers the question, “Which payment method has the higher fare, on average, cash or credit card?”
- (f) Determine a numerical summary that answers the question, “Which payment method has more dispersion in the fare, cash or credit card?”



20. Threaded Problem: Tornado The data set “Tornadoes_2017” located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a) Draw side-by-side boxplots of length of the tornado by month. Which month had the longest tornado? Which month does not have an outlier for the length?
- (b) Draw side-by-side boxplots of length of the tornado for F2 and F3 tornadoes. **Hint:** If using StatCrunch, enter the following in the Where: box of the dialogue window: “F Scale” ≥ 2 and “F Scale” ≤ 3 . Then select “F Scale” under “Group by:”. Which F Scale rating had the longer tornadoes? Which F scale rating had more dispersion? What was the longest tornado among these two F scale ratings?

Retain Your Knowledge

21. Retain Your Knowledge: Decision Making and Hunger

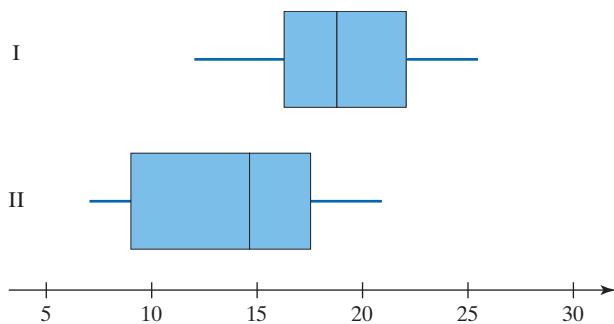
Does hunger improve strategic decision making? That is, if you are hungry are you more likely to make a favorable decision when the outcome of your decision is uncertain (as in business decisions)? To test this theory, researchers randomly divided 30 normal weight individuals into two groups. All subjects were asked to refrain from eating or drinking (except water) from 11 P.M. on the day prior to their 9 A.M. meeting. At 9 A.M., the subjects were randomly assigned to one of two groups. The subjects in Group 1 were fed breakfast while the subjects in Group 2 were not fed. All subjects were administered a computerized version of an exam that assesses complex decision making under uncertain conditions. The assessment consisted of subjects choosing cards from four decks marked A, B, C, and D. Cards in decks A and B had a point value of 100 while cards in decks C and D had point values of 50. However, deck A had penalty cards that deducted points between 150 and 300; deck B had one penalty card of 1250; deck C had penalty cards between 25 and 75 points; deck D had a single penalty card of 250 points. So, decks A and B had stiffer penalties over the long haul than decks C and D, and in the long haul, decks C and D resulted in more points than decks A and B. In total, the subjects would select 100 cards. However, the response variable was the number of cards selected from decks C and D out of the last 60 cards selected. The thinking here is that after 40 card selections, the subjects would be aware of the advantage of decks C and D. The researchers administered a Barret Impulsivity Scale to be sure the two groups did not differ in terms of impulsivity (e.g., “I do things without thinking.”) There was no difference in impulsivity, age, or body mass index between the two groups. Before the exam, subjects were asked to report their level of hunger and it was found that Group 2 was significantly more hungry than Group 1. After analysis of the data, it was determined that the mean number of advantageous cards (decks C and D) selected by the subjects in Group 2 was 33.36 cards while the mean was 25.86 for the subjects in Group 1. The researchers concluded that hunger improves advantageous decision making.

Source: de Ridder, D., Kroese, F., Adriaanse, M., & Evers, C., “Always Gamble on an Empty Stomach: Hunger Is Associated with Advantageous Decision Making,” *PLOS One* 9(10). doi: 10.1371/journal.pone.0111081.

- (a) What type of experimental design is this?
- (b) Identify the experimental units.
- (c) What is the response variable? Is it qualitative or quantitative? If quantitative, is it discrete or continuous?
- (d) What factors that might impact the response variable are cited in the article? Which factor is manipulated? How many levels are there for the manipulated factor?
- (e) What role does randomization play in the study? How do the researchers verify that randomization resulted in similar groups prior to the treatment?
- (f) What are the statistics in the study?
- (g) What is the conclusion of the study?

Explaining the Concepts

22. Which boxplot shown to the right likely has the data with a larger standard deviation? Why?
23. Explain how to determine the shape of a distribution using the boxplot and quartiles.



Chapter 3 Review

Summary

We are still in Part II of the statistical process—describing data.

This chapter concentrated on describing distributions numerically. Measures of central tendency are used to indicate the typical value in a distribution. Three measures of central tendency were discussed. The mean measures the center of gravity of the distribution. It is found by adding the observations in the data set and dividing by the number of observations. Therefore, the data must be quantitative to compute the mean. The median separates the bottom 50% of the data from the top 50%. The data must be at least ordinal (capable of arranging the data in ascending order) to compute the median. The mode measures the most frequent observation. The data can be either quantitative or qualitative to compute the mode. Resistance refers to the idea that extreme observations in the data set do not significantly impact the value of the statistic. The median is a resistant measure of central tendency, whereas the mean is not resistant.

Measures of dispersion describe the spread in the data. The range is the difference between the highest and lowest data values. The standard deviation is based on the average squared deviation about the mean. The variance is the square of the standard deviation. The range, standard deviation, and variance, are not resistant.

The mean and standard deviation are used in many types of statistical inference.

The mean, median, mode, and standard deviation can be approximated from grouped data.

We can determine the relative position of an observation using *z*-scores and percentiles. A *z*-score denotes the number of standard deviations an observation is from the mean. Percentiles determine the percent of observations that lie below a specific observation.

Quartiles are a specific type of percentile. There is the first quartile, which is the 25th percentile; the second quartile, which is the 50th percentile, or median; the third quartile, which is the 75th percentile. The interquartile range, the difference between the third and first quartile, is a resistant measure of spread.

We use quartiles to help identify outliers in a data set. To find outliers, we must determine the lower and upper fences of the data.

The five-number summary provides an idea about the center and spread of a data set through the median and interquartile range. The range of the data can be determined from the smallest and largest data values. The five-number summary is used to construct boxplots. Boxplots can be used to describe the shape of the distribution and to visualize outliers.

Vocabulary

Arithmetic mean (p. 108)

Population arithmetic mean (p. 108)

Sample arithmetic mean (p. 108)

Mean (p. 108)

Median (p. 110)

Resistant (p. 112)

Mode (p. 114)

No mode (p. 114)

Bimodal (p. 115)

Multimodal (p. 115)

Dispersion (p. 121)

Range (p. 122)

Deviation about the mean (p. 123)

Population standard deviation (p. 123)

Conceptual formula (p. 123)

Computational formula (p. 123)

Sample standard deviation (p. 125)

Degrees of freedom (p. 125)

Population variance (p. 126)

Sample variance (p. 126)

Biased (p. 129)

The Empirical Rule (p. 129)

Chebyshev's Inequality (p. 131)

Grouped data (p. 139)

Weighted mean (p. 141)

z-score (p. 146)

*k*th percentile (p. 147)

Quartiles (p. 147)

Interquartile range (p. 149)

Describe the distribution (p. 150)

Outlier (p. 150)

Fences (p. 151)

Exploratory data analysis (p. 155)

Five-number summary (p. 156)

Boxplot (p. 157)

Whiskers (p. 157)

Formulas

Population Mean

$$\mu = \frac{\sum x_i}{N}$$

Sample Mean

$$\bar{x} = \frac{\sum x_i}{n}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N}}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}}$$

Population Variance

$$\sigma^2$$

Sample Variance

$$s^2$$

Range = Largest Data Value – Smallest Data Value

Weighted Mean

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

Population Mean from Grouped Data

$$\mu = \frac{\sum x_i f_i}{\sum f_i}$$

Sample Mean from Grouped Data

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}$$

Population Standard Deviation from Grouped Data

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2 f_i}{\sum f_i}}$$

Sample Standard Deviation from Grouped Data

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{(\sum f_i) - 1}}$$

Population z-Score

$$z = \frac{x - \mu}{\sigma}$$

Sample z-Score

$$z = \frac{x - \bar{x}}{s}$$

Interquartile Range

$$IQR = Q_3 - Q_1$$

Lower and Upper Fences

$$\text{Lower Fence} = Q_1 - 1.5(IQR)$$

$$\text{Upper Fence} = Q_3 + 1.5(IQR)$$

Objectives

Section	You should be able to . . .	Example(s)	Review Exercises
3.1	1 Determine the arithmetic mean of a variable from raw data (p. 108) 2 Determine the median of a variable from raw data (p. 110) 3 Explain what it means for a statistic to be resistant (p. 112) 4 Determine the mode of a variable from raw data (p. 114)	1, 4, 6 2, 3, 4, 6 5 7, 8, 9	1(a), 2(a), 3(a), 4(c), 10(a) 1(a), 2(a), 3(a), 4(c), 10(a) 2(c), 10(h), 10(i), 12 3(a), 4(d)
3.2	1 Determine the range of a variable from raw data (p. 122) 2 Determine the standard deviation of a variable from raw data (p. 123) 3 Determine the variance of a variable from raw data (p. 128) 4 Use the Empirical Rule to describe data that are bell shaped (p. 129) 5 Use Chebyshev's Inequality to describe any set of data (p. 131)	2 3–6 7 8 9	1(b), 2(b), 3(b) 1(b), 2(b), 3(b), 10(d) 1(b) 5(a)–(d) 5(e)–(f)
3.3	1 Approximate the mean of a variable from grouped data (p. 139) 2 Compute the weighted mean (p. 140) 3 Approximate the standard deviation of a variable from grouped data (p. 141)	1, 4 2 3, 4	6(a) 7 6(b)
3.4	1 Determine and interpret z-scores (p. 146) 2 Interpret percentiles (p. 147) 3 Determine and interpret quartiles (p. 147) 4 Determine and interpret the interquartile range (p. 149) 5 Check a set of data for outliers (p. 150)	1 2 3, 4 5 6	8 11 10(b) 2(b), 10(d) 10(e)
3.5	1 Compute the five-number summary (p. 155) 2 Draw and interpret boxplots (p. 157)	1, 2 3, 4	10(c) 9, 10(f)–10(g)

Review Exercises

- 1. Muzzle Velocity** The following data represent the muzzle velocity (in meters per second) of rounds fired from a 155-mm gun.

793.8	793.1	792.4	794.0	791.4
792.4	791.7	792.3	789.6	794.4

Source: Christenson, Ronald, and Blackwood, Larry, "Tests for Precision and Accuracy of Multiple Measuring Devices," *Technometrics*, 35(4): 411–421, 1993.

- (a) Compute the sample mean and median muzzle velocity.
 (b) Compute the range, sample variance, and sample standard deviation.

- 2. Price of Chevy Cobalts** The following data represent the sales price (in dollars) for nine 2-year-old Chevrolet Cobalts in the Los Angeles area.

14050	13999	12999	10995	9980
8998	7889	7200	5500	

Source: cars.com

- (a) Determine the sample mean and median price.
 (b) Determine the range, sample standard deviation, and interquartile range.
 (c) Redo (a) and (b) if the data value 14,050 was incorrectly entered as 41,050. How does this change affect the mean? the median? the range? the standard deviation? the interquartile range? Which of these values is resistant?

- DATA 3. Chief Justices** The following data represent the ages of chief justices of the U.S. Supreme Court when they were appointed.

Justice	Age
John Jay	44
John Rutledge	56
Oliver Ellsworth	51
John Marshall	46
Roger B. Taney	59
Salmon P. Chase	56
Morrison R. Waite	58
Melville W. Fuller	55
Edward D. White	65
William H. Taft	64
Charles E. Hughes	68
Harlan F. Stone	69
Frederick M. Vinson	56
Earl Warren	62
Warren E. Burger	62
William H. Rehnquist	62
John G. Roberts	50

Source: *Information Please Almanac*.

- (a) Determine the population mean, median, and mode ages.
 (b) Determine the range and population standard deviation ages.
 (c) Obtain two simple random samples of size 4, and determine the sample mean and sample standard deviation ages.

- DATA 4. Number of Tickets Issued** As part of a statistics project, a student surveys 30 randomly selected students and asks them how many speeding tickets they have been issued in the past month. The results of the survey are as follows:

1	1	0	1	0	0
0	0	0	1	0	1
0	1	2	0	1	1
0	0	0	0	1	1
0	0	0	0	1	0

- (a) Draw a frequency histogram of the data and describe its shape.
 (b) Based on the shape of the histogram, do you expect the mean to be more than, equal to, or less than the median?
 (c) Determine the mean and the median of the number of tickets issued.
 (d) Determine the mode of the number of tickets issued.

- 5. Chebyshev's Inequality and the Empirical Rule** Suppose that a certain brand of light bulb has a mean life of 600 hours and a standard deviation of 53 hours.

- (a) A histogram of the data indicates the sample data follow a bell-shaped distribution. According to the Empirical Rule, 99.7% of light bulbs have lifetimes between _____ and _____ hours.
 (b) Assuming the data are bell shaped, determine the percentage of light bulbs that will have a life between 494 and 706 hours.
 (c) Assuming the data are bell shaped, what percentage of light bulbs will last between 547 and 706 hours.
 (d) If the company that manufactures the light bulb guarantees to replace any bulb that does not last at least 441 hours, what percentage of light bulbs can the firm expect to have to replace, according to the Empirical Rule?
 (e) Use Chebyshev's Inequality to determine the minimum percentage of light bulbs with a life within 2.5 standard deviations of the mean.
 (f) Use Chebyshev's Inequality to determine the minimum percentage of light bulbs with a life between 494 and 706 hours.

- 6. Travel Time to Work** The frequency distribution listed in the table represents the travel time to work (in minutes) for a random sample of 895 U.S. adults.

Travel Time (minutes)	Frequency
0–9	125
10–19	271
20–29	186
30–39	121
40–49	54
50–59	62
60–69	43
70–79	20
80–89	13

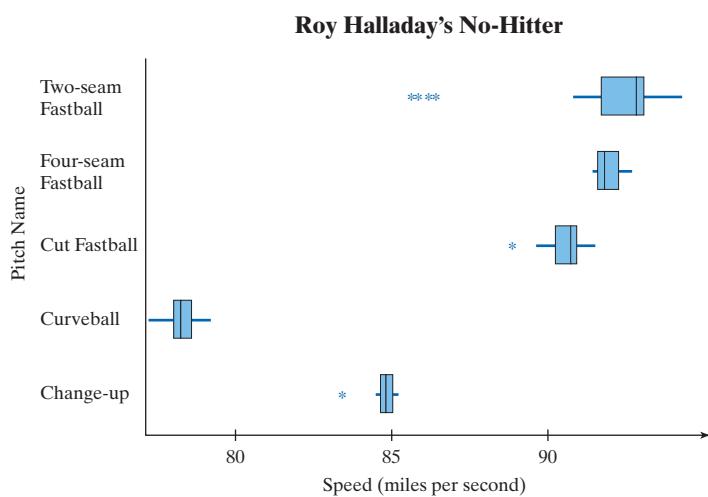
Source: Based on data from the American Community Survey.

- (a) Approximate the mean travel time to work for U.S. adults.
 (b) Approximate the standard deviation travel time to work for U.S. adults.

7. Weighted Mean Michael has just completed his first semester in college. He earned an A in his five-hour calculus course, a B in his four-hour chemistry course, an A in his three-hour speech course, and a C in his three-hour psychology course. Assuming an A equals 4 points, a B equals 3 points, and a C equals 2 points, determine Michael's grade-point average if grades are weighted by class hours.

8. Weights of Males versus Females According to the National Center for Health Statistics, the mean weight of a 20- to 29-year-old female is 156.5 pounds, with a standard deviation of 51.2 pounds. The mean weight of a 20- to 29-year-old male is 183.4 pounds, with a standard deviation of 40.0 pounds. Who is relatively heavier: a 20- to 29-year-old female who weighs 160 pounds or a 20- to 29-year-old male who weighs 185 pounds?

9. Halladay No-No On October 6, 2010, Roy Halladay of the Philadelphia Phillies threw the second post-season no-hitter in Major League history. The side-by-side boxplot shows the pitch speed (in miles per hour) for all of Halladay's pitches during the game.



- (a) Which pitch is typically thrown the fastest?
- (b) Which pitch is most erratic as far as pitch speed goes?
- (c) Which pitch is more consistent as far as pitch speed goes, the cut fastball or the four-seam fastball?
- (d) Are there any outliers for Halladay's cut fastball? If so, approximate the pitch speed of the outlier.
- (e) Describe the shape of the distribution of Halladay's curveball.
- (f) Describe the shape of the distribution of Halladay's four-seam fastball.

10. Presidential Inaugural Addresses Ever wonder how many words are in a typical inaugural address? The data in the next column represent the lengths of all the inaugural addresses (measured in word count) for all presidents up to Donald Trump.

1425	1125	1128	5433	2242	2283
135	1172	1337	1802	2446	1507
2308	3838	2480	1526	2449	2170
1729	8445	2978	3318	1355	1571
2158	4776	1681	4059	1437	2073
1175	996	4388	3801	2130	2406
1209	3319	2015	1883	1668	2137
3217	2821	3967	1807	1087	1433
4467	3634	2217	1340	2463	
2906	698	985	559	2546	

Source: infoplease.com

- (a) Determine the mean and median number of words in a presidential inaugural address.
- (b) Determine and interpret the quartiles for the number of words in a presidential inaugural address.
- (c) Determine the five-number summary for the number of words in a presidential inaugural address.
- (d) Determine the standard deviation and interquartile range for the number of words in a presidential inaugural address.
- (e) Are there any outliers in the data set? If so, what is (are) the value(s)?
- (f) Draw a boxplot of the data.
- (g) Describe the shape of the distribution. Support your position using the boxplot and quartiles.
- (h) Which measure of central tendency do you think better describes the typical number of words in an inaugural address? Why?
- (i) Which measure of dispersion do you think better describes the spread of the typical number of words in an inaugural address? Why?

11. You Explain It! Percentiles According to the National Center for Health Statistics, a 19-year-old female whose height is 67.1 inches has a height that is at the 85th percentile. Explain what this means.

12. Skinfold Thickness Procedure One method of estimating body fat is through skinfold thickness measurement using from three to nine different standard anatomical sites around the body from the right side only (for consistency). The tester pinches the skin at the appropriate site to raise a double layer of skin and the underlying adipose tissue, but not the muscle. Calipers are then applied 1 centimeter below and at right angles to the pinch and a reading is taken 2 seconds later. The mean of two measurements should be taken. If the two measurements differ greatly, a third should be done and then the median value taken. Explain why a median is used as the measure of central tendency when three measures are taken, rather than the mean.

DATA **10. Presidential Inaugural Addresses** Ever wonder how many words are in a typical inaugural address? The data in the next column represent the lengths of all the inaugural addresses (measured in word count) for all presidents up to Donald Trump.

Chapter Test

1. The following data represent the amount of time (in minutes) a random sample of eight students enrolled in Sullivan's Intermediate Algebra course spent on the homework from Section 4.5, Factoring Polynomials.

48	88	57	109
111	93	71	63

Source: MyLabMath.

- (a) Determine the mean amount of time spent doing Section 4.5 homework.
- (b) Determine the median amount of time spent doing Section 4.5 homework.
- (c) Suppose the observation 109 minutes is incorrectly recorded as 1009 minutes. Recompute the mean and the median. What do you notice? What property of the median does this illustrate?

2. The Federal Bureau of Investigation classifies various larcenies. The data below represent the type of larcenies based on a random sample of 15 larcenies. What is the mode type of larceny?

Pocket picking and purse snatching	Bicycles	From motor vehicles
From motor vehicles	From motor vehicles	From buildings
From buildings	Shoplifting	Motor vehicle accessories
From motor vehicles	Shoplifting	From motor vehicles
From motor vehicles	Pocket picking and purse snatching	From motor vehicles

3. Determine the range of the homework data from Problem 1.
4. (a) Determine the standard deviation of the homework data from Problem 1.
(b) By hand, determine and interpret the interquartile range of the homework data from Problem 1.
(c) Which of these two measures of dispersion is resistant? Why?
5. In a random sample of 250 toner cartridges, the mean number of pages a toner cartridge can print is 4302 and the standard deviation is 340.
(a) Suppose a histogram of the data indicates that the sample data follow a bell-shaped distribution. According to the Empirical Rule, 99.7% of toner cartridges will print between _____ and _____ pages.
(b) Assuming that the distribution of the data are bell shaped, determine the percentage of toner cartridges whose print total is between 3622 and 4982 pages.
(c) If the company that manufactures the toner cartridges guarantees to replace any cartridge that does not print at least 3622 pages, what percent of cartridges can the firm expect to be responsible for replacing, according to the Empirical Rule?
(d) Use Chebyshev's inequality to determine the minimum percentage of toner cartridges with a page count within 1.5 standard deviations of the mean.
(e) Use Chebyshev's inequality to determine the minimum percentage of toner cartridges that print between 3282 and 5322 pages.
6. The following data represent the length of time (in minutes) between eruptions of Old Faithful in Yellowstone National Park.

Time (minutes)	Frequency
40–49	8
50–59	44
60–69	23
70–79	6
80–89	107
90–99	11
100–109	1

- (a) Approximate the mean length of time between eruptions.
(b) Approximate the standard deviation length of time between eruptions.

7. Yolanda wishes to develop a new type of meatloaf to sell at her restaurant. She decides to combine 2 pounds of ground sirloin (cost \$2.70 per pound), 1 pound of ground turkey (cost \$1.30 per pound), and $\frac{1}{2}$ pound of ground pork (cost \$1.80 per pound). What is the cost per pound of the meatloaf?

8. An engineer is studying bearing failures for two different materials in aircraft gas turbine engines. The following data are failure times (in millions of cycles) for samples of the two material types.

Material A	Material B
3.17	5.88
4.31	6.91
4.52	8.01
4.66	8.97
5.69	11.92
	8.20
	24.37

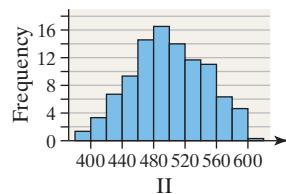
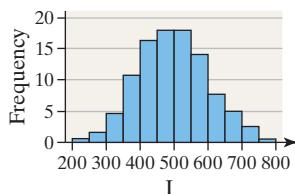
- (a) Determine the sample mean failure time for each material.
(b) By hand, compute the median failure time for each material.
(c) Determine the sample standard deviation of the failure times for each material. Which material has its failure times more dispersed?
(d) By hand, compute the five-number summary for each material.
(e) On the same graph, draw boxplots for the two materials. Annotate the graph with some general remarks comparing the failure times.
(f) Describe the shape of the distribution of each material using the boxplot and quartiles.

9. The following data represent the weights (in grams) of 50 randomly selected quarters. Determine and interpret the quartiles. Does the data set contain any outliers?

5.49	5.58	5.60	5.62	5.68
5.52	5.58	5.60	5.63	5.70
5.53	5.58	5.60	5.63	5.71
5.53	5.58	5.60	5.63	5.71
5.53	5.58	5.60	5.65	5.72
5.56	5.58	5.60	5.66	5.73
5.57	5.59	5.60	5.66	5.73
5.57	5.59	5.61	5.66	5.73
5.57	5.59	5.62	5.67	5.74
5.57	5.59	5.62	5.67	5.84

10. Armando is filling out a college application that requires he supply either his SAT math score or his ACT math score. Armando scored 610 on the SAT math and 27 on the ACT math. Which score should Armando report, given that the mean SAT math score is 515 with a standard deviation of 114, and the mean ACT math score is 21.0 with a standard deviation of 5.1? Why?

- 11.** According to the National Center for Health Statistics, a 10-year-old male whose height is 53.5 inches has a height that is at the 15th percentile. Explain what this means.
- 12.** The distribution of income tends to be skewed to the right. Suppose you are running for a congressional seat and wish to portray that the average income in your district is low. Which measure of central tendency, the mean or the median, would you report? Why?
- 13.** Answer the following based on the histograms shown in the next column.
- (a) Which measure of central tendency would you recommend reporting for the data whose histogram is shown in Figure I? Why?
- (b) Which one has more dispersion? Explain.



- 14.** Explain how the standard deviation measures dispersion. In your explanation, include a discussion of deviation about the mean.

Making an Informed Decision

What Car Should I Buy?

Suppose you are in the market to purchase a used car. To make an informed decision regarding your purchase, you would like to collect as much information as possible. Among the information you might consider are the typical price of the car, the typical number of miles the car should have, its crash test results, insurance costs, and expected repair costs.

1. Make a list of at least three cars that you would consider purchasing. To be fair, the cars should be in the same class (such as compact, midsize, and so on). They should also be of the same age.
2. Collect information regarding the three cars in your list by finding at least eight cars of each type that are for sale. Obtain information such as the asking price and the number of miles the car has. Sources of data include your local newspaper, classified ads, and car websites (such as www.cars.com). Compute summary statistics for asking price, number of miles, and other variables of interest. Using the same scale, draw boxplots of each variable considered.

3. Go to the Insurance Institute for Highway Safety website (www.iihs.org). Select the Ratings link. Choose the make and model for each car you are considering. Obtain information regarding crash testing for each car under consideration. Compare cars in the same class. How does each car compare? Is one car you are considering substantially safer than the others? What about repair costs? Compute summary statistics for crash tests and repair costs.



4. Obtain information about insurance costs. Contact various insurance companies to determine the cost of insuring the cars you are considering. Compute summary statistics for insurance costs and draw boxplots.
5. Write a report supporting your conclusion regarding which car you would purchase.



Describing the Relation between Two Variables

Outline

- 4.1** Scatter Diagrams and Correlation
- 4.2** Least-Squares Regression
- 4.3** The Coefficient of Determination
- 4.4** Contingency Tables and Association

Making an Informed Decision



The world is a very interesting and dynamic place. How do quantitative variables relate to each other on a world scale? A website that allows us to see how the world is changing over time and, in particular, how relationships among variables in our world change over time is www.gapminder.org. See the Decisions Project on page 224.

Putting It Together

In Chapters 2 and 3 we examined data in which a single variable was measured for each individual in the study (**univariate data**), such as the fine for camera and parking violations in New York City (the variable) for various cars (the individuals). We found both graphical and numerical descriptive measures for the variable.

In this chapter, we discuss graphical and numerical methods for describing **bivariate data**, data in which two variables are measured on an individual. For example, we might want to know whether the amount of cola consumed per week is related to one's bone density. The individuals would be the people in the study, and the two variables would be the amount of cola and bone density. In this study, both variables are quantitative. We present methods for describing the relation between two quantitative variables in Sections 4.1–4.3.

Suppose we want to know whether level of education is related to one's employment status (employed or unemployed). Here, both variables are qualitative. We present methods for describing the relation between two qualitative variables in Section 4.4.

Situations may also occur in which one variable is quantitative and the other is qualitative. We have already presented a technique for describing this situation. Look back at Example 4 in Section 3.5, where we considered whether space flight affected red blood cell mass. There, space flight is qualitative (rat sent to space or not), and red blood cell mass is quantitative.

4.1 Scatter Diagrams and Correlation



Preparing for This Section Before getting started, review the following:

- Mean (Section 3.1, pp. 108–110)
- z -scores (Section 3.4, pp. 146–147)
- Standard deviation (Section 3.2, pp. 123–128)
- Confounding, lurking variables, and confounding variables (Section 1.2, pp. 16–17)

Objectives

- ① Draw and interpret scatter diagrams
- ② Describe the properties of the linear correlation coefficient
- ③ Compute and interpret the linear correlation coefficient
- ④ Determine whether a linear relation exists between two variables
- ⑤ Explain the difference between correlation and causation

Before we can represent bivariate data graphically, we must decide which variable will be used to predict the value of the other variable. For example, it seems reasonable to think that as the speed at which a golf club is swung increases, the distance the golf ball travels also increases. Therefore, we might use club-head speed to predict distance. We call distance the *response (or dependent) variable* and club-head speed the *explanatory (or predictor or independent) variable*.

Definition

The **response variable** is the variable whose value can be explained by the value of the **explanatory or predictor variable**.

1

Draw and Interpret Scatter Diagrams

IN OTHER WORDS
We use the term **explanatory variable** because it helps to explain variability in the response variable.

Definition

A **scatter diagram** is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram. The explanatory variable is plotted on the horizontal axis, and the response variable is plotted on the vertical axis.

EXAMPLE 1

Drawing a Scatter Diagram

Table 1

Club-head Speed (mph)	Distance (yards)
100	257
102	264
103	274
101	266
105	277
100	263
99	258
105	275

Source: Paul Stephenson, student at Joliet Junior College.

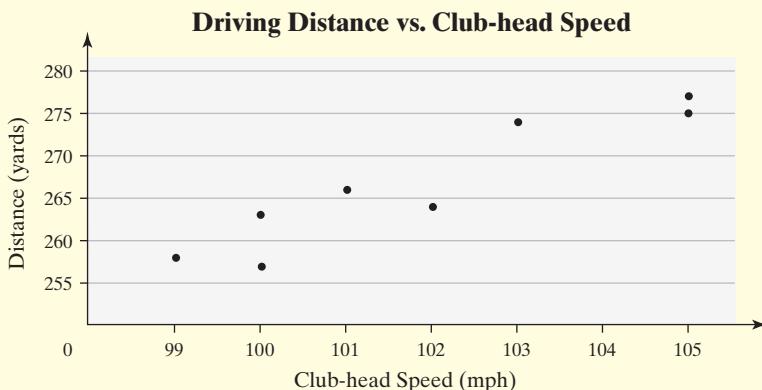
Problem A golf pro wants to investigate the relation between the club-head speed of a golf club (measured in miles per hour) and the distance (in yards) that the ball will travel. He realizes other variables besides club-head speed determine the distance a ball will travel (such as club type, ball type, golfer, and weather conditions). To eliminate the variability due to these variables, the pro uses a single model of club and ball, one golfer, and a clear, 70-degree day with no wind. The pro records the club-head speed, measures the distance the ball travels, and collects the data in Table 1. Draw a scatter diagram of the data.

Approach Because the pro wants to use club-head speed to predict the distance the ball travels, club-head speed is the explanatory variable (horizontal axis) and distance is the response variable (vertical axis). Plot the ordered pairs (100, 257), (102, 264), and so on, in a rectangular coordinate system.

(continued)

Solution The scatter diagram is shown in Figure 1. It appears from the graph that as club-head speed increases, the distance the ball travels increases as well.

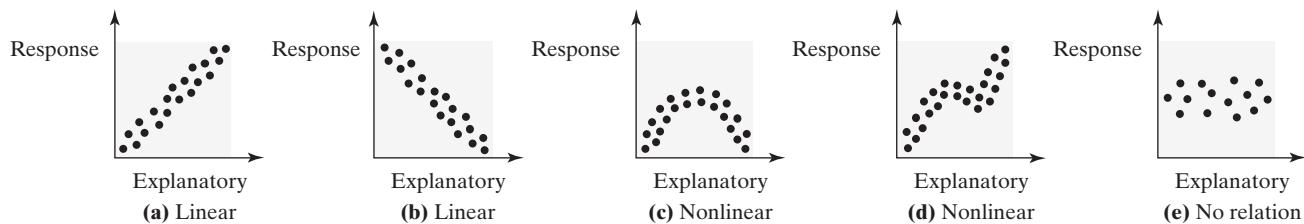
Figure 1

**CAUTION!**

Do not connect points when drawing a scatter diagram.

NW Now Work Problems 27(a) and 27(b)

Figure 2



Notice the difference between Figure 2(a) and Figure 2(b). The data follow a linear pattern that slants upward to the right in Figure 2(a) and downward to the right in Figure 2(b). Figures 2(c) and 2(d) show nonlinear relations. In Figure 2(e), there is no relation between the explanatory and response variables.

Definitions

Two variables that are linearly related are **positively associated** when above-average values of one variable are associated with above-average values of the other variable and below-average values of one variable are associated with below-average values of the other variable. That is, two variables are positively associated if, whenever the value of one variable increases, the value of the other variable also increases.

Two variables that are linearly related are **negatively associated** when above-average values of one variable are associated with below-average values of the other variable. That is, two variables are negatively associated if, whenever the value of one variable increases, the value of the other variable decreases.

IN OTHER WORDS

If two variables are positively associated, then as one goes up the other also tends to go up. If two variables are negatively associated, then as one goes up the other tends to go down.

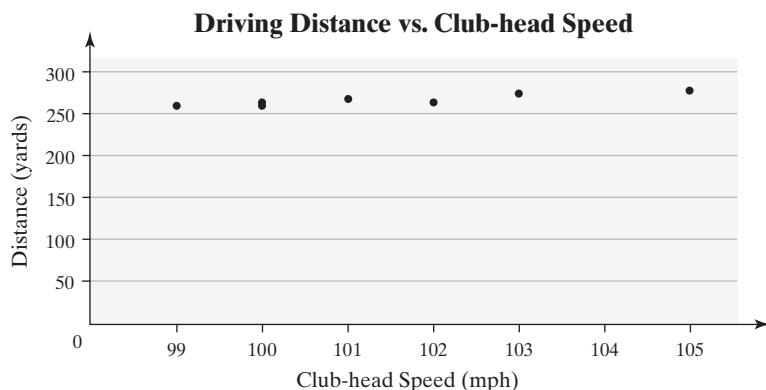
NW Now Work Problem 9

The scatter diagram in Figure 1 implies that club-head speed is positively associated with the distance a golf ball travels.

② Describe the Properties of the Linear Correlation Coefficient

It is dangerous to use only a scatter diagram to determine if two variables are linearly related. In Figure 3, we have redrawn the scatter diagram from Figure 1 using a different vertical scale.

Figure 3

**CAUTION!**

The horizontal and vertical scales of a scatter diagram should be set so that the scatter diagram does not mislead a reader.

It is more difficult to conclude that the variables are related in Figure 3 than in Figure 1. The moral of the story is this: Just as we can manipulate the scale of graphs of univariate data, we can also manipulate the scale of the graphs of bivariate data, possibly resulting in incorrect conclusions. Therefore, numerical summaries of bivariate data should be used in addition to graphs to describe any relation that exists between two variables.

Definition

The **linear correlation coefficient** or **Pearson product moment correlation coefficient** is a measure of the strength and direction of the linear relation between two quantitative variables. The Greek letter ρ (rho) represents the population correlation coefficient, and r represents the sample correlation coefficient. We present only the formula for the sample correlation coefficient.

Sample Linear Correlation Coefficient*

$$r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)}{n - 1} \quad (1)$$

where x_i is the i th observation of the explanatory variable

\bar{x} is the sample mean of the explanatory variable

s_x is the sample standard deviation of the explanatory variable

y_i is the i th observation of the response variable

\bar{y} is the sample mean of the response variable

s_y is the sample standard deviation of the response variable

n is the number of individuals in the sample

*An equivalent computational formula for the linear correlation coefficient is

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right)} \sqrt{\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n} \right)}} \quad (1)$$

The Pearson linear correlation coefficient is named in honor of Karl Pearson (1857–1936). See the Historical Note on page 176.

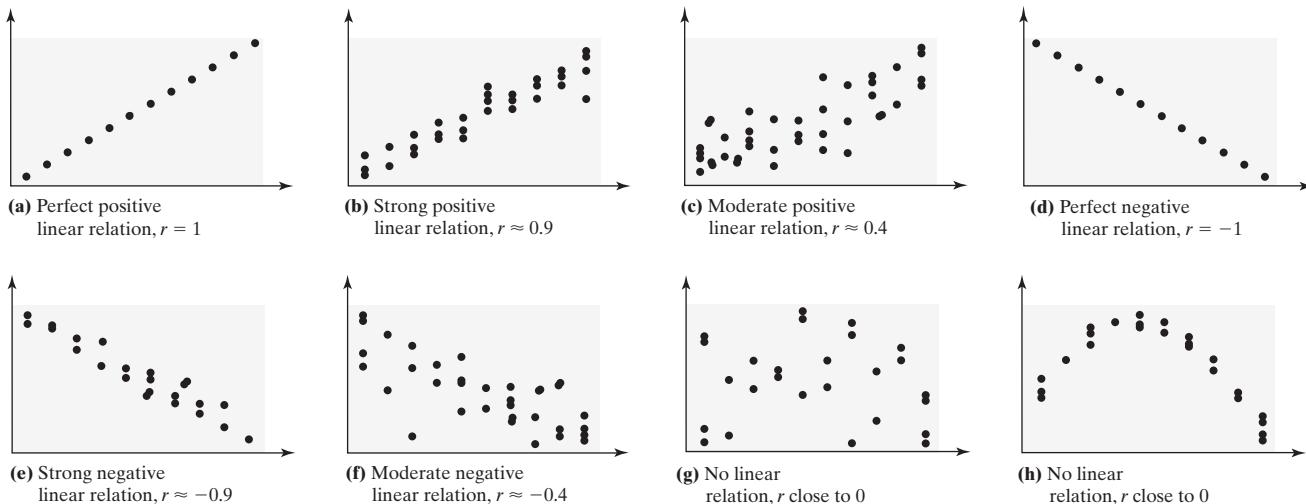
Properties of the Linear Correlation Coefficient

1. The linear correlation coefficient is always between -1 and 1 , inclusive. That is, $-1 \leq r \leq 1$.
2. If $r = +1$, then a perfect positive linear relation exists between the two variables. See Figure 4(a).
3. If $r = -1$, then a perfect negative linear relation exists between the two variables. See Figure 4(d).
4. The closer r is to $+1$, the stronger is the evidence of positive association between the two variables. See Figures 4(b) and 4(c).
5. The closer r is to -1 , the stronger is the evidence of negative association between the two variables. See Figures 4(e) and 4(f).
6. If r is close to 0 , then little or no evidence exists of a *linear* relation between the two variables. So **r close to 0 does not imply no relation, just no linear relation**. See Figures 4(g) and 4(h).
7. The linear correlation coefficient is a unitless measure of association. So the unit of measure for x and y plays no role in the interpretation of r .
8. The correlation coefficient is not resistant. Therefore, an observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.

CAUTION!

A linear correlation coefficient close to 0 does not imply that there is no relation, just no linear relation. For example, although the scatter diagram drawn in Figure 4(h) indicates that the two variables are related, the linear correlation coefficient is close to 0 .

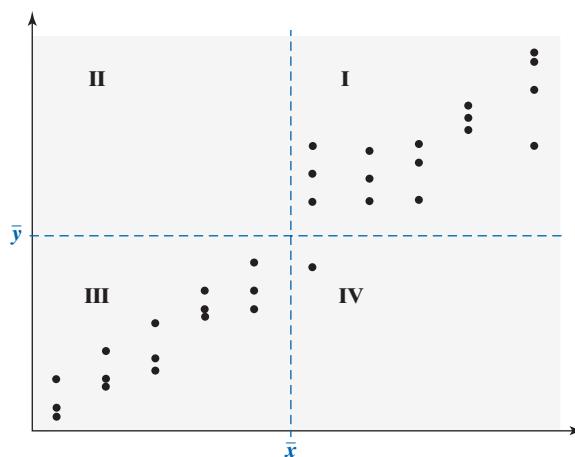
Figure 4



In Formula (1), notice that the numerator is the sum of the products of z -scores for the explanatory (x) and response (y) variables. A positive linear correlation coefficient means that the sum of the products of the z -scores for x and y must be positive. How does this occur? Figure 5 on the following page shows a scatter diagram with positive association between x and y . The vertical dashed line represents the value of \bar{x} , and the horizontal dashed line represents the value of \bar{y} . These two dashed lines divide the scatter diagram into four quadrants, labeled I, II, III, and IV.

Consider the data in quadrants I and III. If a certain x -value is above its mean, \bar{x} , then the corresponding y -value will be above its mean, \bar{y} . If a certain x -value is below its mean, \bar{x} , then the corresponding y -value will be below its mean, \bar{y} . Therefore, for data in quadrant I, we have $\frac{x_i - \bar{x}}{s_x}$ positive and $\frac{y_i - \bar{y}}{s_y}$ positive, so their product is positive. For data in quadrant III, we have $\frac{x_i - \bar{x}}{s_x}$ negative and $\frac{y_i - \bar{y}}{s_y}$ negative, so their product is positive. The sum of these products is positive, so the linear correlation coefficient is positive. A similar argument can be made for negative correlation.

Figure 5



Now suppose that the data are equally dispersed in the four quadrants. Then the negative products (resulting from data in quadrants II and IV) will offset the positive products (resulting from data in quadrants I and III). The result is a linear correlation coefficient close to 0.

NW Now Work Problem 13

3 Compute and Interpret the Linear Correlation Coefficient

In practice, linear correlation coefficients are found using technology. However, we present one computation by hand so that you may gain an appreciation of how the formula measures the strength of linear relation.

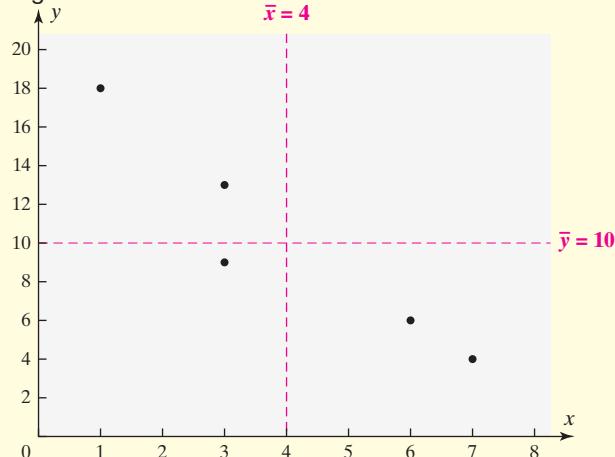
EXAMPLE 2 Computing the Correlation Coefficient by Hand

Problem For the data shown in Table 2, compute the linear correlation coefficient. A scatter diagram of the data is shown in Figure 6. The dashed lines on the scatter diagram represent the mean of x and the mean of y .

Table 2

x	y
1	18
3	13
3	9
6	6
7	4

Figure 6



Approach

Step 1 Compute \bar{x} , s_x , \bar{y} , and s_y .

Step 2 Determine $\frac{x_i - \bar{x}}{s_x}$ and $\frac{y_i - \bar{y}}{s_y}$ for each observation.

Step 3 Compute $\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$ for each observation.

Step 4 Determine $\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$ and substitute this value into Formula (1).

(continued)

Historical Note

Karl Pearson was born March 27, 1857. Pearson's statistical proficiency was recognized early in his life. It is said that his mother told him not to suck his thumb, because otherwise it would wither away. Pearson analyzed the size of each thumb and said to himself, "They look alike to me. I can't see that the thumb I suck is any smaller than the other. I wonder if she could be lying to me."

Karl Pearson graduated from Cambridge University in 1879. From 1893 to 1911, he wrote 18 papers on genetics and heredity. Through this work, he developed ideas regarding correlation and the chi-square test. (See Chapter 12.) In addition, Pearson coined the term *standard deviation*.

Pearson and Ronald Fisher (see page 45) didn't get along. Their dispute was severe enough that Fisher turned down the post of chief statistician at the Galton Laboratory in 1919 because it would have meant working under Pearson.

Pearson died on April 27, 1936.

Solution

Step 1 We find $\bar{x} = 4$, $s_x = 2.44949$, $\bar{y} = 10$, and $s_y = 5.612486$.

Step 2 Columns 1 and 2 of Table 3 contain the data from Table 2. We determine $\frac{x_i - \bar{x}}{s_x}$ and $\frac{y_i - \bar{y}}{s_y}$ in Columns 3 and 4 of Table 3.

Table 3

x	y	$\frac{x_i - \bar{x}}{s_x}$	$\frac{y_i - \bar{y}}{s_y}$	$\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$
1	18	-1.2247	1.4254	-1.7457
3	13	-0.4083	0.5345	-0.2182
3	9	-0.4083	-0.1782	0.0727
6	6	0.8165	-0.7127	-0.5819
7	4	1.2247	-1.0690	-1.3093
$\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = -3.7824$				

Step 3 Multiply the entries in Columns 3 and 4 to obtain the entries in Column 5 of Table 3.

Step 4 Add the entries in Column 5 and substitute this value into Formula (1) to obtain the correlation coefficient.

$$r = \frac{\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n - 1} = \frac{-3.7824}{5 - 1} = -0.946$$

We will agree to round the correlation coefficient to three decimal places. The correlation coefficient suggests a strong negative association between the two variables.

Compare the signs of the entries in Columns 3 and 4 in Table 3. Notice that negative values in Column 3 correspond with positive values in Column 4 and that positive values in Column 3 correspond with negative values in Column 4 (except for the third observation). Look back at the scatter diagram in Figure 6, where the mean of x and the mean of y is drawn. Notice that below-average values of x are associated with above-average values of y , and above-average values of x are associated with below-average values of y . This is why the linear correlation coefficient is negative.

EXAMPLE 3**Determining the Linear Correlation Coefficient Using Technology**

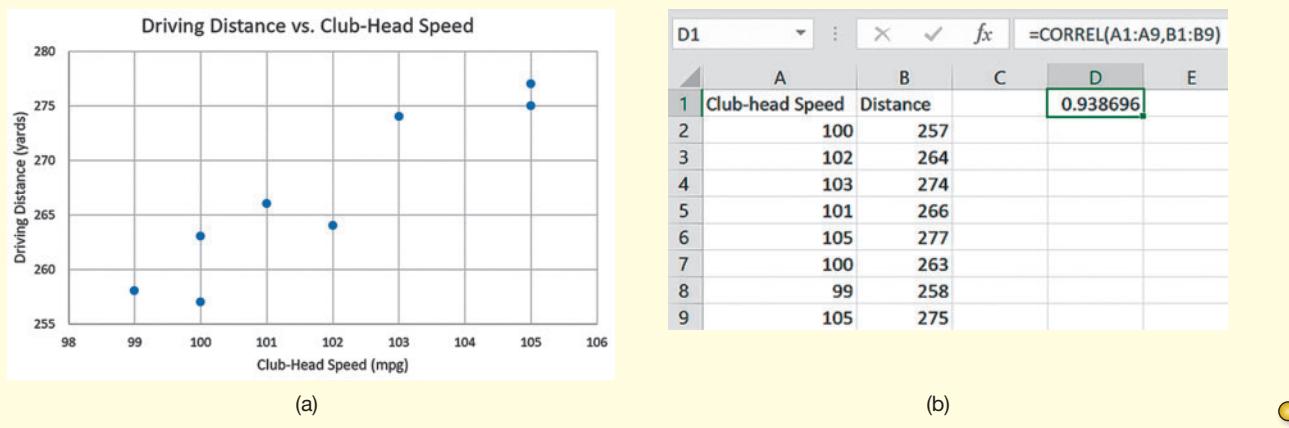
Problem Use a statistical spreadsheet or a graphing calculator with advanced statistical features to draw a scatter diagram of the data in Table 1. Then determine the linear correlation between club-head speed and distance.

Approach We will use Excel to draw the scatter diagram and obtain the linear correlation coefficient. The steps for drawing scatter diagrams and obtaining the linear correlation coefficient using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on pages 179–180.

Solution Figure 7(a) on the following page shows the scatter diagram, and Figure 7(b) shows the linear correlation coefficient of 0.939 obtained from Excel.

Because the linear correlation coefficient is positive, we know above-average values of x , club-head speed, are associated with above-average values of y , driving distance, and below-average values of x are associated with below-average values of y .

Figure 7

**NW Now Work Problem 27(c)**

We stated in Property 8 that the linear correlation coefficient is not resistant. For example, suppose the golfer in Examples 1 and 3 hits one more golf ball. As he swings, a car driving by blows its horn, which distracts the golfer. His swing speed is 105 mph, but the mis-hit golf ball travels only 255 yards. The linear correlation coefficient with this additional observation decreases to 0.535.

④ Determine Whether a Linear Relation Exists between Two Variables

A question you may be asking yourself is “How do I know the correlation between two variables is strong enough for me to conclude that a linear relation exists between them?” Although rigorous tests can answer this question, for now we will be content with a simple comparison test.

Testing for a Linear Relation

Step 1 Determine the absolute value of the correlation coefficient.

Step 2 Find the critical value in Table II from Appendix A for the given sample size.

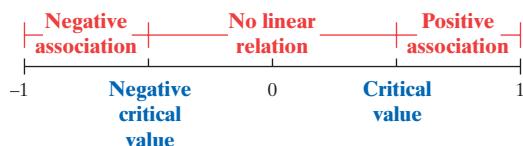
Step 3 If the absolute value of the correlation coefficient is greater than the critical value, then a linear relation exists between the two variables. Otherwise, no linear relation exists.

IN OTHER WORDS

We use two vertical bars to denote absolute value, as in $|5|$ or $|-4|$. Recall, $|5| = 5$, $|-4| = 4$, and $|0| = 0$.

Another way to think about the procedure is to consider Figure 8. If the correlation coefficient is positive and greater than the critical value, the variables are positively associated. If the correlation coefficient is negative and less than the opposite of the critical value, the variables are negatively associated.

Figure 8



EXAMPLE 4 Does a Linear Relation Exist?

Problem Using the results from Example 3, determine whether a linear relation exists between club-head speed and distance.

Approach Follow Steps 1 through 3 given on the previous page.

Solution

Step 1 The linear correlation coefficient between club-head speed and distance is 0.939, so $|0.939| = 0.939$.

Step 2 Table II shows the critical value with $n = 8$ is 0.707.

Step 3 Because $|0.939| = 0.939 > 0.707$, a positive association (positive linear relation) exists between club-head speed and distance. 

NW Now Work Problem 27(d)

⑤ Explain the Difference between Correlation and Causation

In Chapter 1 we stated that there are two types of studies: observational studies and designed experiments. The data examined in Examples 1, 3, and 4 are the result of an experiment. Therefore, we can claim that a faster club-head speed causes the golf ball to travel a longer distance.

CAUTION!

A linear correlation coefficient that implies a strong positive or negative association does not imply causation if it was computed using observational data.

If data used in a study are observational, we cannot conclude the two correlated variables have a causal relationship. For example, the correlation between teenage birthrate and homicide rate since 1993 is 0.9987, but we cannot conclude that higher teenage birthrates cause a higher homicide rate because the data are observational. In fact, time-series data are often correlated because both variables happen to move in the same (or opposite) direction over time. Both teenage birthrates and homicide rates have been declining since 1993, so they have a high positive correlation.

Is there another way two variables can be correlated without there being a causal relation? Yes—through a *lurking variable*. A **lurking variable** is related to both the explanatory variable and response variable. For example, as air-conditioning bills increase, so does the crime rate. Does this mean that folks should turn off their air conditioners so that crime rates decrease? Certainly not! In this case, the lurking variable is air temperature. As air temperatures rise, both air-conditioning bills and crime rates rise.

EXAMPLE 5 Lurking Variables in a Bone Mineral Density Study

Problem Because colas tend to replace healthier beverages and colas contain caffeine and phosphoric acid, researchers Katherine L. Tucker and associates wanted to know whether cola consumption is associated with lower bone mineral density in women. Table 4 on the following page lists the typical number of cans of cola consumed in a week and the femoral neck bone mineral density for a sample of 15 women. The data were collected through a prospective cohort study.

Figure 9 shows the scatter diagram of the data. The correlation between number of colas per week and bone mineral density is -0.806 . The critical value for correlation with $n = 15$ from Table II in Appendix A is 0.514. Because $|-0.806| > 0.514$, we conclude a negative linear relation exists between number of colas consumed and bone mineral density. Can the authors conclude that an increase in the number of colas consumed causes a decrease in bone mineral density? Identify some lurking variables in the study.

Table 4

Number of Colas per Week	Bone Mineral Density (g/cm^2)
0	0.893
0	0.882
1	0.891
1	0.881
2	0.888
2	0.871
3	0.868
3	0.876
4	0.873
5	0.875
5	0.871
6	0.867
7	0.862
7	0.872
8	0.865

Source: Based on data obtained from Katherine L. Tucker et al., "Colas, but not other carbonated beverages, are associated with low bone mineral density in older women: The Framingham Osteoporosis Study." *American Journal of Clinical Nutrition* 2006, 84:936–942.

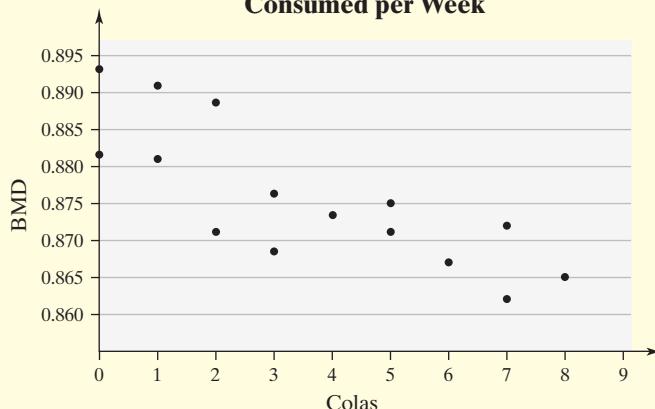
IN OTHER WORDS

Confounding means that any relation that may exist between two variables may be due to a third variable not accounted for in the study.

NW Now Work Problem 43

Figure 9

Bone Mineral Density vs. Colas Consumed per Week



Approach To claim causality, the data must be collected through a designed experiment. Remember, lurking variables are related to both the explanatory and response variables in a study.

Solution In prospective cohort studies, data are collected on a group of subjects through questionnaires and surveys over time. Therefore, the data are observational. So the researchers cannot claim that increased cola consumption causes a decrease in bone mineral density.

In their article, the authors identified a number of lurking variables that could confound the results:

Variables that could potentially confound the relation between cola consumption and bone mineral density . . . included the following: age, body mass index, height, smoking, average daily intakes of alcohol, calcium, caffeine, total energy intake, physical activity, season of measurement, estrogen use, and menopause status.

The authors were careful to say that increased cola consumption is *associated* with lower bone mineral density because of potential lurking variables. They never stated that increased cola consumption *causes* lower bone mineral density.

Technology Step-by-Step

Drawing Scatter Diagrams and Determining the Correlation Coefficient

TI-83/84 Plus

Scatter Diagrams

- Enter the explanatory variable in L1 and the response variable in L2.
- Press 2nd Y= to bring up the StatPlot menu. Select 1: Plot1.
- Turn Plot 1 on by highlighting the On button and pressing ENTER.
- Highlight the scatter diagram icon and press ENTER. Be sure that Xlist is L1 and Ylist is L2.
- Press ZOOM and select 9: ZoomStat.

Correlation Coefficient

- Enter the explanatory variable values in L1 and the response variable values in L2.
- Turn the diagnostics on by selecting the catalog (2nd 0). Scroll down and select DiagnosticOn. Hit ENTER

twice to activate diagnostics (this step needs to be done only once).

- From the HOME screen, press STAT, highlight CALC, and select 4: LinReg(ax + b). Select L1 for Xlist and L2 for Ylist. Highlight Calculate and press ENTER.

Minitab

Scatter Diagrams

- Enter the explanatory variable in C1 and the response variable in C2. You may want to name the variables.
- Select the Graph menu and then Scatterplot . . .
- Highlight the Simple icon and click OK.
- With the cursor in the Y column, select the response variable. With the cursor in the X column, select the explanatory variable. Click OK.

(continued)

Correlation Coefficient

- With the explanatory variable in C1 and the response variable in C2, select the **Stat** menu and highlight **Basic Statistics**. Select **Correlation . . .**.
- Select the variables whose correlation you wish to determine and click **OK**.

Excel**Scatter Diagrams**

- Enter the explanatory variable in column A and the response variable in column B.
- Highlight both sets of data.
- Select Insert. Choose the scatter diagram icon.

Correlation Coefficient

- Select Formulas and then More Functions. Highlight Statistical.
- Select CORREL.
- With the cursor in Array 1, highlight the data containing the explanatory variable. With the cursor in Array 2, highlight the data containing the response variable. Click OK.

StatCrunch**Scatter Diagrams**

- If necessary, enter the explanatory variable in column var1 and the response variable in column var2. Name each column variable.
- Select **Graph** and highlight **Scatter Plot**.
- Choose the explanatory variable for the X column and the response variable for the Y column. Enter the labels for the X-axis and Y-axis. Enter a title for the graph. Click **Compute!**.

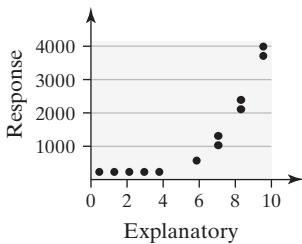
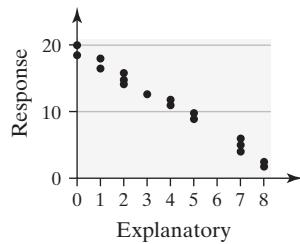
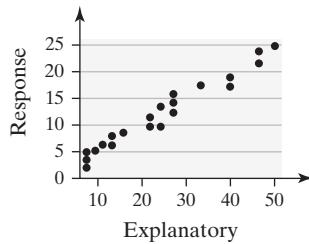
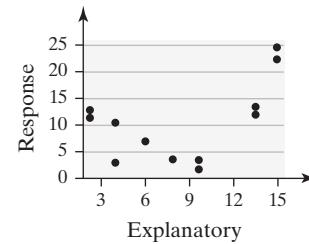
Correlation Coefficient

- If necessary, enter the explanatory variable in column var1 and the response variable in column var2. Name each column variable.
- Select **Stat**, highlight **Summary Stats**, and select **Correlation**.
- Click on the variables whose correlation you want to determine. Click **Compute!**.

**4.1 Assess Your Understanding****Vocabulary and Skill Building**

- What is the difference between univariate data and bivariate data?
- The _____ variable is the variable whose value can be explained by the value of the explanatory variable.
- A _____ is a graph that shows the relation between two quantitative variables.
- What does it mean to say two variables are positively associated? Negatively associated?
- State the properties of the linear correlation coefficient.
- True or False:* If the linear correlation coefficient is close to 0, then the two variables have no relation.
- A _____ variable is a variable that is related to both the explanatory and response variable.
- True or False:* Correlation implies causation.

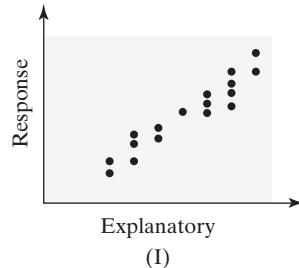
In Problems 9–12, determine whether the scatter diagram indicates that a linear relation may exist between the two variables. If the relation is linear, determine whether it indicates a positive or negative association between the variables.

NW 9.**10.****11.****12.**

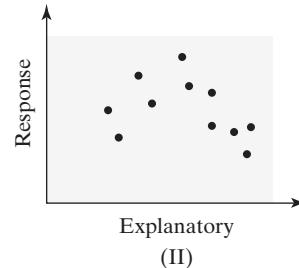
- NW** 13. Match the linear correlation coefficient to the scatter diagram. The scales on the *x*- and *y*-axes are the same for each diagram.

- (a) $r = 0.787$
(c) $r = -0.053$

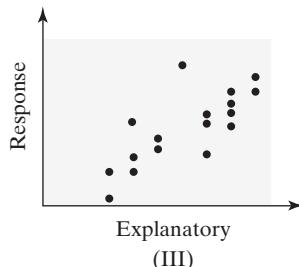
- (b) $r = -0.787$
(d) $r = 0.946$



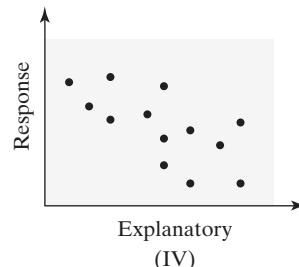
(I)



(II)



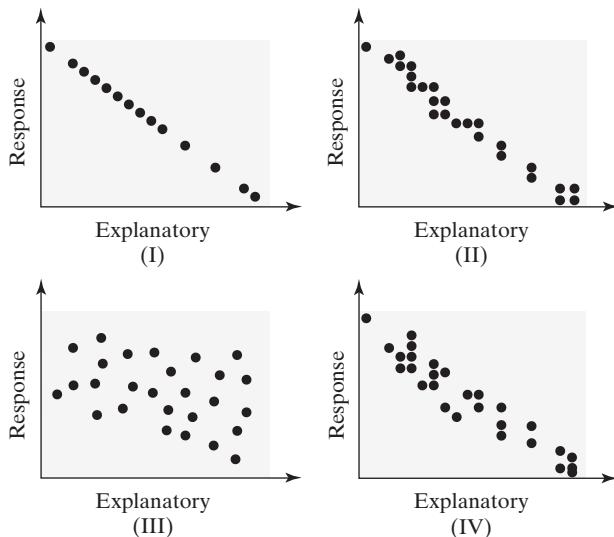
(III)



(IV)

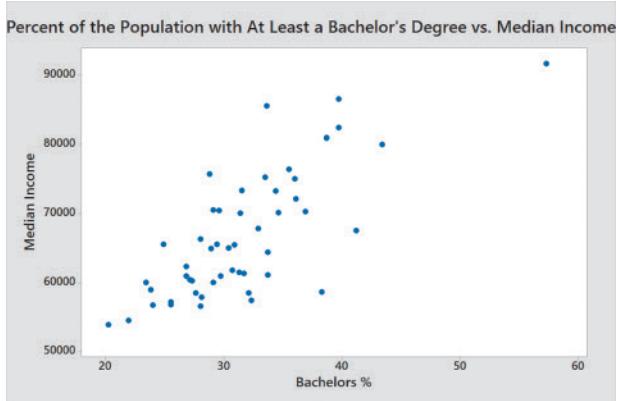
- 14.** Match the linear correlation coefficient to the scatter diagram. The scales on the x - and y -axes are the same for each diagram.

- (a) $r = -0.969$ (b) $r = -0.049$
 (c) $r = -1$ (d) $r = -0.992$

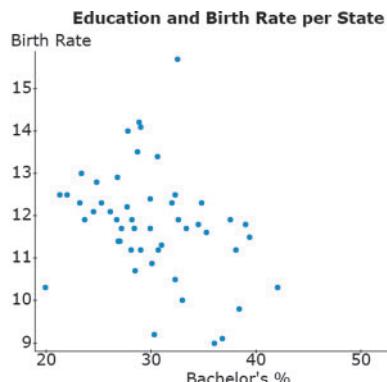


- 15. Does Education Pay?** The scatter diagram drawn in Minitab shows the relation between the percentage of the population of a state plus Washington, DC, that has at least a bachelor's degree and the median income (in dollars) of the state for 2017.

Source: U.S. Census Bureau.



- (a) Describe any relation that exists between level of education and median income.
 (b) One observation appears to stick out from the rest. Which one? This particular observation is for Washington, DC. Can you think of any reasons why Washington, DC, might stick out from the rest of the states?
 (c) The correlation coefficient between the percentage of the population with a bachelor's degree and median income is 0.760. Does a linear relation exist between percent of the population with at least a bachelor's degree and median income?
16. Relation between Education and Birthrate? The scatter diagram in the next column drawn in StatCrunch shows the relation between percent of the population with at least a bachelor's degree in a state and birthrate (births per 1000 women 15 to 44 years old).



- (a) Describe any relation that exists between median income and birthrate.
 (b) The correlation between percent of population with at least a bachelor's degree and birthrate is -0.127 . What does this imply about the relation between median income and birthrate?

In Problems 17–20, (a) draw a scatter diagram of the data, (b) by hand, compute the correlation coefficient, and (c) determine whether there is a linear relation between x and y .

17.

x	2	4	6	6	7
y	4	8	10	13	20

18.

x	2	3	5	6	6
y	10	9	7	4	2

19.

x	2	6	6	7	9
y	8	7	6	9	5

20.

x	0	5	7	8	9
y	3	8	6	9	4

- 21. Name the Relation, Part I** For each of the following statements, explain whether you think the variables will have positive correlation, negative correlation, or no correlation. Support your opinion.

- (a) Number of children in the household under the age of 3 and expenditures on diapers
 (b) Interest rates on car loans and number of cars sold
 (c) Number of hours per week on the treadmill and cholesterol level
 (d) Price of a Big Mac and number of McDonald's french fries sold in a week
 (e) Foot length and IQ

- 22. Name the Relation, Part II** For each of the following statements, explain whether you think the variables will have positive correlation, negative correlation, or no correlation. Support your opinion.

- (a) Number of cigarettes smoked by a pregnant woman each week and birth weight of her baby
 (b) Years of education and annual salary
 (c) Number of doctors on staff at a hospital and number of administrators on staff
 (d) Head circumference and IQ
 (e) Number of moviegoers and movie ticket price

23. Put the following correlation coefficients in order from weakest to strongest in terms of strength of linear association.

$-0.903, 0.339, -0.431, 0.137, 0.869$

24. Put the following correlation coefficients in order from weakest to strongest in terms of strength of linear association.

$-1, 0.377, 0.084, -0.436, 0.444, -0.733$

Applying the Concepts

25. **The TIMMS Exam** The Trends in International Mathematics and Science (TIMMS) is a mathematics and science achievement exam given internationally. On each exam, students are asked to respond to a variety of background questions. For the 41 nations that participated in TIMMS, the correlation between the percentage of items answered in the background questionnaire (used as a proxy for student task persistence) and mean score on the exam was 0.79. Does this suggest there is a linear relation between student task persistence and achievement score? Write a sentence that explains what this result might mean.

26. **The TIMMS Exam Part II** (See Problem 25) For the 41 nations that participated in TIMMS, the correlation between the percentage of students who skipped class at least once in the past month and the mean score on the exam was -0.52 . Does this suggest there is a linear relation between attendance and achievement score? Write a sentence that explains what this result might mean.

- NW 27. An Unhealthy Commute** The Gallup Organization regularly surveys adult Americans regarding their commute time to work. In addition, they administer a Well-Being Survey. According to the Gallup Organization, "The Gallup-Healthways Well-Being Index Composite Score is comprised of six sub-indices: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment and Basic Access." A complete description of the index can be found at <http://www.well-beingindex.com/>. The data in the following table are based on the results of the survey, which represent commute time to work (in minutes) and well-being index score.

Commute Time (in minutes)	Gallup-Healthways Well-Being Index Composite Score
5	69.2
15	68.3
25	67.5
35	67.1
50	66.4
72	66.1
105	63.9

Source: The Gallup Organization.

- (a) Which variable do you believe is likely the explanatory variable and which is the response variable?
(b) Draw a scatter diagram of the data.
(c) Determine the linear correlation coefficient between commute time and well-being index score.
(d) Does a linear relation exist between the commute time and well-being index score?

28. **Credit Scores** Your Fair Isaacs Corporation (FICO) credit score is used to determine your creditworthiness. It is used to help determine whether you qualify for a mortgage or credit and is

even used to determine insurance rates. FICO scores have a range of 300 to 850, with a higher score indicating a better credit history. The given data represent the interest rate (in percent) a bank would offer on a 36-month auto loan for various FICO scores.

Credit Score	Interest Rate (percent)
545	18.982
595	17.967
640	12.218
675	8.612
705	6.680
750	5.150

Source: www.myfico.com

- (a) Which variable do you believe is likely the explanatory variable and which is the response variable?
(b) Draw a scatter diagram of the data.
(c) Determine the linear correlation coefficient between FICO score and interest rate on a 36-month auto loan.
(d) Does a linear relation exist between the FICO score and interest rate?
29. **Height versus Head Circumference** A pediatrician wants to determine the relation that may exist between a child's height and head circumference. She randomly selects eleven 3-year-old children from her practice, measures their heights and head circumference, and obtains the data shown in the table below.

Height (inches)	Head Circumference (inches)	Height (inches)	Head Circumference (inches)
27.75	17.5	26.5	17.3
24.5	17.1	27	17.5
25.5	17.1	26.75	17.3
26	17.3	26.75	17.5
25	16.9	27.5	17.5
27.75	17.6		

Source: Denise Slucki, student at Joliet Junior College.

- (a) If the pediatrician wants to use height to predict head circumference, determine which variable is the explanatory variable and which is the response variable.
(b) Draw a scatter diagram of the data.
(c) Compute the linear correlation coefficient between the height and head circumference of a child.
(d) Does a linear relation exist between height and head circumference?
(e) Convert the data to centimeters (1 inch = 2.54 cm), and recompute the linear correlation coefficient. What effect did the conversion have on the linear correlation coefficient?

30. **American Black Bears** The American black bear (*Ursus americanus*) is one of eight bear species in the world. It is the smallest North American bear and the most common bear species on the planet. In 1969, Dr. Michael R. Pelton of the University of Tennessee initiated a long-term study of the population in the Great Smoky Mountains National Park. One aspect of the study was to develop a model that could be used to predict a bear's weight (since it is not practical to weigh bears in the field). One variable thought to be related to weight is the

length of the bear. The following data represent the lengths and weights of 12 American black bears.

Total Length (cm)	Weight (kg)
139.0	110
138.0	60
139.0	90
120.5	60
149.0	85
141.0	100
141.0	95
150.0	85
166.0	155
151.5	140
129.5	105
150.0	110

Source: fieldtripearth.org

- (a) Which variable is the explanatory variable based on the goals of the research?
- (b) Draw a scatter diagram of the data.
- (c) Determine the linear correlation coefficient between weight and length.
- (d) Does a linear relation exist between the weight of the bear and its length?

DATA 31. Home Runs The following data represent the speed at which a ball was hit (in miles per hour) and the distance it traveled (in feet) for a random sample of home runs in a Major League baseball game.

Speed (mph)	Distance (feet)
107.9	441
110.4	427
103.5	422
105.4	418
105.5	414
101.7	411
103.3	408
101.0	405
103.6	402
101.4	399
100.7	396
101.3	393

Source: baseballsavant.mlb.com

- (a) A researcher wants to determine if the speed with which the ball is hit may be used to predict the distance the ball travels. Based on this research goal, what is the explanatory variable?
- (b) Draw a scatter diagram of the data.
- (c) Compute the linear correlation coefficient between speed and distance.
- (d) Does a linear relation exist between the speed at which a ball is hit and the distance the ball travels?

DATA 32. Hurricanes The data in the next column represent the maximum wind speed (in knots) and atmospheric pressure (in millibars) for a random sample of hurricanes that originated in the Atlantic Ocean.

Atmospheric Pressure (mb)	Wind Speed (knots)	Atmospheric Pressure (mb)	Wind Speed (knots)
993	50	1006	40
995	60	942	120
994	60	1002	40
997	45	986	50
1003	45	983	70
1004	40	994	65
1000	55	940	120
994	55	976	80
942	105	966	100
1006	30	982	55

Source: National Hurricane Center.

- (a) Draw a scatter diagram treating atmospheric pressure as the explanatory variable.
- (b) Compute the linear correlation coefficient between atmospheric pressure and wind speed.
- (c) Does a linear relation exist between atmospheric pressure and wind speed?

DATA 33. CEO Performance The following data represent the total compensation for 12 randomly selected chief executive officers (CEO) and the company's stock performance in 2017.

Company	Compensation (millions of dollars)	Stock Return (%)
Macerich Co	12.8	-2.9
Regency Centers Corp	5.6	3.6
Rockwell Collins Inc	8.1	57
Kohls Corp	11.3	71
Interpublic Group Of Companies, Inc	16.9	-11
Microsoft Corp	20	38
Henry Schein Inc	7.2	-7.9
International Flavors & Fragrances Inc	7.7	32
Hanesbrands Inc	9.6	-0.3
Walgreens Boots Alliance, Inc	14.7	2.8
Cerner Corp	2.6	42
M&T Bank Corp	4.2	11

Source: The Wall Street Journal.

- (a) One would think that a higher stock return would lead to a higher compensation. Based on this, what would likely be the explanatory variable?
- (b) Draw a scatter diagram of the data.
- (c) Determine the linear correlation coefficient between compensation and stock return.
- (d) Does a linear relation exist between compensation and stock return? Does stock performance appear to play a role in determining the compensation of a CEO?

DATA 34. Bear Markets A bear market in the stock market is defined as a condition in which the market declines by 20% or more over the course of at least two months. The following data

represent the number of months and percentage change in the S&P 500 (a group of 500 stocks) for a sample of bear markets.

Months	Percent Change	Months	Percent Change
1.9	-44.57	12.1	-20.57
8.3	-44.29	14.9	-21.63
3.3	-32.86	6.5	-27.97
3.4	-42.54	8	-22.18
6.8	-61.81	18.1	-36.06
5.8	-40.60	21	-48.20
3.1	-29.43	20.7	-27.11
13.4	-31.81	3.4	-33.51
12.9	-54.47	18.2	-36.77
5.1	-24.44	9.3	-33.75
7.6	-28.69	13.6	-51.93
17.9	-34.42	2.1	-27.62
11.8	-28.47		

Source: Gold-Eagle.

- (a) Treating the length of the bear market as the explanatory variable, draw a scatter diagram of the data.
- (b) Determine the linear correlation coefficient between months and percent change.
- (c) Does a linear relation exist between duration of the bear market and market performance?

DATA 35. Does Size Matter? Researchers wondered whether the size of a person's brain was related to the individual's mental capacity. They selected a sample of right-handed introductory psychology students who had SAT scores higher than 1350. The subjects took the Wechsler Adult Intelligence Scale-Revised to obtain their IQ scores. MRI scans were performed at the same facility for the subjects. The scans consisted of 18 horizontal MR images. The computer counted all pixels with a nonzero gray scale in each of the 18 images, and the total count served as an index for brain size.

Gender	MRI Count	IQ	Gender	MRI Count	IQ
Female	816,932	133	Male	949,395	140
Female	951,545	137	Male	1,001,121	140
Female	991,305	138	Male	1,038,437	139
Female	833,868	132	Male	965,353	133
Female	856,472	140	Male	955,466	133
Female	852,244	132	Male	1,079,549	141
Female	790,619	135	Male	924,059	135
Female	866,662	130	Male	955,003	139
Female	857,782	133	Male	935,494	141
Female	948,066	133	Male	949,589	144

Source: L. Willerman, R. Schultz, J. N. Rutledge, and E. Bigler (1991). "In Vivo Brain Size and Intelligence," *Intelligence*, 15, 223–228.

- (a) Draw a scatter diagram treating MRI count as the explanatory variable and IQ as the response variable. Comment on what you see.
- (b) Compute the linear correlation coefficient between MRI count and IQ. Are MRI count and IQ linearly related?

(c) A lurking variable in the analysis is gender. Draw a scatter diagram treating MRI count as the explanatory variable and IQ as the response variable, but use a different plotting symbol for each gender. For example, use a circle for males and a triangle for females. What do you notice?

(d) Compute the linear correlation coefficient between MRI count and IQ for females. Compute the linear correlation coefficient between MRI count and IQ for males. Are MRI count and IQ linearly related? What is the moral?

DATA 36. Male versus Female Drivers The following data represent the number of licensed drivers in various age groups and the number of fatal accidents within the age group by gender.

Age	Number of Male Licensed Drivers (000s)	Number of Fatal Crashes	Number of Female Licensed Drivers (000s)	Number of Fatal Crashes
<16	12	227	12	77
16–20	6,424	5,180	6,139	2,113
21–24	6,941	5,016	6,816	1,531
25–34	18,068	8,595	17,664	2,780
35–44	20,406	7,990	20,063	2,742
45–54	19,898	7,118	19,984	2,285
55–64	14,340	4,527	14,441	1,514
65–74	8,194	2,274	8,400	938
>74	4,803	2,022	5,375	980

Source: National Highway and Traffic Safety Institute.

- (a) On the same graph, draw a scatter diagram for both males and females. Be sure to use a different plotting symbol for each group. For example, use a square (□) for males and a plus sign (+) for females. Treat number of licensed drivers as the explanatory variable.
- (b) Based on the scatter diagrams, do you think that insurance companies are justified in charging different insurance rates for males and females? Why?
- (c) Compute the linear correlation coefficient between number of licensed drivers and number of fatal crashes for males.
- (d) Compute the linear correlation coefficient between number of licensed drivers and number of fatal crashes for females.
- (e) Which gender has the stronger linear relation between number of licensed drivers and number of fatal crashes. Why?

DATA 37. Weight of a Car versus Miles per Gallon An engineer wants to determine how the weight of a car affects gas mileage. The following data represent the weights of various domestic cars and their gas mileages in the city for the 2015 model year.

Car	Weight (lb)	Miles per Gallon
Buick LaCrosse	4724	17
Cadillac XTS	4006	18
Chevrolet Cruze	3097	22
Chevrolet Impala	3555	19
Chrysler 300	4029	19
Dodge Charger	3934	19
Dodge Dart	3242	24
Ford Focus	2960	26
Ford Mustang	3530	19
Lincoln MKZ	3823	18

Source: Each manufacturer's website.

- (a) Determine which variable is the likely explanatory variable and which is the likely response variable.
- (b) Draw a scatter diagram of the data.
- (c) Compute the linear correlation coefficient between the weight of a car and its miles per gallon in the city.
- (d) Does a linear relation exist between the weight of a car and its miles per gallon in the city?
- (e) Suppose that we add the Ford Fusion Hybrid to the data. A Fusion Hybrid weighs 3639 pounds and gets 44 miles per gallon. Redraw the scatter diagram with the Fusion included. What do you notice?
- (f) Recompute the linear correlation coefficient with the Fusion included. How did this new value affect your result?
- (g) Why does this observation not follow the pattern of the data?
- 38. American Black Bears** The website that contained the American black bear data listed in Problem 30 actually had a bear whose length is 141.0 cm and weight is 100 kg, but incorrectly listed its length as 41.0 cm.
- (a) Redraw the scatter diagram with the incorrect entry.
- (b) Recompute the linear correlation coefficient using the data with the incorrect entry.
- (c) Explain how the scatter diagram can be used to identify incorrectly entered data values. Explain how the incorrectly entered data value affects the correlation coefficient.

- 39. Draw Your Data!** Consider the four data sets shown below.

Data Set 1		Data Set 2		Data Set 3		Data Set 4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
7	4.82	7	7.26	7	6.42	8	6.89
5	5.68	5	4.47	5	5.73	19	12.50

Source: Frank Anscombe. "Graphs in Statistical Analysis," *American Statistician* 27: 17–21, 1993.

- (a) Compute the linear correlation coefficient for each data set.
- (b) Draw a scatter diagram for each data set. Conclude that linear correlation coefficients and scatter diagrams must be used together in any statistical analysis of bivariate data.

- 40. Predicting Winning Percentage** The ultimate goal in any sport (besides having fun) is to win. One measure of how well a team does is winning percentage. In baseball, a lot of effort goes into figuring out the variables that help to predict a team's winning percentage. Go to www.pearsonhighered.com/sullivanstats to obtain the data file 4_1_40 using the file format of your choice for the version of the text you are using. The data represent the winning percentages of teams in both the American League (AL) and National League (NL). The difference between the AL and NL is that teams in the AL allow a designated hitter to hit for the pitcher. Pitchers are typically poor hitters, so teams in the NL tend to score fewer runs.

- (a) Which variable is the best predictor of winning percentage? What does the relation between this variable and winning percentage suggest?
- (b) Suppose a manager wants to score lots of runs. What variable best predicts runs scored?
- (c) What does the negative correlation between runs and at-bats per home run suggest?

- 41. Diversification** One basic theory of investing is diversification. The idea is that you want to have a basket of stocks that do not all "move in the same direction." In other words, if one investment goes down, you don't want a second investment in your portfolio that is also likely to go down. One hallmark of a good portfolio is a low correlation between investments. Go to www.pearsonhighered.com/sullivanstats to obtain the data file 4_1_41 using the file format of your choice for the version of the text you are using. The data represent the annual rates of return for various stocks. If you only wish to invest in two stocks, which two would you select if your goal is to have low correlation between the two investments? Which two would you select if your goal is to have one stock go up when the other goes down?

Source: Yahoo!Finance

- 42. Lyme Disease versus Drownings** Lyme disease is an inflammatory disease that results in a skin rash and flu-like symptoms. It is transmitted through the bite of an infected deer tick. The following data represent the number of reported cases of Lyme disease and the number of drowning deaths for a rural county in the United States.

Month	J	F	M	A	M	J	J	A	S	O	N	D
Cases of Lyme Disease	3	2	2	4	5	15	22	13	6	5	4	1
Drowning Deaths	0	1	2	1	2	9	16	5	3	3	1	0

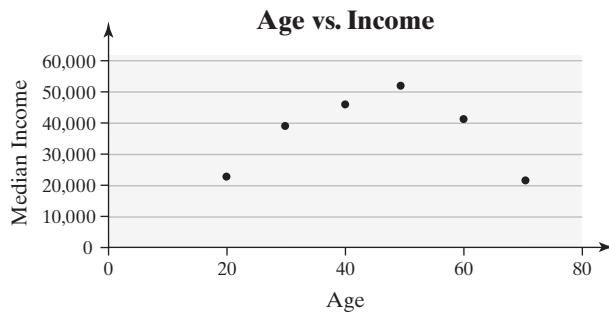
- (a) Draw a scatter diagram of the data using cases of Lyme disease as the explanatory variable.
- (b) Compute the correlation coefficient for the data.
- (c) Based on your results from parts (a) and (b), what type of relation exists between the number of reported cases of Lyme disease and drowning deaths? Do you believe that an increase in cases of Lyme disease causes an increase in drowning deaths? What is a likely lurking variable between cases of Lyme disease and drowning deaths?

- 43. Television Stations and Life Expectancy** Based on data obtained from the *CIA World Factbook*, the linear correlation coefficient between the number of television stations in a country and the life expectancy of residents of the country is 0.599. What does this correlation imply? Do you believe that the more television stations a country has, the longer its population can expect to live? Why or why not? What is a likely lurking variable between number of televisions and life expectancy?

- 44. Obesity** In a study published in the *Journal of the American Medical Association*, researchers found that the length of time a mother breast-feeds is negatively associated with the likelihood a child is obese. In an interview, the head investigator stated, "It's not clear whether breast milk has obesity-preventing properties or the women who are breast-feeding are less likely to have obese kids because they are less likely to be obese themselves." Using the researcher's statement, explain what might be wrong with concluding that breast-feeding prevents obesity. Identify some lurking variables in the study.

45. Crime Rate and Cell Phones The linear correlation between violent crime rate and percentage of the population that has a cell phone is -0.918 for years since 1995. Do you believe that increasing the percentage of the population that has a cell phone will decrease the violent crime rate? What might be a lurking variable between percentage of the population with a cell phone and violent crime rate?

46. Faulty Use of Correlation On the basis of the scatter diagram below, explain what is wrong with the following statement: “Because the linear correlation coefficient between age and median income is 0.012 , there is no relation between age and median income.”



47. Influential Consider the following set of data:

x	2.2	3.7	3.9	4.1	2.6	4.1	2.9	4.7
y	3.9	4.0	1.4	2.8	1.5	3.3	3.6	4.9

- (a) Draw a scatter diagram of the data and compute the linear correlation coefficient.
- (b) Draw a scatter diagram of the data and compute the linear correlation coefficient with the additional data point $(10.4, 9.3)$. Comment on the effect the additional data point has on the linear correlation coefficient. Explain why correlations should always be reported with scatter diagrams.



48. Transformations Consider the following data set:

x	5	6	7	7	8	8	8	8
y	4.2	5	5.2	5.9	6	6.2	6.1	6.9
x	9	9	10	10	11	11	12	12
y	7.2	8	8.3	7.4	8.4	7.8	8.5	9.5

- (a) Draw a scatter diagram with the x -axis starting at 0 and ending at 30 and with the y -axis starting at 0 and ending at 20.
- (b) Compute the linear correlation coefficient.
- (c) Now multiply both x and y by 2.
- (d) Draw a scatter diagram of the new data with the x -axis starting at 0 and ending at 30 and with the y -axis starting at 0 and ending at 20. Compare the scatter diagrams.
- (e) Compute the linear correlation coefficient.
- (f) Conclude that multiplying each value in the data set by a nonzero constant does not affect the correlation between the variables. Explain why this is the case.



49. Graduation Rates Go to www.pearsonhighered.com/sullivanstats to obtain the data file 4_1_49 using the file format of your choice for the version of the text you are using. The data represent the cost, return on investment, and graduation rate for

a random sample of 50 colleges or universities in the United States. The data are from payscale.com. The variable “Cost” represents the four-year cost including tuition, supplies, room and board of attending the school. The variable “Annual ROI” represents the return on investment for graduates of the school. It essentially represents how much you would earn on the investment of attending the school. The variable “Grad Rate” represents the graduation rate of the school.

- (a) Describe the association between “Cost” and “Graduation Rate” both numerically and graphically.
- (b) Describe the association between “Cost” and “Annual ROI” both numerically and graphically.

50. RateMyProfessors.com Professors Theodore Coladarci and Irv Kornfield from the University of Maine found a correlation of 0.68 between responses to questions on the RateMyProfessors.com website and typical in-class evaluations. Use this correlation to make an argument in favor of the validity of RateMyProfessors.com as a legitimate evaluation tool. RateMyProfessors.com used to have a chili pepper icon, which was meant to indicate a “hotness scale” for the professor. This hotness scale served as a proxy for the sexiness of the professor. It was found that the correlation between quality and sexiness is 0.64 . In addition, it was found that the correlation between easiness of the professor and quality is 0.85 for instructors with at least 70 posts. Use this information to make an argument against RateMyProfessors.com as a legitimate evaluation tool.

*Source: Theodore Coladarci and Irv Kornfield. “RateMyProfessors.com versus Formal In-class Student Evaluations of Teaching,” *Practical Assessment, Research, & Evaluation*, 12:6, May, 2007.*

Explaining the Concepts

- 51. What does it mean to say that the linear correlation coefficient between two variables equals 1? What would the scatter diagram look like?
- 52. What does it mean if $r = 0$?
- 53. Explain what is wrong with the following statement: “We have concluded that a high correlation exists between the gender of drivers and rates of automobile accidents.” Suggest a better way to write the sentence.
- 54. Write a paragraph that explains the concept of correlation. Include a discussion of the role that $x_i - \bar{x}$ and $y_i - \bar{y}$ play in the computation.
- 55. Explain the difference between correlation and causation. When is it appropriate to state that the correlation implies causation?
- 56. Draw a scatter diagram that might represent the relation between the number of minutes spent exercising on an elliptical and calories burned. Draw a scatter diagram that might represent the relation between number of hours each week spent on Facebook and grade point average.
- 57. Suppose you work a part-time job and earn \$15 per hour. Draw a scatter diagram that might represent the relation between your gross pay and hours worked. Is this a deterministic relation or a probabilistic relation?
- 58. Suppose that two variables, X and Y , are negatively associated. Does this mean that above-average values of X will always be associated with below-average values of Y ? Explain.

4.2 Least-Squares Regression



Preparing for This Section Before getting started, review the following:

- Lines (Appendix B, pp. B1–B5)

Objectives

- 1 Find the least-squares regression line and use the line to make predictions
- 2 Interpret the slope and the y -intercept of the least-squares regression line
- 3 Compute the sum of squared residuals

Once the scatter diagram and linear correlation coefficient show that two variables have a linear relation, we can find a linear equation that describes this relation. One way to do this is to select two points from the data that appear to provide a good fit (the line drawn appears to describe the relation between the two variables well) and to find the equation of the line through these two points.

EXAMPLE 1

Finding an Equation That Describes Linearly Related Data

Problem The data in Table 5 represent the club-head speed and the distance a golf ball travels for eight swings of the club. We found that these data are linearly related in Section 4.1.

CAUTION!

Example 1 *does not* present the least-squares method. It is used to set up the concepts of determining the line that best fits the data.

Table 5

Club-head Speed (mph) x	Distance (yd) y	(x, y)
100	257	(100, 257)
102	264	(102, 264)
103	274	(103, 274)
101	266	(101, 266)
105	277	(105, 277)
100	263	(100, 263)
99	258	(99, 258)
105	275	(105, 275)

Source: Paul Stephenson, student at Joliet Junior College.

- Find a linear equation that relates club-head speed, x (the explanatory variable), and distance, y (the response variable), by selecting two points and finding the equation of the line containing the points.
- Graph the line on the scatter diagram.
- Use the equation to predict the distance a golf ball will travel if the club-head speed is 104 miles per hour.

Approach

- Perform the following steps:

Step 1 Select two points so that a line drawn through the points appears to give a good fit. Call the points (x_1, y_1) and (x_2, y_2) . See the scatter diagram in Figure 1 on page 172.

Step 2 Find the slope of the line containing the two points using $m = \frac{y_2 - y_1}{x_2 - x_1}$.

Step 3 Use the point-slope formula, $y - y_1 = m(x - x_1)$, to find the equation of the line through the points selected in Step 1. Express the equation in the form $y = mx + b$, where m is the slope and b is the y -intercept.

(continued)

IN OTHER WORDS

A **good fit** means that the line drawn appears to describe the relation between the two variables well.

- (b) Draw a line through the points selected in Step 1 of part (a).
 (c) Let $x = 104$ in the equation found in part (a).

Solution

(a) **Step 1** We select $(x_1, y_1) = (99, 258)$ and $(x_2, y_2) = (105, 275)$, because a line drawn through these two points seems to give a good fit based on the scatter diagram shown in Figure 10.

$$\text{Step 2 } m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{275 - 258}{105 - 99} = \frac{17}{6} = 2.8333$$

Step 3 Use the point-slope formula to find the equation of the line.

$$y - y_1 = m(x - x_1)$$

$$y - 258 = 2.8333(x - 99) \quad m = 2.8333, x_1 = 99, y_1 = 258$$

$$y - 258 = 2.8333x - 280.4967$$

$$y = 2.8333x - 22.4967 \quad (1)$$

CAUTION!

The line found in Step 3 of Example 1 is not the least-squares regression line.

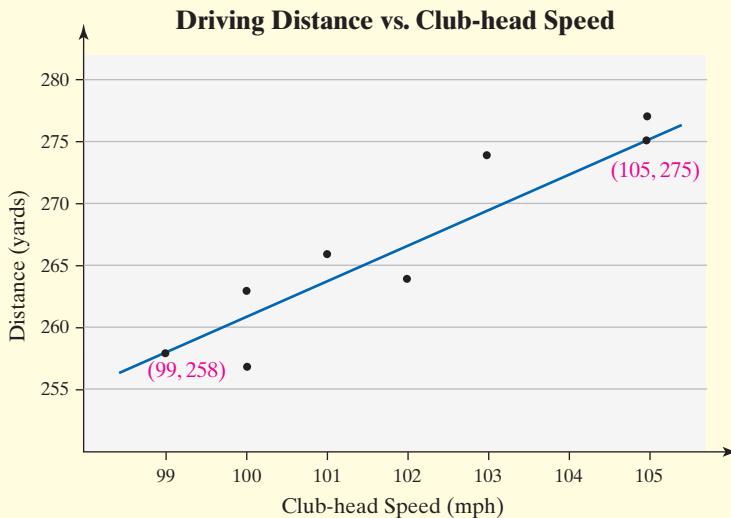
CAUTION!

Unless otherwise noted, we will round the slope and y -intercept to four decimal places. As always, do not round until the last computation.

The slope of the line is 2.8333, and the y -intercept is -22.4967 .

- (b) Figure 10 shows the scatter diagram along with the line drawn through the points $(99, 258)$ and $(105, 275)$.

Figure 10



- (c) Let $x = 104$ in equation (1) to predict the distance.

$$\begin{aligned} y &= 2.8333(104) - 22.4967 \\ &= 272.2 \text{ yards} \end{aligned}$$

We predict that a golf ball will travel 272.2 yards when it is hit with a club-head speed of 104 miles per hour.

NW Now Work Problems 7(a)–(c)



1 Find the Least-Squares Regression Line and Use the Line to Make Predictions

The line that we found in Example 1 appears to describe the relation between club-head speed and distance quite well. However, is there a line that fits the data better? Is there a line that fits the data *best*?

Whenever we attempt to determine the best of something, a criterion is needed for determining best. For example, suppose that we are trying to identify the best domestic car. What do we mean by best? Best gas mileage? Best reliability? When describing the relation between two variables, a criterion is needed for determining the best line.

IN OTHER WORDS

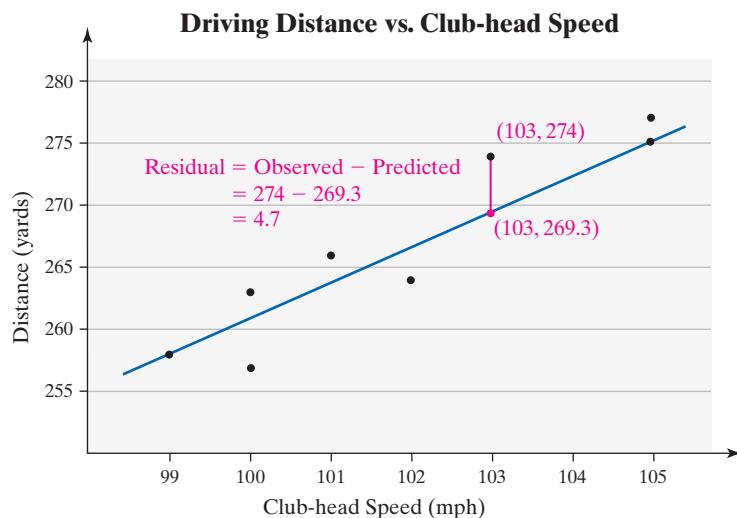
The residual represents how close our prediction comes to the actual observation. The smaller the residual, the better the prediction.

Consider Figure 11. Each y -coordinate on the line corresponds to a predicted distance for a given club-head speed. For example, if club-head speed is 103 miles per hour, the predicted distance is $2.8333(103) - 22.4967 = 269.3$ yards. The observed distance for this club-head speed is 274 yards. The difference between the observed and predicted values of y is the error, or **residual**. For a club-head speed of 103 miles per hour, the residual is

$$\begin{aligned}\text{Residual} &= \text{observed } y - \text{predicted } y \\ &= 274 - 269.3 \\ &= 4.7 \text{ yards}\end{aligned}$$

The residual for a club-head speed of 103 miles per hour is represented in Figure 11 by the length of the vertical line segment drawn between the points $(103, 269.3)$ and $(103, 274)$.

Figure 11



The criterion to determine the line that *best* describes the relation between two variables is based on the residuals. The most popular technique for making the residuals as small as possible is the *method of least squares*, developed by Adrien Marie Legendre.

Definition**Least-Squares Regression Criterion**

The **least-squares regression line** is the line that minimizes the sum of the squared errors (or residuals). This line minimizes the sum of the squared vertical distance between the observed values of y and those predicted by the line, \hat{y} (read “ y -hat”). We represent this as “minimize \sum residuals²”.

The advantage of the least-squares criterion is that it allows for statistical inference on the predicted value and slope (Chapter 14). Another advantage of the least-squares criterion is explained by Legendre in his text *Nouvelles méthodes pour la détermination des orbites des comètes*, published in 1806.

Of all the principles that can be proposed for this purpose, I think there is none more general, more exact, or easier to apply, than that which we have used in this work; it consists of making the sum of squares of the errors a *minimum*. By this method, a kind of equilibrium is established among the errors which, since it prevents the extremes from dominating, is appropriate for revealing the state of the system which most nearly approaches the truth.

The least-squares regression criterion leads to the following formulas for obtaining the *least-squares regression line*.

Historical Note

Adrien Marie Legendre was born on September 18, 1752, into a wealthy family and was educated in mathematics and physics at the College Mazarin in Paris. From 1775 to 1780, he taught at École Militaire. In 1783, Legendre was appointed an adjoint in the Académie des Sciences. He became a member of the committee of the Académie des Sciences in 1791 and was charged with the task of standardizing weights and measures. The committee worked to compute the length of the meter. During the French Revolution, Legendre lost his small fortune. In 1794, Legendre published *Éléments de géométrie*, which was the leading elementary text in geometry for around 100 years. In 1806, Legendre published a book on orbits, in which he developed the theory of least squares. He died on January 10, 1833.

**The Least-Squares Regression Line**

The equation of the least-squares regression line is given by

$$\hat{y} = b_1x + b_0$$

where

$$b_1 = r \cdot \frac{s_y}{s_x} \text{ is the } \mathbf{slope} \text{ of the least-squares regression line*} \quad (2)$$

and

$$b_0 = \bar{y} - b_1\bar{x} \text{ is the } \mathbf{y\text{-intercept}} \text{ of the least-squares regression line} \quad (3)$$

Note: \bar{x} is the sample mean and s_x is the sample standard deviation of the explanatory variable x ; \bar{y} is the sample mean and s_y is the sample standard deviation of the response variable y .

The notation \hat{y} is used in the least-squares regression line to remind us that it is a predicted value of y for a given value of x . The least-squares regression line, $\hat{y} = b_1x + b_0$, always contains the point (\bar{x}, \bar{y}) . This property can be useful when drawing the least-squares regression line by hand.

Since s_y and s_x must both be positive, the sign of the linear correlation coefficient, r , and the sign of the slope of the least-squares regression line, b_1 , are the same. For example, if r is positive, then b_1 will also be positive.

EXAMPLE 2**Finding the Least-Squares Regression Line by Hand****CAUTION!**

In Example 2, the slope of the least-squares regression line was found by hand using rounded values of the correlation coefficient, standard deviation of x , and standard deviation of y . The danger in doing this is that the rounded values may lead to errors in subsequent computations. For example, the slope of the least-squares regression line in the "by hand" computation using rounded values is found to be -2.1676 . However, if the original data (Table 2 on page 175) is used to determine the slope without any rounding in the computation (or using technology), the slope of the least-squares regression line is found to be -2.1667 . The moral of the story is that rounded values should not be used whenever using the values to compute other statistics.

Problem Find the least-squares regression line for the data in Table 2 from Section 4.1.

Approach In Example 2 of Section 4.1 (page 175), we found

$$r = -0.946, \bar{x} = 4, s_x = 2.44949, \bar{y} = 10, \text{ and } s_y = 5.612486$$

Substitute the values in Formulas (2) and (3) to find the slope and intercept of the least-squares regression line.

Solution Substitute $r = -0.946, s_x = 2.44949$, and $s_y = 5.612486$ into Formula (2):

$$b_1 = r \cdot \frac{s_y}{s_x} = -0.946 \cdot \frac{5.612486}{2.44949} = -2.1676$$

Substitute $\bar{x} = 4, \bar{y} = 10$, and $b_1 = -2.1676$ into Formula (3):

$$b_0 = \bar{y} - b_1\bar{x} = 10 - (-2.1676)(4) = 18.6704$$

The least-squares regression line is

$$\hat{y} = -2.1676x + 18.6704$$

*An equivalent formula is

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

Rounding Rule for the Slope and Intercept

Throughout the course, we agree to round the slope and y -intercept of the least-squares regression equation to four decimal places.

In Example 2, we obtained the least-squares regression line by hand. Why? Mainly to see the role that the correlation coefficient, r , the standard deviation of y , and the standard deviation of x play in finding the slope. Plus, from the y -intercept formula we learn that all regression lines travel through the point (\bar{x}, \bar{y}) . In practice, however, the least-squares regression line is found using technology.

Interpretation of Predicted Values The predicted value of y , \hat{y} , has two interpretations.

- (1) It is an estimate of the mean value of the response variable for a particular value of the explanatory variable.
- (2) It is an estimate of the value of the response variable for a particular value of the explanatory variable.

For example, suppose a least-squares regression equation is obtained that relates students' grade point average (GPA) to the number of hours studied each week. If the equation results in a predicted GPA of 3.14 when a student studies 20 hours each week, we would say (1) the mean GPA of *all* students who study 20 hours each week is 3.14, and (2) the predicted GPA of a particular student who studied 20 hours each week is 3.14.

EXAMPLE 3 Finding the Least-Squares Regression Line Using Technology

Problem Use the golf data in Table 5 (page 187).

- (a) Find the least-squares regression line.
- (b) Draw the least-squares regression line on the scatter diagram of the data.
- (c) Predict the mean distance the golf ball will travel when hit with a club-head speed of 103 miles per hour (mph).
- (d) Predict the distance a particular golf ball will travel when hit with a club-head speed of 103 miles per hour.
- (e) Determine the residual for the predicted value found in part (c). Is the distance in Table 5 above average or below average among all balls hit with a swing speed of 103 mph?

Approach Because technology plays a major role in obtaining the least-squares regression line, we will use a TI-84 Plus CE graphing calculator, Minitab, Excel, and StatCrunch to obtain the least-squares regression line. The steps for obtaining regression lines are given in the Technology Step-by-Step on page 195.

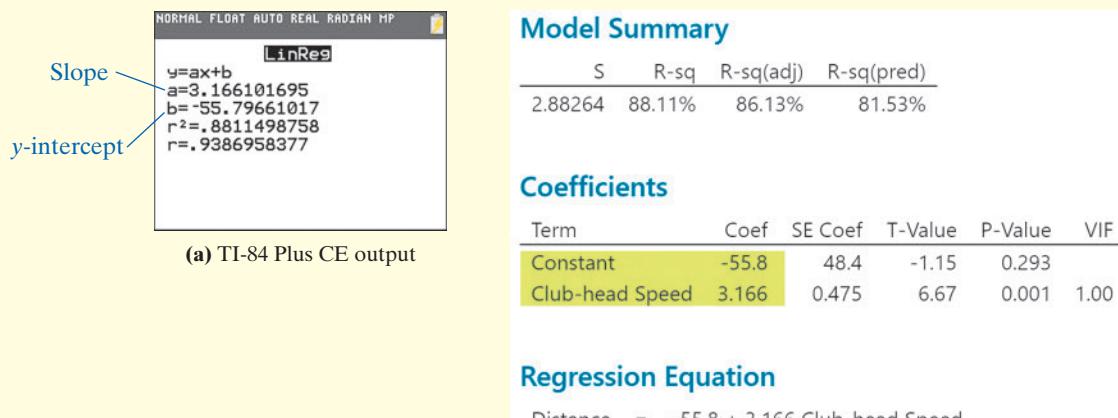
CAUTION!

Throughout the text, we will round the slope and y -intercept values to four decimal places. Predictions will be rounded to one more decimal place than the response variable.

Solution

- (a) Figure 12 shows the output obtained from the various technologies. The least-squares regression line is $\hat{y} = 3.1661x - 55.7966$.

Figure 12



(b) Minitab output

(continued)

	Coefficients	Standard Error	t Stat	P-value
Intercept	-55.79661017	48.37134953	-1.15351	0.292574312
Club-Head Speed	3.166101695	0.47470539	6.669614	0.000549825

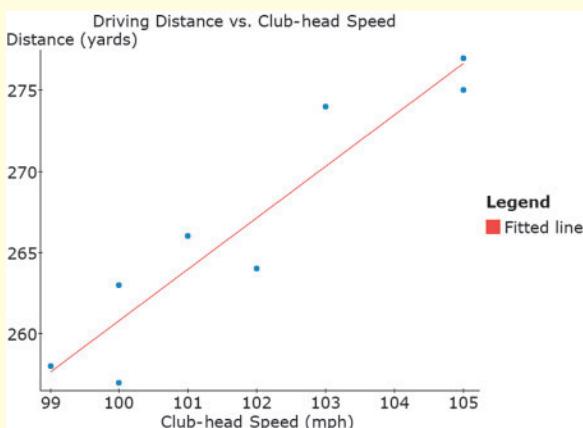
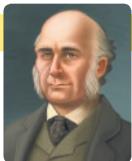
(c) Excel output

Simple linear regression results:
 Dependent Variable: Distance
 Independent Variable: Club-Head Speed
 $\text{Distance} = -55.79661 + 3.1661017 \text{ Club-Head Speed}$
 Sample size: 8
 R (correlation coefficient) = 0.93869584
 $R\text{-sq} = 0.88114988$
 Estimate of error standard deviation: 2.8826385

(d) StatCrunch output

- (b) Figure 13 shows the least-squares regression line drawn on the scatter diagram using StatCrunch.

Figure 13

**Historical Note**

Sir Francis Galton was born on February 16, 1822. Galton came from a wealthy and well-known family. Charles Darwin was his first cousin. Galton studied medicine at Cambridge. After receiving a large inheritance, he left the medical field and traveled the world. He explored Africa from 1850 to 1852. In the 1860s, his study of meteorology led him to discover anticyclones. Influenced by Darwin, Galton always had an interest in genetics and heredity. He studied heredity through experiments with sweet peas. He noticed that the weight of the "children" of the "parent" peas reverted or regressed to the mean weight of all peas—hence, the term *regression analysis*. Galton died January 17, 1911.

- (c) Let $x = 103$ in the least-squares regression equation $\hat{y} = 3.1661x - 55.7966$ to predict the mean distance the ball will travel when hit with a club-head speed of 103 mph.

$$\begin{aligned}\hat{y} &= 3.1661(103) - 55.7966 \\ &= 270.3 \text{ yards}\end{aligned}$$

We predict the mean distance the golf ball will travel when hit at 103 mph is 270.3 yards.

- (d) Again, let $x = 103$ in the least-squares regression equation. The predicted distance a particular ball hit with a club-head speed of 103 mph is 270.3 yards.
 (e) From the data in Table 1 from Section 4.1, the observed distance the ball traveled when the swing speed is 103 mph was 274 yards. So,

$$\begin{aligned}\text{Residual} &= y - \hat{y} & \text{Residual} &= \text{observed } y - \text{predicted } y \\ &= 274 - 270.3 \\ &= 3.7 \text{ yards}\end{aligned}$$

Because the residual is positive (the observed value of 274 yards is greater than the predicted value of 270.3 yards), the distance of 274 yards is above average for a swing speed of 103 mph.

② Interpret the Slope and the y -Intercept of the Least-Squares Regression Line

Interpretation of Slope In algebra, we learned that the definition of slope is $\frac{\text{rise}}{\text{run}}$ or $\frac{\text{change in } y}{\text{change in } x}$. If a line has slope $\frac{2}{3}$, then if x increases by 3, y will increase by 2. Or if the slope of a line is $-4 = \frac{-4}{1}$, then if x increases by 1, y will decrease by 4.

Interpreting slope for least-squares regression lines has a minor twist. Statistical models such as a least-squares regression equation are *probabilistic*. This means that any predictions or interpretations made as a result of the model are based on uncertainty. Therefore, when we interpret the slope of a least-squares regression equation, we do not want to imply that there is 100% certainty behind the interpretation. For example, the slope of the least-squares regression line from Example 3 is 3.1661 yards per mph. In algebra, we would interpret the slope to mean “if x increases by 1 mph, then y will increase by 3.1661 yards.” In statistics, this interpretation is close, but not quite accurate, because increasing the club-head speed by 1 mph does not guarantee the distance the ball travels will increase by 3.1661 yards. Instead, over the course of data we observed, an increase in club-head speed of 1 mph increased distance 3.1661 yards, *on average*—sometimes the ball might travel a shorter additional distance, sometimes a longer additional distance, but on average this is the change in distance. So two interpretations of slope are acceptable:

If club-head speed increases by 1 mile per hour, the distance the golf ball travels increases by 3.1661 yards, on average.

or

If club-head speed increases by 1 mile per hour, the expected distance the golf ball will travel increases by 3.1661 yards.

Interpretation of the y -Intercept The y -intercept of any line is the point where the graph intersects the vertical axis. It is found by letting $x = 0$ in an equation and solving for y . In general, we interpret a y -intercept as the value of the response variable when the value of the explanatory variable is 0. To interpret the y -intercept, we must first ask two questions:

1. Is 0 a reasonable value for the explanatory variable?
2. Do any observations near $x = 0$ exist in the data set?

If the answer to either of these questions is no, we do not interpret the y -intercept. In the regression equation of Example 3, a swing speed of 0 miles per hour does not make sense, so we do not interpret the y -intercept.

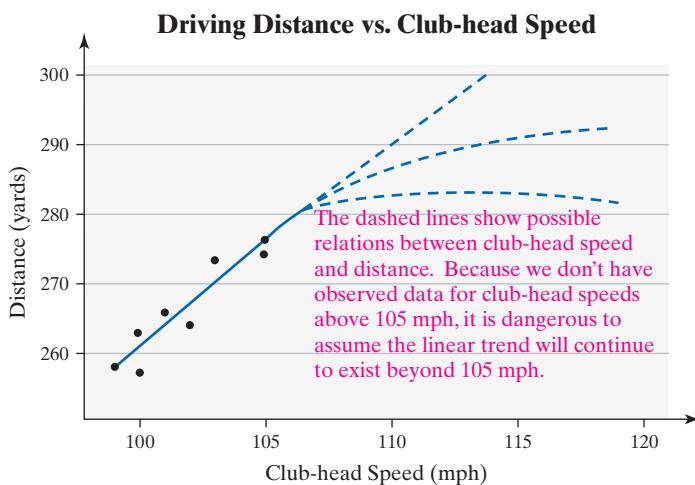
The second condition for interpreting the y -intercept is especially important because we should not use the regression model to make predictions **outside the scope of the model**, meaning we should not use the regression model to make predictions for values of the explanatory variable that are much larger or much smaller than those observed. This is a dangerous practice because we cannot be certain of the behavior of data for which we have no observations.

For example, we should not use the line in Example 3 to predict distance when club-head speed is 140 mph because the highest observed club-head speed is 105 mph. The linear relation between distance and club-head speed might not continue. See Figure 14.

CAUTION!

Be careful when using the least-squares regression line to make predictions for values of the explanatory variable that are much larger or much smaller than those observed.

Figure 14

**NW Now Work Problem 13**

③ Compute the Sum of Squared Residuals

Recall that the least-squares regression line is the line that minimizes the sum of the squared residuals. This means that the sum of the squared residuals, $\Sigma \text{residuals}^2$, is smaller for the least-squares line than for any other line that may describe the relation between the two variables. In particular, the sum of the squared residuals is smaller for the least-squares regression line in Example 3 than for the line obtained in Example 1. It is worthwhile to verify this result.

EXAMPLE 4

Comparing the Sum of Squared Residuals

Problem Compare the sum of squared residuals for the lines in Examples 1 and 3.

Approach Use a table to compute $\Sigma \text{residuals}^2$ using the predicted values of y , \hat{y} , for the equations in Examples 1 and 3.

Solution Table 6 contains the value of the explanatory variable in Column 1. Column 2 contains the corresponding response variable. Column 3 contains the predicted values using the equation found in Example 1: $\hat{y} = 2.8333x - 22.4967$. In Column 4, we compute the residuals for each observation: residual = observed y – predicted y = $y - \hat{y}$. For example, the first residual is $y - \hat{y} = 257 - 260.8 = -3.8$. Column 5 contains the squares of the residuals. Column 6 contains the predicted values using the least-squares regression equation found in Example 3: $\hat{y} = 3.1661x - 55.7966$. Column 7 represents the residuals for each observation, and Column 8 represents the squared residuals.

Table 6

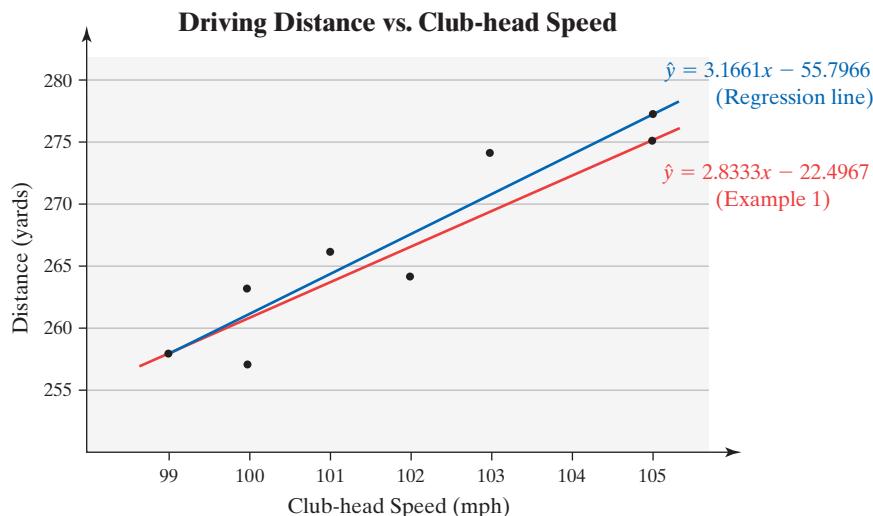
Club-head Speed (mph)	Distance (yd)	Example 1 ($\hat{y} = 2.8333x - 22.4967$)	Residual $y - \hat{y}$	Residual ² $(y - \hat{y})^2$	Example 3 ($\hat{y} = 3.1661x - 55.7966$)	Residual $y - \hat{y}$	Residual ² $(y - \hat{y})^2$	
100	257	260.8	-3.8	14.44	260.8	-3.8	14.44	
102	264	266.5	-2.5	6.25	267.1	-3.1	9.61	
103	274	269.3	4.7	22.09	270.3	3.7	13.69	
101	266	263.7	2.3	5.29	264.0	2.0	4.00	
105	277	275.0	2.0	4.00	276.6	0.4	0.16	
100	263	260.8	2.2	4.84	260.8	2.2	4.84	
99	258	258.0	0.0	0.00	257.6	0.4	0.16	
105	275	275.0	0.0	0.00	276.6	-1.6	2.56	
$\Sigma \text{residual}^2 = 56.91$					$\Sigma \text{residual}^2 = 49.46$			

The sum of the squared residuals for the line in Example 1 is 56.91; the sum of the squared residuals for the least-squares regression line is 49.46. Again, any line other than the least-squares regression line will have a sum of squared residuals that is greater than 49.46.

NW Now Work Problems 7(d)–(h)

We draw the graphs of the two lines obtained in Examples 1 and 3 on the same scatter diagram in Figure 15 to help the reader visualize the difference.

Figure 15



Technology Step-by-Step

Determining the Least-Squares Regression Line

TI-83/84 Plus

Use the same steps that were followed to obtain the correlation coefficient. (See Section 4.1.)

Minitab

- With the explanatory variable in C1 and the response variable in C2, select the **Stat** menu and highlight **Regression**. Select **Fit Regression Model...**
- With the cursor in the Responses cell, select the response variable. With the cursor in the Continuous predictors cell, select the explanatory variable. Click **OK**.

Excel

- Enter the explanatory variable in column A and the response variable in column B.
- Select the **Data** menu and then **Data Analysis**.
- Select the **Regression** option.

- With the cursor in the Y-range cell, highlight the column that contains the response variable. With the cursor in the X-range cell, highlight the column that contains the explanatory variable. Select the Output Range by entering a cell where you want the output to appear. Press **OK**.

StatCrunch

- If necessary, enter the explanatory variable in column var1 and the response variable in column var2. Name each column variable.
- Select **Stat**, highlight **Regression**, and select **Simple Linear**.
- Choose the explanatory variable for the X variable and the response variable for the Y variable. If you want, enter a value of the explanatory variable to Predict Y for X. If you want the least-squares regression line drawn on the scatter diagram, highlight **Fitted line plot** under **Graphs**. Click **Compute!**.



4.2 Assess Your Understanding

Vocabulary and Skill Building

- The difference between the observed and predicted value of y is the error, or _____.
 - If the linear correlation between two variables is negative, what can be said about the slope of the regression line?
 - Which of the following is true of the least-squares regression line $\hat{y} = b_1x + b_0$?
- (a) The predicted value of y , \hat{y} , is an estimate of the mean value of the response variable for a particular value of the explanatory variable.

- (b) The predicted value of y , \hat{y} , is an estimate of the mean value of the explanatory variable for a particular value of the response variable.
- (c) The predicted value of y , \hat{y} , is an estimate of the value of the response variable for a particular value of the explanatory variable.
- (d) The predicted value of y , \hat{y} , is an estimate of the value of the explanatory variable for a particular value of the response variable.
- (e) The sign of the linear correlation coefficient, r , and the sign of the slope of the least-squares regression line, b_1 , are the same.

- (f) The least-squares regression line maximizes the sum of squared residuals.
 (g) The least-squares regression line always contains the point $(0, 0)$.
 (h) The least-squares regression line always contains the point (\bar{x}, \bar{y}) .
 (i) The least-squares regression line minimizes the sum of squared residuals.
4. If the linear correlation coefficient is 0, what is the equation of the least-squares regression line?

DATA 5. For the data set

<i>x</i>	0	2	3	5	6	6
<i>y</i>	5.8	5.7	5.2	2.8	1.9	2.2

- (a) Draw a scatter diagram. Comment on the type of relation that appears to exist between *x* and *y*.
 (b) Given that $\bar{x} = 3.6667$, $s_x = 2.4221$, $\bar{y} = 3.9333$, $s_y = 1.8239$, and $r = -0.9477$, determine the least-squares regression line.
 (c) Graph the least-squares regression line on the scatter diagram drawn in part (a).

6. For the data set

<i>x</i>	2	4	8	8	9
<i>y</i>	1.4	1.8	2.1	2.3	2.6

- (a) Draw a scatter diagram. Comment on the type of relation that appears to exist between *x* and *y*.
 (b) Given that $\bar{x} = 6.2$, $s_x = 3.03315$, $\bar{y} = 2.04$, $s_y = 0.461519$, and $r = 0.957241$, determine the least-squares regression line.
 (c) Graph the least-squares regression line on the scatter diagram drawn in part (a).

In Problems 7–12:

- (a) By hand, draw a scatter diagram treating *x* as the explanatory variable and *y* as the response variable.
 (b) Select two points from the scatter diagram and find the equation of the line containing the points selected.
 (c) Graph the line found in part (b) on the scatter diagram.
 (d) By hand, determine the least-squares regression line.
 (e) Graph the least-squares regression line on the scatter diagram.
 (f) Compute the sum of the squared residuals for the line found in part (b).
 (g) Compute the sum of the squared residuals for the least-squares regression line found in part (d).
 (h) Comment on the fit of the line found in part (b) versus the least-squares regression line found in part (d).

NW 7.

<i>x</i>	3	4	5	7	8
<i>y</i>	4	6	7	12	14

8.

<i>x</i>	3	5	7	9	11
<i>y</i>	0	2	3	6	9

9.

<i>x</i>	-2	-1	0	1	2
<i>y</i>	-4	0	1	4	5

10.

<i>x</i>	-2	-1	0	1	2
<i>y</i>	7	6	3	2	0

11.	<i>x</i>	20	30	40	50	60
	<i>y</i>	100	95	91	83	70

12.	<i>x</i>	5	10	15	20	25
	<i>y</i>	2	4	7	11	18

- NW** 13. **Income and Education** In Problem 15 from Section 4.1, a scatter diagram and correlation coefficient suggested there is a linear relation between the percentage of individuals who have at least a bachelor's degree and median income in the states. In fact, the least-squares regression equation is $\hat{y} = 1103x + 31,955$ where *y* is the median income and *x* is the percentage of individuals 25 years and older with at least a bachelor's degree in the state.

- (a) Predict the median income of a state in which 25% of adults 25 years and older have at least a bachelor's degree.
 (b) In North Dakota, 28% of adults 25 years and older have at least a bachelor's degree. The median income in North Dakota is \$66,321. Is this income higher than what you would predict? Why?
 (c) Interpret the slope.
 (d) Explain why it does not make sense to interpret the intercept.

14. **You Explain It! Study Time and Exam Scores** After the first exam in a statistics course, Professor Katula surveyed 14 randomly selected students to determine the relation between the amount of time they spent studying for the exam and exam score. She found that a linear relation exists between the two variables. The least-squares regression line that describes this relation is $\hat{y} = 6.3333x + 53.0298$.

- (a) Predict the exam score of a student who studied 2 hours.
 (b) Interpret the slope.
 (c) What is the mean score of students who did not study?
 (d) A student who studied 5 hours for the exam scored 81 on the exam. Is this student's exam score above or below average among all students who studied 5 hours?

15. **Age Gap at Marriage** Is there a relation between the age difference between husband/wives and the percent of a country that is literate. Researchers found the least-squares regression between age difference (husband age minus wife age), *y*, and literacy rate (percent of the population that is literate), *x*, is $\hat{y} = -0.0527x + 7.1$. The model applied for $18 \leq x \leq 100$.

Source: Xu Zhang and Solomon W. Polachek, State University of New York at Binghamton "The Husband-Wife Age Gap at First Marriage: A Cross-Country Analysis."

- (a) Interpret the slope.
 (b) Does it make sense to interpret the *y*-intercept? Explain.
 (c) Predict the age difference between husband/wife in a country where the literacy rate is 25%.
 (d) Would it make sense to use this model to predict the age difference between husband/wife in a country where the literacy rate is 10%? Explain.
 (e) The literacy rate in the United States is 99% and the age difference between husbands and wives is 2 years. Is this age difference above or below the average age difference among all countries whose literacy rate is 99%?

16. **You Explain It! CO₂ and Energy Production** The least-squares regression equation $\hat{y} = 1.3491x - 11.2847$ relates the carbon dioxide emissions (in millions of tonnes), *y*, and electricity produced (terrawatt-hours), *x*, for all countries in the world.

Source: www.bp.com

- (a) Interpret the slope.

- (b) The lowest electric energy-producing country is Luxemborg, which produces 2.2 terrawatt-hours of electricity. The highest energy-producing country is China, which produces 6495 terrawatt-hours of energy. Would it be reasonable to use this model to predict the CO₂ emissions of a country if it produces 6394 terrawatt-hours of electricity? Why or why not?
- (c) The United States produces 4282 terrawatt-hours of electricity and emits 5088 million tonnes of carbon dioxide. What is the residual for the U.S.? How would you interpret this residual?

Applying the Concepts

Problems 17–22 use the results from Problems 27–32 in Section 4.1.

- NW 17. An Unhealthy Commute** (Refer to Problem 27, Section 4.1.)

The following data represent commute times (in minutes) and score on a well-being survey.

Commute Time (minutes), x	Gallup-Healthways Well-Being Index Composite Score, y
5	69.2
15	68.3
25	67.5
35	67.1
50	66.4
72	66.1
105	63.9

Source: The Gallup Organization.

- (a) Find the least-squares regression line treating the commute time, x , as the explanatory variable and the index score, y , as the response variable.
- (b) Interpret the slope and y -intercept, if appropriate.
- (c) Predict the well-being index of a person whose commute is 30 minutes.
- (d) Suppose Barbara has a 20-minute commute and scores 67.3 on the survey. Is Barbara more “well-off” than the typical individual who has a 20-minute commute?

- DATA 18. Credit Scores** (Refer to Problem 28, Section 4.1.) An economist wants to determine the relation between one’s FICO score, x , and the interest rate of a 36-month auto loan, y . The given data represent the interest rate (in percent) a bank would offer on a 36-month auto loan for various FICO scores.

Credit Score, x	Interest Rate (percent), y
545	18.982
595	17.967
640	12.218
675	8.612
705	6.680
750	5.150

Source: www.myfico.com

- (a) Find the least-squares regression line treating the FICO score, x , as the explanatory variable and the interest rate, y , as the response variable.
- (b) Interpret the slope and y -intercept, if appropriate. Note: Credit scores have a range of 300 to 850.
- (c) Predict the interest rate a person would pay if her FICO score were the median score of 723.
- (d) Suppose Bob has a FICO score of 680 and he is offered an interest rate of 8.3%. Is this a good offer? Why?

- DATA 19. Height versus Head Circumference** (Refer to Problem 29, Section 4.1.) A pediatrician wants to determine the relation that exists between a child’s height, x , and head circumference, y . She randomly selects 11 children from her practice, measures their heights and head circumferences, and obtains the following data.

Height (inches), x	Head Circumference (inches), y	Height (inches), x	Head Circumference (inches), y
27.75	17.5	26.5	17.3
24.5	17.1	27	17.5
25.5	17.1	26.75	17.3
26	17.3	26.75	17.5
25	16.9	27.5	17.5
27.75	17.6		

Source: Denise Slucki, student at Joliet Junior College.

- (a) Find the least-squares regression line treating height as the explanatory variable and head circumference as the response variable.
- (b) Interpret the slope and y -intercept, if appropriate.
- (c) Use the regression equation to predict the head circumference of a child who is 25 inches tall.
- (d) Compute the residual based on the observed head circumference of the 25-inch-tall child in the table. Is the head circumference of this child above average or below average?
- (e) Draw the least-squares regression line on the scatter diagram of the data and label the residual from part (d).
- (f) Notice that two children are 26.75 inches tall. One has a head circumference of 17.3 inches; the other has a head circumference of 17.5 inches. How can this be?
- (g) Would it be reasonable to use the least-squares regression line to predict the head circumference of a child who was 32 inches tall? Why?

- DATA 20. American Black Bears** (Refer to Problem 30, Section 4.1.)

The American black bear (*Ursus americanus*) is one of eight bear species in the world. It is the smallest North American bear and the most common bear species on the planet. In 1969, Dr. Michael R. Pelton of the University of Tennessee initiated a long-term study of the population in Great Smoky Mountains National Park. One aspect of the study was to develop a model that could be used to predict a bear’s weight (since it is not practical to weigh bears in the field). One variable thought to be related to weight is the length of the bear. The data below represent the lengths of 12 American black bears.

Total Length (cm), x	Weight (kg), y
139.0	110
138.0	60
139.0	90
120.5	60
149.0	85
141.0	100
141.0	95
150.0	85
166.0	155
151.5	140
129.5	105
150.0	110

Source: www.fieldtripearth.org

- (a) Find the least-squares regression line, treating total length as the explanatory variable and weight as the response variable.
 (b) Interpret the slope and y -intercept, if appropriate.
 (c) Suppose a 149.0-cm bear is captured in the field. Use the least-squares regression line to predict the weight of the bear.
 (d) What is the residual of the 149.0-cm bear? Is this bear's weight above or below average for a bear of this length?

DATA **21. Home Runs** (Refer to Problem 31, Section 4.1.) The following data represent the speed at which a ball was hit (in miles per hour) and the distance it traveled (in feet) for a random sample of home runs in a Major League baseball game.

Speed (mph)	Distance (feet)
107.9	441
110.4	427
103.5	422
105.4	418
105.5	414
101.7	411
103.3	408
101.0	405
103.6	402
101.4	399
100.7	396
101.3	393

Source: baseballsavant.mlb.com

- (a) Find the least-squares regression line treating speed at which the ball was hit as the explanatory variable and distance the ball traveled as the response variable.
 (b) Interpret the slope and y -intercept, if appropriate.
 (c) Predict the mean distance of all home runs hit at 105 mph.
 (d) If a ball is hit with a speed of 105 miles per hour, predict how far it will travel.
 (e) Christian Yelich hit a home run 398 feet. The speed at which the ball was hit was 106.2 mph. Did this ball travel farther than you would have predicted? Explain.
 (f) Would you feel comfortable using the least-squares regression model on home runs where the speed of the ball was 122 mph? Explain.

DATA **22. Hurricanes** (Refer to Problem 32, Section 4.1.) The following data represent the maximum wind speed (in knots) and atmospheric pressure (in millibars) for a random sample of hurricanes that originated in the Atlantic Ocean.

Atmospheric Pressure (mb)	Wind Speed (knots)	Atmospheric Pressure (mb)	Wind Speed (knots)
993	50	1006	40
995	60	942	120
994	60	1002	40
997	45	986	50
1003	45	983	70
1004	40	994	65
1000	55	940	120
994	55	976	80
942	105	966	100
1006	30	982	55

Source: National Hurricane Center.

- (a) Find the least-squares regression line treating atmospheric pressure as the explanatory variable.
 (b) Interpret the slope.
 (c) Is it reasonable to interpret the y -intercept? Why?
 (d) One hurricane had an atmospheric pressure of 997 mb. Is this hurricane's wind speed above or below average for a hurricane with this level of atmospheric pressure?

23. Cola Consumption vs. Bone Density Example 5 in Section 4.1 on page 178 discussed the effect of cola consumption on bone mineral density in the femoral neck of women.

- (a) Find the least-squares regression line treating cola consumption per week as the explanatory variable.
 (b) Interpret the slope.
 (c) Interpret the intercept.
 (d) Predict the bone mineral density of the femoral neck of a woman who consumes four colas per week.
 (e) The researchers found a woman who consumed four colas per week to have a bone mineral density of 0.873 g/cm^2 . Is this woman's bone mineral density above or below average among all women who consume four colas per week?
 (f) Would you recommend using the model found in part (a) to predict the bone mineral density of a woman who consumes two cans of cola per day? Why?

DATA **24. Weight of a Car versus Miles per Gallon** (Refer to Problem 37, Section 4.1.) An engineer wants to determine how the weight of a car, x , affects gas mileage, y . The following data represent the weights of various domestic cars and their miles per gallon in the city for the 2015 model year.

Car	Weight (lb)	Miles per Gallon
Buick LaCrosse	4724	17
Cadillac XTS	4006	18
Chevrolet Cruze	3097	22
Chevrolet Impala	3555	19
Chrysler 300	4029	19
Dodge Charger	3934	19
Dodge Dart	3242	24
Ford Focus	2960	26
Ford Mustang	3530	19
Lincoln MKZ	3823	18

Source: Each manufacturer's website.

- (a) Find the least-squares regression line treating weight as the explanatory variable and miles per gallon as the response variable.
 (b) Interpret the slope and y -intercept, if appropriate.
 (c) A Cadillac CTS weighs 3649 pounds and gets 20 miles per gallon. Is the miles per gallon of a CTS above average or below average for cars of this weight?
 (d) Would it be reasonable to use the least-squares regression line to predict the miles per gallon of a Toyota Prius, a hybrid gas and electric car? Why or why not?

25. CEO Performance Explain why it does not make sense to find a least-squares regression line for the CEO Performance data from Problem 33 in Section 4.1.

26. Bear Markets Explain why it does not make sense to find a least-squares regression line for the Bear Market data from Problem 34 in Section 4.1.

- DATA 27. Male vs. Female Drivers** (Refer to Problem 36, Section 4.1.) The following data represent the number of licensed drivers in various age groups and the number of fatal accidents within the age group by gender.

Age	Number of Male Licensed Drivers (000s)	Number of Fatal Crashes	Number of Female Licensed Drivers (000s)	Number of Fatal Crashes
<16	12	227	12	77
16–20	6,424	5,180	6,139	2,113
21–24	6,941	5,016	6,816	1,531
25–34	18,068	8,595	17,664	2,780
35–44	20,406	7,990	20,063	2,742
45–54	19,898	7,118	19,984	2,285
55–64	14,340	4,527	14,441	1,514
65–74	8,194	2,274	8,400	938
>74	4,803	2,022	5,375	980

Source: National Highway and Traffic Safety Institute.

- (a) Find the least-squares regression line for males treating number of licensed drivers as the explanatory variable, x , and number of fatal crashes, y , as the response variable. Repeat this procedure for females.
- (b) Interpret the slope of the least-squares regression line for each gender. How might an insurance company use this information?
- (c) Was the number of fatal accidents for 16- to 20-year-old males above or below average? Was the number of fatal accidents for 21- to 24-year-old-males above or below average? Was the number of fatal accidents for males greater than 74 years old above or below average? How might an insurance company use this information? Does the same relationship hold for females?

- DATA 28. Graduation Rates** Go to www.pearsonhighered.com/sullivanstats to obtain the data file 4_2_28 using the file format of your choice for the version of the text you are using. The variable “Cost” represents the four-year cost including tuition, supplies, room and board. The variable “Annual ROI” represents the return on investment for graduates of the school. It essentially represents how much you would earn on the investment of attending the school. The variable “Grad Rate” represents the graduation rate of the school.

- (a) In Problem 49 from Section 4.1, we saw that a scatter diagram between Cost and Grad Rate treating Cost as the explanatory variable suggested a positive association between the two variables. Find the least-squares regression line treating Cost as the explanatory variable. Round the slope to six decimal places.
- (b) Interpret the slope.
- (c) Washington University in St. Louis has a four-year cost of \$266,000 and a graduation rate of 94%. Is Washington University’s graduation rate above or below average among schools that cost \$266,000?
- (d) A scatter diagram between cost and return on investment (treating cost as the explanatory variable) suggested a negative association between the two variables. Find the

least-squares regression line treating cost as the explanatory variable. Round the slope to eight decimal places.

- (e) Interpret the slope.
 - (f) Washington University’s return on investment is 5.2%. Is this above or below average among all schools that cost \$266,000?
- DATA 29. Putting It Together: Housing Prices** One of the biggest factors in determining the value of a home is the square footage. The following data represent the square footage and selling price (in thousands of dollars) for a random sample of homes for sale in Naples, Florida in January 2017.

Square Footage, x	Selling Price (\$000s), y
2204	379.9
3183	375
1128	189.9
1975	338
3101	619.9
2769	370
4113	627.7
2198	375
2609	425
1708	298.1
1786	271
3813	690.1

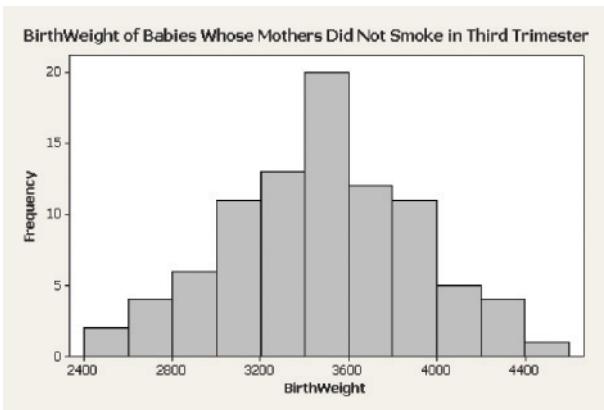
Source: zillow.com

- (a) Which variable is the explanatory variable?
- (b) Draw a scatter diagram of the data.
- (c) Determine the linear correlation coefficient between square footage and asking price.
- (d) Is there a linear relation between the square footage and asking price?
- (e) Find the least-squares regression line treating square footage as the explanatory variable.
- (f) Interpret the slope.
- (g) Is it reasonable to interpret the y -intercept? Why?
- (h) One home that is 1465 square feet sold for \$285,000. Is this home’s price above or below average for a home of this size? What might be some reasons for this price?

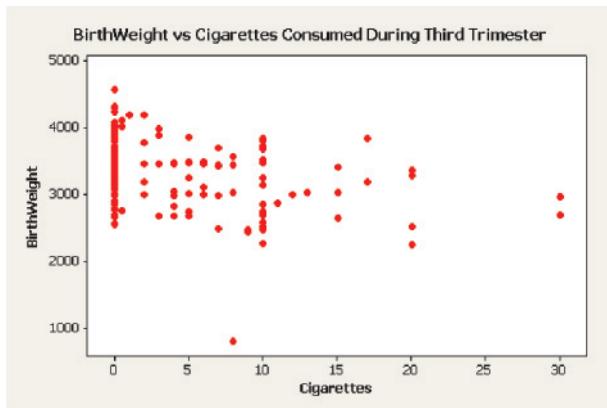
- 30. Putting It Together: Smoking and Birth Weight** It is well known that women should not smoke while pregnant, but what is the effect of smoking on a baby’s birth weight? Researchers Ira M. Bernstein and associates “sought to estimate how the pattern of maternal smoking throughout pregnancy influences newborn size.” To conduct this study, 160 pregnant, smoking women were enrolled in a prospective study. During the third trimester of pregnancy, the woman self-reported the number of cigarettes smoked. Urine samples were collected to measure cotinine levels (to assess nicotine levels). Birth weights (in grams) of the babies were obtained upon delivery.

Source: Ira M. Bernstein et. al. “Maternal Smoking and Its Association with Birthweight.” *Obstetrics & Gynecology* 106 (Part 1) 5, 2005.

- (a) The histogram on the next page drawn in Minitab, shows the birth weight of babies whose mothers did not smoke in the third trimester of pregnancy (but did smoke prior to the third trimester). Describe the shape of the distribution. What is the class width of the histogram?



- (b) Is this an observational study or a designed experiment?
 (c) What does it mean for the study to be prospective?
 (d) Why would the researchers conduct a urinalysis to measure cotinine levels?
 (e) What is the explanatory variable in the study? What is the response variable? Is the explanatory variable qualitative or quantitative? Is the response variable qualitative or quantitative?
 (f) The scatter diagram of the data drawn using Minitab is shown below. What type of relation appears to exist between cigarette consumption in the third trimester and birth weight?



- (g) Use the regression output from Minitab to report the least-squares regression line between cigarette consumption and birth weight.

Regression Analysis: BirthWeight versus Cigarettes

The regression equation is

$$\text{BirthWeight} = 3456 - 31.0 \text{ Cigarettes}$$

Predictor	Coef	SE Coef	T	P
Constant	3456.04	45.44	76.07	0.000
Cigarettes	-31.014	6.498	-4.77	0.000

$$S = 482.603 \quad R - Sq = 12.6\% \quad R - Sq(\text{adj}) = 12.0\%$$

- (h) Interpret the slope.
 (i) Interpret the y-intercept.
 (j) Would you recommend using this model to predict the birth weight of a baby whose mother smoked 10 cigarettes per day during the third trimester? Why?

- (k) Does this study demonstrate that smoking during the third trimester causes lower-birth-weight babies?
 (l) Cite some lurking variables that may confound the results of the study.

Explaining the Concepts

31. What is a residual? What does it mean when a residual is positive?
 32. Explain the phrase *outside the scope of the model*. Why is it dangerous to make predictions outside the scope of the model?
 33. Explain the meaning of Legendre's quote given on page 189.
 34. Explain what each point on the least-squares regression line represents.
 35. Mark Twain, in his book *Life on the Mississippi* (1884), makes the following observation:

Therefore, the Mississippi between Cairo and New Orleans was twelve hundred and fifteen miles long one hundred and seventy-six years ago. It was eleven hundred and eighty after the cut-off of 1722. It was one thousand and forty after the American Bend cut-off. It has lost sixty-seven miles since. Consequently its length is only nine hundred and seventy-three miles at present.

Now, if I wanted to be one of those ponderous scientific people, and "let on" to prove what had occurred in the remote past by what had occurred in a given time in the recent past, or what will occur in the far future by what has occurred in late years, what an opportunity is here! Geology never had such a chance, nor such exact data to argue from! Nor "development of species," either! Glacial epochs are great things, but they are vague—vague. Please observe:

In the space of one hundred and seventy-six years the Lower Mississippi has shortened itself two hundred and forty-two miles. That is an average of a trifle over one mile and a third per year. Therefore, any calm person, who is not blind or idiotic, can see that in the Old Oolitic Silurian Period, just a million years ago next November, the Lower Mississippi River was upwards of one million three hundred thousand miles long, and stuck out over the Gulf of Mexico like a fishing-rod. And by the same token any person can see that seven hundred and forty-two years from now the Lower Mississippi will be only a mile and three-quarters long, and Cairo and New Orleans will have joined their streets together, and be plodding comfortably along under a single mayor and a mutual board of aldermen. There is something fascinating about Science. One gets such wholesale returns of conjecture out of such a trifling investment of fact.

Discuss how Twain's observation relates to the material presented in this section.

4.3 The Coefficient of Determination



Preparing for This Section Before getting started, review the following:

- Outliers (Section 3.4, pp. 150–151)

Objective ① Compute and interpret the coefficient of determination

① Compute and Interpret the Coefficient of Determination

Consider the club-head speed versus distance data introduced in Section 4.1. How could we predict the distance of a randomly selected shot? Our best guess might be the mean distance of all shots from the sample data given in Table 1, $\bar{y} = 266.75$ yards.

Now suppose this particular shot resulted from a swing with a club-head speed of 103 mph. Knowing that a linear relation exists between club-head speed and distance, we can improve our estimate of the distance of the shot by using the least-squares regression line to adjust our guess to $\hat{y} = 3.1661(103) - 55.7966 = 270.3$ yards. In statistical terms, we say that some of the variation in distance is explained by the linear relation between club-head speed and distance.

The proportion of variation in the response variable that is explained by the least-squares regression line is called the *coefficient of determination*.

Definition

The **coefficient of determination**, R^2 , measures the proportion of total variation in the response variable that is explained by the least-squares regression line.

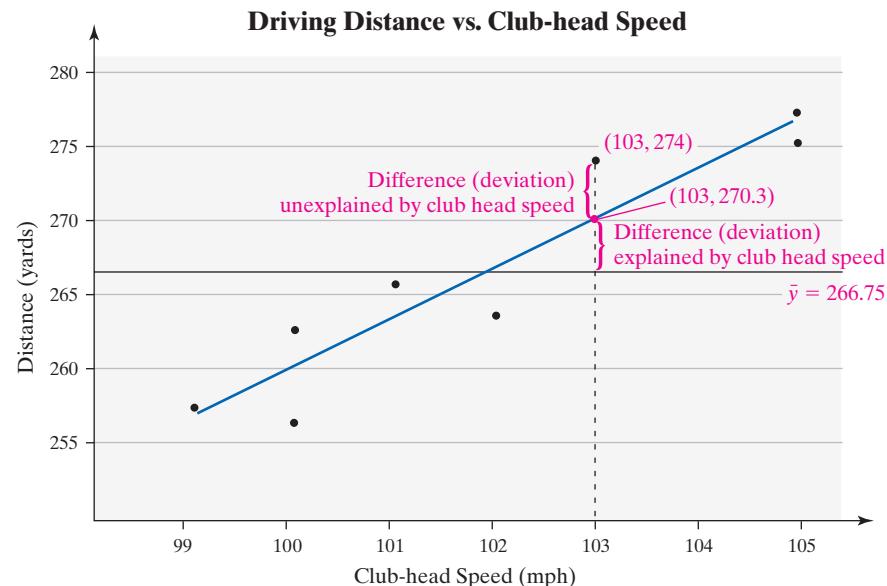
IN OTHER WORDS

The coefficient of determination is a measure of how well the least-squares regression line describes the relation between the explanatory and response variables. The closer R^2 is to 1, the better the line describes how changes in the explanatory variable affect the value of the response variable.

The coefficient of determination is a number between 0 and 1, inclusive. That is, $0 \leq R^2 \leq 1$. If $R^2 = 0$, the least-squares regression line has no explanatory value. If $R^2 = 1$, the least-squares regression line explains 100% of the variation in the response variable.

In Figure 16, a horizontal line is drawn at $\bar{y} = 266.75$, the predicted distance of a shot without any knowledge of club-head speed. If we know that the club-head speed is 103 miles per hour, we increase our guess to 270.3 yards. The difference between the predicted distance of 266.75 yards and the predicted distance of 270.3 yards is due to

Figure 16



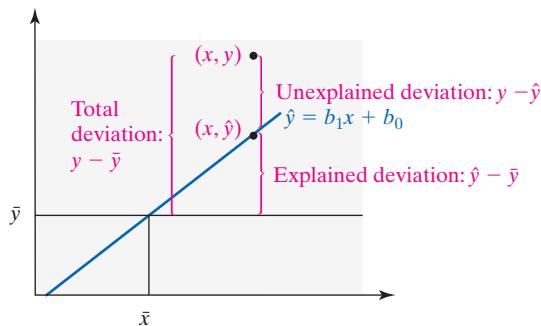
IN OTHER WORDS

The word **deviations** comes from deviate. To deviate means "to stray."

the fact that the club-head speed is 103 miles per hour. In other words, the difference between the prediction of $\hat{y} = 270.3$ and $\bar{y} = 266.75$ is explained by the linear relation between club-head speed and distance. The observed distance when club-head speed is 103 miles per hour is 274 yards (see Table 5 on page 187). The difference between our predicted value, $\hat{y} = 270.3$, and the actual value, $y = 274$, is due to factors (variables) other than the club-head speed (wind speed, position of the ball on the club face, and so on) and also to random error. The differences just discussed are called **deviations**.

The deviation between the observed and mean values of the response variable is called the **total deviation**, so total deviation = $y - \bar{y}$. The deviation between the predicted and mean values of the response variable is called the **explained deviation**, so explained deviation = $\hat{y} - \bar{y}$. Finally, the deviation between the observed and predicted values of the response variable is called the **unexplained deviation**, so unexplained deviation = $y - \hat{y}$. See Figure 17.

Figure 17



The figure illustrates that

$$\text{Total deviation} = \text{unexplained deviation} + \text{explained deviation}$$

$$y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$$

Although beyond the scope of this text, it can be shown that

$$\text{Total variation} = \text{unexplained variation} + \text{explained variation}$$

$$\sum(y - \bar{y})^2 = \sum(y - \hat{y})^2 + \sum(\hat{y} - \bar{y})^2$$

Dividing both sides by total variation, we obtain

$$1 = \frac{\text{unexplained variation}}{\text{total variation}} + \frac{\text{explained variation}}{\text{total variation}}$$

Subtracting $\frac{\text{unexplained variation}}{\text{total variation}}$ from both sides, we get

$$R^2 = \frac{\text{explained variation}}{\text{total variation}}$$

$$= 1 - \frac{\text{unexplained variation}}{\text{total variation}}$$

Unexplained variation is found by summing the squares of the residuals,

$\sum \text{residuals}^2 = \sum (y - \hat{y})^2$. So the smaller the sum of squared residuals, the smaller the unexplained variation and, therefore, the larger R^2 will be. Therefore, the closer the observed y 's are to the regression line (the predicted \hat{y} 's), the larger R^2 will be.

To find the coefficient of determination, R^2 , for the least-squares regression model $\hat{y} = b_1x + b_0$ (that is, a single explanatory variable to the first degree), square the linear correlation coefficient. That is, $R^2 = r^2$. This method does not work in general, however.

CAUTION!

Squaring the linear correlation coefficient to obtain the coefficient of determination works only for the least-squares linear regression model

$$\hat{y} = b_1x + b_0$$

The method does not work in general.

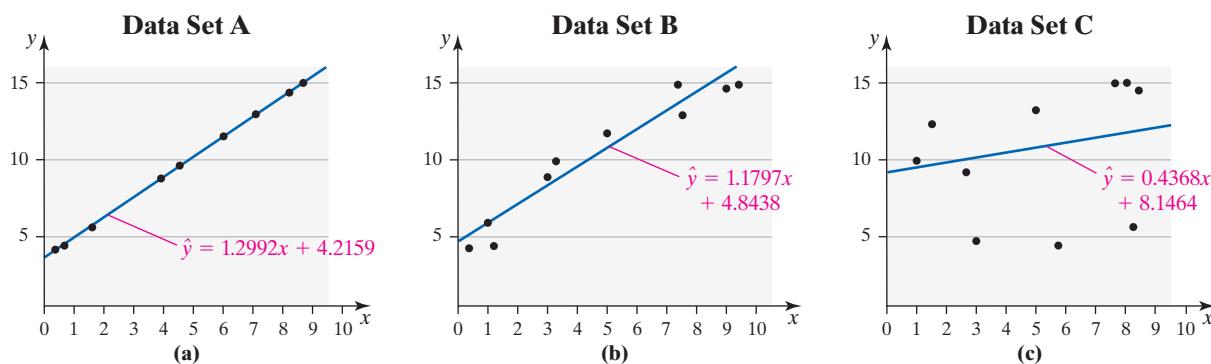
To reinforce the concept of the coefficient of determination, consider the three data sets in Table 7.

Table 7

Data Set A		Data Set B		Data Set C	
x	y	x	y	x	y
3.6	8.9	3.1	8.9	2.8	8.9
8.3	15.0	9.4	15.0	8.1	15.0
0.5	4.8	1.2	4.8	3.0	4.8
1.4	6.0	1.0	6.0	8.3	6.0
8.2	14.9	9.0	14.9	8.2	14.9
5.9	11.9	5.0	11.9	1.4	11.9
4.3	9.8	3.4	9.8	1.0	9.8
8.3	15.0	7.4	15.0	7.9	15.0
0.3	4.7	0.1	4.7	5.9	4.7
6.8	13.0	7.5	13.0	5.0	13.0

Figures 18(a), (b), and (c) represent the scatter diagrams of data sets A, B, and C, respectively.

Figure 18



Notice that the y -values in each of the three data sets are the same. The variance of y is 17.49. In Figure 18(a), almost 100% of the variability in y can be explained by the least-squares regression line because the data lie almost perfectly on a straight line. In Figure 18(b), a high percentage of the variability in y can be explained by the least-squares regression line because the data have a strong linear relation. Finally, in Figure 18(c), a low percentage of the variability in y is explained by the least-squares regression line. If x increases, we cannot easily predict the change in y . If we compute the coefficient of determination, R^2 , for the three data sets in Table 7 we obtain the following results:

Data Set	Coefficient of Determination, R^2	Interpretation
A	99.99%	99.99% of the variability in y is explained by the least-squares regression line.
B	94.7%	94.7% of the variability in y is explained by the least-squares regression line.
C	9.4%	9.4% of the variability in y is explained by the least-squares regression line.

Notice that, as the explanatory ability of the line decreases, the coefficient of determination, R^2 , also decreases.

EXAMPLE 1 Determining the Coefficient of Determination

Problem Determine and interpret the coefficient of determination, R^2 , for the club-head speed versus distance data shown in Table 5 on page 187.

(continued)

By Hand Approach

To compute R^2 , square the linear correlation coefficient, r , found in Example 3 from Section 4.1 on page 176.

By Hand Solution

$$\begin{aligned} R^2 &= r^2 = 0.939^2 \\ &= 0.882 \\ &= 88.2\% \end{aligned}$$

Technology Approach

We will use Excel to determine R^2 . The steps for obtaining the coefficient of determination using the TI-83/84 Plus graphing calculator, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step below.

Technology Solution

Figure 19 shows the results from Excel. The coefficient of determination is highlighted.

Figure 19

Summary Output	
Regression Statistics	
Multiple R	0.938695838
R Square	0.881149876
Adjusted R	0.861341522
Square	
Standard Error	2.882638465
Observations	8

Interpretation 88.2% [Tech: 88.1%] of the variation in distance is explained by the least-squares regression line, and 11.8% of the variation in distance is explained by other factors. ☐

NW Now Work Problems 5 and 7**Technology Step-by-Step****Determining R^2** **TI-83/84 Plus**

Use the same steps that were followed to obtain the correlation coefficient to obtain R^2 . Diagnostics must be on.

Minitab

This is provided in the standard regression output.

Excel

This is provided in the standard regression output.

StatCrunch

Follow the same steps used to obtain the least-squares regression line. The coefficient of determination is given as part of the output (R-sq).

**4.3 Assess Your Understanding****Vocabulary and Skill Building**

1. The _____, R^2 , measures the proportion of total variation in the response variable that is explained by the least-squares regression line.

2. Total deviation = _____ deviation + _____ deviation

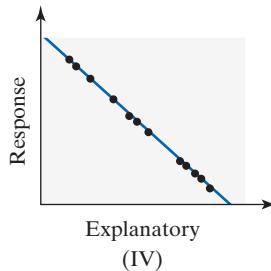
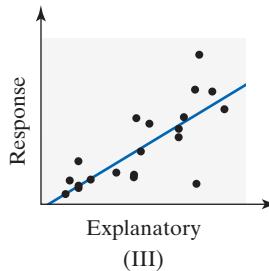
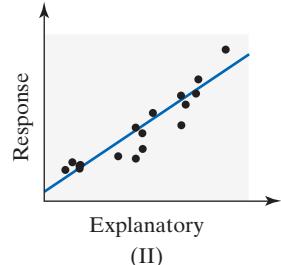
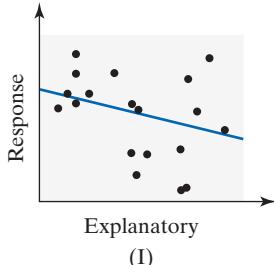
3. Match the coefficient of determination to the scatter diagram. The scales on the horizontal and vertical axis are the same for each scatter diagram.

(a) $R^2 = 0.58$

(c) $R^2 = 1$

(b) $R^2 = 0.90$

(d) $R^2 = 0.12$



4. Use the linear correlation coefficient given to determine the coefficient of determination, R^2 . Interpret each R^2 .

(a) $r = -0.32$

(c) $r = 0.40$

(b) $r = 0.13$

(d) $r = 0.93$

Applying the Concepts

- NW 5. The Other Old Faithful** Perhaps you are familiar with the famous Old Faithful geyser in Yellowstone National Park. Another Old Faithful geyser is located in Calistoga in California's Napa Valley. The following data represent the time (in minutes) between eruptions and the length of eruption for 9 randomly selected eruptions. The coefficient of determination is 83.0%. Provide an interpretation of this value.

Time between Eruptions, x	Length of Eruption, y	Time between Eruptions, x	Length of Eruption, y
12.17	1.88	11.70	1.82
11.63	1.77	12.27	1.93
12.03	1.83	11.60	1.77
12.15	1.83	11.72	1.83
11.30	1.70		

Source: Ladonna Hansen, Park Curator.

- DATA 6. Concrete** As concrete cures, it gains strength. The following data represent the 7-day and 28-day strength (in pounds per square inch) of a certain type of concrete. The coefficient of determination is 57.5%. Provide an interpretation of this value.

7-Day Strength, x	28-Day Strength, y	7-Day Strength, x	28-Day Strength, y
2300	4070	2480	4120
3390	5220	3380	5020
2430	4640	2660	4890
2890	4620	2620	4190
3330	4850	3340	4630

Problems 7–12 use the results from Problems 27–32 in Section 4.1 and Problems 17–22 in Section 4.2.

- NW 7. An Unhealthy Commute** Use the results from Problem 27 in Section 4.1 and Problem 17 in Section 4.2 to:

- (a) Determine the coefficient of determination, R^2 .
 (b) Interpret the coefficient of determination and comment on the adequacy of the linear model.

- 8. Credit Scores** Use the results from Problem 28 in Section 4.1 and Problem 18 in Section 4.2 to:

- (a) Determine the coefficient of determination, R^2 .
 (b) Interpret the coefficient of determination and comment on the adequacy of the linear model.

- 9. Height versus Head Circumference** Use the results from Problem 29 in Section 4.1 and Problem 19 in Section 4.2 to:

- (a) Compute the coefficient of determination, R^2 .
 (b) Interpret the coefficient of determination and comment on the adequacy of the linear model.

- 10. American Black Bears** Use the results from Problem 30 in Section 4.1 and Problem 20 in Section 4.2 to:

- (a) Compute the coefficient of determination, R^2 .
 (b) Interpret the coefficient of determination and comment on the adequacy of the linear model.

- 11. Home Runs** Use the results from Problem 31 in Section 4.1 and Problem 21 in Section 4.2.

- (a) What proportion of the variability in distance is explained by the relation between speed at which the ball is hit and distance?
 (b) Interpret the coefficient of determination and comment on the adequacy of the linear model.

- 12. Hurricanes** Use the results from Problem 32 in Section 4.1 and Problem 22 in Section 4.2.

- (a) What proportion of the variability in the wind speed is explained by the relation between the atmospheric pressure and wind speed?
 (b) Interpret the coefficient of determination and comment on the adequacy of the linear model.

- 13. Weight of a Car versus Miles per Gallon** Suppose that we add the Dodge Viper to the data in Problem 24 in Section 4.2. A Dodge Viper weighs 3425 pounds and gets 11 miles per gallon. Compute the coefficient of determination of the expanded data set. What effect does the addition of the Viper to the data set have on R^2 ?

- 14. American Black Bears** Suppose that we find a bear that is 205 cm long and weighs 187 kg and add the bear to the data in Problem 20 from Section 4.2. Compute the coefficient of determination of the expanded data set. What effect does the additional bear have on R^2 ?

- 15. Threaded Problem: Tornado** Is the width of a tornado related to the amount of distance for which the tornado is on the ground? Go to www.pearsonhighered.com/sullivanstats to obtain the data file 4_3_15. The data represent the width (in yards) and length (in miles) for tornadoes in Louisiana in 2017 and are from the file “Tornadoes_2017” that we have been analyzing throughout the course.

- (a) What is the explanatory variable?
 (b) Explain why this data should be analyzed as bivariate quantitative data.
 (c) Draw a scatter diagram of the data. What type of relation appears to exist between the width and length of a tornado?
 (d) Determine the correlation coefficient between width and length.
 (e) Is there a linear relation between a tornado’s width and its length on the ground?
 (f) Find the least-squares regression line.
 (g) Predict the length of a tornado whose width is 500 yards.
 (h) Was the tornado whose width was 600 yards and length was 10.09 miles on the ground longer than would be expected?
 (i) Interpret the slope.
 (j) Explain why it does not make sense to interpret the intercept.
 (k) What proportion of the variability in tornado length is explained by the width of the tornado?

- DATA 16. Putting It Together: Exam Scores** The data on the next page represent scores earned by students in Sullivan’s Elementary Algebra course for Chapter 2 (Linear Equations and Inequalities in One Variable) and Chapter 3 (Linear Equations and Inequalities in Two Variables). Completely summarize the relation between Chapter 2 and Chapter 3 exam scores, treating Chapter 2 exam scores as the explanatory variable. Write a report detailing the results of the analysis. What does the relationship say about the role Chapter 2 plays in a student’s understanding of Chapter 3?

Chapter 2 Score	Chapter 3 Score	Chapter 2 Score	Chapter 3 Score
71.4	76.1	95.2	87.0
76.2	82.6	85.7	73.9
60.3	60.9	71.4	74.6
88.1	91.3	33.3	45.7
95.2	78.3	78.0	37.3
82.5	100	83.3	88.0
100	95.7	100	100
87.3	81.5	81.0	76.1
71.4	50.7	76.2	63.0
95.2	81.5		

Source: Michael Sullivan.



17. Putting It Together: Cigarette Smuggling Go to www.pearsonhighered.com/sullivanstats to obtain the data file 4_3_17. The data represent the 2015 tax rate per pack of cigarettes and the percent of cigarettes smuggled as a percentage of total consumption. A negative value of consumption represents a net outflow of cigarettes while a positive value represents an inflow of cigarettes. For example, in Alabama, 7.5% of all cigarettes purchased leave the state. In Arizona, 44.8% of all cigarettes consumed are smuggled into the state. Alaska, Hawaii, North Carolina, and the District of Columbia are not included in the analysis. Describe the data and write an article that discusses the impact that cigarette taxes may have on smuggling.

Source: The Tax Foundation.

Retain Your Knowledge

18. Sullivan Survey II Go to www.pearsonhighered.com/sullivanstats to obtain the data file SullivanStatsSurveyII using the file format of your choice for the version of the text you are using. One question asked in this survey is, “What percent of income do you believe individuals should pay in federal income tax?”

- (a) Draw a relative frequency histogram of the data in the column “Tax Rate” using a lower class limit of the first class equal to 0 and a class width of 5. Comment on the shape of the distribution. What is the class with the highest relative frequency?
- (b) Determine the mean and median tax rate.
- (c) Determine the standard deviation and interquartile range of tax rate.
- (d) Determine the lower and upper fences. Are there any outliers?
- (e) In this chapter, we have focused on bivariate quantitative data. However, side-by-side boxplots may be used to see association between a qualitative and quantitative variable. Draw side-by-side boxplots of tax rate by gender. Comment on any interesting features. Which gender appears to prefer the higher tax rate?
- (f) Draw side-by-side boxplots of tax rate by political philosophy. Comment on any features in the graph.

4.4 Contingency Tables and Association



Preparing for This Section Before getting started, review the following:

- Side-by-side bar graphs (Section 2.1, pp. 66–67)

Objectives

- ① Compute the marginal distribution of a variable
- ② Use the conditional distribution to identify association among categorical data
- ③ Explain Simpson’s Paradox

In Sections 4.1 to 4.3, we looked at techniques for summarizing relations between two quantitative variables. We now look at techniques for summarizing relations between two qualitative variables.

Consider the data (measured in thousands) in Table 8, which represent the employment status and level of education of all U.S. residents 25 years old or older in

Table 8

Employment Status	Level of Education			
	Did Not Finish High School	High School Graduate	Some College	Bachelor's Degree or Higher
Employed	9607	34,625	36,370	57,102
Unemployed	570	1274	1170	1305
Not in the labor force	11,662	26,426	19,861	20,841

Source: Bureau of Labor Statistics.

November 2018. By definition, an individual is unemployed if he or she is actively seeking but is unable to find work. An individual is not in the labor force if he or she is not employed and not actively seeking employment.

Table 8 is referred to as a **contingency table**, or a **two-way table**, because it relates two categories of data. The **row variable** is employment status, because each row in the table describes the employment status of a group. The **column variable** is level of education. Each box inside the table is called a **cell**. For example, the cell corresponding to employed high school graduates is in the first row, second column. Each cell contains the frequency of the category: In November 2018, 9607 thousand employed individuals had not finished high school.

The data in Table 8 describe two characteristics regarding the population of U.S. residents who are 25 years or older: their employment status and their level of education. We want to investigate whether the two variables are associated. For example, are individuals who have a higher level of education more likely to be employed? Just as we did in Sections 4.1 to 4.3, we discuss both numerical and graphical methods for summarizing the data.

① Compute the Marginal Distribution of a Variable

The first step in summarizing data in a contingency table is to determine the distribution of each variable separately. To do so, we create *marginal distributions*.

IN OTHER WORDS

The distributions are **Definition** called **marginal distributions** because they appear in the right and the bottom margin of the contingency table.

A **marginal distribution** of a variable is a frequency or relative frequency distribution of either the row or column variable in the contingency table.

A marginal distribution removes the effect of either the row variable or the column variable in the contingency table.

To create a marginal distribution for a variable, calculate the row and column totals for each category of the variable. The row totals represent the distribution of the row variable. The column totals represent the distribution of the column variable.

EXAMPLE 1

Determining Frequency Marginal Distributions

Problem Find the frequency marginal distributions for employment status and level of education from the data in Table 8.

Approach Find the row total for the category “employed” by adding the number of employed individuals who did not finish high school, who finished high school, and so on. Repeat this process for each category of employment status.

Find the column total for the category “did not finish high school” by adding the number of employed individuals, unemployed individuals, and individuals not in the labor force who did not finish high school. Repeat this process for each level of education.

Solution In Table 9 on the next page, the blue entries represent the marginal distribution of the row variable “employment status.” For example, there were $9607 + 34,625 + 36,370 + 57,102 = 137,704$ thousand employed individuals in November 2018. The red entries represent the marginal distribution of the column variable “level of education.”

The marginal distribution for employment status removes the effect of level of education; the marginal distribution for level of education removes the effect of employment status. The marginal distribution of level of education shows there were more Americans with a bachelor’s degree or higher than there were Americans who were high school graduates (79,248 thousand versus 62,325 thousand) in November 2018. The marginal distribution of employment status shows that 137,704 thousand Americans were employed. The table also indicates that there were 220,813 thousand U.S. residents 25 years old or older.

(continued)

Table 9

Level of Education					
Employment Status	Did Not Finish High School	High School Graduate	Some College	Bachelor's Degree or Higher	Totals
Employed	9607	34,625	36,370	57,102	137,704
Unemployed	570	1274	1170	1305	4319
Not in the Labor Force	11,662	26,426	19,861	20,841	78,790
Totals	21,839	62,325	57,401	79,248	220,813

NW Now Work Problem 9(a)

We can use the row and column totals obtained in Example 1 to calculate the relative frequency marginal distribution for level of education and employment status.

EXAMPLE 2

Determining Relative Frequency Marginal Distributions

Problem Determine the relative frequency marginal distribution for level of education and employment status from the data in Table 9.

CAUTION!

For relative frequency marginal distributions (such as in Table 10), row or column totals might not sum exactly to 1 due to rounding.

Approach The relative frequency marginal distribution for the row variable, employment status, is found by dividing the row total for each employment status by the table total, 220,813. The relative frequency marginal distribution for the column variable, level of education, is found by dividing the column total for each level of education by the table total.

Solution Table 10 represents the relative frequency marginal distribution for each variable.

Table 10

Level of Education					
Employment Status	Did Not Finish High School	High School Graduate	Some College	Bachelor's Degree or Higher	Totals
Employed	9607	34,625	36,370	57,102	$\frac{137,704}{220,813} = 0.624$
Unemployed	570	1274	1170	1305	$\frac{4319}{220,813} = 0.020$
Not in the Labor Force	11,662	26,426	19,861	20,841	$\frac{78,790}{220,813} = 0.357$
Totals	$\frac{21,839}{220,813} = 0.099$	$\frac{62,325}{220,813} = 0.282$	$\frac{57,401}{220,813} = 0.260$	$\frac{79,248}{220,813} = 0.359$	1

Table 10 shows that 9.9% of U.S. residents 25 years old or older did not graduate from high school, and 62.4% of U.S. residents 25 years old or older were employed in November 2018.

NW Now Work Problems 9(b) and (c)

② Use the Conditional Distribution to Identify Association among Categorical Data

As we look at the information in Tables 9 and 10, we might ask whether a higher level of education is associated with a higher likelihood of being employed. If level of education does not play any role, we would expect the relative frequencies for employment status at

each level of education to be close to 0.624, the relative frequency marginal distribution for employment status given in blue in Table 10. So we would expect 62.4% of individuals who did not finish high school, 62.4% of individuals who finished high school, 62.4% of individuals with some college, and 62.4% of individuals with at least a bachelor's degree to be employed. If the relative frequencies for these various levels of education are different, we would associate this difference with the level of education.

CAUTION!

To describe the association between two categorical variables, relative frequencies must be used because there are different numbers of observations for the categories.

The marginal distributions in Tables 9 and 10 allow us to see the distribution of either the row variable (employment status) or the column variable (level of education) but we do not get a sense of association from these tables.

When describing any association between two categories of data, we must use relative frequencies instead of frequencies, because frequencies are difficult to compare when there are different numbers of observations for the categories of a variable.

EXAMPLE 3

Comparing Two Categories of a Variable

Problem What proportion of the following groups of individuals is employed?

- (a) Those who did not finish high school (b) High school graduates
- (c) Those who finished some college (d) Those who have at least a bachelor's degree

Approach In part (a), we are asking, “Of the individuals who did not finish high school, what proportion is employed?” To determine this proportion, divide the number of employed individuals who did not finish high school by the number of people who did not finish high school. Repeat this process to answer parts (b)–(d).

Solution

- (a) In November 2018, 21,839 thousand individuals 25 years old or older did not finish high school. (See Table 9.) Of this number, 9607 thousand were employed. Therefore, $\frac{9607}{21,839} = 0.440$ represents the proportion of individuals who did not finish high school who are employed.
- (b) In November 2018, 62,325 thousand individuals 25 years old or older were high school graduates. Of this number, 34,625 thousand were employed. Therefore, $\frac{34,625}{62,325} = 0.556$ represents the proportion of individuals who graduated from high school who are employed.
- (c) In November 2018, 57,401 thousand individuals 25 years old or older had some college. Of this number, 36,370 thousand were employed. Therefore, $\frac{36,370}{57,401} = 0.634$ were employed.
- (d) In November 2018, 79,248 thousand individuals 25 years old or older had at least a bachelor's degree. Of this number, 57,102 thousand were employed. Therefore, $\frac{57,102}{79,248} = 0.721$ were employed.

We can see from these relative frequencies that as the level of education increases, the proportion of individuals employed increases.

IN OTHER WORDS

Since we are finding the conditional distribution of employment status by level of education, the level of education is the **explanatory variable** and the employment status is the **response variable**.

The results in Example 3 are only partial. In general, we create a relative frequency distribution for each value of the explanatory variable. For our example, this means we would construct a relative frequency distribution for individuals who did not finish high school, a second relative frequency distribution for individuals who are high school graduates, and so on. These relative frequency distributions are called *conditional distributions*.

Definition

A **conditional distribution** lists the relative frequency of each category of the response variable, given a specific value of the explanatory variable in the contingency table.

An example should help to solidify our understanding of the definition.

EXAMPLE 4 Constructing a Conditional Distribution

Problem Find the conditional distribution of the response variable employment status by level of education, the explanatory variable, for the data in Table 8 on page 206. What is the association between level of education and employment status?

Approach First compute the relative frequency for each employment status, given that the individual did not finish high school. Next, compute the relative frequency for each employment status, given that the individual is a high school graduate. Continue computing the relative frequency for each employment status for the next two levels of education. This is the same approach taken in Example 3 for employed individuals.

Solution First use the number of individuals who did not finish high school (21,839 thousand) as the denominator in computing the relative frequencies for each employment status.

Then use the number of individuals who graduated from high school as the denominator in computing relative frequencies for each employment status.

Next, compute the relative frequency for each employment status, given that the individual had some college and then given that the individual had at least a bachelor's degree. We obtain Table 11. Notice that the "Employed" row in Table 11 shows the results from Example 3.

Table 11

Employment Status	Level of Education			
	Did Not Finish High School	High School Graduate	Some College	Bachelor's Degree or Higher
Employed	$\frac{9607}{21,839} = 0.440$	$\frac{34,625}{62,325} = 0.556$	$\frac{36,370}{57,401} = 0.634$	$\frac{57,102}{79,248} = 0.721$
Unemployed	$\frac{570}{21,839} = 0.026$	$\frac{1274}{62,325} = 0.020$	$\frac{1170}{57,401} = 0.020$	$\frac{1305}{79,248} = 0.016$
Not in the Labor Force	$\frac{11,662}{21,839} = 0.534$	$\frac{26,426}{62,325} = 0.424$	$\frac{19,861}{57,401} = 0.346$	$\frac{20,841}{79,248} = 0.263$
Totals	1	1	1	1

Looking at the conditional distributions of employment status by level of education, associations become apparent. Read the information in Table 11 from left to right. As the amount of schooling (the explanatory variable) increases, the proportion employed within each category also increases. As the amount of schooling increases, the proportion not in the labor force decreases. The proportion unemployed with at least a bachelor's degree is lower than those unemployed in the other three levels of education.

Information about individuals' levels of education provides insight into their employment status. For example without any information about level of education, we might predict the number of employed individuals out of 100 to be about 62 because the employment rate for the entire United States is 62.4%. (See Table 10.) However, we would change our "guess" to 72 if we knew the 100 individuals had at least a bachelor's degree. Do you see why? It is because 72.1% of individuals with at least a bachelor's degree are employed. The association between level of education and employment status allows us to adjust our predictions.

NW Now Work Problem 9(d)

In Example 4, we were able to see how a change in level of education affected employment status. Therefore, we are treating level of education (the column variable) as the explanatory variable and employment status as the response variable.

We could also construct a conditional distribution of level of education by employment status. In this situation, we would be treating employment status (the row

variable) as the explanatory variable and level of education as the response variable. The procedure is the same, except the distribution uses the rows instead of the columns because now the explanatory variable, employment status, is the row variable.

As is usually the case, a graph can provide a powerful depiction of the data.

EXAMPLE 5

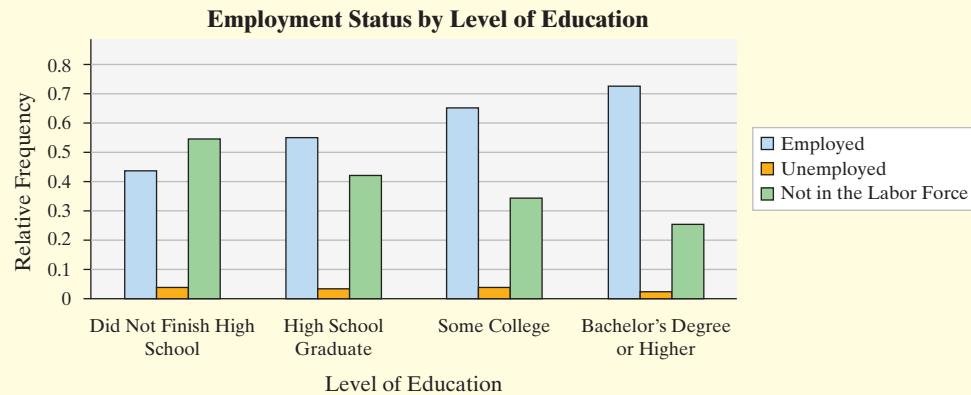
Drawing a Bar Graph of a Conditional Distribution

Problem Using the results of Example 4, draw a bar graph that represents the conditional distribution of employment status by level of education.

Approach When drawing conditional bar graphs, label the values of the explanatory variable on the horizontal axis, and use different colored bars for each value of the response variable. So, draw three bars for each level of education. Let the horizontal axis represent level of education and the vertical axis represent the relative frequency. Each “grouping” represents the distribution of employment status at that level of education.

Solution See Figure 20. It is clear that as level of education increases, the proportion employed also increases. As the level of education increases, the proportion not in the labor force decreases.

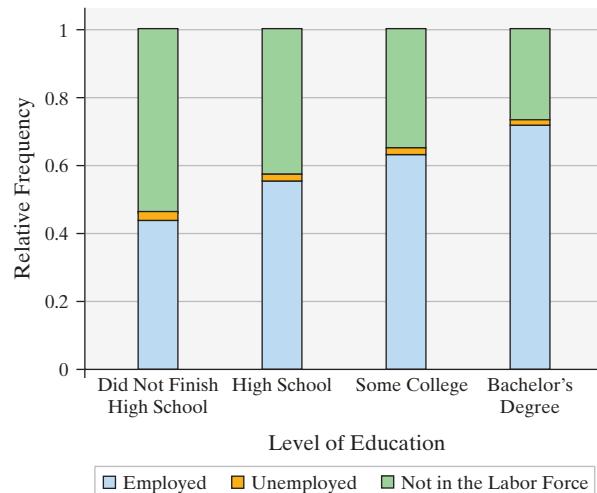
Figure 20



Instead of drawing a conditional bar graph as shown in Figure 20, we can draw *stacked* (or *segmented*) *bar graphs*. In these graphs, there is one bar for each value of the explanatory variable. Each bar is then divided into segments such that the height of each segment within a bar represents the proportion of observations corresponding to the response variable. Figure 21 shows the stacked bar graph for the data in Table 11. The height of the blue portion in each segment is the proportion employed for each level of

Figure 21

Employment Status by Level of Education



**NW Now Work Problems 9(e)
and (f)**

education, the height of the orange portion in each segment is the proportion unemployed for each level of education, and the height of the green bar is the proportion not in the labor force for each level of education. Again, we can see as level of education increases, the proportion employed increases while the proportion not in the labor force decreases.

The methods presented in this section for identifying the association between two categorical variables are different from the methods for measuring association between two quantitative variables. The measure of association in this section is based on whether there are differences in the relative frequencies of the response variable (employment status) for the different categories of the explanatory variable (level of education). If differences exist, we might attribute these differences to the explanatory variable.

In addition, because the data in Example 1 are observational, we do not make any statements regarding causation. Level of education is not said to be a cause of employment status, because a controlled experiment was not conducted.

③ Explain Simpson's Paradox

In Section 4.1, we discussed how a lurking variable can cause two quantitative variables to be correlated even though they are unrelated. This same phenomenon exists when exploring the relation between two qualitative variables.

EXAMPLE 6 Gender Bias at the University of California, Berkeley

Problem The data in Table 12 show the admission status and gender of students who applied to the University of California, Berkeley. From the data in Table 12, the proportion of accepted applications is $\frac{1748}{4425} = 0.395$. The proportion of accepted men is $\frac{1191}{2590} = 0.460$ and the proportion of accepted women is $\frac{557}{1835} = 0.304$. On the basis of these proportions, a gender bias suit was brought against the university. The university was shocked and claimed that program of study is a lurking variable that created the apparent association between admission status and gender. The university supplied Table 13 in its defense. Develop a conditional distribution by program of study to defend the university's admission policies.

Source: P. J. Bickel, E. A. Hammel, and J. W. O'Connell. "Sex Bias in Graduate Admissions: Data from Berkeley." *Science* 187(4175): 398–404, 1975.

Table 12

	Accepted (A)	Not Accepted (NA)	Total
Men	1191	1399	2590
Women	557	1278	1835
Total	1748	2677	4425

**Table 13 Admission Status (Accepted, A, or Not Accepted, NA),
for Six Programs of Study (A, B, C, D, E, F) by Gender**

	A	B	C	D	E	F						
Men	A 511	NA 314	A 353	NA 207	A 120	NA 205	A 138	NA 279	A 53	NA 138	A 16	NA 256
Women	A 89	NA 19	A 17	NA 8	A 202	NA 391	A 131	NA 244	A 94	NA 299	A 24	NA 317

Approach Determine the proportion of accepted men for each program of study and separately determine the proportion of accepted women for each program of study. A significant difference between the proportions of men and women accepted within each program of study may be evidence of discrimination; otherwise, the university should be exonerated.

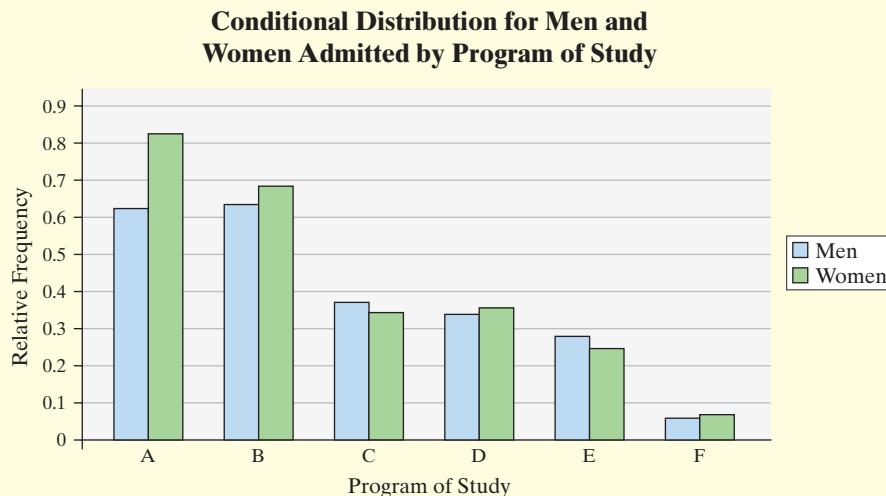
Solution The proportion of men who applied to program A and were accepted is $\frac{511}{511 + 314} = 0.619$. The proportion of women who applied to program A and were accepted is $\frac{89}{89 + 19} = 0.824$. So, within program A, a higher proportion of women were accepted. Table 14 shows the conditional distribution for the remaining programs.

Table 14 Conditional Distribution of Applicants Admitted for Men and Women by Program of Admission

	A	B	C	D	E	F
Men	$\frac{511}{825} = 0.619$	0.630	0.369	0.331	0.277	0.059
Women	0.824	0.680	0.341	0.349	0.239	0.070

Figure 22 shows the bar graph of the conditional distribution in Table 14. The blue bars represent the proportion of men admitted for each program; the green bars represent the proportion of women admitted for each program.

Figure 22



Four of the six programs actually had a higher proportion of women accepted! And the proportion of men accepted in programs C and E is not much higher than the proportion of women.

What caused the overall proportion of accepted men to be so much higher than the overall proportion of accepted women within the entire university, when, within each program, the proportions either differ very little or may imply that women are accepted at a higher rate? The initial analysis did not account for the lurking variable, program of study. There were many more male applicants in programs A and B than female applicants, and these two programs happen to have higher acceptance rates. The higher acceptance rates in these programs led to the false conclusion that the University of California, Berkeley, was biased against gender in its admissions.

In Example 6, the association between gender and admission reverses when the lurking variable program of study is accounted for in the analysis. This illustrates a phenomenon known as *Simpson's Paradox*.

Definition

Simpson's Paradox describes a situation in which an association between two variables inverts or goes away when a third variable is introduced to the analysis.

Technology Step-by-Step

Contingency Tables and Association

Minitab

Determining Marginal and Conditional Distributions

- Enter the values of the row variable in column C1 and the corresponding values of the column variable in C2. The frequency for the cell is entered in C3. For example, the data in Table 8 would be entered as follows:

	C1-T	C2-T	C3
	Employment Status	Level of Education	Frequency
1	Employed	No High School	9607
2	Unemployed	No High School	570
3	Not in Labor Force	No High School	11662
4	Employed	HS	34625
5	Unemployed	HS	1274
6	Not in Labor Force	HS	26426
7	Employed	Some College	36370
8	Unemployed	Some College	1170
9	Not in Labor Force	Some College	19861
10	Employed	Bachelors	57102
11	Unemployed	Bachelors	1305
12	Not in Labor Force	Bachelors	20841

- Select the **Stat** menu and highlight **Tables**. Then select **Descriptive Statistics...**
- In the cell “For Rows:” enter C1. In the cell “For Columns:” enter C2. In the cell “Frequencies are in:” enter C3. Click the Options button and make sure the radio button for Display marginal statistics for Rows and columns is checked. Click OK. Click the Categorical Variables button and then select the summaries you desire. Click OK twice.

Drawing Bar Graphs of Conditional Distributions

- Enter the contingency table into the spreadsheet. The first column should be the row variable. For example, for the data in Table 8, the first column would be employment status. Each subsequent column would be the counts of each category of the column variable (level of education). Title each column.
- Select **Graph**. Highlight **Bar Chart...**
- Select “Values from a table” in the drop-down menu “Bars represent:”. Select either Cluster or Stack under Two-way table. Click OK.
- With the cursor in “Graph variables”, select the column variables. With the cursor in “Row labels:”, select the column containing the row labels (C1). If the column variable is the explanatory variable, select “Columns are...” under Table Arrangement; otherwise, select “Rows are ...”. If you are graphing a stacked bar graph, check the box “Stack innermost...”. Click Chart Option... Click “Show Y as Percent”. Click “Within categories...” under Take Percent. Click OK twice.

StatCrunch

Determining Marginal and Conditional Distributions

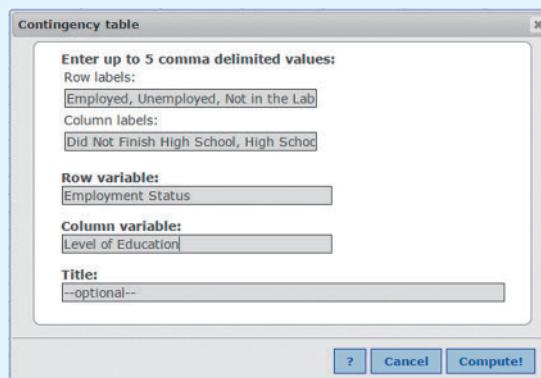
- Enter the contingency table into the spreadsheet. The first column should be the row variable. For example, for the data in Table 8, the first column

would be employment status. Each subsequent column would be the counts of each category of the column variable. For the data in Table 8, enter the counts for each level of education. Title each column (including the first column indicating the row variable).

- Select **Stat**, highlight **Tables**, and then **Contingency**, then select **With Summary**.
- Select the column variables. Then select the label of the row variable. For example, the data in Table 8 has four column variables (Did Not Finish High School, and so on) and the row label is employment status. To obtain relative frequency marginal distributions and conditional distributions, choose row percent and column percent under Display. Click Compute!.

Drawing Stacked Bar Graphs of Conditional Distributions

- Select **Applets**, highlight **Contingency Table**.
- Under Row labels, enter the values of the variables in the rows of the contingency table. Under Column labels, enter the values of the variables in the columns of the contingency table. For the data in Table 8, enter the labels as shown below. Click Compute!.



- Enter the counts in the contingency table. If you want the conditional distribution by the row variable, select “Row distribution.” If you want the conditional distribution by the column variable, select “Column distribution.”

Excel

Determining Marginal and Conditional Distributions

- Enter the given contingency table into a worksheet, including the column and row labels. Select **XLSTAT** > **Correlation/Association tests** > **Tests on contingency tables (Chi-square...)**
- General Tab: Fill in the dialogue box as follows:
 - Contingency table: Highlight the contingency table cell range (such as A1:D4)
 - Data format: Select the contingency table option.
 - Range/Sheet/Workbook: Select the Sheet option.
 - Labels included: Check this box.
- Outputs tab: Check the boxes for Observed frequencies, Proportions/Row, Proportion/Column, Proportions/Total and select Proportions. Click OK.



4.4 Assess Your Understanding

Vocabulary and Skill Building

- What is meant by a marginal distribution? What is meant by a conditional distribution?
- Refer to Table 8. Is constructing a conditional distribution by level of education different from constructing a conditional distribution by employment status? If they are different, explain the difference.
- Explain why we use the term *association* rather than *correlation* when describing the relation between two qualitative variables.
- Explain the idea behind Simpson's Paradox.

In Problems 5 and 6,

- Construct a frequency marginal distribution.
- Construct a relative frequency marginal distribution.
- Construct a conditional distribution by x .
- Draw a bar graph of the conditional distribution found in part (c).

5.

	x_1	x_2	x_3
y_1	20	25	30
y_2	30	25	50

6.

	x_1	x_2	x_3
y_1	35	25	20
y_2	65	75	80

Applying the Concepts

- Made in America** In a recent Harris Poll, a random sample of adult Americans (18 years and older) was asked, "When you see an ad emphasizing that a product is 'Made in America,' are you more likely to buy it, less likely to buy it, or neither more nor less likely to buy it?" The results of the survey, by age group, are presented in the contingency table below.

	18–34	35–44	45–54	55 +	Total
More likely	238	329	360	402	1329
Less likely	22	6	22	16	66
Neither more nor less likely	282	201	164	118	765
Total	542	536	546	536	2160

Source: The Harris Poll.

- How many adult Americans were surveyed? How many were 55 and older?
- Construct a relative frequency marginal distribution.
- What proportion of Americans are more likely to buy a product when the ad says "Made in America"?
- Construct a conditional distribution of likelihood to buy "Made in America" by age. That is, construct a conditional distribution treating age as the explanatory variable.
- Draw a bar graph of the conditional distribution found in part (d).
- Write a couple of sentences explaining any relation between likelihood to buy and age.

- Desirability Traits** In a recent Harris Poll, a random sample of adult Americans (18 years and older) was asked, "Given a choice of the following, which one would you most want to be?" Results of the survey, by gender, are given in the contingency table.

	Richer	Thinner	Smarter	Younger	None of these	Total
Male	520	158	159	181	102	1120
Female	425	300	144	81	92	1042
Total	945	458	303	262	194	2162

Source: The Harris Poll.

- How many adult Americans were surveyed? How many males were surveyed?
- Construct a relative frequency marginal distribution.
- What proportion of adult Americans want to be richer?
- Construct a conditional distribution of desired trait by gender. That is, construct a conditional distribution treating gender as the explanatory variable.
- Draw a bar graph of the conditional distribution found in part (d).
- Write a couple sentences explaining any relation between desired trait and gender.

- NW 9. Economic Mobility** A college education is certainly a factor in determining one's income. Researchers at payscale.com wanted to determine if there was an association between mid-career income and household income while an individual is in college. The idea was to determine if individuals from low-income households (while in college) earned a lower mid-career income than those from high-income households. Researchers surveyed 725 individuals in the middle of their careers and asked them to disclose their income class while in college and disclose their current income. The following data are based on their results.

Household Income in College	Current Mid-Career Income			
	Bottom Quartile	2nd Quartile	3rd Quartile	Top Quartile
Bottom Quartile	72	62	46	40
2nd or 3rd Quartile	70	78	75	67
Top Quartile	32	47	52	84

- Construct a frequency marginal distribution.
- Construct a relative frequency marginal distribution.
- What proportion of households are in the 2nd quartile in mid-career?
- Construct a conditional distribution by household income in college.
- Draw a bar graph of the conditional distribution found in part (d).
- Write a couple of sentences explaining any relation between household income in college and mid-career income.

- Feelings on Abortion** The Pew Research Center for the People and the Press conducted a poll in which it asked individuals to disclose their level of education and feelings about the availability of abortion. The table is based on the results of the survey.

	High School or Less	Some College	College Graduate
Generally available	90	72	113
Allowed, but more limited	51	60	77
Illegal, with few exceptions	125	94	69
Never permitted	51	14	17

Source: Pew Research Center for the People and the Press.

- (a) Construct a frequency marginal distribution.
- (b) Construct a relative frequency marginal distribution.
- (c) What proportion of college graduates feel that abortion should never be permitted?
- (d) Construct a conditional distribution of people's feelings about the availability of abortion by level of education.
- (e) Draw a bar graph of the conditional distribution found in part (d).
- (f) Is level of education associated with opinion on the availability of abortion? If so, how?

11. Health and Happiness The General Social Survey asks questions about one's happiness and health. One would think that health plays a role in one's happiness. Use the data in the table to determine whether healthier people tend to also be happier. Treat level of health as the explanatory variable.

	Poor	Fair	Good	Excellent	Total
Not too happy	696	1,386	1,629	732	4,443
Pretty happy	950	3,817	9,642	5,195	19,604
Very happy	350	1,382	4,520	5,095	11,347
Total	1,996	6,585	15,791	11,022	35,394

Source: General Social Survey.

12. Happy in Your Marriage? The General Social Survey asks questions about one's happiness in marriage. Is there an association between gender and happiness in marriage? Use the data in the table to determine if gender is associated with happiness in marriage. Treat gender as the explanatory variable.

	Male	Female	Total
Very happy	7,609	7,942	15,551
Pretty happy	3,738	4,447	8,185
Not too happy	259	460	719
Total	11,606	12,849	24,455

Source: General Social Survey.

NW 13. Smoking Is Healthy? Could it be that smoking actually increases survival rates among women? The following data represent the 20-year survival status and smoking status of 1314 English women who participated in a cohort study from 1972 to 1992.

Smoking Status			
	Smoker (S)	Nonsmoker (NS)	Total
Dead	139	230	369
Alive	443	502	945
Total	582	732	1314

Source: David R. Appleton et al. "Ignoring a Covariate: An Example of Simpson's Paradox." *American Statistician* 50(4), 1996.

- (a) What proportion of the smokers was dead after 20 years? What proportion of the nonsmokers was dead after 20 years? What does this imply about the health consequences of smoking?

The data in the table above do not take into account a variable that is strongly related to survival status, age. The data shown next give the survival status of women and their age at the beginning of the study. For example,

14 women who were 35 to 44 at the beginning of the study were smokers and dead after 20 years.

Age Group								
	18–24	25–34	35–44	45–54	55–64	65–74	75 or older	
	S	NS	S	NS	S	NS	S	NS
Dead	2	1	3	5	14	7	27	12
Alive	53	61	121	152	95	114	103	66
					64	81	7	28
							0	0

- (b) Determine the proportion of 18- to 24-year-old smokers who were dead after 20 years. Determine the proportion of 18- to 24-year-old nonsmokers who were dead after 20 years.
- (c) Repeat part (b) for the remaining age groups to create a conditional distribution of survival status by smoking status for each age group.
- (d) Draw a bar graph of the conditional distribution from part (c).
- (e) Write a short report detailing your findings.

14. Death Sentence The following data represent the sentences imposed on offenders convicted of murder by race.

	Jail Time	Death Sentence	Total
Black Offender	2498	28	2526
White Offender	2323	49	2372
Total	4821	77	4898

Source: John Blume, Theodore Eisenberg, and Martin T. Wells. "Explaining Death Row's Population and Racial Composition," *Journal of Empirical Legal Studies*, 1(1), 165–207, March, 2004.

- (a) Which race appears to get a death sentence more frequently? Why?
- The data in the table above do not consider the race of the victim. The data below show the sentence of the offender by race of the victim.

	Black Victim		White Victim	
	Jail Time	Death Sentence	Jail Time	Death Sentence
Black Offender	2139	12	359	16
White Offender	100	0	2223	49

- (b) Determine the proportion of black offenders who were given a death sentence by race of the victim. Determine the proportion of white offenders who were given a death sentence by race of the victim.
- (c) Repeat part (b) for offenders given jail time for each race of the offender. Build a conditional distribution by race of the victim.
- (d) Draw a conditional bar graph of the conditional distribution from part (c).
- (e) Write a report detailing your findings.

15. Putting It Together: Sullivan Survey II Go to www.pearsonhighered.com/sullivanstats to obtain the data file *SullivanStatsSurveyII*.

- (a) Create a relative frequency distribution for political philosophy. What percent of the respondents are moderate?
- (b) Draw a relative frequency bar graph for political philosophy.

- (c) The column “GenderIncomeInequality” represents the response to the question, “Do you believe there is income inequality between males and females when each has the same experience and education?” Draw a pie chart for “GenderIncomeInequality.” What do the results suggest?
- (d) Is one’s gender associated with the response to the gender income inequality question? Build a contingency table for

these two variables treating gender as the row variable. Show the relative marginal distribution for both the column and row variables.

- (e) Construct a conditional distribution by gender. What do you notice?
- (f) Draw a bar graph of the conditional distribution from part (e).



Chapter 4 Review

Summary

This is the last chapter in Part II—Describe the Data. In Chapters 2 and 3, we described univariate data. That is, we summarized data where a single variable was measured on each individual. For example, we might consider the weight of a random sample of 30 students. In this chapter, we described bivariate data. That is, we measured two variables (such as height and weight) on each individual in the study. In particular, we looked at describing the relation between two quantitative variables (Sections 4.1–4.3) and between two qualitative variables (Section 4.4).

The first step in describing the relation between two quantitative variables is to draw a scatter diagram. The explanatory variable is plotted on the horizontal axis and the corresponding response variable on the vertical axis. The scatter diagram can be used to discover whether the relation between the explanatory and the response variables is linear. In addition, for linear relations, we can judge whether the linear relation shows positive or negative association.

A numerical measure for the strength of linear relation between two quantitative variables is the linear correlation coefficient. It is a number between -1 and 1 , inclusive. Values of the correlation coefficient near -1 are indicative of a negative linear relation between the two variables. Values of the correlation coefficient near $+1$ indicate a positive linear relation between the two variables. If the correlation coefficient is near 0 , then little *linear* relation exists between the two variables.

Be careful! Just because the correlation coefficient between two quantitative variables indicates that the variables are linearly related, it does not mean that

a change in one variable *causes* a change in a second variable. It could be that the correlation is the result of a lurking variable.

Once a linear relation between the two variables has been discovered, we describe the relation by finding the least-squares regression line. We can use the least-squares regression line to predict a value of the response variable for a given value of the explanatory variable.

The coefficient of determination, R^2 , measures the percent of variation in the response variable that is explained by the least-squares regression line. It is a measure between 0 and 1 , inclusive. The closer R^2 is to 1 , the more explanatory value the line has.

Section 4.4 introduced methods that allow us to describe any association that might exist between two qualitative variables. This is done through contingency tables. Both marginal and conditional distributions allow us to describe the effect one variable might have on the other variable in the study. We also construct bar graphs to see the association between the two variables in the study. Again, just because two qualitative variables are associated does not mean that a change in one variable *causes* a change in a second variable. We also looked at Simpson’s Paradox, which represents situations in which an association between two variables inverts or goes away when a third (lurking) variable is introduced into the analysis.

Vocabulary

- Univariate data (p. 170)
- Bivariate data (p. 170)
- Response variable (p. 171)
- Explanatory variable (p. 171)
- Predictor variable (p. 171)
- Scatter diagram (p. 171)
- Positively associated (p. 172)
- Negatively associated (p. 172)
- Linear correlation coefficient (p. 173)
- Lurking variable (p. 178)

- Residual (p. 189)
- Least-squares regression line (p. 189)
- Slope (p. 190)
- y-intercept (p. 190)
- Outside the scope of the model (p. 193)
- Coefficient of determination (p. 201)
- Deviation (p. 202)
- Total deviation (p. 202)
- Explained deviation (p. 202)

- Unexplained deviation (p. 202)
- Contingency (or two-way) table (p. 207)
- Row variable (p. 207)
- Column variable (p. 207)
- Cell (p. 207)
- Marginal distribution (p. 207)
- Conditional distribution (p. 209)
- Simpson’s Paradox (p. 213)

Formulas

Correlation Coefficient

$$r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)}{n - 1}$$

Equation of the Least-Squares Regression Line

$$\hat{y} = b_1 x + b_0$$

where

\hat{y} is the predicted value of the response variable

$b_1 = r \cdot \frac{s_y}{s_x}$ is the slope of the least-squares regression line

$b_0 = \bar{y} - b_1 \bar{x}$ is the y -intercept of the least-squares regression line

Coefficient of Determination, R^2

$$R^2 = \frac{\text{explained variation}}{\text{total variation}}$$

$$= 1 - \frac{\text{unexplained variation}}{\text{total variation}}$$

$= r^2$ for the least-squares regression model $\hat{y} = b_1 x + b_0$

Objectives

Section	You should be able to ...	Example(s)	Review Exercises
4.1	1 Draw and interpret scatter diagrams (p. 171) 2 Describe the properties of the linear correlation coefficient (p. 173) 3 Compute and interpret the linear correlation coefficient (p. 175) 4 Determine whether a linear relation exists between two variables (p. 177) 5 Explain the difference between correlation and causation (p. 178)	1, 3 page 174 2, 3 4 5	2(b), 3(a), 6(a) 13 2(c), 3(b) 2(d), 3(c) 9, 12
4.2	1 Find the least-squares regression line and use the line to make predictions (p. 188) 2 Interpret the slope and y -intercept of the least-squares regression line (p. 193) 3 Compute the sum of squared residuals (p. 194)	2, 3 page 193 4	1(a), 1(b), 4(a), 4(d), 5(a), 5(c), 6(d), 14(c) 1(c), 1(d), 4(c), 5(b), 14(b) 6(f), 6(g)
4.3	1 Compute and interpret the coefficient of determination (p. 201)	1	1(e), 7, 8
4.4	1 Compute the marginal distribution of a variable (p. 207) 2 Use the conditional distribution to identify association among categorical data (p. 208) 3 Explain Simpson's Paradox (p. 212)	1 and 2 3–5 6	10(b) 10(d), 10(e), 10(f) 11

Review Exercises

- 1. Basketball Spreads** In sports betting, Las Vegas sports books establish winning margins for a team that is favored to win a game. An individual can place a wager on the game and will win if the team bet upon wins after accounting for the spread. For example, if Team A is favored by 5 points and wins the game by 7 points, then a bet on Team A is a winning bet. However, if Team A wins the game by only 3 points, then a bet on Team A is a losing bet. For NCAA Division I basketball games, a least-squares regression with explanatory variable home team Las Vegas spread, x , and response variable home team winning margin, y , is $\hat{y} = 1.007x - 0.012$.

Source: Justin Wolfers. "Point Shaving: Corruption in NCAA Basketball."

- (a) Predict the winning margin if the home team is favored by 3 points.
- (b) Predict the winning margin (of the visiting team) if the visiting team is favored by 7 points (this is equivalent to the home team being favored by -7 points).
- (c) Interpret the slope.
- (d) Interpret the y -intercept.
- (e) The coefficient of determination is 0.39. Interpret this value.

-  **2. Fat and Calories in Cheeseburgers** A nutritionist was interested in developing a model that describes the relation between the amount of fat (in grams) in cheeseburgers at

fast-food restaurants and the number of calories. She obtains the following data from the websites of the companies.

Sandwich (Restaurant)	Fat Content (g)	Calories
Quarter-pound Single with Cheese (Wendy's)	20	430
Whataburger (Whataburger)	39	750
Cheeseburger (In-n-Out)	27	480
Big Mac (McDonald's)	29	540
Quarter-pounder with cheese (McDonald's)	26	510
Whopper with cheese (Burger King)	47	760
Jumbo Jack (Jack in the Box)	35	690
Double Steakburger with cheese (Steak 'n Shake)	38	632

Source: Each company's website.

- (a) The researcher wants to use fat content to predict calories. Which is the explanatory variable?
- (b) Draw a scatter diagram of the data.
- (c) Compute the linear correlation coefficient between fat content and calories.
- (d) Does a linear relation exist between fat content and calories in fast-food restaurant sandwiches?

DATA 3. Apartments The following data represent the square footage and rent for apartments in the borough of Queens and Nassau County, New York.

Queens (New York City)		Nassau County (Long Island)	
Square Footage, x	Rent per Month, y	Square Footage, x	Rent per Month, y
500	650	1100	1875
588	1215	588	1075
1000	2000	1250	1775
688	1655	556	1050
825	1250	825	1300
460	1805	743	1475
1259	2700	660	1315
650	1200	975	1400
560	1250	1429	1900
1073	2350	800	1650
1452	3300	1906	4625
1305	3100	1077	1395

Source: apartments.com

- (a) On the same graph, draw a scatter diagram for both Queens and Nassau County apartments treating square footage as the explanatory variable. Be sure to use a different plotting symbol for each group.
- (b) Compute the linear correlation coefficient between square footage and rent for each location.
- (c) Does a linear relation exist between the two variables for each location?
- (d) Does location appear to be a factor in rent?

4. Using the data and results from Problem 2, do the following:

- (a) Find the least-squares regression line treating fat content as the explanatory variable.
- (b) Draw the least-squares regression line on the scatter diagram.
- (c) Interpret the slope and y -intercept, if appropriate.
- (d) Predict the number of calories in a sandwich that has 30 grams of fat.

(e) A cheeseburger from Sonic has 700 calories and 42 grams of fat. Is the number of calories for this sandwich above or below average among all sandwiches with 42 grams?

5. Using the Queens data and results from Problem 3, do the following:

- (a) Find the least-squares regression line, treating square footage as the explanatory variable.
- (b) Interpret the slope and y -intercept, if appropriate.
- (c) Is the rent on the 825-square-foot apartment in the data above or below average among 825-square-foot apartments?

6.

x	10	14	17	18	21
y	105	94	82	76	63

(a) Draw a scatter diagram treating x as the explanatory variable and y as the response variable.

(b) Select two points from the scatter diagram, and find the equation of the line containing the points selected.

(c) Graph the line found in part (b) on the scatter diagram.

(d) Determine the least-squares regression line.

(e) Graph the least-squares regression line on the scatter diagram.

(f) Compute the sum of the squared residuals for the line found in part (b).

(g) Compute the sum of the squared residuals for the least-squares regression line found in part (d).

(h) Comment on the fit of the line found in part (b) versus the least-squares regression line found in part (d).

7. Use the results from Problems 2 and 4 to compute and interpret R^2 .

8. Use Queens data and the results from Problems 3 and 5 to compute and interpret R^2 .

9. **Shark Attacks** The correlation between the number of visitors to the state of Florida and the number of shark attacks since 1990 is 0.946. Should the number of visitors to Florida be reduced in an attempt to reduce shark attacks? Explain your reasoning.

Source: Florida Museum of Natural History.

10. **New versus Used Car Satisfaction** Are you more likely to be satisfied with your automobile purchase when it is new or used? The following data represent the level of satisfaction of the buyer for both new and used cars.

	New	Used	Total
Not too satisfied	11	25	36
Pretty satisfied	78	79	157
Extremely satisfied	118	85	203
Total	207	189	396

Source: General Social Survey.

(a) How many were extremely satisfied with their automobile purchase?

(b) Construct a relative frequency marginal distribution.

- (c) What proportion of consumers was extremely satisfied with their automobile purchase?
 (d) Construct a conditional distribution of satisfaction by purchase type (new or used).
 (e) Draw a bar graph of the conditional distribution found in part (d).
 (f) Do you think that the purchase type (new versus used) is associated with satisfaction?
-

11. Unemployment Rates Recessions are defined as two consecutive quarters of reduced national output. One measure to assess the severity of a recession is the rate of unemployment. The table shows the number of employed and unemployed residents of the United States at the peak of each recession (in thousands).

	Recession of 1982	Recession of 2009
Employed	98.9	130.1
Unemployed	11.3	14.5

Source: Bureau of Labor Statistics.

- (a) Determine the unemployment rate for each recession. Which recession appears worse as measured by unemployment rate?

Note: Unemployment rate = unemployed/(employed + unemployed).

The data in the table above do not account for level of education. The following data show the unemployment rate by level of education for each recession.

	Recession of 1982			Recession of 2009				
	Less than high school	High school	Bachelor's degree or higher	Less than high school	High school	Bachelor's degree or higher		
	Employed	20.3	58.2	20.4	Unemployed	10.0	76.7	43.4
	Employed	3.9	6.6	0.8	Unemployed	2.0	10.3	2.2

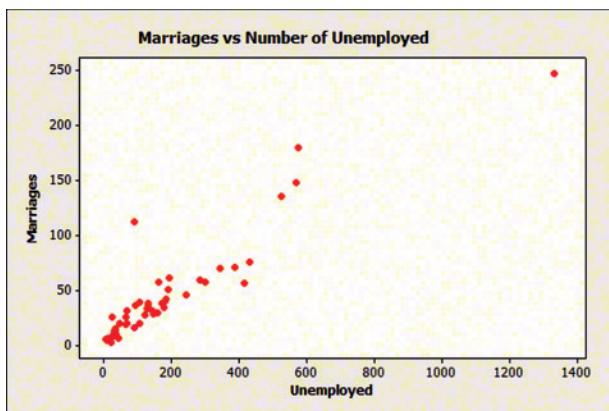
Source: Bureau of Labor Statistics.

- (b) Determine the unemployment rate for each level of education for both recessions.

- (c) Draw a bar graph of the conditional distribution from part (b).

- (d) Write a report that suggests the recession of 2009 is worse than that of 1982.

12. (a) The correlation between number of married residents and number of unemployed residents for the 50 states and Washington, DC, is 0.922. A scatter diagram of the data is shown. What type of relation appears to exist between number of marriages and number unemployed?



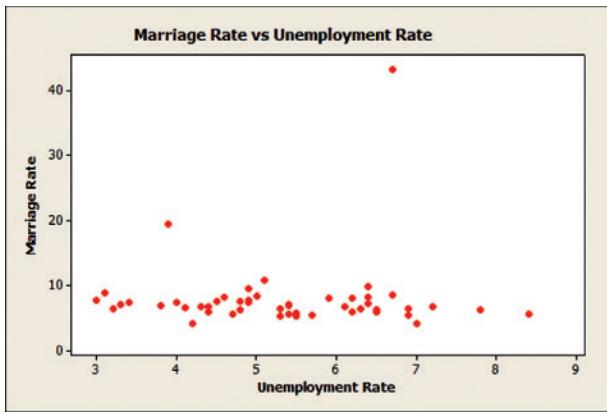
- (b)** Use the following correlation matrix to explain how population may be a lurking variable in the relation presented in part (a).

Correlations: Unemployed, Population, Marriages

	Unemployed	Population
Population	0.979	
Marriages	0.922	0.947

Cell Contents: Pearson correlation

- (c) The correlation between unemployment rate (number unemployed divided by population size) and marriage rate (number married divided by population size) for the 50 states and Washington, DC, is 0.050. A scatter diagram between unemployment rate and marriage rate is shown next. What type of relation appears to exist between unemployment rate and marriage rate?

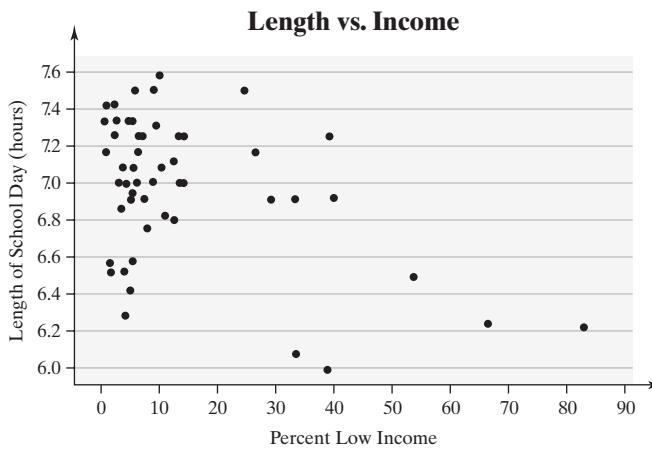


- (d) Write a few sentences to explain the danger in using correlation to conclude that a relation exists between two variables without considering lurking variables.

13. List the eight properties of the linear correlation coefficient.

14. Analyzing a Newspaper Article In a newspaper article written in the *Chicago Tribune*, it was claimed that poorer school districts have shorter school days.

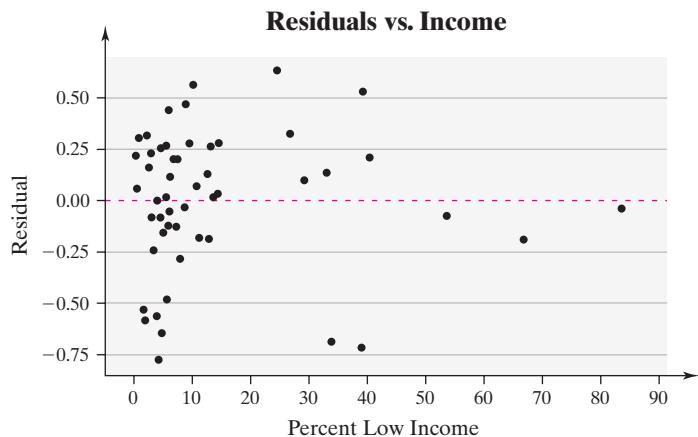
- (a) The following scatter diagram was drawn using the data supplied in the article. In this scatter diagram, the response variable is length of the school day (in hours) and the explanatory variable is percent of the population that is low income. The correlation between length and income is -0.461 . Do you think that the scatter diagram and correlation coefficient support the position of the article?



- (b) The least-squares regression line between length, y , and income, x , is $\hat{y} = -0.0102x + 7.11$. Interpret the slope of this regression line. Does it make sense to interpret the y -intercept? If so, interpret the y -intercept.

- (c) Predict the length of the school day for a district in which 20% of the population is low income by letting $x = 20$.

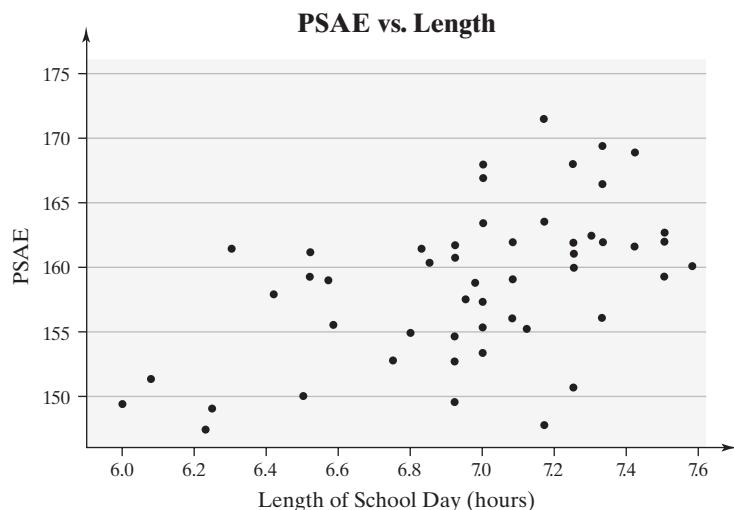
- (d) Based on the following residual plot, do you think a linear model is appropriate for describing the relation between length of school day and income? Why?



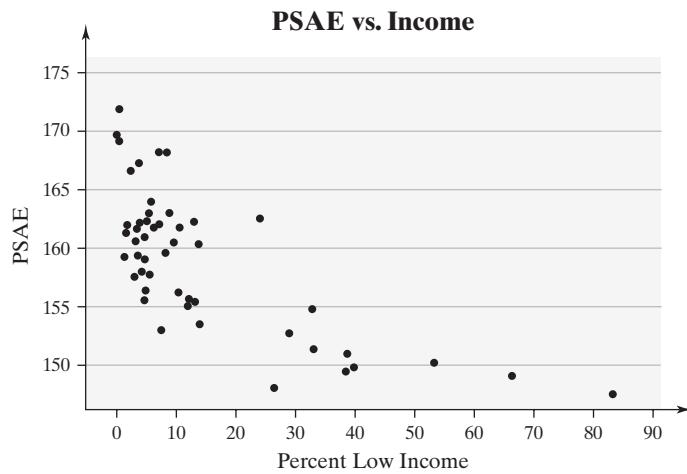
- (e) Three observations may be influential. Based on the scatter diagram in part (a), which three observations do you think might be influential?

- (f) We should not remove influential observations just because they are influential. What suggestions would you make to the author of the article to improve the study and therefore give more credibility to the conclusions?

- (g) This same article included average Prairie State Achievement Examination (PSAE) scores for each district. The article implied that shorter school days result in lower PSAE scores. The correlation between PSAE score and length of school day is 0.517. A scatter diagram treating PSAE as the response variable is shown next. Do you believe that a longer school day is positively associated with a higher PSAE score?



- (h) The correlation between percentage of the population that is low income and PSAE score is -0.720 . A scatter diagram treating PSAE score as the response variable is shown below. Do you believe that percentage of the population that is low income is negatively associated with PSAE score?



- (i) Can you think of any lurking variables that are playing a role in this study?
15. In studies of monozygotic (identical) twins, the correlation of intelligence (IQ) scores is 0.85.
- (a) What or who are the individuals in this scenario?
- (b) What are the variables?
- (c) What proportion of the variability in one twin's IQ is explained by the other twin's IQ?



Chapter Test

- DATA** 1. Crickets make a chirping noise by sliding their wings rapidly over each other. Perhaps you have noticed that the number of chirps seems to increase with the temperature. The following data list the temperature (in degrees Fahrenheit) and the number of chirps per second for the striped ground cricket.

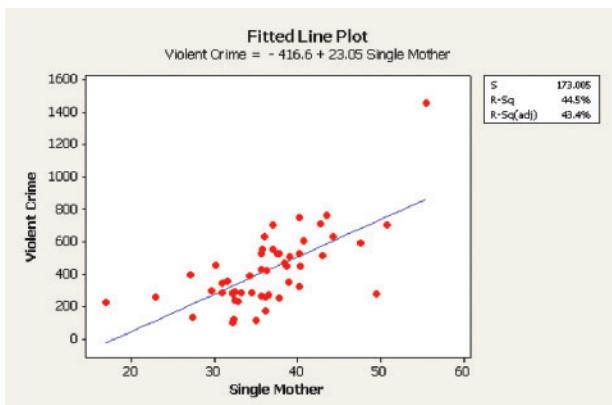
Temperature, x	Chirps per Second, y	Temperature, x	Chirps per Second, y
88.6	20.0	71.6	16.0
93.3	19.8	84.3	18.4
80.6	17.1	75.2	15.5
69.7	14.7	82.0	17.1
69.4	15.4	83.3	16.2
79.6	15.0	82.6	17.2
80.6	16.0	83.5	17.0
76.3	14.4		

Source: George W. Pierce. *The Songs of Insects*. Cambridge, MA: Harvard University Press, 1949, pp. 12–21.

- (a) What is the likely explanatory variable in these data? Why?
- (b) Draw a scatter diagram of the data.
- (c) Compute the linear correlation coefficient between temperature and chirps per second.

- (d) Does a linear relation exist between temperature and chirps per second?
2. Use the data from Problem 1.
- (a) Find the least-squares regression line treating temperature as the explanatory variable and chirps per second as the response variable.
- (b) Interpret the slope and y -intercept, if appropriate.
- (c) Predict the chirps per second if it is 83.3°F .
- (d) A cricket chirps 15 times per second when the temperature is 82°F . Is this rate of chirping above or below average at this temperature?
- (e) It is 55°F outside. Would you recommend using the linear model found in part (a) to predict the number of chirps per second of a cricket? Why or why not?
3. Use the results from Problems 1 and 2 to compute and interpret R^2 .
4. A researcher collects data regarding the percent of all births to unmarried women and the number of violent crimes for the 50 states and Washington, DC. The scatter diagram along with the least-squares regression line obtained from Minitab is shown on the next page. The correlation between percent of births to unmarried women and the number of violent crimes is 0.667 . A politician looks over the data and claims that, for each 1% decrease in births to single mothers, violent crimes will decrease by 23. Therefore, percent of births to single mothers needs

to be reduced in an effort to decrease violent crimes. Is there anything wrong with the reasoning of the politician? Explain.



5. What is the relationship between education and belief in Heaven? The following data represent the highest level of education and belief in Heaven for a random sample of adult Americans.

	Yes, Definitely	Yes, Probably	No, Probably Not	No, Definitely Not	Total
Less than high school	316	66	21	9	412
High school	956	296	122	65	1439
Bachelor's	267	131	62	64	524
Total	1539	493	205	138	2375

Source: General Social Survey.

- (a) Construct a relative frequency marginal distribution.
- (b) What proportion of adult Americans in the survey definitely believe in Heaven?
- (c) Construct a conditional distribution of belief in Heaven by level of education.
- (d) Draw a bar graph of the conditional distribution found in part (c).
- (e) Is there an association between level of education and belief in Heaven?

6. Consider the following contingency table, which relates the number of applicants accepted to a college and gender.

	Accepted	Denied
Male	98	522
Female	90	200

- (a) Construct a conditional distribution of acceptance status by gender.
 - (b) What proportion of males was accepted? What proportion of females was accepted?
 - (c) What might you conclude about the admittance policies of the school?
- A lurking variable is the type of school applied to. This particular college has two programs of study: business and social work. The following table shows applications by type of school.

	Business School		Social Work School	
	Accepted	Denied	Accepted	Denied
Male	90	510	8	12
Female	10	60	80	140

- (d) What proportion of males who applied to the business school was accepted? What proportion of females who applied to the business school was accepted?
- (e) What proportion of males who applied to the social work school was accepted? What proportion of females who applied to the social work school was accepted?
- (f) Explain carefully how the bias disappears when type of school is considered.
- 7. If the slope of a least-squares regression line is negative, what could be said about the correlation between the explanatory and response variable?
- 8. What does it mean if a linear correlation coefficient is close to zero? Draw two scatter diagrams for which the linear correlation coefficient is close to zero.
- 9. What would you say about a set of quantitative bivariate data whose linear correlation coefficient is -1 ? What would a scatter diagram of the data look like?

Making an Informed Decision

Relationships among Variables on a World Scale

The purpose of this Decisions Project is to decide on two quantitative variables, see how the relationship between these variables has changed over time, and determine the most current relationship between the variables.

1. Watch the video by Hans Rosling titled “Let My Dataset Change Your Mindset” at http://www.ted.com/talks/lang/eng/hans_rosling_at_state.html.
 - (a) Describe the “We” versus “Them” discussion in the video.
 - (b) In the video, there is a scatter diagram drawn between life expectancy and children per woman. Which variable is the explanatory variable in the scatter diagram? What type of relation exists between the two variables? How has the relationship changed over time when considering “We” versus “Them”?
2. Click the Data tab on the GapMinder website. Download the data for life expectancy and children per woman. Draw a scatter diagram of the data. Determine the linear correlation coefficient between the two variables. Be careful; some countries may not have both variables measured, so they will need to be excluded from the analysis.
3. Find the least-squares regression line between life expectancy and children per woman. Interpret the slope of the regression line.
4. Is the life expectancy of the United States above or below average given the number of children per woman?
5. Choose any two variables in the GapMinder library and write a report detailing the relationship between the variables over time. Does the data reflect your intuition about the relationship?



PART



Probability and Probability Distributions

We now take a break from the statistical process. Why? In Chapter 1, we mentioned that inferential statistics uses methods that generalize results obtained from a sample to the population and measures their reliability. But how can we measure their reliability? It turns out that the methods we use to generalize results from a sample to a population are based on probability and probability models. Probability is a measure of the likelihood that something occurs. This part of the course will focus on methods for determining probabilities.

CHAPTER 5 Probability

CHAPTER 6 Discrete Probability Distributions

CHAPTER 7 The Normal Probability Distribution



Probability

Outline

- 5.1** Probability Rules
- 5.2** The Addition Rule and Complements
- 5.3** Independence and the Multiplication Rule
- 5.4** Conditional Probability and the General Multiplication Rule
- 5.5** Counting Techniques
- 5.6** Simulating Probability Experiments
- 5.7** Putting It Together: Which Method Do I Use?

Making an Informed Decision



You are at a party and everyone is having a good time. Unfortunately, a few of the party-goers are having a little too much to drink. If they decide to drive home, what are the risks? Perhaps a view of some scary statistics about the effects of alcohol on one's driving skills will convince more people not to drink and drive. See the Decisions project on page 298.

Putting It Together

In Chapter 1, we learned methods for collecting data. In Chapters 2 through 4, we learned how to summarize raw data using tables, graphs, and numbers. As far as the statistical process goes, we have discussed the collecting, organizing, and summarizing parts of the process.

Before we begin to analyze data, we introduce probability, which forms the basis of inferential statistics. Why? Well, we can think of the probability of an outcome as the likelihood of observing that outcome. If something has a high likelihood of happening, it has a high probability (close to 1). If something has a small chance of happening, it has a low probability (close to 0). For example, it is unlikely that we would roll five straight sixes when rolling a single die, so this result has a low probability. In fact, the probability of rolling five straight sixes is 0.0001286. If we were playing a game in which a player threw five sixes in a row with a single die, we would consider the player to be lucky (or a cheater) because it is such an unusual occurrence. Statisticians use probability in the same way. If something occurs that has a low probability, we investigate to find out "what's up."

5.1 Probability Rules



Preparing for This Section Before getting started, review the following:

- Relative frequency (Section 2.1, p. 64)

Objectives

- ① Understand random processes and the Law of Large Numbers
- ② Apply the rules of probabilities
- ③ Compute and interpret probabilities using the empirical method
- ④ Compute and interpret probabilities using the classical method
- ⑤ Recognize and interpret subjective probabilities

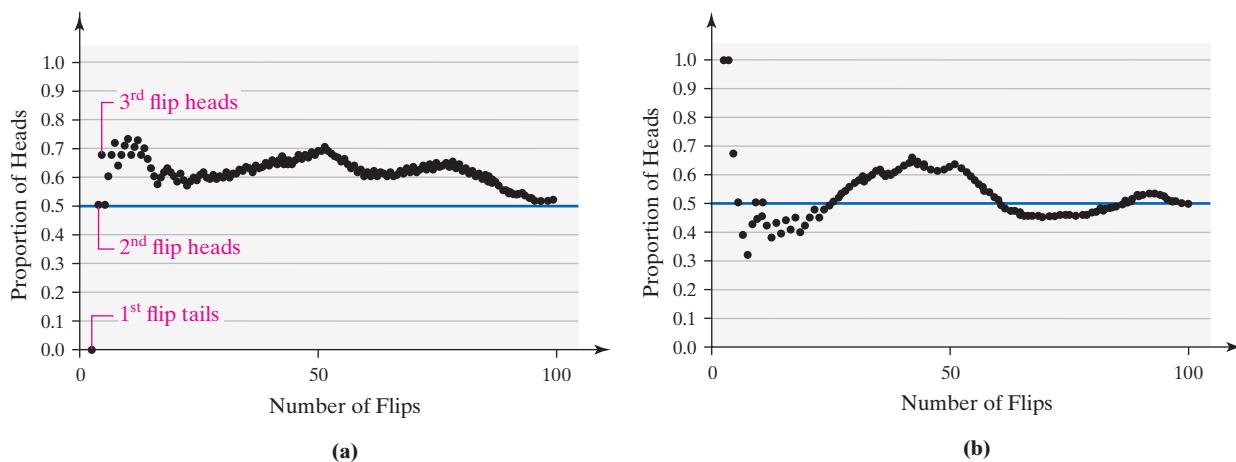
① Understand Random Processes and the Law of Large Numbers



The word *random* suggests an unpredictable result (or *outcome*). Predicting outcomes while facing uncertainty is challenging. For example, it would be difficult to predict whether the outcome of flipping a fair coin would be heads or tails for one particular flip. However, if a coin is flipped many times, we may be able to determine the long-run proportion of times a head is observed. The process of flipping a coin many times is a *simulation*. **Simulation** is a technique used to re-create a random event. Simulations can be tactile (as in physically flipping a coin) or virtual (using a computer to pretend it is flipping a coin). In both instances, the goal of the simulation is to measure how often a certain outcome (such as a head in coin flipping) is observed.

To see this idea, suppose we flip a coin 100 times and compute the proportion of heads observed after each toss of the coin. The first flip is tails, so the proportion of heads is $\frac{0}{1} = 0$; the second flip is heads, so the proportion of heads is $\frac{1}{2} = 0.5$; the third flip is heads, so the proportion of heads is $\frac{2}{3} = 0.667$, and so on. Plot the proportion of heads versus the number of flips and obtain the graph in Figure 1(a). We repeat this experiment with the results shown in Figure 1(b).

Figure 1



The graphs in Figures 1(a) and 1(b) illustrate a *random process*.

Definition

A **random process** represents scenarios where the outcome of any particular trial of an experiment is unknown, but the proportion (or relative frequency) a particular outcome is observed approaches a specific value as the number of trials increases.

IN OTHER WORDS

Probability describes how likely it is that some event will happen. If we look at the proportion of times an event has occurred over a long period of time (or over a large number of trials), we can be more certain of the likelihood of its occurrence.

CAUTION!

Probability is a measure of the likelihood of events that have yet to occur. Prior to flipping a coin, we say the probability of observing a head is $\frac{1}{2}$. However, once the coin is flipped, the probability is no longer $\frac{1}{2}$ since the outcome has been determined.

The graphs in Figures 1(a) and (b) show that in the short term (fewer flips) the observed proportion of heads is different and unpredictable for each experiment. As the number of flips increases, however, both graphs tend toward a proportion of 0.5. This is the basic premise of probability. **Probability** is a measure of the likelihood of a random phenomenon or chance behavior occurring. Probability deals with experiments that yield random short-term results or **outcomes** yet reveal long-term predictability. **The long-term proportion in which a certain outcome is observed is the probability of that outcome.** So we say that the probability of observing a head is $\frac{1}{2}$ or 50% or 0.5 because, as we flip the coin more times, the proportion of heads tends toward $\frac{1}{2}$. This phenomenon, which is illustrated in Figure 1, is referred to as the *Law of Large Numbers*.

The Law of Large Numbers

As the number of repetitions of a probability experiment increases, the proportion with which a certain outcome is observed gets closer to the probability of the outcome.

Jakob Bernoulli (a major contributor to the field of probability) believed that the Law of Large Numbers was common sense. This is evident in the following quote from his text *Ars Conjectandi*: “For even the most stupid of men, by some instinct of nature, by himself and without any instruction, is convinced that the more observations have been made, the less danger there is of wandering from one’s goal.”

The Law of Averages?

The Law of Large Numbers is often interpreted as a nonexistent law called the *Law of Averages* by folks who misunderstand the Law of Large Numbers. For example, in baseball you may hear an announcer say that a certain player is *due* to get a hit because he has gone a number of at-bats without getting a hit. Or, you may hear a mother of four boys say she is more likely to have a girl now. In both instances, the confusion is between what happens in the long run (the Law of Large Numbers) and what might happen on the next trial of a probability experiment. That is, given that you had four boys are you due for a girl with your fifth child? No, the likelihood of a girl is the same on the fifth child as it was for the first child. Think of it this way: the biology of determining the gender of a child does not look back at the first four trials (first four children) and say, “okay, it’s time for a girl.” Another way to think about this is that the trials are “memoryless.” That is, the trials don’t recall what has happened in the past—they only consider the next trial.

To see this, let’s conduct a random process 30,000 times in which five children are born. In this random process, we assume the likelihood of having a boy is the same as having a girl. In Figure 2, each row represents a family of five, where 0 is a girl and 1 is a boy. The column “Number of Boys” represents the number of boys the family had out of the first four children. For example, in Row 127, the first child is a boy (1), the second child is a girl (0), the third child is a boy (1), and the fourth child is a boy (1). Therefore, the Number of Boys equals 3 for the family in Row 127. For the families in Rows 130 and 138 (highlighted), notice that the first four children were all boys. For the family in Row 130, the fifth child is a girl (0), while for the family in row 138, the fifth child is a boy (1). In the 1921 times that the first four children were all boys, 940 resulted in the fifth child being a girl (48.9%) and 981 resulted in the fifth child being a boy (51.1%). So, just because the first four children are boys does not mean the family is “due” to have a girl should they choose to have a fifth child. This illustrates that the “Law of Averages” is not a law at all.

Figure 2

Row	First Child	Second Child	Third Child	Fourth Child	Fifth Child	Number of Boys	var7	var8	var9	var10	var11	var12
127	1	0	1	1	0	3						
128	1	0	0	0	1	1						
129	0	0	0	1	0	1						
130	1	1	1	1	0	4						
131	0	1	0	0	1	1						
132	0	0	1	1	0	2						
133	1	0	1	0	1	2						
134	0	1	1	1	1	3						
135	0	0	0	0	1	0						
136	1	1	0	0	0	2						
137	0	1	0	0	1	1						
138	1	1	1	1	1	4						
139	0	1	0	0	0	1						
140	n	n	1	n	1	1						

Frequency table results for Fifth Child:
Where: "Number of Boys"=4
Count = 1921

Fifth Child	Frequency	Relative Frequency
0	940	0.48932847
1	981	0.51067153

In probability, an **experiment** is any process with uncertain results that can be repeated. The result of any single trial of the experiment is not known ahead of time. However, the results of the experiment over many trials produce regular patterns that allow accurate predictions. For example, an insurance company cannot know whether a particular 16-year-old driver will have an accident over the course of a year. However, based on historical records, the company can be fairly certain that about three out of every ten 16-year-old male drivers will have a traffic accident during the course of a year. Therefore, of 825,000 male 16-year-old drivers (825,000 repetitions of the experiment), the insurance company is fairly confident that about 30%, or 247,500, will have an accident. This prediction helps to establish insurance rates for any particular 16-year-old male driver.

We now introduce some terminology that we will need to study probability.

Definitions

IN OTHER WORDS

An outcome is the result of one trial of a probability experiment. The sample space is a list of all possible results of a probability experiment.

The **sample space**, S , of a probability experiment is the collection of all possible outcomes.

An **event** is any collection of outcomes from a probability experiment. An event consists of one outcome or more than one outcome. We will denote events with one outcome, sometimes called *simple events*, e_i . In general, events are denoted using capital letters such as E .

EXAMPLE 1

Identifying Events and the Sample Space of a Probability Experiment



A **fair die** is one in which each possible outcome is equally likely. For example, rolling a 2 is just as likely as rolling a 5. We contrast this with a **loaded die**, in which a certain outcome is more likely. For example, if rolling a 1 is more likely than rolling a 2, 3, 4, 5, or 6, the die is loaded.

NW Now Work Problem 15

Problem A probability experiment consists of rolling a single *fair* die.

- (a) Identify the outcomes of the probability experiment.
- (b) Determine the sample space.
- (c) Define the event E = “roll an even number.”

Approach The outcomes are the possible results of the experiment. The sample space is a list of all possible outcomes.

Solution

- (a) The outcomes from rolling a single fair die are e_1 = “rolling a one” = {1}, e_2 = “rolling a two” = {2}, e_3 = “rolling a three” = {3}, e_4 = “rolling a four” = {4}, e_5 = “rolling a five” = {5}, and e_6 = “rolling a six” = {6}.
- (b) The set of all possible outcomes forms the sample space, S = {1, 2, 3, 4, 5, 6}. There are 6 outcomes in the sample space.
- (c) The event E = “roll an even number” = {2, 4, 6}.



2 Apply the Rules of Probabilities

In the following probability rules, the notation $P(E)$ means “the probability that event E occurs.”

IN OTHER WORDS

Rule 1 states that probabilities less than 0 or greater than 1 are not possible. Therefore, probabilities such as -0.3 or 1.32 are not possible. Rule 2 states when the probabilities of all outcomes are added, the sum must be 1.

Rules of Probabilities

1. The probability of any event E , $P(E)$, must be greater than or equal to 0 and less than or equal to 1. That is, $0 \leq P(E) \leq 1$.
2. The sum of the probabilities of all outcomes must equal 1. That is, if the sample space $S = \{e_1, e_2, \dots, e_n\}$, then

$$P(e_1) + P(e_2) + \dots + P(e_n) = 1$$

A **probability model** lists the possible outcomes of a probability experiment and each outcome's probability. A probability model must satisfy Rules 1 and 2 of the rules of probabilities.

EXAMPLE 2

A Probability Model

Table 1

Color	Probability
Brown	0.13
Yellow	0.14
Red	0.13
Blue	0.24
Orange	0.20
Green	0.16

Source: M&Ms.

NW Now Work Problem 3

The color of a plain M&M milk chocolate candy can be brown, yellow, red, blue, orange, or green. Suppose a candy is randomly selected from a bag. Table 1 shows each color and the probability of drawing that color.

To verify that this is a probability model, we must show that Rules 1 and 2 are satisfied.

Each probability is greater than or equal to 0 and less than or equal to 1, so Rule 1 is satisfied.

Because

$$0.13 + 0.14 + 0.13 + 0.24 + 0.20 + 0.16 = 1$$

Rule 2 is also satisfied. The table is an example of a probability model.



If an event is **impossible**, the probability of the event is 0. If an event is a **certainty**, the probability of the event is 1.

The closer a probability is to 1, the more likely the event will occur. The closer a probability is to 0, the less likely the event will occur. For example, an event with probability 0.8 is more likely to occur than an event with probability 0.75. An event with probability 0.8 will occur about 80 times out of 100 repetitions of the experiment, whereas an event with probability 0.75 will occur about 75 times out of 100.

Be careful of this interpretation. An event with a probability of 0.75 does not have to occur 75 times out of 100. Rather, we *expect* the number of occurrences to be close to 75 in 100 trials. The more repetitions of the probability experiment, the closer the proportion with which the event occurs will be to 0.75 (the Law of Large Numbers).

One goal of this course is to learn how probabilities can be used to identify *unusual events*.

Definition

An **unusual event** is an event that has a low probability of occurring.

IN OTHER WORDS

An unusual event is an event that is not likely to occur.

Typically, an event with a probability less than 0.05 (or 5%) is considered unusual, but this *cutoff point* is not set in stone. The researcher and the context of the problem determine the probability that separates unusual events from *not so unusual events*.

For example, suppose that the probability of being wrongly convicted of a capital crime punishable by death is 3%. Even though 3% is below our 5% cutoff point, this

probability is too high in light of the consequences (death for the wrongly convicted), so the event is not unusual (unlikely) enough. We would want this probability to be much closer to zero.

Now suppose that you are planning a picnic on a day having a 3% chance of rain. In this context, you would consider “rain” an unusual (unlikely) event and proceed with the picnic plans.

CAUTION!

A probability of 0.05 should not always be used to separate unusual events from not so unusual events.

The point is this: Selecting a probability that separates unusual events from not so unusual events is subjective and depends on the situation. Statisticians typically use cutoff points of 0.01, 0.05, and 0.10.

Next, we introduce three methods for determining the probability of an event: the empirical method, the classical method, and the subjective method.

③ Compute and Interpret Probabilities Using the Empirical Method

Probabilities deal with the likelihood that a particular outcome will be observed. For this reason, we begin our discussion of determining probabilities using the idea of relative frequency. Probabilities computed in this manner rely on empirical evidence, that is, evidence based on the outcomes of a probability experiment.

Approximating Probabilities Using the Empirical Approach

The probability of an event E is approximately the number of times event E is observed divided by the number of repetitions of the experiment.

$$P(E) \approx \text{relative frequency of } E = \frac{\text{frequency of } E}{\text{number of trials of experiment}} \quad (1)$$

When we find probabilities using the empirical approach, the result is approximate because different trials of the experiment lead to different outcomes and, therefore, different estimates of $P(E)$. Consider flipping a coin 20 times and recording the number of heads. Use the results of the experiment to estimate the probability of obtaining a head. Now repeat the experiment. Because the results of the second run of the experiment do not necessarily yield the same results, we cannot say the probability *equals* the relative frequency; rather we say the probability is *approximately* the relative frequency. As we increase the number of trials of a probability experiment, our estimate becomes more accurate (again, the Law of Large Numbers).

EXAMPLE 3 Using Relative Frequencies to Approximate Probabilities

An insurance agent currently insures 182 teenage drivers (ages 16 to 19). Last year, 24 of the teenagers had to file a claim on their auto policy. Based on these results, the probability that a teenager will file a claim on his or her auto policy in a given year is

$$\frac{24}{182} \approx 0.132$$

So, for every 100 insured teenage drivers, we expect about 13 to have a claim on their auto policy.

NOTE

To interpret probabilities using relative frequencies, convert the decimal to a fraction where the denominator is a multiple of 10 (such as 100 or 1000). For example, $0.132 \approx 0.13$ and $0.13 = \frac{13}{100}$.

Surveys are probability experiments. Why? Each time a survey is conducted, a different random sample of individuals is selected. Therefore, the results of a survey are likely to be different each time the survey is conducted because different people are included.

EXAMPLE 4**Building a Probability Model from Survey Data****Table 2**

Means of Travel	Frequency
Drive alone	153
Carpool	22
Public transportation	10
Walk	5
Other means	3
Work at home	7

Table 3

Means of Travel	Probability
Drive alone	0.765
Carpool	0.11
Public transportation	0.05
Walk	0.025
Other means	0.015
Work at home	0.035

NW Now Work Problem 31

Problem The data in Table 2 represent the results of a survey in which 200 people were asked their means of travel to work.

- (a) Use the survey data to build a probability model for means of travel to work.
- (b) Estimate the probability that a randomly selected individual car pools to work. Interpret this result.
- (c) Would it be unusual to randomly select an individual who walks to work?

Approach To build a probability model, estimate the probability of each outcome by determining its relative frequency.

Solution

- (a) There are $153 + 22 + \dots + 7 = 200$ individuals in the survey. The individuals can be thought of as trials of the probability experiment. The relative frequency for “drive alone” is $\frac{153}{200} = 0.765$. We compute the relative frequency of the other outcomes similarly and obtain the probability model in Table 3.
- (b) From Table 3, we estimate the probability to be 0.11 that a randomly selected individual car pools to work. Because $0.11 = \frac{11}{100}$, we interpret this result by saying, “If we were to survey 100 individuals, we would expect about 11 to car pool to work.”
- (c) The probability that an individual walks to work is approximately 0.025. This means if we survey 1000 individuals, we would expect about 25 to walk to work (because $0.025 = \frac{25}{1000}$). Therefore, it is unusual to randomly choose a person who walks to work.

4**Compute and Interpret Probabilities Using the Classical Method**

The empirical method gives an approximate probability of an event by conducting a probability experiment.

The classical method of computing probabilities does not require that a probability experiment actually be performed. Rather, it relies on counting techniques.

The classical method of computing probabilities requires *equally likely outcomes*. An experiment has **equally likely outcomes** when each outcome has the same probability of occurring. For example, when a fair die is thrown once, each of the six outcomes in the sample space, $\{1, 2, 3, 4, 5, 6\}$, has an equal chance of occurring. Contrast this situation with a loaded die in which a five or six is twice as likely to occur as a one, two, three, or four.

Computing Probability Using the Classical Method

If an experiment has n equally likely outcomes and if the number of ways that an event E can occur is m , then the probability of E , $P(E)$, is

$$P(E) = \frac{\text{number of ways that } E \text{ can occur}}{\text{number of possible outcomes}} = \frac{m}{n} \quad (2)$$

So, if S is the sample space of this experiment,

$$P(E) = \frac{N(E)}{N(S)} \quad (3)$$

where $N(E)$ is the number of outcomes in E , and $N(S)$ is the number of outcomes in the sample space.

EXAMPLE 5**Computing Probabilities Using the Classical Method**

Problem A pair of fair dice is rolled. Fair die are die where each outcome is equally likely.

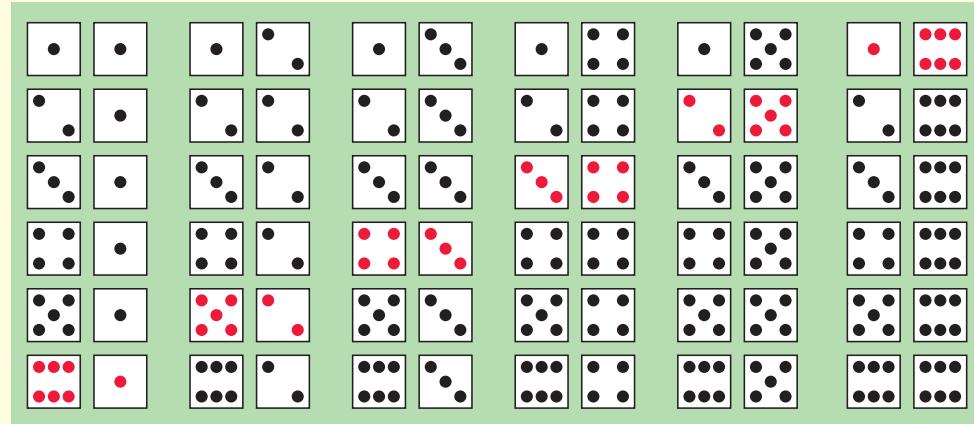
- Compute the probability of rolling a seven.
- Compute the probability of rolling “snake eyes”; that is, compute the probability of rolling a two.
- Comment on the likelihood of rolling a seven versus rolling a two.

Approach To compute probabilities using the classical method, count the number of outcomes in the sample space and count the number of ways the event can occur. Then, divide the number of outcomes by the number of ways the event can occur.

Solution

- There are 36 equally likely outcomes in the sample space, as shown in Figure 3.

Figure 3



So, $N(S) = 36$. The event $E = \text{“roll a seven”} = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$ has six outcomes, so $N(E) = 6$. Using Formula (3),

$$P(E) = P(\text{roll a seven}) = \frac{N(E)}{N(S)} = \frac{6}{36} = \frac{1}{6}$$

- The event $F = \text{“roll a two”} = \{(1, 1)\}$ has one outcome, so $N(F) = 1$.

$$P(F) = P(\text{roll a two}) = \frac{N(F)}{N(S)} = \frac{1}{36} \quad \text{Use Formula (3).}$$

- Since $P(\text{roll a seven}) = \frac{1}{6}$ and $P(\text{roll a two}) = \frac{1}{36}$, rolling a seven is six times as likely as rolling a two. In other words, in 36 rolls of the dice, we *expect* to observe about 6 sevens and only 1 two.

Historical Note

Girolamo Cardano (in English Jerome Cardan) was born in Pavia, Italy, on September 24, 1501. He was an illegitimate child whose father was Fazio Cardano, a lawyer in Milan. Fazio was a part-time mathematician and taught Girolamo. In 1526, Cardano earned his medical degree. Shortly thereafter, his father died. Unable to maintain a medical practice, Cardano spent his inheritance and turned to gambling to help support himself. Cardano developed an understanding of probability that helped him to win. Eventually, Cardano became a lecturer of mathematics at the Piatti Foundation. This position allowed him to practice medicine and develop a favorable reputation as a doctor. In 1545, he published his greatest work, *Ars Magna*. His booklet on probability, *Liber de Ludo Alaeæ*, was not printed until 1663, 87 years after his death. The booklet is a practical guide to gambling, including cards, dice, and cheating.

We just saw that the classical probability of rolling a seven is $\frac{1}{6} \approx 0.167$. Suppose a pit boss at a casino rolls a pair of dice 100 times and obtains 15 sevens. From this empirical evidence, we would assign the probability of rolling a seven as $\frac{15}{100} = 0.15$. If the dice are fair, we would expect the relative frequency of sevens to get closer to 0.167 as the number of rolls of the dice increases. In other words, the empirical probability will get closer to the classical probability as the number of trials of the experiment increases due to the Law of Large Numbers. If the two probabilities do not get closer together, we may suspect that the dice are not fair.

In simple random sampling, each individual has the same chance of being selected. Therefore, we can use the classical method to compute the probability of obtaining a specific sample.

EXAMPLE 6**Computing Probabilities Using Equally Likely Outcomes**

Problem Sophia has three tickets to a concert, but Yolanda, Michael, Kevin, and Marissa all want to go to the concert with her. To be fair, Sophia randomly selects the two people who can go with her.

- Determine the sample space of the experiment. In other words, list all possible simple random samples of size $n = 2$.
- Compute the probability of the event “Michael and Kevin attend the concert.”
- Compute the probability of the event “Marissa attends the concert.”
- Interpret the probability in part (c).

Approach First, determine the outcomes in the sample space by making a table. The probability of an event is the number of outcomes in the event divided by the number of outcomes in the sample space.

Solution

- (a) The sample space is listed in Table 4.
- (b) We have $N(S) = 6$, and there is one way the event “Michael and Kevin attend the concert” can occur. Therefore, the probability that Michael and Kevin attend the concert is $\frac{1}{6}$.
- (c) We have $N(S) = 6$, and there are three ways the event “Marissa attends the concert” can occur. The probability that Marissa will attend is $\frac{3}{6} = 0.5 = 50\%$.
- (d) If we conducted this experiment 1000 times, about 500 of the experiments would result in Marissa attending the concert.

Table 4

Yolanda, Michael	Yolanda, Kevin
Yolanda, Marissa	Michael, Kevin
Michael, Marissa	Kevin, Marissa

NW Now Work Problems 25 and 43

**EXAMPLE 7****Comparing the Classical Method and Empirical Method**

Problem Suppose that a survey asked 500 families with three children to disclose the gender of their children and found that 180 of the families had two boys and one girl.

- Estimate the probability of having two boys and one girl in a three-child family using the empirical method.
- Compute and interpret the probability of having two boys and one girl in a three-child family using the classical method, assuming boys and girls are equally likely.

Approach To answer part (a), determine the relative frequency of the event “two boys and one girl.” To answer part (b), count the number of ways the event “two boys and one girl” can occur and divide this by the number of possible outcomes for this experiment.

Solution

- (a) The empirical probability of the event $E = \text{“two boys and one girl”}$ is

$$P(E) \approx \text{relative frequency of } E = \frac{180}{500} = 0.36$$

The probability that a family of three children will have two boys and one girl is approximately 0.36.

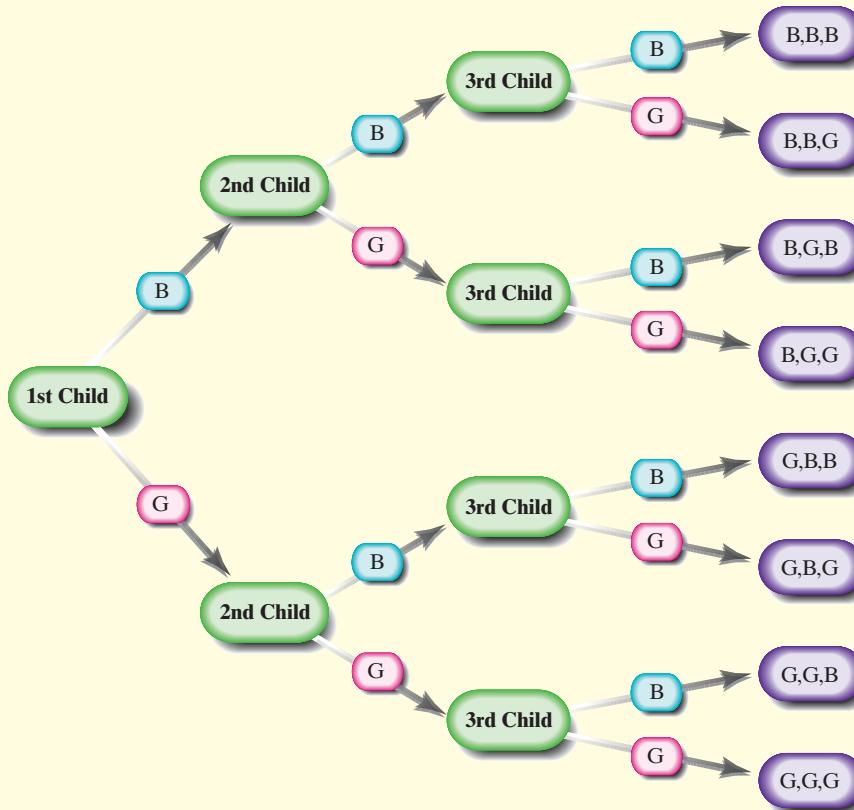
Historical Note

Pierre de Fermat was born into a wealthy family. His father was a leather merchant and second consul of Beaumont-de-Lomagne. Fermat attended the University of Toulouse. By 1631, Fermat was a lawyer and government official. He rose quickly through the ranks because of deaths from the plague. In fact, in 1653, Fermat's death was incorrectly reported. In 1654, Fermat received a correspondence from Blaise Pascal in which Pascal asked Fermat to confirm his ideas on probability. Fermat and Pascal discussed the problem of how to divide the stakes in a game that is interrupted before completion, knowing how many points each player needs to win. Their short correspondence laid the foundation for the theory of probability. They are regarded as joint founders of the subject.

Mathematics was Fermat's passionate hobby and true love. He is most famous for his Last Theorem, which states that the equation $x^n + y^n = z^n$ has no nonzero integer solutions for $n > 2$. The theorem was scribbled in the margin of a book by Diophantus, a Greek mathematician. Fermat stated, "I have discovered a truly marvelous proof of this theorem, which, however, the margin is not large enough to contain." The status of Fermat's Last Theorem baffled mathematicians until Andrew Wiles proved it to be true in 1994.

- (b) To determine the sample space, construct a **tree diagram** to list the equally likely outcomes of the experiment. Draw two branches corresponding to the two possible outcomes (boy or girl) for the first repetition of the experiment (the first child). For the second child, draw four branches, and so on. See Figure 4, where B stands for boy and G stands for girl.

Figure 4



Find the sample space S of this experiment by following each branch to identify all the possible outcomes of the experiment:

$$S = \{ \text{BBB}, \text{BBG}, \text{BGB}, \text{BGG}, \text{GBB}, \text{GBG}, \text{GGB}, \text{GGG} \}$$

$$\text{So, } N(S) = 8.$$

For the event $E = \text{"two boys and a girl"} = \{\text{BBG}, \text{BGB}, \text{GBB}\}$, we have $N(E) = 3$. Since the outcomes are equally likely (for example, BBG is just as likely as BGB), we have

$$P(E) = \frac{N(E)}{N(S)} = \frac{3}{8} = 0.375$$

The probability that a family of three children will have two boys and one girl is 0.375. If we repeat this experiment 1000 times and the outcomes are equally likely (having a girl is just as likely as having a boy), we would expect about 375 of the trials to result in two boys and one girl.

In comparing the results of Examples 7(a) and 7(b), notice that the two probabilities are slightly different. Empirical probabilities and classical probabilities often differ in value, but, as the number of repetitions of a probability experiment increases, the empirical probability should get closer to the classical probability. However, it is possible that the two probabilities differ because having a boy or having a girl are not equally likely events. (Maybe the probability of having a boy is 0.505 and the probability of having a girl is 0.495.)

Historical Note

Blaise Pascal was born in 1623, in Clermont, France. Pascal's father felt that Blaise should not be taught mathematics before age 15. Pascal couldn't resist studying mathematics on his own, and at the age of 12 started to teach himself geometry. In December 1639, the Pascal family moved to Rouen, where Pascal's father had been appointed as a tax collector. Between 1642 and 1645, Pascal worked on developing a calculator to help his father collect taxes. In correspondence with Fermat, he helped develop the theory of probability. This correspondence consisted of five letters written in the summer of 1654. They considered the dice problem and the problem of points. The dice problem deals with determining the expected number of times a pair of dice must be thrown before a pair of sixes is observed. The problem of points asks how to divide the stakes if a game of dice is incomplete. They solved the problem of points for a two-player game, but did not solve it for three or more players.

5**Recognize and Interpret Subjective Probabilities**

If a sports reporter is asked what the chances are for the Boston Red Sox to play in the World Series, the reporter would likely process information about the Red Sox (their pitching staff, lead-off hitter, and so on) and then make an educated guess of the likelihood. The reporter may respond that there is a 20% chance the Red Sox will play in the World Series. This forecast is a probability although it is not based on relative frequencies. We cannot, after all, repeat the experiment of playing a season under the same circumstances (same players, schedule, and so on) over and over. Nonetheless, the forecast of 20% does satisfy the criterion that a probability be between 0 and 1, inclusive. This forecast is known as a *subjective probability*.

Definition

A **subjective probability** of an outcome is a probability obtained on the basis of personal judgment.

It is important to understand that subjective probabilities are perfectly legitimate and are often the only method of assigning likelihood to an outcome. As another example, a financial reporter may ask an economist about the likelihood the economy will fall into recession next year. Again, we cannot conduct an experiment n times to obtain a relative frequency. The economist must use knowledge of the current conditions of the economy and make an educated guess about the likelihood of recession.

**5.1 Assess Your Understanding****Vocabulary and Skill Building**

1. Define each of the following.

- | | |
|-----------------------------|----------------------|
| (a) Probability | (b) Experiment |
| (c) Event | (d) Sample space |
| (e) Equally likely outcomes | (f) Impossible event |
| (g) Unusual event | |

2. *True or False:* In a probability model, the sum of the probabilities of all outcomes must equal 1.

- NW** 3. Verify that the following is a probability model. What do we call the outcome “blue”?

Color	Probability
Red	0.3
Green	0.15
Blue	0
Brown	0.15
Yellow	0.2
Orange	0.2

4. Verify that the following is a probability model. If the model represents the colors of M&Ms in a bag of milk chocolate M&Ms, explain what the model implies.

Color	Probability
Red	0
Green	0
Blue	0
Brown	0
Yellow	1
Orange	0

5. Why is the following not a probability model?

Color	Probability
Red	0.3
Green	-0.3
Blue	0.2
Brown	0.4
Yellow	0.2
Orange	0.2

6. Why is the following not a probability model?

Color	Probability
Red	0.1
Green	0.1
Blue	0.1
Brown	0.4
Yellow	0.2
Orange	0.3

7. Which of the following numbers could be the probability of an event?

$$0, 0.01, 0.35, -0.4, 1, 1.4$$

8. Which of the following numbers could be the probability of an event?

$$1.5, \frac{1}{2}, \frac{3}{4}, \frac{2}{3}, 0, -\frac{1}{4}$$

9. According to Nate Silver, the probability of a senate candidate winning his/her election with a 5% lead in an average of polls with a week until the election is 0.89. Interpret this probability.

Source: fivethirtyeight.com

10. In seven-card stud poker, a player is dealt seven cards. The probability that the player is dealt two cards of the same value and five other cards of different value so that the player has a pair is 0.44. Explain what this probability means. If you play seven-card stud 100 times, will you be dealt a pair exactly 44 times? Why or why not?

11. Suppose that you toss a coin 100 times and get 95 heads and five tails. Based on these results, what is the estimated probability that the next flip results in a head?

12. Suppose that you roll a die 100 times and get six 80 times. Based on these results, what is the estimated probability that the next roll results in six?

13. Bob is asked to construct a probability model for rolling a pair of fair dice. He lists the outcomes as 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. Because there are 11 outcomes, he reasoned the probability of rolling a two must be $\frac{1}{11}$. What is wrong with Bob's reasoning?

14. **Blood Types** A person can have one of four blood types: A, B, AB, or O. If a person is randomly selected, is the probability they have blood type A equal to $\frac{1}{4}$? Why or why not?

- NW 15. If a person rolls a six-sided die and then flips a coin, describe the sample space of possible outcomes using 1, 2, 3, 4, 5, 6 for the die outcomes and H, T for the coin outcomes.

16. If a basketball player shoots three free throws, describe the sample space of possible outcomes using S for a made free throw and F for a missed free throw.

17. According to the U.S. Department of Education, 42.8% of 3-year-olds are enrolled in day care. What is the probability that a randomly selected 3-year-old is enrolled in day care?

18. According to the American Veterinary Medical Association, the proportion of households owning a dog is 0.372. What is the probability that a randomly selected household owns a dog?

For Problems 19–22, let the sample space be $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Suppose the outcomes are equally likely.

19. Compute the probability of the event $E = \{1, 2, 3\}$.
20. Compute the probability of the event $F = \{3, 5, 9, 10\}$.
21. Compute the probability of the event $E = \text{"an even number less than 9"}$.
22. Compute the probability of the event $F = \text{"an odd number"}$.

Applying the Concepts

23. **Play Sports?** A survey of 500 randomly selected high school students determined that 288 played organized sports.

- (a) What is the probability that a randomly selected high school student plays organized sports?

- (b) Interpret this probability.

24. **Volunteer?** In a survey of 1100 female adults (18 years of age or older), it was determined that 341 volunteered at least once in the past year.

- (a) What is the probability that a randomly selected adult female volunteered at least once in the past year?

- (b) Interpret this probability.

- NW 25. **Home Runs** The *Wall Street Journal* regularly publishes an article entitled "The Count." In one article, The Count looked at 1000 randomly selected home runs in Major League Baseball. Source: *Wall Street Journal*, September 24, 2014.

- (a) Of the 1000 home runs, it was found that 85 were caught by fans. What is the probability that a randomly selected home run is caught by a fan?

- (b) Of the 1000 home runs, it was found that 296 were dropped when a fan had a legitimate play on the ball. What is the probability that a randomly selected home run is dropped?

- (c) Of the 85 caught balls, it was determined that 34 were barehanded catches, 49 were caught with a glove, and two were caught in a hat. What is the probability a randomly selected caught ball was caught in a hat? Interpret this probability.

- (d) Of the 296 dropped balls, it was determined that 234 were barehanded attempts, 54 were dropped with a glove, and eight were dropped with a failed hat attempt. What is the probability a randomly selected dropped ball was a failed hat attempt? Interpret this probability.

26. **Police Complaints** The *Chicago Tribune* analyzed 17,713 complaints by citizens against Chicago police officers.

Source: *Chicago Tribune*, June 14, 2015.

- (a) Of the 17,713 complaints against police officers, it was found that 7296 were accompanied by a signed affidavit, which is required by state law for the complaint against the officer to proceed. What is the probability that a randomly selected complaint is accompanied by a signed affidavit?

- (b) Of the 7296 complaints with an affidavit, 794 were found to be legitimate. What is the probability that a complaint with an affidavit is found to be legitimate?

- (c) Of the 794 complaints with an affidavit that were found to be legitimate, 297 resulted in a suspension of the officer. What is the probability that a legitimate complaint with an affidavit results in the suspension of the officer?

- (d) Of the 794 complaints with an affidavit that were found to be legitimate, 12 resulted in the officer being dismissed. Is it unusual for a legitimate complaint with an affidavit to result in the officer being dismissed?

- 27. Roulette** In the game of roulette, a wheel consists of 38 slots numbered 0, 00, 1, 2, . . . , 36. (See the photo.) To play the game, a metal ball is spun around the wheel and is allowed to fall into one of the numbered slots.



- (a) Determine the sample space.
 - (b) Determine the probability that the metal ball falls into the slot marked eight. Interpret this probability.
 - (c) Determine the probability that the metal ball lands in an odd slot. Interpret this probability.
- 28. Birthdays** Exclude leap years from the following calculations and assume each birthday is equally likely:
- (a) Determine the probability that a randomly selected person has a birthday on the 1st day of a month. Interpret this probability.
 - (b) Determine the probability that a randomly selected person has a birthday on the 31st day of a month. Interpret this probability.
 - (c) Determine the probability that a randomly selected person was born in December. Interpret this probability.
 - (d) Determine the probability that a randomly selected person has a birthday on November 8. Interpret this probability.
 - (e) If you just met somebody and she asked you to guess her birthday, are you likely to be correct?
 - (f) Do you think it is appropriate to use the methods of classical probability to compute the probability that a person is born in December?

29. Genetics A gene is composed of two alleles, either dominant or recessive. Suppose that a husband and wife, who are both carriers of the sickle-cell anemia allele but do not have the disease, decide to have a child. Because both parents are carriers of the disease, each has one dominant normal-cell allele (S) and one recessive sickle-cell allele (s). Therefore, the genotype of each parent is Ss . Each parent contributes one allele to his or her offspring, with each allele being equally likely.

- (a) Genes are always written with the dominant gene first. Therefore, there are two instances the offspring could have genotype Ss (one if the mother contributes the dominant allele and the father contributes the nondominant allele, and vice versa). List the possible genotypes of their offspring.
- (b) What is the probability that the offspring will have sickle-cell anemia? In other words, what is the probability that the offspring will have genotype ss ? Interpret this probability.
- (c) What is the probability that the offspring will not have sickle-cell anemia but will be a carrier? In other words, what is the probability that the offspring will have one dominant normal-cell allele and one recessive sickle-cell allele? Interpret this probability.

30. More Genetics In Problem 29, we learned that for some diseases, such as sickle-cell anemia, an individual will get the disease only if he or she receives both recessive alleles. This is not always the case. For example, Huntington's disease only requires one dominant gene for an individual to contract the disease. Suppose that a husband and wife, who both have a dominant Huntington's disease allele (S) and a normal recessive allele (s), decide to have a child.

- (a) Genes are always written with the dominant gene first. Therefore, there are two instances the offspring could have genotype Ss (one if the mother contributes the dominant allele and the father contributes the nondominant allele, and vice versa). List the possible genotypes of their offspring.
- (b) What is the probability that the offspring will not have Huntington's disease? In other words, what is the probability that the offspring will have genotype ss ? Interpret this probability.
- (c) What is the probability that the offspring will have Huntington's disease?

NW 31. College Survey In a national survey conducted by the Centers for Disease Control to determine college students' health-risk behaviors, college students were asked, "How often do you wear a seatbelt when riding in a car driven by someone else?" The frequencies appear in the following table:

Response	Frequency
Never	125
Rarely	324
Sometimes	552
Most of the time	1257
Always	2518

- (a) Construct a probability model for seatbelt use by a passenger.
- (b) Would you consider it unusual to find a college student who never wears a seatbelt when riding in a car driven by someone else? Why?

32. College Survey In a national survey conducted by the Centers for Disease Control to determine college students' health-risk behaviors, college students were asked, "How often do you wear a seatbelt when driving a car?" The frequencies appear in the following table:

Response	Frequency
Never	118
Rarely	249
Sometimes	345
Most of the time	716
Always	3093

- (a) Construct a probability model for seatbelt use by a driver.
 - (b) Is it unusual for a college student to never wear a seatbelt when driving a car? Why?
- 33. Larceny Theft** A police officer randomly selected 642 police records of larceny thefts. The following data represent the number of offenses for various types of larceny thefts.

Type of Larceny Theft	Number of Offenses
Pocket picking	4
Purse snatching	6
Shoplifting	133
From motor vehicles	219
Motor vehicle accessories	90
Bicycles	42
From buildings	143
From coin-operated machines	5

Source: U.S. Federal Bureau of Investigation.

- (a) Construct a probability model for type of larceny theft.

(b) Are purse snatching larcenies unusual?

(c) Are bicycle larcenies unusual?

- 34. Multiple Births** The following data represent the number of live multiple-delivery births (three or more babies) in 2017 for women 15 to 54 years old.

Age	Number of Multiple Births
15–19	43
20–24	365
25–29	964
30–34	1442
35–39	837
40–44	197
45–54	69

Source: National Vital Statistics Report.

- (a) Construct a probability model for number of multiple births.

(b) In the sample space of all multiple births, are multiple births for 15- to 19-year-old mothers unusual?

(c) In the sample space of all multiple births, are multiple births for 40- to 44-year-old mothers unusual?



- 35. Walt Disney Stock** The table shows the movement of Walt Disney stock for 30 randomly selected trading days. “Up” means the stock price increased in value for the day. “Down” means the stock price decreased in value for the day, and “No Change” means the stock price closed at the same price it closed for the previous day.

Down	Up	Up	Down	Down	Up
Down	Up	Down	Up	Down	Up
Down	Down	Up	Up	Up	Up
Down	Down	Down	Up	Down	Up
No Change	Up	Down	Down	No Change	Down

Source: Yahoo! Finance.

- (a) Construct a probability model for stock movement of Walt Disney stock.

(b) Are the probabilities in part (a) empirical, classical, or subjective?

(c) What is the probability that Walt Disney stock is up for a randomly selected day?

(d) Is it unusual for Walt Disney stock to close at the same price it closed on the previous day?

(e) Would the estimate of the probability of Walt Disney stock price movement improve if we considered 60 randomly selected days instead? Explain.



- 36. Favorite Day to Order Takeout** A survey was conducted by Wakefield Research in which participants were asked to disclose their favorite night to order takeout for dinner. The following data are based on their results.

Thursday	Saturday	Friday	Friday	Sunday
Wednesday	Saturday	Friday	Tuesday	Friday
Saturday	Monday	Friday	Friday	Sunday
Friday	Tuesday	Wednesday	Saturday	Friday
Wednesday	Monday	Wednesday	Wednesday	Friday
Friday	Wednesday	Thursday	Tuesday	Friday
Tuesday	Saturday	Friday	Tuesday	Friday
Saturday	Saturday	Saturday	Sunday	Friday

Source: Based on results from Wakefield Research.

- (a) Construct a probability model for favorite night to order takeout.

(b) What is the probability a randomly selected individual would choose Friday as their favorite night to order takeout?

(c) Would it be unusual for an individual to state their favorite night to order takeout is Tuesday?



- 37. A Random Process: Trains** Your daily commute to work requires that you cross railroad tracks. At this particular railroad crossing the trains tend to be long and slow. So, getting stopped by a train will likely make you late for work. You start recording data to determine the likelihood of arriving at the tracks while a train is there. Open data set 5_1_37 at www.pearsonhighered.com/sullivanstats, which contains the day number and whether a train was present, or not, for 200 consecutive days in which you drove to work. The column “Train” shows a series of 0s and 1s. In that column, a 0 indicates there was no train present and a 1 indicates that a train was present. The column “Aggregate Train” represents the cumulative number of times a train was present. The column “Aggregate Proportion Train” represents the cumulative proportion of times a train was present.

(a) Explain why getting stuck by the train is a random process.

(b) What proportion of the days was a train present after 8 days?

(c) What proportion of the days was a train present after 20 days?

(d) Were you stopped by a train on the 30th day?

(e) Graph the proportion of days a train was present against the number days (number of days should be on the horizontal axis).

(f) What is the estimate of the probability of being stuck by a train during your commute?



- 38. A Random Process: Green Lights** On your drives to school each day you feel like there is a light that is always red when you reach it. You decide to record data to determine the likelihood of arriving at the light while it is red. Open the data set 5_1_38 at www.pearsonhighered.com/sullivanstats, which contains the day number and whether the light was red (1), or not (0), for 120 consecutive days in which you drove to school.

(a) Explain what makes this a random process.

(b) What proportion of the days were you stuck by a red light after 15 days?

(c) What proportion of the days were you stuck by a red light after 40 days?

(d) Were you stuck by a red light on the 50th day?

- (e) Graph the proportion of days the light was red against the number of days (number of days should be on the horizontal axis).
 (f) What is the estimate of the probability of the light being red when you reach the intersection?

In Problems 39–42, use the given table, which lists six possible assignments of probabilities for tossing a coin twice, to answer the following questions.

Assignments	Sample Space			
	HH	HT	TH	TT
A	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
B	0	0	0	1
C	$\frac{3}{16}$	$\frac{5}{16}$	$\frac{5}{16}$	$\frac{3}{16}$
D	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$
E	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$
F	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{4}{9}$

39. Which of the assignments of probabilities are consistent with the definition of a probability model?
 40. Which of the assignments of probabilities should be used if the coin is known to be fair?
 41. Which of the assignments of probabilities should be used if the coin is known to always come up tails?
 42. Which of the assignments of probabilities should be used if tails is twice as likely to occur as heads?

NW 43. Going to Disney World John, Roberto, Clarice, Dominique, and Marco work for a publishing company. The company wants to send two employees to a statistics conference in Orlando. To be fair, the company decides that the two individuals who get to attend will have their names randomly drawn from a hat.

- (a) Determine the sample space of the experiment. That is, list all possible simple random samples of size $n = 2$.
 (b) What is the probability that Clarice and Dominique attend the conference?
 (c) What is the probability that Clarice attends the conference?
 (d) What is the probability that John stays home?

44. Six Flags In 2011, Six Flags St. Louis had 10 roller coasters: The Screamin' Eagle, The Boss, River King Mine Train, Batman the Ride, Mr. Freeze, Ninja, Tony Hawk's Big Spin, Evel Knievel, Xcalibur, and Sky Screamer. Of these, The Boss, The Screamin' Eagle, and Evel Knievel are wooden coasters. Ethan wants to ride two more roller coasters before leaving the park (not the same one twice) and decides to select them by drawing names from a hat.

- (a) Determine the sample space of the experiment. That is, list all possible simple random samples of size $n = 2$.
 (b) What is the probability that Ethan will ride Mr. Freeze and Evel Knievel?
 (c) What is the probability that Ethan will ride the Screamin' Eagle?
 (d) What is the probability that Ethan will ride two wooden roller coasters?
 (e) What is the probability that Ethan will not ride any wooden roller coasters?

45. Classifying Probability Determine whether the probabilities below are computed using classical methods, empirical methods, or subjective methods.

- (a) The probability of having eight girls in an eight-child family is 0.00390625.
 (b) On the basis of a survey of 1000 families with eight children, the probability of a family having eight girls is 0.005.
 (c) According to a sports analyst, the probability that the Chicago Bears will win their next game is about 0.30.
 (d) On the basis of clinical trials, the probability of efficacy of a new drug is 0.75.

46. Checking for Loaded Dice You suspect a 6-sided die to be loaded and conduct a probability experiment by rolling the die 400 times. The outcome of the experiment is listed in the table below. Do you think the die is loaded? Why?

Value of Die	Frequency
1	105
2	47
3	44
4	49
5	51
6	104

47. Conduct a survey in your school by randomly asking 50 students whether they drive to school. Based on the results of the survey, approximate the probability that a randomly selected student drives to school.

- 48. (a)** In 2017, the median income of families in the United States was \$60,336. What is the probability that a randomly selected family has an income greater than \$60,336?
(b) The middle 50% of enrolled freshmen at Washington University in St. Louis had SAT math scores in the range 700–780. What is the probability that a randomly selected freshman at Washington University has an SAT math score of 700 or higher?

49. Threaded Problem: Tornadoes The data set “Tornadoes_2017” located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a)** Construct a probability model for month in which the tornado occurred.
(b) What is the probability a randomly selected tornado from 2017 occurred in April?
(c) Is it unusual for a 2017 tornado to occur in December?
(d) Construct a probability model for the F scale.
(e) Is it unusual to observe a tornado whose F scale is 4?

50. NFL Combine Each year the National Football League (NFL) runs a combine in which players who wish to be considered for the NFL draft must participate in a variety of activities. Go to www.pearsonhighered.com/sullivanstats to obtain the data file 5_1_50 using the file format of your choice for the version of the text you are using. The data represent results of the 2015 NFL combine. Construct a probability model for the variable “POS,” which represents the position of the player. For example, Nelson Agholor plays wide receiver (WR). If a player is randomly selected from the 2015 NFL combine, what position has the highest probability of being selected? Would you be surprised if a center (C) was randomly selected? Why?

DATA **51. The Law of Averages?** Let's say a player typically gets a hit in 3 out of every 10 at-bats (for a 0.300 batting average).

Suppose the player has not had a hit in his previous four at-bats. In baseball, you will often hear an announcer say, "This player is due for a hit." What does the announcer mean by this? Is the announcer's statement true? The announcer is saying that if a player gets a hit in 3 of every 10 at-bats and has not had a hit in his previous 4 at-bats, the nonexistent law of averages must apply. To see why the announcer's statement is not true, use the data set 5_1_51 found at www.pearsonhighered.com/sullivanstats and answer the following questions. In the data, the first four columns contain a 0 or 1. If the entry is 0, then the player does not get a hit, and if the entry is 1, then the player gets a hit. Each row represents a sequence of four at-bats. For example, in row 4, the player got a hit in his first and third at-bat but did not get a hit in his second and fourth at-bat.

- (a) What was the outcome of the player's at-bats based on the entries in the first four columns of row 1? Row 5?
- (b) Column 5 contains the number of hits in the player's previous four at-bats. We are only interested in those entries where the player had 0 hits in his previous four at-bats. Column 6 indicates whether the player got a hit in the fifth at-bat. For example, in row 1, the player did not get a hit in at-bat 5, but the player did get a hit in at-bat 5 in row 2. Among those instances where the player did not get a hit in his first four at-bats, determine the proportion of times in at-bat 5 the player gets a hit. To do this in StatCrunch, select Stats > Tables > Frequency. Under Select columns choose AtBat5. Under Where enter "Hits Previous 4" = 0. [Note: The quotes around "Hits Previous 4" are necessary]. Click Compute!. What proportion of fifth at-bats results in a hit when the previous four at-bats did not result in a hit? What does this tell you?

52. Putting It Together: Drug Side Effects In placebo-controlled clinical trials for the drug Viagra, 734 subjects received Viagra and 725 subjects received a placebo (subjects did not know which treatment they received). The table below summarizes reports of various side effects that were reported.

Adverse Effect	Viagra (<i>n</i> = 734)	Placebo (<i>n</i> = 725)
Headache	117	29
Flushing	73	7
Dyspepsia	51	15
Nasal congestion	29	15
Urinary tract infection	22	15
Abnormal vision	22	0
Diarrhea	22	7
Dizziness	15	7
Rash	15	7

- (a) Is the variable "adverse effect" qualitative or quantitative?
- (b) Which type of graph would be appropriate to display the information in the table? Construct the graph.
- (c) What is the estimated probability that a randomly selected subject from the Viagra group reported flushing as an adverse effect? Would this be unusual?

- (d) What is the estimated probability that a subject receiving a placebo would report flushing as an adverse effect? Would this be unusual?
- (e) If a subject reports flushing after receiving a treatment, what might you conclude?
- (f) What type of experimental design is this?

Explaining the Concepts

53. The following is a quote by Pierre-Simon Laplace: "To discover the best treatment to use in curing a disease, it is sufficient to test each treatment on the same number of patients, while keeping all circumstances perfectly similar. The superiority of the most beneficial treatment will become more and more evident as this number is increased, and the calculus will yield the corresponding probability of its benefit and of the ratio by which it is greater than the others"

Source: Laplace, P.-S. (1825) *Essay on Probabilities*, 5th ed. Paris: Bachelier.

- (a) What does Laplace mean when he says, "while keeping all circumstances perfectly similar"?
- (b) Explain the meaning of "the superiority of the most beneficial treatment will become more and more evident as this number is increased." What law does this statement utilize?

54. A friend of yours regularly plays the lottery but has never won. She says that she feels really good about this weekend's drawing because she is due for a winning ticket. Explain the flaw in your friend's reasoning.

55. Explain the Law of Large Numbers. How does this law apply to gambling casinos?

56. In computing classical probabilities, all outcomes must be equally likely. Explain what this means.

57. Describe what an unusual event is. Should the same cutoff always be used to identify unusual events? Why or why not?

58. You are planning a trip to a water park tomorrow and the weather forecaster says there is a 70% chance of rain. Explain what this result means.

59. Describe the difference between classical and empirical probability.

60. Ask Marilyn In a September 19, 2010, article in *Parade Magazine* written to *Ask Marilyn*, Marilyn vos Savant was asked the following: Four identical sealed envelopes are on a table, one of which contains \$100. You are to select one of the envelopes. Then the game host discards two of the remaining three envelopes and informs you that they do not contain the \$100. In addition, the host offers you the opportunity to switch envelopes. What should you do?

- (a) Keep your envelope
- (b) switch
- (c) it does not matter.

61. Suppose you live in a town with two hospitals—one large and the other small. On a given day in one of the hospitals, 60% of the babies who were born were girls. Which one do you think it is? Or, is it impossible to tell. Support your decision?

62. Suppose that a probability is approximated to be zero based on empirical results. Does this mean the event is impossible? Explain.

5.2 The Addition Rule and Complements



Preparing for This Section Before getting started, review the following:

- Contingency Tables (Section 4.4, p. 207)

Objectives

- ① Use the Addition Rule for Disjoint Events
- ② Use the General Addition Rule
- ③ Compute the probability of an event using the Complement Rule

1 Use the Addition Rule for Disjoint Events

Before presenting more rules for computing probabilities, we discuss *disjoint events*.

Definition

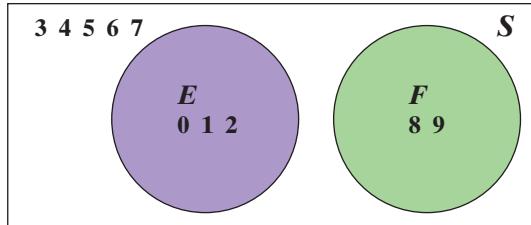
Two events are **disjoint** if they have no outcomes in common. Another name for disjoint events is **mutually exclusive** events.

IN OTHER WORDS

Two events are disjoint if they cannot occur at the same time.

We can use **Venn diagrams** to represent events as circles enclosed in a rectangle. The rectangle represents the sample space, and each circle represents an event. For example, suppose we randomly select chips from a bag. Each chip is labeled 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Let E represent the event “choose a number less than or equal to 2,” and let F represent the event “choose a number greater than or equal to 8.” Because E and F have no outcomes in common, they are disjoint. Figure 5 shows a Venn diagram of these disjoint events.

Figure 5



Notice that the outcomes in event E are inside circle E , and the outcomes in event F are inside circle F . All outcomes in the sample space that are not in E or F are outside the circles, but inside the rectangle. From this diagram, we know that

$$P(E) = \frac{N(E)}{N(S)} = \frac{3}{10} = 0.3 \text{ and } P(F) = \frac{N(F)}{N(S)} = \frac{2}{10} = 0.2. \text{ In addition, } P(E \text{ or } F) =$$

$$\frac{N(E \text{ or } F)}{N(S)} = \frac{5}{10} = 0.5 \text{ and } P(E \text{ or } F) = P(E) + P(F) = 0.3 + 0.2 = 0.5. \text{ This result}$$

occurs because of the *Addition Rule for Disjoint Events*.

IN OTHER WORDS

The Addition Rule for Disjoint Events states that, if you have two events that have no outcomes in common, the probability that one or the other occurs is the sum of their probabilities.

Addition Rule for Disjoint Events

If E and F are disjoint (or mutually exclusive) events, then

$$P(E \text{ or } F) = P(E) + P(F)$$

The Addition Rule for Disjoint Events can be extended to more than two disjoint events. In general, if E, F, G, \dots each have no outcomes in common (they are pairwise disjoint), then

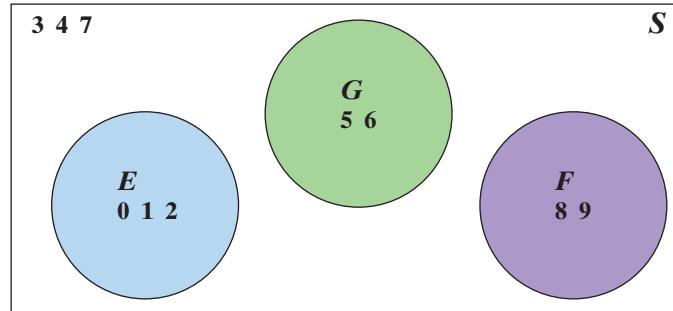
$$P(E \text{ or } F \text{ or } G \text{ or } \dots) = P(E) + P(F) + P(G) + \dots$$

Let event G represent “the number is a 5 or 6.” The Venn diagram in Figure 6 illustrates the Addition Rule for more than two disjoint events using the chip example. Notice that no pair of events has any outcomes in common. So, from the Venn diagram, we can see that

$$P(E) = \frac{N(E)}{N(S)} = \frac{3}{10} = 0.3, P(F) = \frac{N(F)}{N(S)} = \frac{2}{10} = 0.2, \text{ and } P(G) = \frac{N(G)}{N(S)} = \frac{2}{10} = 0.2.$$

In addition, $P(E \text{ or } F \text{ or } G) = P(E) + P(F) + P(G) = 0.3 + 0.2 + 0.2 = 0.7$.

Figure 6



EXAMPLE 1 Benford's Law and the Addition Rule for Disjoint Events

Problem Our number system consists of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. Because we do not write numbers such as 12 as 012, the first significant digit in any number must be 1, 2, 3, 4, 5, 6, 7, 8, or 9. We may think that each digit appears with equal frequency so that each digit has a $\frac{1}{9}$ probability of being the first significant digit, but this is not true. In

1881, Simon Newcomb discovered that digits do not occur with equal frequency. The physicist Frank Benford discovered the same result in 1938. After studying lots and lots of data, he assigned probabilities of occurrence for each of the first digits, as shown in Table 5. The probability model is now known as *Benford's Law* and plays a major role in identifying fraudulent data on tax returns and accounting books.

Table 5

Digit	1	2	3	4	5	6	7	8	9
Probability	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Source: The First Digit Phenomenon, T. P. Hill, *American Scientist*, July–August, 1998.

- (a) Verify that Benford's Law is a probability model.
- (b) Use Benford's Law to determine the probability that a randomly selected first digit is 1 or 2.
- (c) Use Benford's Law to determine the probability that a randomly selected first digit is at least 6.

Approach For part (a), verify that each probability is between 0 and 1 and that the sum of all probabilities equals 1. For parts (b) and (c), use the Addition Rule for Disjoint Events.

Solution

- (a) Each probability in Table 5 is between 0 and 1. In addition, the sum of all the probabilities, $0.301 + 0.176 + 0.125 + \dots + 0.046$, is 1. Because Rules 1 and 2 are satisfied, Table 5 represents a probability model.

(continued)

(b)

$$\begin{aligned} P(1 \text{ or } 2) &= P(1) + P(2) \\ &= 0.301 + 0.176 \\ &= 0.477 \end{aligned}$$

If we looked at 100 numbers, we would expect about 48 to begin with 1 or 2.

IN OTHER WORDS

The phrase “at least” means “greater than, or equal to.”

(c)

$$\begin{aligned} P(\text{at least } 6) &= P(6 \text{ or } 7 \text{ or } 8 \text{ or } 9) \\ &= P(6) + P(7) + P(8) + P(9) \\ &= 0.067 + 0.058 + 0.051 + 0.046 \\ &= 0.222 \end{aligned}$$

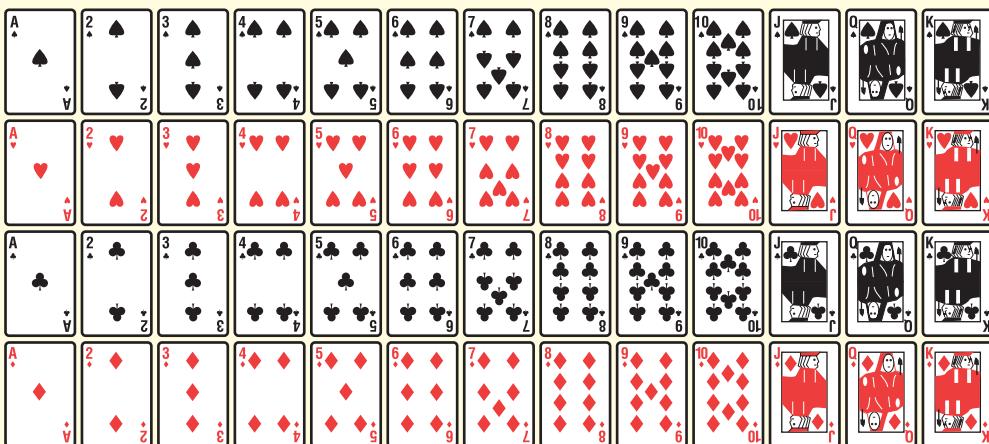
If we looked at 100 numbers, we would expect about 22 to begin with 6, 7, 8, or 9.



EXAMPLE 2 A Deck of Cards and the Addition Rule for Disjoint Events

Problem Suppose that a single card is selected from a standard 52-card deck, such as the one shown in Figure 7.

Figure 7



- (a) Compute the probability of the event E = “drawing a king.”
- (b) Compute the probability of the event E = “drawing a king” or F = “drawing a queen” or G = “drawing a jack.”

Approach Use the classical method for computing the probabilities because the outcomes are equally likely and easy to count. Use the Addition Rule for Disjoint Events to compute the probability in part (b) because the events are mutually exclusive. For example, you cannot simultaneously draw a king and a queen.

Solution The sample space consists of the 52 cards in the deck, so $N(S) = 52$.

- (a) A standard deck of cards has four kings, so $N(E) = 4$. Therefore,

$$P(\text{king}) = P(E) = \frac{N(E)}{N(S)} = \frac{4}{52} = \frac{1}{13}$$

- (b) A standard deck of cards has four kings, four queens, and four jacks. Because events E , F , and G are mutually exclusive, use the Addition Rule for Disjoint Events extended to two or more disjoint events. So

$$\begin{aligned} P(\text{king or queen or jack}) &= P(E \text{ or } F \text{ or } G) \\ &= P(E) + P(F) + P(G) \\ &= \frac{4}{52} + \frac{4}{52} + \frac{4}{52} = \frac{12}{52} = \frac{3}{13} \end{aligned}$$

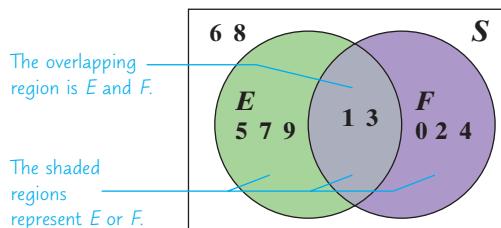


② Use the General Addition Rule

What happens when you need to compute the probability of two events that are not disjoint?

Suppose we are randomly selecting chips from a bag. Each chip is labeled 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9. Let E represent the event “choose an odd number,” and let F represent the event “choose a number less than or equal to 4.” Because $E = \{1, 3, 5, 7, 9\}$ and $F = \{0, 1, 2, 3, 4\}$ have the outcomes 1 and 3 in common, the events are not disjoint. Figure 8 shows a Venn diagram of these events.

Figure 8



We can compute $P(E \text{ or } F)$ directly by counting because each outcome is equally likely. There are 8 outcomes in E or F and 10 outcomes in the sample space, so

$$P(E \text{ or } F) = \frac{N(E \text{ or } F)}{N(S)} = \frac{8}{10} = \frac{4}{5}$$

Notice that using the Addition Rule for Disjoint Events to find $P(E \text{ or } F)$ would be *incorrect*:

$$P(E \text{ or } F) \neq P(E) + P(F) = \frac{5}{10} + \frac{5}{10} = 1$$

This implies that the chips labeled 6 and 8 will never be selected, which contradicts our assumption that all the outcomes are equally likely. Our result is incorrect because we counted the outcomes 1 and 3 twice: once for event E and once for event F . To avoid this double counting, we have to subtract the probability corresponding to the overlapping region, E and F . That is, subtract $P(E \text{ and } F) = \frac{2}{10}$ from the result and obtain

$$\begin{aligned} P(E \text{ or } F) &= P(E) + P(F) - P(E \text{ and } F) \\ &= \frac{5}{10} + \frac{5}{10} - \frac{2}{10} \\ &= \frac{8}{10} = \frac{4}{5} \end{aligned}$$

which agrees with the result we found by counting. The following rule generalizes these results.

The General Addition Rule

For any two events E and F ,

$$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$$

EXAMPLE 3 Computing Probabilities for Events That Are Not Disjoint

Problem Suppose a single card is selected from a standard 52-card deck. Compute the probability of the event E = “drawing a king” or F = “drawing a diamond.”

Approach The events are not disjoint because the outcome “king of diamonds” is in both events, so use the General Addition Rule.

(continued)

Solution

$$\begin{aligned}
 P(E \text{ or } F) &= P(E) + P(F) - P(E \text{ and } F) \\
 P(\text{king or diamond}) &= P(\text{king}) + P(\text{diamond}) - P(\text{king of diamonds}) \\
 &= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} \\
 &= \frac{16}{52} = \frac{4}{13}
 \end{aligned}$$

NW Now Work Problem 31

Consider the data shown in Table 6, which represent the marital status of males and females 15 years old or older in the United States in 2017.

Table 6

	Males (in millions)	Females (in millions)
Never married	46.0	40.1
Married	63.3	62.0
Widowed	3.3	11.9
Divorced	12.1	16.1
Separated	2.2	3.1

Source: U.S. Census Bureau, American Community Survey.

Table 6 is called a **contingency table** or **two-way table**, because it relates two categories of data. The **row variable** is marital status, because each row in the table describes the marital status of an individual. The **column variable** is gender. Each box inside the table is called a **cell**. For example, the cell corresponding to married individuals who are male is in the second row, first column. Each cell contains the frequency of the category: There were 63.3 million married males in the United States in 2017. Put another way, in the United States in 2017, there were 63.3 million individuals who were male *and* married.

EXAMPLE 4**Using the Addition Rule with Contingency Tables**

Problem Using the data in Table 6,

- Determine the probability that a randomly selected U.S. resident 15 years old or older is male.
- Determine the probability that a randomly selected U.S. resident 15 years old or older is widowed.
- Determine the probability that a randomly selected U.S. resident 15 years old or older is widowed or divorced.
- Determine the probability that a randomly selected U.S. resident 15 years old or older is male or widowed.

Approach Add the entries in each row and column to get the total number of people in each category. Then determine the probabilities using either the Addition Rule for Disjoint Events or the General Addition Rule.

Solution First, add the entries in each column. For example, the “male” column shows there are $46.0 + 63.3 + 3.3 + 12.1 + 2.2 = 126.9$ million males 15 years old or older in the United States. Add the entries in each row. For example, in the “never married” row we find there are $46.0 + 40.1 = 86.1$ million U.S. residents 15 years old or older who have never married. Adding the row totals or column totals, we find there are $126.9 + 133.2 = 86.1 + 125.3 + 15.2 + 28.2 + 5.3 = 260.1$ million U.S. residents 15 years old or older.

- (a) There are 126.9 million males 15 years old or older and 260.1 million U.S. residents 15 years old or older. The probability that a randomly selected U.S. resident 15 years old or older is male is $\frac{126.9}{260.1} = 0.488$.
- (b) There are 15.2 million U.S. residents 15 years old or older who are widowed. The probability that a randomly selected U.S. resident 15 years old or older is widowed is $\frac{15.2}{260.1} = 0.058$.
- (c) The events widowed and divorced are disjoint. Do you see why? We use the Addition Rule for Disjoint Events.

$$\begin{aligned} P(\text{widowed or divorced}) &= P(\text{widowed}) + P(\text{divorced}) \\ &= \frac{15.2}{260.1} + \frac{28.2}{260.1} \\ &= \frac{43.4}{260.1} = 0.167 \end{aligned}$$

- (d) The events male and widowed are not mutually exclusive. In fact, there are 3.3 million males who are widowed in the United States. Therefore, we use the General Addition Rule to compute $P(\text{male or widowed})$:

$$\begin{aligned} P(\text{male or widowed}) &= P(\text{male}) + P(\text{widowed}) - P(\text{male and widowed}) \\ &= \frac{126.9}{260.1} + \frac{15.2}{260.1} - \frac{3.3}{260.1} \\ &= \frac{138.8}{260.1} = 0.534 \end{aligned}$$

NW Now Work Problem 39



③ Compute the Probability of an Event Using the Complement Rule

Suppose that the probability of an event E is known and we would like to determine the probability that E does not occur. This can easily be accomplished using the idea of *complements*.

Definition

Complement of an Event

Let S denote the sample space of a probability experiment and let E denote an event. The **complement of E** , denoted E^c , is all outcomes in the sample space S that are not outcomes in the event E .

Because E and E^c are mutually exclusive,

$$P(E \text{ or } E^c) = P(E) + P(E^c) = P(S) = 1$$

Subtracting $P(E)$ from both sides, we obtain the following result.

IN OTHER WORDS

For any event, either the event happens or it doesn't. Use the Complement Rule when you know the probability that some event will occur and you want to know the chance it will not occur.

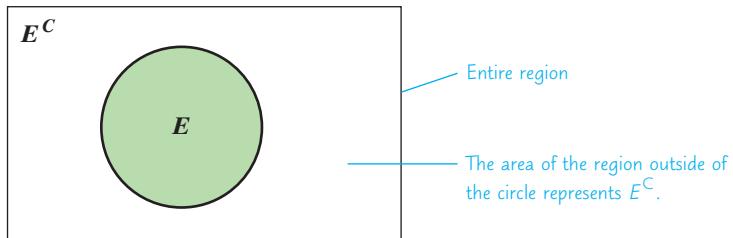
Complement Rule

If E represents any event and E^c represents the complement of E , then

$$P(E^c) = 1 - P(E)$$

Figure 9 illustrates the Complement Rule using a Venn diagram.

Figure 9



EXAMPLE 5 Computing Probabilities Using Complements

Problem According to the National Gambling Impact Study Commission, 52% of Americans have played state lotteries. What is the probability that a randomly selected American has not played a state lottery?

Approach Not playing a state lottery is the complement of playing a state lottery. Compute the probability using the Complement Rule.

Solution $P(\text{not played state lottery}) = 1 - P(\text{played state lottery}) = 1 - 0.52 = 0.48$

The probability of randomly selecting an American who has not played a state lottery is 0.48. 

EXAMPLE 6 Computing Probabilities Using Complements

Problem The data in Table 7 represent the income distribution of households in the United States in 2017.

Table 7

Annual Income	Number (in thousands)	Annual Income	Number (in thousands)
Less than \$10,000	7804	\$50,000 to \$74,999	21,131
\$10,000 to \$14,999	5403	\$75,000 to \$99,999	15,008
\$15,000 to \$24,999	11,166	\$100,000 to \$149,999	17,529
\$25,000 to \$34,999	10,926	\$150,000 to \$199,999	7564
\$35,000 to \$49,999	15,248	\$200,000 or more	8284

Source: U.S. Census Bureau.

Compute the probability that a randomly selected household earned the following incomes in 2017:

- (a) \$200,000 or more
- (b) Less than \$200,000
- (c) At least \$10,000

Approach Determine the probabilities by finding the relative frequency of each event.

Solution

- (a) There were a total of $7804 + 5403 + \dots + 8284 = 120,063$ thousand households in the United States in 2017 and 8284 thousand of them earned \$200,000 or more, so $P(\text{earned } \$200,000 \text{ or more}) = \frac{8284}{120,063} = 0.069$.

The probability of randomly selecting a household that earned \$200,000 or more in 2017 is 0.069. If we randomly selected 1000 households in 2017, we would expect 69 to have household incomes of \$200,000 or more.

NOTE

$$0.069 = \frac{69}{1000}$$

- (b)** We could compute the probability of randomly selecting a household that earned less than \$200,000 in 2017 by adding the relative frequencies of each category less than \$200,000, but it is easier to use complements. The complement of earning less than \$200,000 is earning \$200,000 or more. Therefore,

$$\begin{aligned} P(\text{less than } \$200,000) &= 1 - P(\$200,000 \text{ or more}) \\ &= 1 - 0.069 = 0.931 \end{aligned}$$

The probability of randomly selecting a household that earned less than \$200,000 in 2017 is 0.931. If we randomly selected 1000 households in 2017, we would expect 931 to have household incomes less than \$200,000.

- (c)** The phrase *at least* means greater than or equal to. The complement of at least \$10,000 is less than \$10,000. In 2017, 7804 thousand households earned less than \$10,000. The probability of randomly selecting a household that earned at least \$10,000 is

$$\begin{aligned} P(\text{at least } \$10,000) &= 1 - P(\text{less than } \$10,000) \\ &= 1 - \frac{7804}{120,063} = 0.935 \end{aligned}$$

The probability of randomly selecting a household that earned at least \$10,000 in 2017 is 0.935. If we randomly selected 1000 households in 2017, we would expect 935 to have household income of at least \$10,000.

**NW Now Work Problems 25(d)
and 29**



5.2 Assess Your Understanding

Vocabulary and Skill Building

1. What does it mean when two events are disjoint?
2. If E and F are disjoint events, then $P(E \text{ or } F) = \underline{\hspace{2cm}}$.
3. If E and F are not disjoint events, then $P(E \text{ or } F) = \underline{\hspace{2cm}}$.
4. What does it mean when two events are complements?

In Problems 5–12, a probability experiment is conducted in which the sample space of the experiment is $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Let event $E = \{2, 3, 4, 5, 6, 7\}$, event $F = \{5, 6, 7, 8, 9\}$, event $G = \{9, 10, 11, 12\}$, and event $H = \{2, 3, 4\}$. Assume that each outcome is equally likely.

5. List the outcomes in E and F . Are E and F mutually exclusive?
6. List the outcomes in F and G . Are F and G mutually exclusive?
7. List the outcomes in F or G . Now find $P(F \text{ or } G)$ by counting the number of outcomes in F or G . Determine $P(F \text{ or } G)$ using the General Addition Rule.
8. List the outcomes in E or H . Now find $P(E \text{ or } H)$ by counting the number of outcomes in E or H . Determine $P(E \text{ or } H)$ using the General Addition Rule.
9. List the outcomes in E and G . Are E and G mutually exclusive?
10. List the outcomes in F and H . Are F and H mutually exclusive?
11. List the outcomes in E^c . Find $P(E^c)$.
12. List the outcomes in F^c . Find $P(F^c)$.

In Problems 13–18, find the probability of the indicated event if $P(E) = 0.25$ and $P(F) = 0.45$.

13. $P(E \text{ or } F)$ if $P(E \text{ and } F) = 0.15$
14. $P(E \text{ and } F)$ if $P(E \text{ or } F) = 0.6$
15. $P(E \text{ or } F)$ if E and F are mutually exclusive
16. $P(E \text{ and } F)$ if E and F are mutually exclusive
17. $P(E^c)$
18. $P(F^c)$
19. If $P(E) = 0.60$, $P(E \text{ or } F) = 0.85$, and $P(E \text{ and } F) = 0.05$, find $P(F)$.
20. If $P(F) = 0.30$, $P(E \text{ or } F) = 0.65$, and $P(E \text{ and } F) = 0.15$, find $P(E)$.

In Problems 21–24, a golf ball is selected at random from a golf bag. If the golf bag contains 9 Titleists, 8 Maxflis, and 3 Top Flites, find the probability that the golf ball is:

21. A Titleist or Maxfli.
22. A Maxfli or Top Flite.
23. Not a Titleist.
24. Not a Top Flite.

Applying the Concepts

- NW 25. Baseball Injuries** The following probability model shows the distribution of injuries of youth baseball players, ages 5–14, according to researchers at SportsMedBC.

Injury Location	Probability
Head	0.11
Face	0.33
Wrist	0.05
Hand	0.16
Knee/Ankle	0.11
Other	0.24

Source: <https://sportsmedbc.com/article/baseball-injuries>

- (a) Verify that this is a probability model.
 (b) What is the probability that a randomly selected baseball injury to a 5–14-year-old is the head or face? Interpret this probability.
 (c) What is the probability that a randomly selected baseball injury to a 5–14-year-old is the head, face, or wrist? Interpret this probability.
 (d) What is the probability that a randomly selected baseball injury to a 5–14-year-old is something other than the face? Interpret this probability.

26. Family Structure The following probability model shows the distribution of family structure among families with at least one child younger than 18 years of age.

Family Structure	Probability
Two married parents, first marriage	0.46
Two married parents, one or both remarried	0.15
Single parent	0.26
Cohabiting parents	0.08
No parent at home	0.05

Source: Pew Research.

- (a) Verify that this is a probability model.
 (b) What is the probability that a randomly selected family with at least one child younger than 18 years of age has two married parents in their first marriage? Interpret this probability.
 (c) What is the probability that a randomly selected family with at least one child younger than 18 years of age has two married parents? Interpret this probability.
 (d) What is the probability that a randomly selected family with at least one child younger than 18 years of age has at least one parent at home? Interpret this probability.

27. If events E and F are disjoint and the events F and G are disjoint, must the events E and G necessarily be disjoint? Give an example to illustrate your opinion.

28. Draw a Venn diagram like that in Figure 8 that expands the General Addition Rule to three events. Use the diagram to write the General Addition Rule for three events.

NW 29. Medicare Fines In an effort to reduce the number of hospital-acquired conditions (such as infection resulting from the hospital stay), Medicare officials score hospitals on a 10-point scale with a lower score representing a better patient track record. The federal government reduces Medicare payments to those hospitals with the worst scores. The following data represent the scores received by Illinois hospitals.

Score	Frequency
1–1.9	3
2–2.9	12
3–3.9	16
4–4.9	23
5–5.9	23
6–6.9	21
7–7.9	17
8–8.9	5
9–10	5
Total	125

Source: Center for Medicare and Medicaid Services.

- (a) Determine the probability that a randomly selected hospital in Illinois has a score between 5 and 5.9.
 (b) Determine the probability that a randomly selected hospital in Illinois has a score that is not between 5 and 5.9.
 (c) Determine the probability that a randomly selected hospital in Illinois has a score less than 9.
 (d) Suppose Medicare reduces Medicare payments to any hospital with a score of 8 or higher. What is the probability a randomly selected hospital in Illinois will experience reduced Medicare payments? Interpret this result. Is it unusual?

30. Housing The following data represents the number of rooms in a random sample of U.S. housing units.

Rooms	Frequency
One	5
Two	11
Three	88
Four	183
Five	230
Six	204
Seven	123
Eight or more	156

Source: U.S. Census Bureau.

- (a) What is the probability that a randomly selected housing unit has four or more rooms? Interpret this probability.
 (b) What is the probability that a randomly selected housing unit has fewer than eight rooms? Interpret this probability.
 (c) What is the probability that a randomly selected housing unit has from four to six (inclusive) rooms? Interpret this probability.
 (d) What is the probability that a randomly selected housing unit has at least two rooms? Interpret this probability.

NW 31. A Deck of Cards A standard deck of cards contains 52 cards, as shown in Figure 7. One card is randomly selected from the deck.

- (a) Compute the probability of randomly selecting a heart or club from a deck of cards.
 (b) Compute the probability of randomly selecting a heart or club or diamond from a deck of cards.
 (c) Compute the probability of randomly selecting an ace or heart from a deck of cards.

32. A Deck of Cards A standard deck of cards contains 52 cards, as shown in Figure 7. One card is randomly selected from the deck.

- (a) Compute the probability of randomly selecting a two or three from a deck of cards.
 (b) Compute the probability of randomly selecting a two or three or four from a deck of cards.
 (c) Compute the probability of randomly selecting a two or club from a deck of cards.

33. Birthdays Exclude leap years from the following calculations:

- (a) Compute the probability that a randomly selected person does not have a birthday on November 8.
 (b) Compute the probability that a randomly selected person does not have a birthday on the 1st day of a month.
 (c) Compute the probability that a randomly selected person does not have a birthday on the 31st day of a month.
 (d) Compute the probability that a randomly selected person was not born in December.

34. Roulette In the game of roulette, a wheel consists of 38 slots numbered 0, 00, 1, 2, . . . , 36. The odd-numbered slots are red, and the even-numbered slots are black. The numbers 0 and 00 are green. To play the game, a metal ball is spun around the wheel and is allowed to fall into one of the numbered slots.

- (a) What is the probability that the metal ball lands on green or red?
- (b) What is the probability that the metal ball does not land on green?

35. Health Problems According to the Centers for Disease Control, the probability that a randomly selected citizen of the United States has hearing problems is 0.151. The probability that a randomly selected citizen of the United States has vision problems is 0.093. Can we compute the probability of randomly selecting a citizen of the United States who has hearing problems or vision problems by adding these probabilities? Why or why not?

36. Visits to the Doctor A National Ambulatory Medical Care Survey administered by the Centers for Disease Control found that the probability a randomly selected patient visited the doctor for a blood pressure check is 0.593. The probability a randomly selected patient visited the doctor for urinalysis is 0.064. Can we compute the probability of randomly selecting a patient who visited the doctor for a blood pressure check or urinalysis by adding these probabilities? Why or why not?

37. Language Spoken at Home According to the U.S. Census Bureau, the probability that a randomly selected household speaks only English at home is 0.784. The probability that a randomly selected household speaks only Spanish at home is 0.123.

- (a) What is the probability that a randomly selected household speaks only English or only Spanish at home?
- (b) What is the probability that a randomly selected household speaks a language other than only English or only Spanish at home?
- (c) What is the probability that a randomly selected household speaks a language other than only English at home?
- (d) Can the probability that a randomly selected household speaks only Polish at home equal 0.103? Why or why not?

38. Getting to Work According to the U.S. Census Bureau, the probability that a randomly selected worker primarily drives a car to work is 0.764. The probability that a randomly selected worker primarily takes public transportation to work is 0.051.

- (a) What is the probability that a randomly selected worker primarily drives a car or takes public transportation to work?
- (b) What is the probability that a randomly selected worker neither drives a car nor takes public transportation to work?
- (c) What is the probability that a randomly selected worker does not take public transportation to work?
- (d) Can the probability that a randomly selected worker works at home equal 0.15? Why or why not?

NW 39. Cigar Smoking The data in the following table show the results of a national study of 137,243 U.S. men that investigated the association between cigar smoking and death from cancer. **Note:** Current cigar smoker means cigar smoker at time of death.

	Died from Cancer	Did Not Die from Cancer
Never smoked cigars	782	120,747
Former cigar smoker	91	7,757
Current cigar smoker	141	7,725

Source: Shapiro, Jacobs, and Thun. "Cigar Smoking in Men and Risk of Death from Tobacco-Related Cancers," *Journal of the National Cancer Institute*, February 16, 2000.

- (a) If an individual is randomly selected from this study, what is the probability that he died from cancer?
- (b) If an individual is randomly selected from this study, what is the probability that he was a current cigar smoker?
- (c) If an individual is randomly selected from this study, what is the probability that he died from cancer and was a current cigar smoker?
- (d) If an individual is randomly selected from this study, what is the probability that he died from cancer or was a current cigar smoker?

40. Working Couples A guidance counselor at a middle school collected the following information regarding the employment status of married couples within his school's boundaries.

Worked	Number of Children under 18 Years Old			
	0	1	2 or More	Total
Husband only	172	79	174	425
Wife only	94	17	15	126
Both spouses	522	257	370	1149
Total	788	353	559	1700

- (a) What is the probability that, for a married couple selected at random, both spouses work?
- (b) What is the probability that, for a married couple selected at random, the couple has one child under the age of 18?
- (c) What is the probability that, for a married couple selected at random, the couple has two or more children under the age of 18 and both spouses work?
- (d) What is the probability that, for a married couple selected at random, the couple has no children or only the husband works?
- (e) Would it be unusual to select a married couple at random for which only the wife works?

41. The Placebo Effect A company is testing a new medicine for migraine headaches. In the study, 150 women were given the new medicine and 100 women were given a placebo. Each participant was directed to take the medicine when the first symptoms of a migraine occurred and then to record whether the headache went away within 45 minutes or lingered. The results are recorded in the following table:

	Headache Went Away	Headache Lingered
Given medicine	132	18
Given placebo	56	44

- (a) If a study participant is selected at random, what is the probability she was given the placebo?
- (b) If a study participant is selected at random, what is the probability her headache went away within 45 minutes?

- (c) If a study participant is selected at random, what is the probability she was given the placebo and her headache went away within 45 minutes?
- (d) If a study participant is selected at random, what is the probability she was given the placebo or her headache went away within 45 minutes?

42. Social Media Harris Interactive conducted a survey in which they asked adult Americans (18 years or older) whether they used social media (Facebook, Twitter, and so on) regularly. The following table is based on the results of the survey.

	18–34	35–44	45–54	55 +	Total
Use social media	117	89	83	49	338
Do not use social media	33	36	57	66	192
Total	150	125	140	115	530

Source: Harris Interactive.

- (a) If an adult American is randomly selected, what is the probability he or she uses social media?
- (b) If an adult American is randomly selected, what is the probability he or she is 45 to 54 years of age?
- (c) If an adult American is randomly selected, what is the probability he or she is a 35- to 44-year-old social media user?
- (d) If an adult American is randomly selected, what is the probability he or she is 35 to 44 years old or uses social media?

43. Driver Fatalities The following data represent the number of drivers involved in a fatal crash in 2016 in various light and weather conditions.

Weather	Light Condition				
	Daylight	Dark, but Lighted	Dark	Dawn/ Dusk	Other
Normal	14,307	5875	8151	1183	65
Rain	875	497	681	87	8
Snow/Sleet	219	51	156	17	2
Other	125	54	220	40	9
Unknown	810	255	548	71	133

Source: Fatality Analysis Reporting System.

- (a) Determine the probability that a randomly selected fatal crash in 2016 occurred in normal weather.
- (b) Determine the probability that a randomly selected fatal crash in 2016 occurred in daylight.
- (c) Determine the probability that a randomly selected fatal crash in 2016 occurred in normal weather and in daylight.
- (d) Determine the probability that a randomly selected fatal crash in 2016 occurred in normal weather or in daylight.
- (e) Would it be unusual for a fatal crash in 2016 to occur while it is dark outside (without light) and raining? Why might this result be considered misleading?



44. Putting It Together: Speeding Tickets Go to www.pearsonhighered.com/sullivanstats to obtain the data file SullivanStatsSurveyI using the file format of your choice for the version of the text you are using. The data represent the results of a survey conducted by the author. The variable “Text while Driving” represents the response to the question, “Have you ever texted while driving?” The variable “Tickets” represents the response to the question, “How many speeding tickets have you received in the past 12 months?” Treat the individuals in the survey as a random sample of all U.S. drivers.

- (a) Build a contingency table treating “Text while Driving” as the row variable and “Tickets” as the column variable.
- (b) Determine the marginal relative frequency distribution for both the row and column variable.
- (c) What is the probability a randomly selected U.S. driver texts while driving?
- (d) What is the probability a randomly selected U.S. driver received three speeding tickets in the past 12 months?
- (e) What is the probability a randomly selected U.S. driver texts while driving or received three speeding tickets in the past 12 months?
- (f) Interpret the marginal relative frequency distribution for the column variable, “Tickets.”

45. Putting It Together: Red Light Cameras In a study of the feasibility of a red-light camera program in the city of Milwaukee, the data below summarize the projected number of crashes at 13 selected intersections over a five-year period.

Crash Type	Current System	With Red-Light Cameras
Reported injury	289	221
Reported property damage only	392	333
Unreported injury	78	60
Unreported property damage only	362	308
Total	1121	922

Source: Krig, Moran, Regan. “An Analysis of a Red-Light Camera Program in the City of Milwaukee,” Spring 2006, prepared for the city of Milwaukee Budget and Management Division.

- (a) Identify the variables presented in the table.
- (b) State whether each variable is qualitative or quantitative. If quantitative, state whether it is discrete or continuous.
- (c) Construct a relative frequency distribution for each system.
- (d) Construct a side-by-side relative frequency bar graph for the data.
- (e) Determine the mean number of crashes per intersection in the study, if possible. If not possible, explain why.
- (f) Determine the standard deviation number of crashes, if possible. If not possible, explain why.
- (g) Based on the data shown, does it appear that the red-light camera program will be beneficial in reducing crashes at the intersections? Explain.
- (h) For the current system, what is the probability that a crash selected at random will have reported injuries?
- (i) For the camera system, what is the probability that a crash selected at random will have only property damage?

The study classified crashes further by indicating whether they were red-light running crashes or rear-end crashes. The results are as follows:

Crash Type	Rear End		Red-Light Running	
	Current	Cameras	Current	Cameras
Reported injury	67	77	222	144
Reported property damage only	157	180	235	153
Unreported injury	18	21	60	39
Unreported property damage only	145	167	217	141
Total	387	445	734	477

- (j) Using Simpson's Paradox, explain how the additional classification affects your response to part (g).
 (k) What recommendation would you make to the city council regarding the implementation of the red-light camera program? Would you need any additional information before making your recommendation? Explain.

Retain Your Knowledge



- 46. Exam Scores** The following data represent the homework scores for the material on Polynomial and Rational Functions in Sullivan's College Algebra course.

37	67	76	82	89
48	70	77	83	90
54	72	77	84	90
59	73	77	84	91
61	75	77	84	92
65	75	78	85	95
65	76	80	87	95
67	76	81	88	98

Source: Michael Sullivan, MyMathLab.

- (a) Construct a relative frequency distribution with a lower class limit of the first class equal to 30 and a class width of 10.
- (b) Draw a relative frequency histogram of the data.
- (c) Determine the mean and median score.
- (d) Draw a boxplot of the data. Are there any outliers?
- (e) Describe the shape of the distribution based on the results from parts (b) through (d).
- (f) Determine the standard deviation and interquartile range.
- (g) What is the probability a randomly selected student fails the homework (scores less than 60)?
- (h) What is the probability a randomly selected student earns an A or B on the homework (scores 80 or higher)?
- (i) What is the probability a randomly selected student scores less than 30 on the homework?

5.3 Independence and the Multiplication Rule



Objectives

- ① Identify independent events
- ② Use the Multiplication Rule for Independent Events
- ③ Compute at-least probabilities

1 Identify Independent Events

The Addition Rule for Disjoint Events is used to compute the probability of observing an outcome in event E or event F . We now describe a probability rule for computing the probability that E and F both occur.

Before we can present this rule, we must discuss the idea of *independent events*.

Definitions

Two events E and F are **independent** if the occurrence of event E in a probability experiment does not affect the probability of event F . Two events are **dependent** if the occurrence of event E in a probability experiment affects the probability of event F .

Think about flipping a fair coin twice. Does the fact that you obtained a head on the first toss have any effect on the likelihood of obtaining a head on the second toss? Not unless you are a master coin flipper who can manipulate the outcome of a coin flip! For this reason, the outcome from the first flip is independent of the outcome from the second flip. Let's look at other examples.

EXAMPLE 1 Independent or Not?**IN OTHER WORDS**

In determining whether two events are independent, ask yourself whether the probability of one event is affected by the other event. For example, what is the probability that a 29-year-old male has high cholesterol? What is the probability that a 29-year-old male has high cholesterol, given that he eats fast food four times a week? Does the fact that the individual eats fast food four times a week change the likelihood that he has high cholesterol? If yes, the events are not independent.

NW Now Work Problem 7**CAUTION!**

Two events that are disjoint are not independent.

- (a) Suppose you flip a coin and roll a die. The events “obtain a head” and “roll a 5” are independent because the results of the coin flip do not affect the results of the die toss.
- (b) Are the events “earned a bachelor’s degree” and “earn more than \$100,000 per year” independent? No, because knowing that an individual has a bachelor’s degree affects the likelihood that the individual is earning more than \$100,000 per year.
- (c) Two 24-year-old male drivers who live in the United States are randomly selected. The events “male 1 gets in a car accident during the year” and “male 2 gets in a car accident during the year” are independent because the males were randomly selected. This means what happens with one of the drivers has nothing to do with what happens to the other driver.



In Example 1(c), we are able to conclude that the events “male 1 gets in an accident” and “male 2 gets in an accident” are independent because the individuals are randomly selected. By randomly selecting the individuals, it is reasonable to conclude that the individuals are not related in any way (related in the sense that they do not live in the same town, attend the same school, and so on). If the two individuals did have a common link between them (such as they both lived on the same city block), then knowing that one male had a car accident may affect the likelihood that the other male had a car accident. After all, they could hit each other!

Disjoint Events versus Independent Events Disjoint events and independent events are different concepts. Recall that two events are disjoint if they have no outcomes in common, that is, if knowing that one of the events occurs, we know the other event did not occur. Independence means that one event occurring does not affect the probability of the other event occurring. Therefore, knowing two events are disjoint means that the events are not independent.

Consider the experiment of rolling a single die. Let E represent the event “roll an even number,” and let F represent the event “roll an odd number.” We can see that E and F are mutually exclusive (disjoint) because they have no outcomes in common.

In addition, $P(E) = \frac{1}{2}$ and $P(F) = \frac{1}{2}$. However, if we are told that the roll of the die is going to be an even number, then what is the probability of event F ? Because the outcome will be even, the probability of event F is now 0 (and the probability of event E is now 1). So knowledge of event E changes the likelihood of observing event F .

② Use the Multiplication Rule for Independent Events

Suppose that you flip a fair coin twice. What is the probability that you obtain a head on both flips, that is, a head on the first flip *and* you obtain a head on the second flip? If H represents the outcome “heads” and T represents the outcome “tails,” the sample space of this experiment is

$$S = \{\text{HH, HT, TH, TT}\}$$

There is one outcome with both heads. Because each outcome is equally likely, we have

$$\begin{aligned} P(\text{heads on Flip 1 and heads on Flip 2}) &= \frac{N(\text{heads on Flip 1 and heads on Flip 2})}{N(S)} \\ &= \frac{1}{4} \end{aligned}$$

We may have intuitively figured this out by recognizing $P(\text{head}) = \frac{1}{2}$ for each flip. So it seems reasonable that

$$\begin{aligned}
 P(\text{heads on Flip 1 and heads on Flip 2}) &= P(\text{heads on Flip 1}) \cdot P(\text{heads on Flip 2}) \\
 &= \frac{1}{2} \cdot \frac{1}{2} \\
 &= \frac{1}{4}
 \end{aligned}$$

Because both approaches result in the same answer, $\frac{1}{4}$, we conjecture that $P(E \text{ and } F) = P(E) \cdot P(F)$, which is true.

Multiplication Rule for Independent Events

If E and F are independent events, then

$$P(E \text{ and } F) = P(E) \cdot P(F)$$

EXAMPLE 2

Computing Probabilities of Independent Events



Problem In the game of roulette, the wheel has slots numbered 0, 00, and 1 through 36. A metal ball rolls around a wheel until it falls into one of the numbered slots. What is the probability that the ball will land in the slot numbered 21 two times in a row?

Approach The sample space of the experiment has 38 outcomes. We use the classical method of computing probabilities because the outcomes are equally likely. In addition, we use the Multiplication Rule for Independent Events. The events “21 on Spin 1” and “21 on Spin 2” are independent because the ball does not remember it landed on 21 on the first spin, so this cannot affect the probability of landing on 21 on the second spin.

Solution There are 38 possible outcomes, so the probability of landing on 21 is $\frac{1}{38}$. Because the events “21 on Spin 1” and “21 on Spin 2” are independent, we have

$$\begin{aligned}
 P(21 \text{ on Spin 1 and } 21 \text{ on Spin 2}) &= P(21 \text{ on Spin 1}) \cdot P(21 \text{ on Spin 2}) \\
 &= \frac{1}{38} \cdot \frac{1}{38} = \frac{1}{1444} \approx 0.0006925
 \end{aligned}$$

NOTE

$$0.00006925 \approx 0.00007 \text{ and } 0.00007 = \frac{7}{10,000}$$

We expect the ball to land on 21 twice in a row about 7 times in 10,000 trials.

We can extend the Multiplication Rule for three or more independent events.

Multiplication Rule for n Independent Events

If events $E_1, E_2, E_3, \dots, E_n$ are independent, then

$$P(E_1 \text{ and } E_2 \text{ and } E_3 \text{ and } \dots \text{ and } E_n) = P(E_1) \cdot P(E_2) \cdot \dots \cdot P(E_n)$$

EXAMPLE 3

Life Expectancy

Problem The probability that a randomly selected 24-year-old male will survive the year is 0.9986 according to the *National Vital Statistics Report*, Vol. 56, No. 9.

- (a) What is the probability that three randomly selected 24-year-old males will survive the year?
- (b) What is the probability that 20 randomly selected 24-year-old males will survive the year?

(continued)

IN OTHER WORDS

In Example 3, if two of the males lived in the same house, a house fire could kill both males and we lose independence. (Knowledge that one male died in a house fire certainly affects the probability that the other died.) By randomly selecting the males, we minimize the chances that they are related in any way.

Approach It is safe to assume that the outcomes of the probability experiment are independent, because there is no indication that the survival of one male affects the survival of the others.

Solution

$$\begin{aligned}\text{(a)} \quad P(\text{all three males survive}) &= P(\text{1st survives and 2nd survives and 3rd survives}) \\ &= P(\text{1st survives}) \cdot P(\text{2nd survives}) \cdot P(\text{3rd survives}) \\ &= (0.9986)(0.9986)(0.9986) \\ &= 0.9958\end{aligned}$$

If we randomly selected three 24-year-old males 1000 different times, we would expect all three to survive one year in 996 of the samples.

$$\begin{aligned}\text{(b)} \quad P(\text{all 20 males survive}) &= P(\text{1st survives and 2nd survives and ... and 20th survives}) \\ &= P(\text{1st survives}) \cdot P(\text{2nd survives}) \cdots \cdot P(\text{20th survives}) \\ &\quad \text{Multiply 0.9986 by itself 20 times.} \\ &= (0.9986) \cdot (0.9986) \cdots \cdot (0.9986) \\ &= (0.9986)^{20} \\ &= 0.9724\end{aligned}$$

NW Now Work Problems 17(a) and (b)

If we randomly selected twenty 24-year-old males 1000 different times, we would expect all twenty to survive one year in 972 of the samples.



③ Compute At-Least Probabilities

Usually probabilities involving the phrase *at least* use the Complement Rule.

The phrase *at least* means “greater than or equal to.” For example, a person must be at least 17 years old to see an R-rated movie. This means that the person’s age must be greater than or equal to 17 to watch the movie.

EXAMPLE 4

Computing At-Least Probabilities

Problem Compute the probability that at least one male out of 1000 aged 24 years will die during the course of the year if the probability that a randomly selected 24-year-old male survives the year is 0.9986.

Approach The phrase *at least* means “greater than or equal to,” so we wish to know the probability that 1 or 2 or 3 or ... or 1000 males will die during the year. These events are mutually exclusive, so

$$\begin{aligned}P(\text{1 or 2 or 3 or ... or 1000 die}) &= P(\text{1 dies}) + P(\text{2 die}) + P(\text{3 die}) \\ &\quad + \cdots + P(\text{1000 die})\end{aligned}$$

Computing these probabilities would be very time consuming. However, notice that the complement of “at least one dying” is “none die,” or all 1000 survive. Use the Complement Rule to compute the probability.

Solution

$$\begin{aligned}P(\text{at least one dies}) &= 1 - P(\text{none die}) \\ &= 1 - P(\text{1st survives and 2nd survives and ... and 1000th survives}) \\ &= 1 - P(\text{1st survives}) \cdot P(\text{2nd survives}) \cdots \cdot P(\text{1000th survives}) \\ &= 1 - (0.9986)^{1000} \\ &\quad \text{Independent events} \\ &= 1 - 0.2464 \\ &= 0.7536\end{aligned}$$

If we randomly selected 1000 males 24 years of age 100 different times, we would expect at least one to die in 75 of the samples.



NW Now Work Problem 17(c)

Summary: Rules of Probability

1. The probability of any event must be between 0 and 1, inclusive. If we let E denote any event, then $0 \leq P(E) \leq 1$.
2. The sum of the probabilities of all outcomes in the sample space must equal 1. That is, if the sample space $S = \{e_1, e_2, \dots, e_n\}$, then $P(e_1) + P(e_2) + \dots + P(e_n) = 1$.
3. If E and F are disjoint events, then $P(E \text{ or } F) = P(E) + P(F)$. If E and F are not disjoint events, then $P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$.
4. If E represents any event and E^c represents the complement of E , then $P(E^c) = 1 - P(E)$.
5. If E and F are independent events, then $P(E \text{ and } F) = P(E) \cdot P(F)$.

Notice that *or* probabilities use the Addition Rule, whereas *and* probabilities use the Multiplication Rule. Accordingly, *or* probabilities imply addition, while *and* probabilities imply multiplication.



5.3 Assess Your Understanding

Vocabulary and Skill Building

1. Two events E and F are _____ if the occurrence of event E in a probability experiment does not affect the probability of event F .
 2. The word *and* in probability implies that we use the _____ Rule.
 3. The word *or* in probability implies that we use the _____ Rule.
 4. *True or False:* When two events are disjoint, they are also independent.
 5. If two events E and F are independent, $P(E \text{ and } F) = _____$.
 6. Suppose events E and F are disjoint. What is $P(E \text{ and } F)$?
- NW** 7. Determine whether the events E and F are independent or dependent. Justify your answer.
- (a) E : Speeding on the interstate.
 F : Being pulled over by a police officer.
 - (b) E : You gain weight.
 F : You eat fast food for dinner every night.
 - (c) E : You get a high score on a statistics exam.
 F : The Boston Red Sox win a baseball game.
8. Determine whether the events E and F are independent or dependent. Justify your answer.
- (a) E : The battery in your cell phone is dead.
 F : The batteries in your calculator are dead.
 - (b) E : Your favorite color is blue.
 F : Your friend's favorite hobby is fishing.
 - (c) E : You are late for school.
 F : Your car runs out of gas.
9. Suppose that events E and F are independent, $P(E) = 0.3$ and $P(F) = 0.6$. What is the $P(E \text{ and } F)$?
10. Suppose that events E and F are independent, $P(E) = 0.7$ and $P(F) = 0.9$. What is the $P(E \text{ and } F)$?

Applying the Concepts

11. **Flipping a Coin** What is the probability of obtaining five heads in a row when flipping a fair coin? Interpret this probability.

12. **Rolling a Die** What is the probability of obtaining 4 ones in a row when rolling a fair, six-sided die? Interpret this probability.

13. **Southpaws** About 13% of the population is left-handed. If two people are randomly selected, what is the probability that both are left-handed? What is the probability that at least one is right-handed?

14. **Double Jackpot** Shawn lives near the border of Illinois and Missouri. One weekend he decides to play \$1 in both state lotteries in hopes of hitting two jackpots. The probability of winning the Missouri Lotto is about 0.00000028357 and the probability of winning the Illinois Lotto is about 0.000000098239.

- (a) Explain why the two lotteries are independent.
- (b) Find the probability that Shawn will win both jackpots.

15. **False Positives** The ELISA is a test to determine whether the HIV antibody is present. The test is 99.5% effective, which means that the test will come back negative if the HIV antibody is not present 99.5% of the time. The probability of a test coming back positive when the antibody is not present (a false positive) is 0.005. Suppose that the ELISA is given to five randomly selected people who do not have the HIV antibody.

- (a) What is the probability that the ELISA comes back negative for all five people?
- (b) What is the probability that the ELISA comes back positive for at least one of the five people?

16. **Christmas Lights** Christmas lights are often designed with a series circuit. This means that when one light burns out the entire string of lights goes black. Suppose that the lights are designed so that the probability a bulb will last 2 years is 0.995. The success or failure of a bulb is independent of the success or failure of other bulbs.

- (a) What is the probability that in a string of 100 lights all 100 will last 2 years?
- (b) What is the probability that at least one bulb will burn out in 2 years?

NW 17. Life Expectancy The probability that a randomly selected 40-year-old male will live to be 41 years old is 0.99757, according to the *National Vital Statistics Report*, Vol. 56, No. 9.

- (a) What is the probability that two randomly selected 40-year-old males will live to be 41 years old?
- (b) What is the probability that five randomly selected 40-year-old males will live to be 41 years old?
- (c) What is the probability that at least one of five randomly selected 40-year-old males will not live to be 41 years old? Would it be unusual if at least one of five randomly selected 40-year-old males did not live to be 41 years old?

18. Earn More Than Your Parents? In 1970, 92% of American 30-year-olds earned more than their parents did at age 30 (adjusted for inflation). In 2014, only 51% of American 30-year-olds earned more than their parents did at age 30.

Source: Wall Street Journal, December 8, 2016.

- (a) What is the probability a randomly selected 30-year-old in 1970 earned more than his or her parents at age 30?
- (b) What is the probability that two randomly selected 30-year-olds in 1970 earned more than their parents at age 30?
- (c) What is the probability that out of ten randomly selected 30-year-olds in 1970, at least one did not earn more than his or her parents at age 30?
- (d) What is the probability that out of ten randomly selected 30-year-olds in 2014, at least one did not earn more than his or her parents at age 30?

19. Derivatives In finance, a derivative is a financial asset whose value is determined (derived) from a bundle of various assets, such as mortgages. Suppose a randomly selected mortgage has a probability of 0.01 of default.

- (a) What is the probability a randomly selected mortgage will not default (that is, pay off)?
- (b) What is the probability a bundle of five randomly selected mortgages will not default assuming the likelihood any one mortgage being paid off is independent of the others?
- Note:** A derivative might be an investment in which all five mortgages do not default.
- (c) What is the probability the derivative becomes worthless? That is, at least one of the mortgages defaults?
- (d) In part (b), we made the assumption that the likelihood of default is independent. Do you believe this is a reasonable assumption? Explain.

20. Quality Control Suppose that a company selects two people who work independently inspecting two-by-four timbers. Their job is to identify low-quality timbers. Suppose that the probability that an inspector does not identify a low-quality timber is 0.20.

- (a) What is the probability that both inspectors do not identify a low-quality timber?
- (b) How many inspectors should be hired to keep the probability of not identifying a low-quality timber below 1%?
- (c) Interpret the probability from part (a).

21. Reliability and Redundancy In airline applications, failure of a component can result in catastrophe. As a result, many airline components utilize something called *triple modular redundancy*. This means that a critical component has two backup components that may be utilized should the initial component fail. Suppose a certain critical airline component has a probability of failure of 0.006 and the

system that utilizes the component is part of a triple modular redundancy.

- (a) Assuming each component's failure/success is independent of the others, what is the probability all three components fail, resulting in disaster for the flight?
- (b) What is the probability at least one of the components does not fail?

22. Reliability For a parallel structure of identical components, the system can succeed if at least one of the components succeeds. Assume that components fail independently of each other and that each component has a 0.15 probability of failure.

- (a) Would it be unusual to observe one component fail? Two components?
- (b) What is the probability that a parallel structure with 2 identical components will succeed?
- (c) How many components would be needed in the structure so that the probability the system will succeed is greater than 0.9999?

23. Reliability and Redundancy, Part II

See Problem 21. Suppose a particular airline component has a probability of failure of 0.03 and is part of a triple modular redundancy system.

- (a) What is the probability the system does not fail?
- (b) Engineers decide the probability of failure is too high for this system. Use trial and error to determine the number of components that should be included in the system to result in a system that has greater than a 0.99999999 probability of not failing.

24. Cold Streaks Players in sports are said to have "hot streaks" and "cold streaks." For example, a batter in baseball might be considered to be in a slump, or cold streak, if he has made 10 outs in 10 consecutive at-bats. Suppose that a hitter successfully reaches base 30% of the time he comes to the plate.

- (a) Find and interpret the probability that the hitter makes 10 outs in 10 consecutive at-bats, assuming that at-bats are independent events. **Hint:** The hitter makes an out 70% of the time.
- (b) Are cold streaks unusual?
- (c) Find the probability the hitter makes five consecutive outs and then reaches base safely.
- (d) Discuss the assumption of independence in consecutive at-bats.

25. Bowling Suppose that Ralph gets a strike when bowling 30% of the time.

- (a) What is the probability that Ralph gets two strikes in a row?
- (b) What is the probability that Ralph gets a turkey (three strikes in a row)?
- (c) When events are independent, their complements are independent as well. Use this result to determine the probability that Ralph gets a turkey, but fails to get a clover (four strikes in a row).

26. Driving under the Influence Among 21- to 25-year-olds, 29% say they have driven while under the influence of alcohol. Suppose that three 21- to 25-year-olds are selected at random. *Source: U.S. Department of Health and Human Services, reported in USA Today.*

- (a) What is the probability that all three have driven while under the influence of alcohol?
- (b) What is the probability that at least one has not driven while under the influence of alcohol?

- (c) What is the probability that none of the three has driven while under the influence of alcohol?
- (d) What is the probability that at least one has driven while under the influence of alcohol?

27. Defense System Suppose that a satellite defense system is established in which four satellites acting independently have a 0.9 probability of detecting an incoming ballistic missile. What is the probability that at least one of the four satellites detects an incoming ballistic missile? Would you feel safe with such a system?

28. Audits and Pet Ownership According to Internal Revenue Service records, 6.42% of all household tax returns are audited. According to the Humane Society, 39% of all households own a dog. Assuming dog ownership and audits are independent events, what is the probability a randomly selected household is audited and owns a dog?

29. Weight Gain and Gender According to the National Vital Statistics Report, 20.1% of all pregnancies result in weight gain in excess of 40 pounds (for singleton births). In addition, 49.5% of all pregnancies result in the birth of a baby girl. Assuming gender and weight gain are independent, what is the probability a randomly selected pregnancy results in a girl and weight gain in excess of 40 pounds?

30. Stocks Suppose your financial advisor recommends three stocks to you. He claims the likelihood that the first stock will increase in value at least 10% within the next year is 0.7, the likelihood the second stock will increase in value at least 10% within the next year is 0.55, and the likelihood the third stock will increase at least 10% within the next year is 0.20. Would it be unusual for all three stocks to increase at least 10%, assuming the stocks behave independently of each other?

31. Betting on Sports According to a Gallup Poll, about 17% of adult Americans bet on professional sports. Census data indicate that 48.4% of the adult population in the United States is male.

- (a) Assuming that betting is independent of gender, compute the probability that an American adult selected at random is male and bets on professional sports.
- (b) Using the result in part (a), compute the probability that an American adult selected at random is male or bets on professional sports.
- (c) The Gallup poll data indicated that 10.6% of adults in the United States are males and bet on professional sports. What does this indicate about the assumption in part (a)?
- (d) How will the information in part (c) affect the probability you computed in part (b)?

32. Fingerprints Fingerprints are now widely accepted as a form of identification. In fact, many computers today use fingerprint identification to link the owner to the computer. In 1892, Sir Francis Galton explored the use of fingerprints to uniquely identify an individual. A fingerprint consists of ridgelines. Based on empirical evidence, Galton estimated the probability that a square consisting of six ridgelines that covered a fingerprint could be filled in accurately by an experienced fingerprint analyst as $\frac{1}{2}$.

- (a) Assuming that a full fingerprint consists of 24 of these squares, what is the probability that all 24 squares could be filled in correctly, assuming that success or failure in filling

in one square is independent of success or failure in filling in any other square within the region? (This value represents the probability that two individuals would share the same ridgeline features within the 24-square region.)

- (b) Galton further estimated that the likelihood of determining the fingerprint type (e.g., arch, left loop, whorl, etc.) as $\left(\frac{1}{2}\right)^4$ and the likelihood of the occurrence of the correct number of ridges entering and exiting each of the 24 regions as $\left(\frac{1}{2}\right)^8$. Assuming that all three probabilities are independent, compute Galton's estimate of the probability that a particular fingerprint configuration would occur in nature (that is, the probability that a fingerprint match occurs by chance).

33. You Explain It! Independence Suppose a mother already has three girls from three separate pregnancies. Does the fact that the mother already has three girls affect the likelihood of having a fourth girl? Explain.

34. You Explain It! Independence Ken and Dorothy like to fly to Colorado for ski vacations. Sometimes, however, they are late for their flight. On the air carrier they prefer to fly, the probability luggage gets lost is 0.012 for luggage checked at least one hour prior to departure. However, the probability luggage gets lost is 0.043 for luggage checked within one hour of departure. Are the events "luggage check time" and "lost luggage" independent? Explain.

35. A Random Process—The Lady Tasting Tea Ronald Fisher is considered the father of experimental design. Being of English descent, he was having afternoon tea with a colleague. The colleague's wife entered the room as Fisher was pouring tea. Fisher offered tea to the lady. She politely accepted and requested milk with her tea. Fisher started to pour milk into the tea cup first, but the lady indicated that she preferred her tea be poured first, then the milk. Fisher did not believe that the lady could tell the difference between "milk first" versus "milk second" tea, but the lady insisted she could tell the difference. Being the consummate scientist, Fisher suggested an experiment in which he randomly put milk into the tea first in some instances, and milk into the tea second in others. It turns out, the lady tasting tea was correct in all eight trials.

- (a) If we assume that the lady was simply guessing on whether the milk was first or not, what is the probability she would guess correctly on any given cup?
- (b) Assuming guessing correctly on one cup is independent of guessing correctly on any other cup, what is the probability of guessing correctly on eight consecutive cups?
- (c) Explain how a coin could be used to simulate the random process of tasting eight cups of tea.
- (d) Use a coin to simulate Fisher's experiment at least 2000 times. Based on the simulation, what is the probability of guessing correctly on eight consecutive cups of tea? Compare this result to that of part (b). **Note:** You might consider using the coin-flipping applet available at www.pearsonhighered.com/sullivanstats to conduct the simulation.
- (e) What do the probabilities from parts (b) and (d) suggest about the lady tasting tea?

5.4 Conditional Probability and the General Multiplication Rule



Objectives ① Compute conditional probabilities

② Compute probabilities using the General Multiplication Rule

1 Compute Conditional Probabilities

In the last section, we learned that when two events are independent the occurrence of one event has no effect on the probability of the second event. However, we cannot always assume that two events will be independent. Will the probability of being in a car accident change depending on driving conditions? We would expect that the probability of an accident will be higher for driving on icy roads than for driving on dry roads.

According to data from the Centers for Disease Control, 33.3% of adult men in the United States are obese. So the probability is 0.333 that a randomly selected U.S. adult male is obese. However, 28% of adult men aged 20–39 are obese compared to 40% of adult men aged 40–59. The probability is 0.28 that an adult male is obese, *given* that he is aged 20–39. The probability is 0.40 that an adult male is obese, *given* that he is aged 40–59. The probability that an adult male is obese changes depending on his age group. Therefore, obesity and age are not independent. This is called *conditional probability*.

Definition

Conditional Probability

The notation $P(F|E)$ is read “the probability of event F given event E .” It is the probability that the event F occurs, given that the event E has occurred.

For example,

$$P(\text{obese} | 20 \text{ to } 39) = 0.28 \quad \text{and} \quad P(\text{obese} | 40 \text{ to } 59) = 0.40.$$

EXAMPLE 1

An Introduction to Conditional Probability

Problem Suppose a single die is rolled. What is the probability that the die comes up three? Now suppose that the die is rolled a second time, but we are told the outcome will be an odd number. What is the probability that the die comes up three?

Approach Assume that the die is fair. This means that the outcomes are equally likely, so we use the classical method of computing probabilities.

Solution In the first instance, there are six possibilities in the sample space, $S = \{1, 2, 3, 4, 5, 6\}$, so $P(3) = \frac{1}{6}$. In the second instance, there are three possibilities in the sample space, because the only possible outcomes are odd, so $S = \{1, 3, 5\}$.

This probability is expressed symbolically as $P(3 | \text{outcome is odd}) = \frac{1}{3}$, which is read “the probability of rolling a 3, given that the outcome is odd, is one-third.” Notice that the conditional probability reduces the size of the sample space under consideration (from six outcomes to three outcomes).

The data in Table 8 represent the marital status of males and females 15 years old or older in the United States in 2017.

Table 8

	Males (in millions)	Females (in millions)	Totals (in millions)
Never married	46.0	40.1	86.1
Married	63.3	62.0	125.3
Widowed	3.3	11.9	15.2
Divorced	12.1	16.1	28.2
Separated	2.2	3.1	5.3
Totals (in millions)	126.9	133.2	260.1

Source: U.S. Census Bureau, American Community Survey.

To find the probability that a randomly selected individual 15 years old or older is widowed, divide the number of widowed individuals by the total number of individuals who are 15 years old or older.

$$\begin{aligned} P(\text{widowed}) &= \frac{15.2}{260.1} \\ &= 0.058 \end{aligned}$$

Suppose that we know the individual is female. Does this change the probability that the individual is widowed? The sample space now consists only of females, so the probability that the individual is widowed, given that the individual is female, is

$$\begin{aligned} P(\text{widowed} | \text{female}) &= \frac{N(\text{widowed females})}{N(\text{females})} \\ &= \frac{11.9}{133.2} = 0.089 \end{aligned}$$

So, knowing that the individual is female increases the likelihood that the individual is widowed. This leads to the following rule.

Conditional Probability Rule

If E and F are any two events, then

$$P(F|E) = \frac{P(E \text{ and } F)}{P(E)} = \frac{N(E \text{ and } F)}{N(E)} \quad (1)$$

The probability of event F occurring, given the occurrence of event E , is found by dividing the probability of E and F by the probability of E , or by dividing the number of outcomes in E and F by the number of outcomes in E .

EXAMPLE 2 Conditional Probabilities on Marital Status and Gender

Problem The data in Table 8 represent the marital status and gender of the residents of the United States aged 15 years old or older in 2017.

- (a) Compute the probability that a randomly selected individual has never married given the individual is male.
- (b) Compute the probability that a randomly selected individual is male given the individual has never married.

Approach

- (a) Since the randomly selected person is male, concentrate on the male column. There are 126.9 million males and 46.0 million males who never married, so $N(\text{male}) = 126.9$ million and $N(\text{male and never married}) = 46.0$ million. Compute the probability using the Conditional Probability Rule.

(continued)

- (b)** Since the randomly selected person has never married, concentrate on the never married row. There are 86.1 million people who have never married and 46.0 million males who have never married, so $N(\text{never married}) = 86.1$ million and $N(\text{male and never married}) = 46.0$ million. Compute the probability using the Conditional Probability Rule.

Solution

- (a)** Substituting into Formula (1), we obtain

$$P(\text{never married} \mid \text{male}) = \frac{N(\text{male and never married})}{N(\text{male})} = \frac{46.0}{126.9} = 0.362$$

The probability that the randomly selected individual has never married, given that the individual is male, is 0.362.

- (b)** Substituting into Formula (1), we obtain

$$P(\text{male} \mid \text{never married}) = \frac{N(\text{male and never married})}{N(\text{never married})} = \frac{46.0}{86.1} = 0.534$$

The probability that the randomly selected individual is male, given that the individual has never married, is 0.534. 

NW Now Work Problem 17

What is the difference between the results of Examples 2(a) and (b)? In Example 2(a), we found that 36.2% of males have never married, whereas in Example 2(b) we found that 53.4% of individuals who have never married are male.

EXAMPLE 3 Birth Weights of Preterm Babies

Problem Suppose that 12.7% of all births are preterm. (The gestation period of the pregnancy is less than 37 weeks.) Also 0.22% of all births resulted in a preterm baby who weighed 8 pounds, 13 ounces or more. What is the probability that a randomly selected baby weighs 8 pounds, 13 ounces or more, given that the baby is preterm? Is this unusual? *Source:* Vital Statistics Reports.

Approach We want to know the probability that the baby weighs 8 pounds, 13 ounces or more, given that the baby was preterm. Because 0.22% of all babies weigh 8 pounds, 13 ounces or more and are preterm, $P(\text{weighs 8 lb, 13 oz or more and preterm}) = 0.0022$. Since 12.7% of all births are preterm, $P(\text{preterm}) = 0.127$. The phrase “given that” suggests we use the Conditional Probability Rule to compute the probability.

Solution $P(\text{weighs 8 lb, 13 oz or more} \mid \text{preterm})$

$$\begin{aligned} &= \frac{P(\text{weighs 8 lb, 13 oz or more and preterm})}{P(\text{preterm})} \\ &= \frac{0.0022}{0.127} \approx 0.0173 \end{aligned}$$

If 100 preterm babies were randomly selected, we would expect about two to weigh 8 pounds, 13 ounces or more. This is an unusual outcome. 

NW Now Work Problem 13**2 Compute Probabilities Using the General Multiplication Rule**

If we solve the Conditional Probability Rule for $P(E \text{ and } F)$, we obtain the General Multiplication Rule.

General Multiplication Rule

The probability that two events E and F both occur is

$$P(E \text{ and } F) = P(E) \cdot P(F | E)$$

In words, the probability of E and F is the probability of event E occurring times the probability of event F occurring, given the occurrence of event E .

EXAMPLE 4 Using the General Multiplication Rule

Problem The probability that a driver who is speeding gets pulled over is 0.8. The probability that a driver gets a ticket, given that he or she is pulled over, is 0.9. What is the probability that a randomly selected driver who is speeding gets pulled over and gets a ticket?

Approach Let E represent the event “driver who is speeding gets pulled over,” and let F represent the event “driver gets a ticket.” Use the General Multiplication Rule to compute $P(E \text{ and } F)$.

Solution $P(\text{driver who is speeding gets pulled over and gets a ticket}) = P(E \text{ and } F) = P(E) \cdot P(F | E) = 0.8(0.9) = 0.72$. The probability that a driver who is speeding gets pulled over and gets a ticket is 0.72.

NW Now Work Problem 31



EXAMPLE 5 Acceptance Sampling

Problem Suppose that of 100 circuits sent to a manufacturing plant, 5 are defective. The plant manager receiving the circuits randomly selects 2 and tests them. If both circuits work, she will accept the shipment. Otherwise, the shipment is rejected. What is the probability that the plant manager discovers at least 1 defective circuit and rejects the shipment?

Approach To determine the probability that at least one of the tested circuits is defective, consider four possibilities. Neither of the circuits is defective, the first is defective while the second is not, the first is not defective while the second is defective, or both circuits are defective. Note that the outcomes are not equally likely. To find the probability the manager discovers at least 1 defective circuit, we may use one of two approaches.

Approach I: Use a tree diagram to list all possible outcomes and the General Multiplication Rule to compute the probability for each outcome. Then determine the probability of at least 1 defective by adding the probability that only the first is defective, only the second is defective, or both are defective, using the Addition Rule (because they are disjoint).

Approach II: Compute the probability that both circuits are not defective and use the Complement Rule to determine the probability of at least 1 defective.

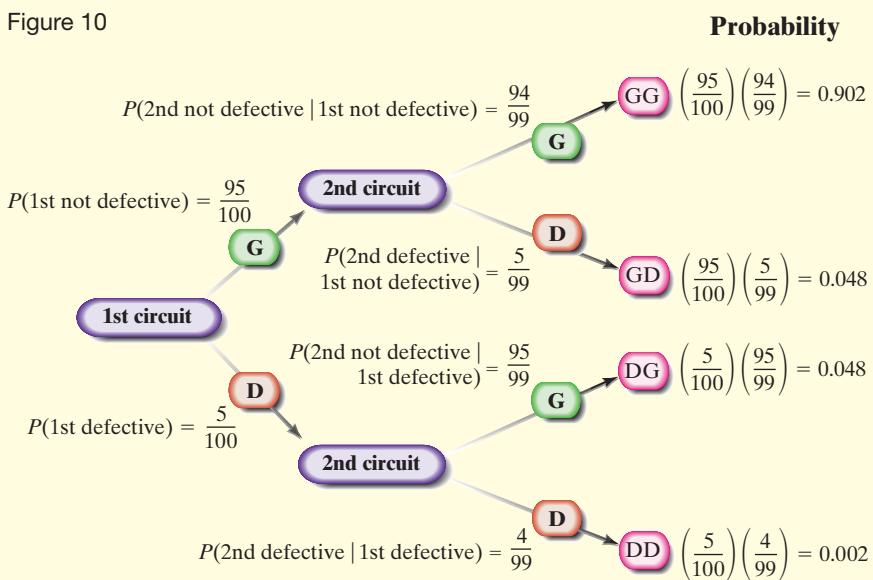
We will illustrate both approaches.

Solution Of the 100 circuits, 5 are defective, so 95 are not defective.

Approach I: Construct a tree diagram to determine the possible outcomes for the experiment. See Figure 10 on the following page, where D stands for defective and G stands for good (not defective). Because the outcomes are not equally likely, we include the probabilities in our diagram to show how the probability of each outcome is obtained. (Multiply the individual probabilities along the corresponding path in the diagram.)

(continued)

Figure 10



From the tree diagram and the Addition Rule, we can write

$$\begin{aligned}
 P(\text{at least 1 defective}) &= P(\text{GD}) + P(\text{DG}) + P(\text{DD}) \\
 &= 0.048 + 0.048 + 0.002 \\
 &= 0.098
 \end{aligned}$$

The probability that the shipment will not be accepted is 0.098.

Approach II: Compute the probability that both circuits are not defective and use the Complement Rule to determine the probability of at least 1 defective.

$$\begin{aligned}
 P(\text{at least 1 defective}) &= 1 - P(\text{none defective}) \\
 &= 1 - P(\text{1st not defective}) \cdot P(\text{2nd not defective} | \text{1st not defective}) \\
 &= 1 - \left(\frac{95}{100}\right) \cdot \left(\frac{94}{99}\right) \\
 &= 1 - 0.902 \\
 &= 0.098
 \end{aligned}$$

NW Now Work Problem 21 The probability that the shipment will not be accepted is 0.098. ()

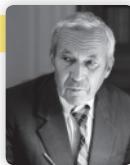
Whenever a small random sample is taken from a large population, it is reasonable to compute probabilities of events assuming independence. Consider the following example.

EXAMPLE 6

Favorite Other

Problem In a study to determine whether preferences for self are more or less prevalent than preferences for others, researchers first asked individuals to identify the person who is most valuable and likeable to you, or your favorite other. Of the 1519 individuals surveyed, 42 had chosen themselves as their favorite other. *Source: Gebauer JE, et al. Self-Love or Other-Love? Explicit Other-Preference but Implicit Self-Preference. PLoS ONE 7(7): e41789. doi:10.1371/journal.pone.0041789*

- (a) Suppose we randomly select 1 of the 1519 individuals surveyed. What is the probability that he or she chose himself or herself as their favorite other?

Historical Note

Andrei Nikolaevich Kolmogorov was born on April 25, 1903, in Tambov, Russia. His parents were not married, and he was raised by his aunt. He graduated from Moscow State University in 1925. That year he published eight papers, including his first on probability. By the time he received his doctorate, in 1929, he already had 18 publications. He became a professor at Moscow State University in 1931. Kolmogorov is quoted as saying, "The theory of probability as a mathematical discipline can and should be developed from axioms in exactly the same way as Geometry and Algebra." Kolmogorov helped to educate gifted children. It did not bother him if the students did not become mathematicians; he simply wanted them to be happy. Andrei Kolmogorov died on October 20, 1987.

- (b) If two individuals from this group are randomly selected, what is the probability that both chose themselves as their favorite other?
- (c) Compute the probability of randomly selecting two individuals from this group who selected themselves as their favorite other assuming independence.

Approach Let event E = "themselves favorite other," so $P(E)$ = number of individuals who select themselves as favorite other divided by 1519, the number of individuals in the survey. To answer part (b), let E_1 = "first person selects themselves as favorite other" and E_2 = "second person selects themselves as favorite other." Then compute $P(E_1 \text{ and } E_2) = P(E_1) \cdot P(E_2|E_1)$. To answer part (c), use $P(E)$ in the Multiplication Rule for Independent Events.

Solution

(a) If one individual is selected, $P(E) = \frac{42}{1519} = 0.02765$.

- (b) Using the Multiplication Rule,

$$P(E_1 \text{ and } E_2) = P(E_1) \cdot P(E_2|E_1) = \frac{42}{1519} \cdot \frac{41}{1518} \approx 0.0007468$$

Notice that $P(E_2|E_1) = \frac{41}{1518}$ because we are sampling without replacement, so after event E_1 occurs there is one less person who considers themselves their favorite other and one less person in the sample space.

- (c) The assumption of independence means that the outcome of the first trial of the experiment does not affect the probability of the second trial. (It is like sampling with replacement.) Therefore, we assume that

$$P(E_1) = P(E_2) = \frac{42}{1519}$$

Then

$$P(E_1 \text{ and } E_2) = P(E_1) \cdot P(E_2) = \frac{42}{1519} \cdot \frac{42}{1519} \approx 0.0007645$$

The probabilities in Examples 6(b) and 6(c) are extremely close in value. Based on these results, we infer the following principle.

If small random samples are taken from large populations without replacement, it is reasonable to assume independence of the events. As a rule of thumb, if the sample size is less than 5% of the population size, we treat the events as independent.

For example, in Example 6, we can compute the probability of randomly selecting two individuals who consider themselves their favorite other assuming independence because the sample size, 2, is only $\frac{2}{1519}$, or 0.13% of the population size, 1519.

We can now express independence using conditional probabilities.

Definition

Two events E and F are independent if $P(E|F) = P(E)$ or, equivalently, if $P(F|E) = P(F)$.

NW Now Work Problem 39

If either condition in our definition is true, the other is as well. In addition, for independent events,

$$P(E \text{ and } F) = P(E) \cdot P(F)$$

So the Multiplication Rule for Independent Events is a special case of the General Multiplication Rule.

Look back at Table 8 on page 261. Because $P(\text{widowed}) = 0.058$ does not equal $P(\text{widowed} | \text{female}) = 0.089$, the events “widowed” and “female” are not independent. In fact, knowing an individual is female increases the likelihood that the individual is also widowed.

NW Now Work Problem 41



5.4 Assess Your Understanding

Vocabulary and Skill Building

1. The notation $P(F|E)$ means the probability of event _____ given event _____.
2. If $P(E) = 0.6$ and $P(E|F) = 0.34$, are events E and F independent?
3. Suppose that E and F are two events and that $P(E \text{ and } F) = 0.6$ and $P(E) = 0.8$. What is $P(F|E)$?
4. Suppose that E and F are two events and that $P(E \text{ and } F) = 0.21$ and $P(E) = 0.4$. What is $P(F|E)$?
5. Suppose that E and F are two events and that $N(E \text{ and } F) = 420$ and $N(E) = 740$. What is $P(F|E)$?
6. Suppose that E and F are two events and that $N(E \text{ and } F) = 380$ and $N(E) = 925$. What is $P(F|E)$?
7. Suppose that E and F are two events and that $P(E) = 0.8$ and $P(F|E) = 0.4$. What is $P(E \text{ and } F)$?
8. Suppose that E and F are two events and that $P(E) = 0.4$ and $P(F|E) = 0.6$. What is $P(E \text{ and } F)$?
9. According to the U.S. Census Bureau, the probability that a randomly selected head of household in the United States earns more than \$100,000 per year is 0.202. The probability that a randomly selected head of household in the United States earns more than \$100,000 per year, given that the head of household has earned a bachelor’s degree, is 0.412. Are the events “earn more than \$100,000 per year” and “earned a bachelor’s degree” independent?
10. The probability that a randomly selected individual in the United States 25 years and older has at least a bachelor’s degree is 0.094. The probability that an individual in the United States 25 years and older has at least a bachelor’s degree, given that the individual lives in Washington DC, is 0.241. Are the events “bachelor’s degree” and “lives in Washington, DC,” independent? *Source: American Community Survey, 2013.*

Applying the Concepts

11. **Drawing a Card** Suppose that a single card is selected from a standard 52-card deck. What is the probability that the card drawn is a club? Now suppose that a single card is drawn from a standard 52-card deck, but we are told that the card is black. What is the probability that the card drawn is a club?
12. **Drawing a Card** Suppose that a single card is selected from a standard 52-card deck. What is the probability that the card drawn is a king? Now suppose that a single card is drawn from a standard 52-card deck, but we are told that the card is a

heart. What is the probability that the card drawn is a king? Did the knowledge that the card is a heart change the probability that the card was a king? What term is used to describe this result?

13. **Marriage** According to Pew Research, in 27% of marriages the woman has a bachelor’s degree and the marriage lasts at least 20 years. According to the Census Bureau, 35% of women have a bachelor’s degree. What is the probability a randomly selected marriage will last at least 20 years if the woman has a bachelor’s degree? **Note:** 52% of all marriages last at least 20 years.
14. **Cause of Death** According to the U.S. National Center for Health Statistics, 0.15% of deaths in the United States are 25- to 34-year-olds whose cause of death is cancer. In addition, 1.71% of all those who die are 25–34 years old. What is the probability that a randomly selected death is the result of cancer if the individual is known to have been 25–34 years old?
15. **High School Dropouts** According to the U.S. Census Bureau, 8.0% of 16- to 24-year-olds are high school dropouts. In addition, 2.1% of 16- to 24-year-olds are high school dropouts and unemployed. What is the probability that a randomly selected 16- to 24-year-old is unemployed, given he or she is a dropout?
16. **Income by Region** According to the U.S. Census Bureau, 17.9% of U.S. households are in the Northeast. In addition, 5.4% of U.S. households earn \$100,000 per year or more and are located in the Northeast. Determine the probability that a randomly selected U.S. household earns more than \$100,000 per year, given that the household is located in the Northeast.
17. **Made in America** In a recent Harris Poll, a random sample of adult Americans (18 years and older) was asked, “When you see an ad emphasizing that a product is ‘Made in America,’ are you more likely to buy it, less likely to buy it, or neither more nor less likely to buy it?” The results of the survey, by age group, are presented in the following contingency table.

	18–34	35–44	45–54	55 +	Total
More likely	238	329	360	402	1329
Less likely	22	6	22	16	66
Neither more nor less likely	282	201	164	118	765
Total	542	536	546	536	2160

Source: The Harris Poll.

- (a) What is the probability that a randomly selected individual is 35–44 years of age, given the individual is more likely to buy a product emphasized as “Made in America”?
- (b) What is the probability that a randomly selected individual is more likely to buy a product emphasized as “Made in America,” given the individual is 35–44 years of age?
- (c) Are 18- to 34-year-olds more likely to buy a product emphasized as “Made in America” than individuals in general?

18. Social Media Adult Americans (18 years or older) were asked whether they used social media (Facebook, Twitter, and so on) regularly. The following table is based on the results of the survey.

	18–34	35–44	45–54	55 +	Total
Use social media	117	89	83	49	338
Do not use social media	33	36	57	66	192
Total	150	125	140	115	530

Source: Harris Interactive.

- (a) What is the probability that a randomly selected adult American uses social media, given the individual is 18–34 years of age?
- (b) What is the probability that a randomly selected adult American is 18–34 years of age, given the individual uses social media?
- (c) Are 18- to 34-year olds more likely to use social media than individuals in general? Why?

19. Driver Fatalities The following data represent the number of drivers involved in a fatal crash in 2016 in various light and weather conditions.

Light Condition					
Weather	Daylight	Dark, but Lighted			
		Dark	Dawn/Dusk	Other	
Normal	14,307	5875	8151	1183	65
Rain	875	497	681	87	8
Snow/Sleet	219	51	156	17	2
Other	125	54	220	40	9
Unknown	810	255	548	71	133

Source: Fatality Analysis Reporting System.

- (a) Among fatal crashes in normal weather, what is the probability that a randomly selected fatal crash occurs at dawn/dusk?
- (b) Among dawn/dusk fatal crashes, what is the probability that a randomly selected fatal crash occurs during normal weather?
- (c) Is the dark (without light) more dangerous in normal weather or in rain? Explain.

20. Speeding Tickets Use the results of Problem 44 in Section 5.2 to answer the following.

- (a) Among those who text while driving, what is the probability that a randomly selected individual was issued no tickets last year?
- (b) Among those who were issued no tickets last year, what is the probability the individual texts while driving?

- (c) Based on the results of this survey, does it appear to be the case that individuals who text while driving are less likely to be issued 0 speeding tickets than those who do not text while driving?

NW 21. Acceptance Sampling Suppose that you just received a shipment of six televisions and two are defective. If two televisions are randomly selected, compute the probability that both televisions work. What is the probability that at least one does not work?

22. Committee A committee consisting of four women and three men will randomly select two people to attend a conference in Hawaii. Find the probability that both are women.

23. Suppose that two cards are randomly selected from a standard 52-card deck.

- (a) What is the probability that the first card is a king and the second card is a king if the sampling is done without replacement?

- (b) What is the probability that the first card is a king and the second card is a king if the sampling is done with replacement?

24. Suppose that two cards are randomly selected from a standard 52-card deck.

- (a) What is the probability that the first card is a club and the second card is a club if the sampling is done without replacement?

- (b) What is the probability that the first card is a club and the second card is a club if the sampling is done with replacement?

25. Board Work This past semester, I had a small business calculus section. The students in the class were Mike, Neta, Jinita, Kristin, and Dave. Suppose that I randomly select two people to go to the board to work problems. What is the probability that Dave is the first person chosen to go to the board and Neta is the second?

26. Party My wife has organized a monthly neighborhood party. Five people are involved in the group: Yolanda (my wife), Lorrie, Laura, Kim, and Anne Marie. They decide to randomly select the first and second home that will host the party. What is the probability that my wife hosts the first party and Lorrie hosts the second? **Note:** Once a home has hosted, it cannot host again until all other homes have hosted.

27. Playing Music on Random Setting Suppose that a Spotify playlist has 13 tracks. After listening to all the songs, you decide that you like 5 of them. With the random feature on the playlist, each of the 13 songs is played once in random order. Find the probability that among the first two songs played

- (a) You like both of them. Would this be unusual?

- (b) You like neither of them.

- (c) You like exactly one of them.

- (d) Redo (a)–(c) if a song can be replayed before all 13 songs are played (if, for example, track 2 can play twice in a row).

28. Packaging Error Because of a manufacturing error, three cans of regular soda were accidentally filled with diet soda and placed into a 12-pack. Suppose that two cans are randomly selected from the 12-pack.

- (a) Determine the probability that both contain diet soda.

- (b) Determine the probability that both contain regular soda. Would this be unusual?

- (c) Determine the probability that exactly one is diet and one is regular?

29. Planting Tulips A bag of 30 tulip bulbs purchased from a nursery contains 12 red tulip bulbs, 10 yellow tulip bulbs, and 8 purple tulip bulbs. Use a tree diagram like the one in Example 5 to answer the following:

- What is the probability that two randomly selected tulip bulbs are both red?
- What is the probability that the first bulb selected is red and the second yellow?
- What is the probability that the first bulb selected is yellow and the second is red?
- What is the probability that one bulb is red and the other yellow?

30. Golf Balls The local golf store sells an “onion bag” that contains 35 “experienced” golf balls. Suppose that the bag contains 20 Titleists, 8 Maxflis, and 7 Top Flites. Use a tree diagram like the one in Example 5 to answer the following:

- What is the probability that two randomly selected golf balls are both Titleists?
- What is the probability that the first ball selected is a Titleist and the second is a Maxfli?
- What is the probability that the first ball selected is a Maxfli and the second is a Titleist?
- What is the probability that one golf ball is a Titleist and the other is a Maxfli?

NW 31. Smokers According to the National Center for Health Statistics, there is a 20.3% probability that a randomly selected resident of the United States aged 18 years or older is a smoker. In addition, there is a 44.5% probability that a randomly selected resident of the United States aged 18 years or older is female, given that he or she smokes. What is the probability that a randomly selected resident of the United States aged 18 years or older is female and smokes? Would it be unusual to randomly select a resident of the United States aged 18 years or older who is female and smokes?

32. Multiple Jobs According to the U.S. Bureau of Labor Statistics, there is a 4.9% probability that a randomly selected employed individual has more than one job (a multiple-job holder). Also, there is a 46.6% probability that a randomly selected employed individual is male, given that he has more than one job. What is the probability that a randomly selected employed individual is a multiple-job holder and male? Would it be unusual to randomly select an employed individual who is a multiple-job holder and male?

33. The Birthday Problem Determine the probability that at least 2 people in a room of 10 people share the same birthday, ignoring leap years and assuming each birthday is equally likely, by answering the following questions:

- Compute the probability that 10 people have 10 different birthdays. **Hint:** The first person’s birthday can occur 365 ways, the second person’s birthday can occur 364 ways, because he or she cannot have the same birthday as the first person, the third person’s birthday can occur 363 ways, because he or she cannot have the same birthday as the first or second person, and so on.
- The complement of “10 people have different birthdays” is “at least 2 share a birthday.” Use this information to compute the probability that at least 2 people out of 10 share the same birthday.

34. The Birthday Problem Using the procedure given in Problem 33, compute the probability that at least 2 people in a room of 23 people share the same birthday.

35. Teen Communication The following data represent the number of different communication activities (e.g., cell phone, text messaging, e-mail, Instagram, and so on) used by a random sample of teenagers over the past 24 hours.

Activities	0	1–2	3–4	5+	Total
Male	21	81	60	38	200
Female	21	52	56	71	200
Total	42	133	116	109	400

- Are the events “male” and “0 activities” independent? Justify your answer.
- Are the events “female” and “5+ activities” independent? Justify your answer.
- Are the events “1–2 activities” and “3–4 activities” mutually exclusive? Justify your answer.
- Are the events “male” and “1–2 activities” mutually exclusive? Justify your answer.

36. Party Affiliation The following data represent political party by age from a random sample of registered Iowa Voters.

	17–29	30–44	45–64	65+	Total
Republican	224	340	1075	561	2200
Democrat	184	384	773	459	1800
Total	408	724	1848	1020	4000

- Are the events “Republican” and “30–44” independent? Justify your answer.
- Are the events “Democrat” and “65+” independent? Justify your answer.
- Are the events “17–29” and “45–64” mutually exclusive? Justify your answer.
- Are the events “Republican” and “45–64” mutually exclusive? Justify your answer.

37. False Positives Mammograms are used to detect breast cancer. Suppose a mammogram is known to be 80% accurate, which means $P(\text{positive mammogram} \mid \text{cancer}) = 0.80$. Suppose in the United States, 40 million women are tested for breast cancer with mammograms and, of these, 200,000 have breast cancer (regardless of the results of the mammogram).

- Suppose the probability of receiving a positive mammogram given cancer is not present is 0.08. This is a false positive. Express this result as a conditional probability.
- Use the information given to fill in the following table.

	Positive Mammogram	Negative Mammogram	Total
Cancer			200,000
No Cancer			
Total			40,000,000

- Use the entries in the table to compute $P(\text{cancer} \mid \text{positive mammogram})$.

38. A Flush A flush in the card game of poker occurs if a player gets five cards that are all the same suit (clubs, diamonds, hearts, or spades). Answer the following questions to obtain the probability of being dealt a flush in five cards.

- (a) We initially concentrate on one suit, say clubs. There are 13 clubs in a deck. Compute $P(\text{five clubs}) = P(\text{first card is clubs and second card is clubs and third card is clubs and fourth card is clubs and fifth card is clubs})$.
- (b) A flush can occur if we get five clubs or five diamonds or five hearts or five spades. Compute $P(\text{five clubs or five diamonds or five hearts or five spades})$. Note that the events are mutually exclusive.
- (c) A royal flush in the game of poker occurs if the player gets the cards Ten, Jack, Queen, King, and Ace all in the same suit. Use the procedure given in parts (a) and (b) to compute the probability of being dealt a royal flush.

NW 39. Independence in Small Samples from Large Populations

Suppose that a computer chip company has just shipped 10,000 computer chips to a computer company. Unfortunately, 50 of the chips are defective.

- (a) Compute the probability that two randomly selected chips are defective using conditional probability.
- (b) There are 50 defective chips out of 10,000 shipped. The probability that the first chip randomly selected is defective is $\frac{50}{10,000} = 0.005$. Compute the probability that two randomly selected chips are defective under the assumption of independent events. Compare your results to part (a). Conclude that, when small samples are taken from large populations without replacement, the assumption of independence does not significantly affect the probability.

40. Independence in Small Samples from Large Populations

Suppose that a poll is being conducted in the village of Lemont. The pollster identifies her target population as all residents of Lemont 18 years old or older. This population has 6494 people.

- (a) Compute the probability that the first resident selected to participate in the poll is Roger Cummings and the second is Rick Whittingham.
- (b) The probability that any particular resident of Lemont is the first person picked is $\frac{1}{6494}$. Compute the probability that Roger is selected first and Rick is selected second, assuming independence. Compare your results to part (a). Conclude that, when small samples are taken from large populations

without replacement, the assumption of independence does not significantly affect the probability.

NW 41. Independent? Refer to the contingency table in Problem 17 that relates age and likelihood to buy American. Determine $P(45\text{--}54 \text{ years old})$ and $P(45\text{--}54 \text{ years old} | \text{more likely})$. Are the events “45–54 years old” and “more likely” independent?

42. Independent? Refer to the contingency table in Problem 18 that relates social media use and age. Determine $P(\text{uses social media})$ and $P(\text{uses social media} | 35\text{--}44 \text{ years})$. Are the events “uses social media” and “35–44 years” independent?

43. Putting It Together: Success Sequence Is there a “path” to success? Brookings scholars Ron Haskins and Isabel Sawhill suggest the path to success is: education, followed by work, followed by marriage, followed by children. Sociologists Wendy Wang and W. Bradford Wilcox tracked a cohort of young millennial adults from their teenage years to early adulthood (ages 28 to 34) and recorded information about their education, marital status, child-rearing, and income.

- (a) Why is this a cohort study?
- (b) According to the article, 53% of millennials who had failed to complete the path to success were below the poverty rate. Express this as a conditional probability.
- (c) The article states that 31% of millennials who completed high school were below the poverty rate; 16% of millennials who had a high school diploma and a full-time job were below the poverty rate; and 3% of millennials who had a high school diploma, a full-time job, and put marriage before having children were below the poverty rate. Express each of these probabilities as a conditional probability.
- (d) Among those from low-income backgrounds, 80% made it to middle- or upper-income when they followed the path to success while only 44% made it to middle- or upper-income when they missed at least one stage in the path to success. Express each of these probabilities as a conditional probability.
- (e) According to researchers at Payscale.com, among those from low-income backgrounds in early adulthood, 28% make it to middle- or upper-income by mid-life. Do you believe that upward mobility and the “path to success” are independent? Explain.

Source: “‘The Sequence’ Is the Secret to Success,” Wendy Wang, *Wall Street Journal*, March 28, 2018.

5.5 Counting Techniques



Objectives

- ① Solve counting problems using the Multiplication Rule
- ② Solve counting problems using permutations
- ③ Solve counting problems using combinations
- ④ Solve counting problems involving permutations with nondistinct items
- ⑤ Compute probabilities involving permutations and combinations

① Solve Counting Problems Using the Multiplication Rule

Counting plays a major role in many diverse areas, including probability. In this section, we look at special types of counting problems and develop general techniques for solving them.

We begin with an example that demonstrates a general counting principle.

EXAMPLE 1 Counting the Number of Possible Meals

Problem The fixed-price dinner at Mabenka Restaurant provides the following choices:

Appetizer: soup or salad

Entrée: baked chicken, broiled beef patty, baby beef liver, or roast beef au jus

Dessert: ice cream or cheesecake

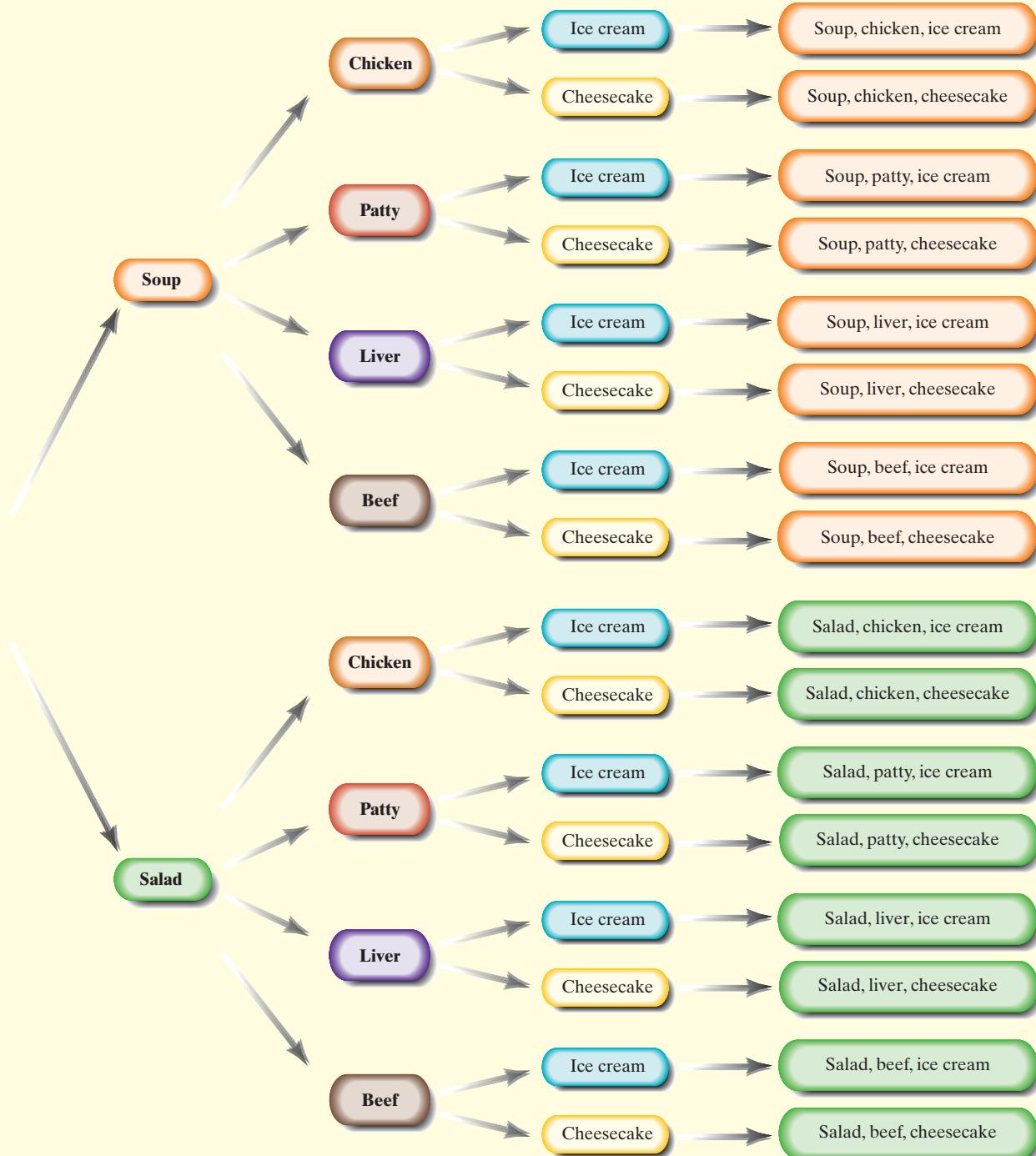
How many different meals can be ordered?

Approach Ordering such a meal requires three separate decisions:

Choose an Appetizer	Choose an Entrée	Choose a Dessert
2 choices	4 choices	2 choices

Figure 11 is a tree diagram that lists the possible meals that can be ordered.

Figure 11



Solution Look at the tree diagram in Figure 11. For each choice of appetizer, we have 4 choices of entrée, and for each of these $2 \cdot 4 = 8$ choices, there are 2 choices for dessert. A total of $2 \cdot 4 \cdot 2 = 16$ different meals can be ordered.

Example 1 illustrates a general counting principle.

Multiplication Rule of Counting

If a task consists of a sequence of choices in which there are p selections for the first choice, q selections for the second choice, r selections for the third choice, and so on, then the task of making these selections can be done in

$$p \cdot q \cdot r \cdots$$

different ways.

EXAMPLE 2 Counting Airport Codes (Repetition Allowed)

Problem The International Air Transport Association (IATA) assigns three-letter codes to represent airport locations. For example, the code for Fort Lauderdale International Airport is FLL. How many different airport codes are possible?

Approach We are choosing three letters from 26 letters and arranging them in order. Notice that repetition of letters is allowed. Use the Multiplication Rule of Counting, recognizing we have 26 ways to choose the first letter, 26 ways to choose the second letter, and 26 ways to choose the third letter.

Solution By the Multiplication Rule,

$$26 \cdot 26 \cdot 26 = 26^3 = 17,576$$

different airport codes are possible.

In the following example, repetition is not allowed, unlike Example 2.

EXAMPLE 3 Counting without Repetition

Problem Three members from a 14-member committee are to be randomly selected to serve as chair, vice-chair, and secretary. The first person selected is the chair, the second is the vice-chair, and the third is the secretary. How many different committee structures are possible?

Approach The task consists of making three selections. The first selection requires choosing from 14 members. Because a member cannot serve in more than one capacity, the second selection requires choosing from the 13 remaining members. The third selection requires choosing from the 12 remaining members. (Do you see why?) We use the Multiplication Rule to determine the number of possible committee structures.

Solution By the Multiplication Rule,

$$14 \cdot 13 \cdot 12 = 2184$$

different committee structures are possible.



The Factorial Symbol We now introduce a special symbol.

Definition

If $n \geq 0$ is an integer, the **factorial symbol**, $n!$, is defined as follows:

$$\begin{aligned} n! &= n(n - 1) \cdots 3 \cdot 2 \cdot 1 \\ 0! &= 1 \quad 1! = 1 \end{aligned}$$

 **Using Technology**

Your calculator has a factorial key. Use it to see how fast factorials increase in value. Find the value of $69!$. What happens when you try to find $70!$? In fact, $70!$ is larger than 10^{100} (a *googol*), the largest number most calculators can display.

For example, $2! = 2 \cdot 1 = 2$, $3! = 3 \cdot 2 \cdot 1 = 6$, $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$, and so on. Table 9 lists the values of $n!$ for $0 \leq n \leq 6$.

Table 9

n	0	1	2	3	4	5	6
$n!$	1	1	2	6	24	120	720

EXAMPLE 4

The Traveling Salesperson

Problem You have just been hired as a book representative for Pearson Education. On your first day, you must travel to seven schools to introduce yourself. How many different routes are possible?

Approach Call the seven schools A , B , C , D , E , F , and G . School A can be visited first, second, third, fourth, fifth, sixth, or seventh. So, there are seven choices for school A . There are six choices for school B , five choices for school C , and so on. Use the Multiplication Rule and the factorial to find the solution.

NW Now Work Problems 5 and 33

Solution $7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 7! = 5040$ different routes are possible.



2

Solve Counting Problems Using Permutations

Examples 3 and 4 illustrate a type of counting problem referred to as a *permutation*.

Definition

A **permutation** is an ordered arrangement in which r objects are chosen from n distinct (different) objects so that $r \leq n$ and repetition is not allowed. The symbol ${}_nP_r$, represents the number of permutations of r objects selected from n objects.

The solution in Example 3 could be represented as

$${}_nP_r = {}_{14}P_3 = 14 \cdot 13 \cdot 12 = 2184$$

and the solution to Example 4 as

$${}_7P_7 = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5040$$

To arrive at a formula for ${}_nP_r$, note that there are n choices for the first selection, $n - 1$ choices for the second selection, $n - 2$ choices for the third selection, \dots , and $n - (r - 1)$ choices for the r th selection. By the Multiplication Rule,

$$\begin{aligned} {}nP_r &= n \cdot (n - 1) \cdot (n - 2) \cdots [n - (r - 1)] \\ &= n \cdot (n - 1) \cdot (n - 2) \cdots (n - r + 1) \end{aligned}$$

This formula for nP_r can be written in **factorial notation**:

$$\begin{aligned}nP_r &= n \cdot (n - 1) \cdot (n - 2) \cdots \cdots (n - r + 1) \\&= n \cdot (n - 1) \cdot (n - 2) \cdots \cdots (n - r + 1) \cdot \frac{(n - r) \cdots \cdots 3 \cdot 2 \cdot 1}{(n - r) \cdots \cdots 3 \cdot 2 \cdot 1} \\&= \frac{n!}{(n - r)!}\end{aligned}$$

Number of Permutations of n Distinct Objects Taken r at a Time

The number of arrangements of r objects chosen from n objects, in which

1. the n objects are distinct,
2. repetition of objects is not allowed, and
3. order is important,

is given by the formula

IN OTHER WORDS

“Order is important” means that **ABC** is different from **BCA**.

$$nP_r = \frac{n!}{(n - r)!} \quad (1)$$

EXAMPLE 5 Computing Permutations

Problem Evaluate:

(a) ${}_7P_5$ (b) ${}_5P_5$

By Hand Approach

Use Formula (1): $nP_r = \frac{n!}{(n - r)!}$

By Hand Solution

$$\begin{aligned}(a) \quad {}_7P_5 &= \frac{7!}{(7 - 5)!} = \frac{7!}{2!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2!}{2!} \\&= \underbrace{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3}_{\text{5 factors}} \\&= 2520\end{aligned}$$

$$\begin{aligned}(b) \quad {}_5P_5 &= \frac{5!}{(5 - 5)!} = \frac{5!}{0!} = 5! \\&= 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \\&= 120\end{aligned}$$

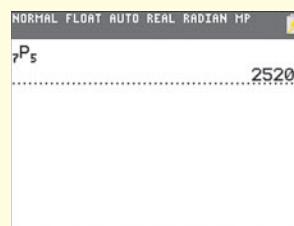
Technology Approach

We will use a TI-84 Plus CE graphing calculator in part (a) and Excel in part (b) to evaluate each permutation. The steps for determining permutations using the TI-83/84 Plus graphing calculator, Excel, and StatCrunch can be found in the Technology Step-by-Step on page 279.

Technology Solution

- (a) Figure 12(a) shows the results on a TI-84 Plus CE calculator, so ${}_7P_5 = 2520$.
(b) Figure 12(b) shows the results from Excel. Enter “=permut(5,5)” in any cell, so ${}_5P_5 = 120$.

Figure 12



D	E
	120

NW Now Work Problem 11

(a)

10

(b)

EXAMPLE 6**Betting on the Trifecta**

Problem In how many ways can horses in a ten-horse race finish first, second, and third?

Approach The ten horses are distinct. Once a horse crosses the finish line, that horse will not cross the finish line again, and, in a race, finishing order is important. We have a permutation of ten objects taken three at a time.

Solution The top three horses can finish a ten-horse race in

$$\begin{aligned} {}^{10}P_3 &= \frac{10!}{(10-3)!} = \frac{10!}{7!} = \frac{10 \cdot 9 \cdot 8 \cdot 7!}{7!} \\ &= \underbrace{10 \cdot 9 \cdot 8}_{\text{3 factors}} \\ &= 720 \text{ ways} \end{aligned}$$

NW Now Work Problem 45



③ Solve Counting Problems Using Combinations

In a permutation, order is important. For example, the arrangements ABC and BAC are considered different arrangements of the letters A , B , and C . If order is unimportant, we do not distinguish ABC from BAC . In poker, the order in which the cards are received does not matter. The *combination* of the cards is what matters.

Definition

A **combination** is a collection, without regard to order, in which r objects are chosen from n distinct objects with $r \leq n$ and without repetition. The symbol ${}_nC_r$ represents the number of combinations of n distinct objects taken r at a time.

EXAMPLE 7**Listing Combinations**

Problem Roger, Ken, Tom, and Jay are going to play golf. They will randomly select teams of two players each. List all possible team combinations. That is, list all the combinations of the four people Roger, Ken, Tom, and Jay taken two at a time. What is ${}_4C_2$?

Approach List the possible teams. Note that order is unimportant, so $\{\text{Roger, Ken}\}$ is the same as $\{\text{Ken, Roger}\}$.

Solution The list of all such teams (combinations) is

Roger, Ken Roger, Tom Roger, Jay Ken, Tom Ken, Jay Tom, Jay

So,

$${}_4C_2 = 6$$

There are six ways of forming teams of two from a group of four players.



We can find a formula for ${}_nC_r$ by noting that the only difference between a permutation and a combination is that we disregard order in combinations. To determine ${}_nC_r$, eliminate from the formula for ${}_nP_r$ the number of permutations that were rearrangements of a given set of r objects. In Example 7, for example, selecting $\{\text{Roger, Ken}\}$ was the same as selecting $\{\text{Ken, Roger}\}$, so there were $2! = 2$ rearrangements of the two objects. This can be determined from the formula for ${}_nP_r$ by calculating ${}_rP_r = r!$. So, if we divide ${}_nP_r$ by $r!$, we will have the desired formula for ${}_nC_r$:

$${}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{r!(n-r)!}$$

Number of Combinations of n Distinct Objects Taken r at a Time

The number of different arrangements of r objects chosen from n objects, in which

1. the n objects are distinct
2. repetition of objects is not allowed, and
3. order is not important

is given by the formula

$${}_nC_r = \frac{n!}{r!(n-r)!} \quad (2)$$

Using Formula (2) to solve the problem presented in Example 7, we obtain

$${}_4C_2 = \frac{4!}{2!(4-2)!} = \frac{4!}{2!2!} = \frac{4 \cdot 3 \cdot 2!}{2 \cdot 1 \cdot 2!} = \frac{12}{2} = 6$$

EXAMPLE 8 Computing Combinations

Problem Evaluate:

(a) ${}_4C_1$

(b) ${}_6C_4$

(c) ${}_6C_2$

By Hand Approach

Use Formula (2): ${}_nC_r = \frac{n!}{r!(n-r)!}$

By Hand Solution

$$(a) {}_4C_1 = \frac{4!}{1!(4-1)!} = \frac{4!}{1! \cdot 3!} = \frac{4 \cdot 3!}{1 \cdot 3!} = 4$$

$$(b) {}_6C_4 = \frac{6!}{4!(6-4)!} = \frac{6!}{4! \cdot 2!} = \frac{6 \cdot 5 \cdot 4!}{4! \cdot 2 \cdot 1} = \frac{30}{2} = 15$$

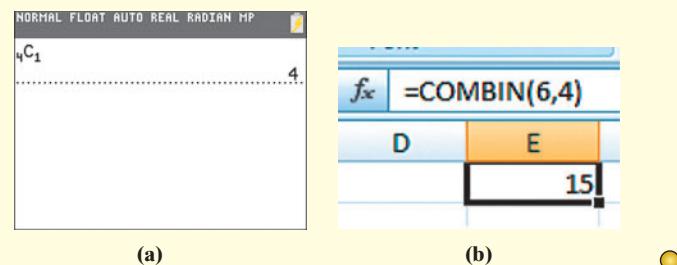
$$(c) {}_6C_2 = \frac{6!}{2!(6-2)!} = \frac{6!}{2! \cdot 4!} = \frac{6 \cdot 5 \cdot 4!}{2 \cdot 1 \cdot 4!} = \frac{30}{2} = 15$$

Technology Approach We will use a TI-84 Plus CE graphing calculator in part (a) and Excel in parts (b) and (c) to evaluate each combination. The steps for determining combinations using the TI-83/84 Plus graphing calculator, Excel, and StatCrunch can be found in the Technology Step-by-Step on page 279.

Technology Solution

- (a) Figure 13(a) shows the results on a TI-84 Plus CE calculator. So, ${}_4C_1 = 4$.
- (b) Figure 13(b) shows the results from Excel. Enter “=combin(6,4)” in any cell. So ${}_6C_4 = 15$.
- (c) If we enter “=combin(6,2)” into any cell in Excel, we obtain a result of 15.

Figure 13



NW Now Work Problem 19

Notice in Examples 8(b) and (c) that ${}_6C_4 = {}_6C_2$. This result can be generalized as

$${}_nC_r = {}_nC_{n-r}$$

EXAMPLE 9 Simple Random Samples

Problem How many different simple random samples of size 4 can be obtained from a population whose size is 20?

Approach The 20 individuals in the population are distinct. In addition, the order in which individuals are selected is unimportant. Thus, the number of simple random samples of size 4 from a population of size 20 is a combination of $n = 20$ objects taken $r = 4$ at a time.

Solution Use Formula (2) with $n = 20$ and $r = 4$:

$${}_{20}C_4 = \frac{20!}{4!(20-4)!} = \frac{20!}{4!16!} = \frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16!}{4 \cdot 3 \cdot 2 \cdot 1 \cdot 16!} = \frac{116,280}{24} = 4845$$

There are 4845 different simple random samples of size 4 from a population whose size is 20. 

NW Now Work Problem 51

④ Solve Counting Problems Involving Permutations with Nondistinct Items

Sometimes we want to arrange objects in order, but some of the objects are not distinguishable.

EXAMPLE 10 DNA Sequence

Problem A DNA sequence consists of a series of letters representing a DNA strand that spells out the genetic code. There are four possible letters (A, C, G, and T), each representing a specific nucleotide base in the DNA strand (adenine, cytosine, guanine, and thymine, respectively). How many distinguishable sequences can be formed using two As, two Cs, three Gs, and one T?

Approach Each sequence formed will have eight letters. To construct each sequence, we need to fill in eight positions with the eight letters:

1 2 3 4 5 6 7 8

The process of forming a sequence consists of four tasks:

Task 1: Choose the positions for the two As.

Task 2: Choose the positions for the two Cs.

Task 3: Choose the positions for the three Gs.

Task 4: Choose the position for the one T.

Task 1 can be done in ${}_8C_2$ ways because we are choosing the 2 positions for A, but order does not matter (because we cannot distinguish the two As). This leaves 6 positions to be filled, so task 2 can be done in ${}_6C_2$ ways. This leaves 4 positions to be filled, so task 3 can be done in ${}_4C_3$ ways. The last position can be filled in ${}_1C_1$ way.

Solution By the Multiplication Rule, the number of possible sequences that can be formed is

$$\begin{aligned} {}_8C_2 \cdot {}_6C_2 \cdot {}_4C_3 \cdot {}_1C_1 &= \frac{8!}{2! \cdot 6!} \cdot \frac{6!}{2! \cdot 4!} \cdot \frac{4!}{3! \cdot 1!} \cdot \frac{1!}{1! \cdot 0!} \\ &= \frac{8!}{2! \cdot 2! \cdot 3! \cdot 1! \cdot 0!} \\ &= 1680 \end{aligned}$$

There are 1680 possible distinguishable sequences that can be formed. 

Example 10 suggests a general result. Had the letters in the sequence each been different, ${}_8P_8 = 8!$ possible sequences would have been formed. This is the numerator of the answer. The presence of two As, two Cs, and three Gs reduces the number of different sequences, as the entries in the denominator illustrate. We are led to the following result:

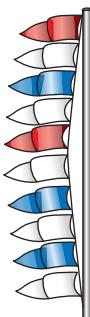
Permutations with Nondistinct Items

The number of permutations of n objects of which n_1 are of one kind, n_2 are of a second kind, \dots , and n_k are of a k th kind is given by

$$\frac{n!}{n_1! \cdot n_2! \cdots \cdot n_k!} \quad (3)$$

where $n = n_1 + n_2 + \cdots + n_k$.

EXAMPLE 11 Arranging Flags



NW Now Work Problem 55

Problem How many different vertical arrangements are there of 10 flags if 5 are white, 3 are blue, and 2 are red?

Approach Because there are nondistinct items and order matters, use the formula for finding the number of permutations with nondistinct items. We seek the number of permutations of $n = 10$ objects, of which $n_1 = 5$ are of one kind (white), $n_2 = 3$ are of a second kind (blue), and $n_3 = 2$ are of a third kind (red).

Solution Using Formula (3), we find that there are

$$\begin{aligned} \frac{10!}{5! \cdot 3! \cdot 2!} &= \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5!}{5! \cdot 3! \cdot 2!} \\ &= 2520 \text{ different vertical arrangements} \end{aligned}$$

Summary

One of the challenges in solving counting problems is selecting the appropriate formula for the given situation. Table 10 below reviews the situations in which each counting problem applies.

Table 10

	Description	Formula
Combination	The selection of r objects from a set of n different objects when the order in which the objects are selected does not matter (so AB is the same as BA) and an object cannot be selected more than once (repetition is not allowed)	${}_nC_r = \frac{n!}{r!(n-r)!}$
Permutation of Distinct Items with Replacement	The selection of r objects from a set of n different objects when the order in which the objects are selected matters (so AB is different from BA) and an object may be selected more than once (repetition is allowed)	n^r
Permutation of Distinct Items without Replacement	The selection of r objects from a set of n different objects when the order in which the objects are selected matters (so AB is different from BA) and an object cannot be selected more than once (repetition is not allowed)	${}_nP_r = \frac{n!}{(n-r)!}$
Permutation of Nondistinct Items without Replacement	The number of ways n objects can be arranged (order matters) in which there are n_1 of one kind, n_2 of a second kind, \dots , and n_k of a k th kind, where $n = n_1 + n_2 + \cdots + n_k$	$\frac{n!}{n_1!n_2!\cdots n_k!}$

5 Compute Probabilities Involving Permutations and Combinations

The counting techniques presented in this section can be used along with the classical method to compute certain probabilities. Recall that this method stated the probability of an event E is the number of ways event E can occur divided by the number of different possible outcomes of the experiment provided each outcome is equally likely.

EXAMPLE 12

Winning the Lottery



Problem In the Illinois Lottery, an urn contains balls numbered 1–52. From this urn, six balls are randomly chosen without replacement. For a \$1 bet, a player chooses two sets of six numbers. To win, all six numbers must match those chosen from the urn. The order in which the balls are picked does not matter. What is the probability of winning the lottery?

Approach The probability of winning is given by the number of ways a ticket could win divided by the size of the sample space. Each ticket has two sets of six numbers and therefore two chances of winning. The size of the sample space S is the number of ways 6 objects can be selected from 52 objects without replacement and without regard to order, so $N(S) = {}_{52}C_6$.

Solution The size of the sample space is

$$N(S) = {}_{52}C_6 = \frac{52!}{6! \cdot (52 - 6)!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 \cdot 47 \cdot 46!}{6! \cdot 46!} = 20,358,520$$

Each ticket has two chances of winning. If E is the event “winning ticket,” then $N(E) = 2$ and

$$P(E) = \frac{2}{20,358,520} = 0.000000098$$

There is about a 1 in 10,000,000 chance of winning the Illinois Lottery!



EXAMPLE 13

Acceptance Sampling

Problem A shipment of 120 fasteners that contains 4 defective fasteners was sent to a manufacturing plant. The plant’s quality-control manager randomly selects and inspects 5 fasteners. What is the probability that exactly 1 of the inspected fasteners is defective?

Approach Find the probability that exactly 1 fastener is defective by calculating the number of ways of selecting exactly 1 defective fastener in 5 fasteners and dividing this result by the number of ways of selecting 5 fasteners from 120 fasteners. To choose exactly 1 defective in the 5 requires choosing 1 defective from the 4 defectives and 4 nondefectives from the 116 nondefectives. The order in which the fasteners are selected does not matter, so we use combinations.

Solution The number of ways of choosing 1 defective fastener from 4 defective fasteners is ${}_4C_1$. The number of ways of choosing 4 nondefective fasteners from 116 nondefectives is ${}_{116}C_4$. Using the Multiplication Rule, we find that the number of ways of choosing 1 defective and 4 nondefective fasteners is

$$({}_4C_1) \cdot ({}_{116}C_4) = 4 \cdot 7,160,245 = 28,640,980$$

The number of ways of selecting 5 fasteners from 120 fasteners is ${}_{120}C_5 = 190,578,024$. The probability of selecting exactly 1 defective fastener is

$$P(1 \text{ defective fastener}) = \frac{(4C_1)(116C_4)}{{}_{120}C_5} = \frac{28,640,980}{190,578,024} = 0.1503$$

The probability of randomly selecting exactly one defective fastener is 0.1503. If we selected 5 fasteners, 100 different times, we would expect about 15 of the samples to have exactly one defective fastener.

NW Now Work Problem 61



Technology Step-by-Step

Factorials, Permutations, and Combinations

TI-83/84 Plus

Factorials

1. To compute $7!$, type 7 on the HOME screen.
2. Press MATH, then highlight PRB (or PROB on a TI-84 Plus CE), and then select 4: ! With 7! on the HOME screen, press ENTER again.

Permutations and Combinations

1. To compute ${}_7P_3$, type 7 on the HOME screen.
2. Press MATH, then highlight PRB (or PROB on a TI-84 Plus CE), and then select 2: ${}_nP_r$.
3. Type 3 on the HOME screen, and press ENTER.

Note: To compute ${}_7C_3$, select 3: ${}_nC_r$ instead of 2: ${}_nP_r$.

Excel

Factorials In any cell, enter “=fact(n)” where n is the factorial desired.

Permutations In any cell, enter “=permut(n, r)” where n is the number of distinct objects and r is the number of objects to be selected.

Combinations In any cell, enter “=combin(n, r)” where n is the number of distinct objects and r is the number of objects to be selected.

StatCrunch

1. Place the cursor in any cell in the StatCrunch spreadsheet. Select **Data** and choose **Compute**, then **Expression**.
2. **Factorials** In the Expression box, type fact(n) to determine $n!$. For example, to find $5!$, type fact(5). Click Compute!.

Permutations In the Expression box, type perm(n, r) to determine ${}_nP_r$. For example, to compute ${}_{10}P_4$, type perm(10,4). Click Compute!.

Combinations In the Expression box, type comb(n, r) to determine ${}_nC_r$. **Note:** To compute ${}_{10}C_4$, type comb(10,4). Click Compute!.



5.5 Assess Your Understanding

Vocabulary and Skill Building

1. A _____ is an ordered arrangement of r objects chosen from n distinct objects without repetition.
2. A _____ is an arrangement of r objects chosen from n distinct objects without repetition and without regard to order.
3. *True or False:* In a combination problem, order is not important.
4. The factorial symbol, $n!$, is defined as $n! = \underline{\hspace{2cm}}$ and $0! = \underline{\hspace{2cm}}$.

In Problems 5–10, find the value of each factorial.

- NW** 5. $5!$ 6. $7!$
 7. $10!$ 8. $12!$
 9. $0!$ 10. $1!$

In Problems 11–18, find the value of each permutation.

- NW** 11. $6P_2$ 12. $7P_2$
 13. $4P_4$ 14. $7P_7$
 15. $5P_0$ 16. $4P_0$
 17. $8P_3$ 18. $9P_4$

In Problems 19–26, find the value of each combination.

- NW** 19. $8C_3$ 20. $9C_2$
 21. $10C_2$ 22. $12C_3$
 23. $52C_1$ 24. $40C_{40}$
 25. $48C_3$ 26. $30C_4$
 27. List all the permutations of five objects a, b, c, d , and e taken two at a time without repetition. What is ${}_5P_2$?

- 28.** List all the permutations of four objects a, b, c , and d taken two at a time without repetition. What is ${}_4P_2$?
- 29.** List all the combinations of five objects a, b, c, d , and e taken two at a time. What is ${}_5C_2$?
- 30.** List all the combinations of four objects a, b, c , and d taken two at a time. What is ${}_4C_2$?

Applying the Concepts

- NW 31. Clothing Options** A man has six shirts and four ties. Assuming that they all match, how many different shirt-and-tie combinations can he wear?
- 32. Clothing Options** A woman has five blouses and three skirts. Assuming that they all match, how many different outfits can she wear?
- NW 33. Arranging Songs** Suppose Dan is going to create a Spotify playlist with 12 songs. In how many ways can the 12 songs be played without repetition?
- 34. Arranging Students** In how many ways can 15 students be lined up?
- 35. Traveling Salesperson** A salesperson must travel to eight cities to promote a new marketing campaign. How many different trips are possible if any route between cities is possible?
- 36. Randomly Playing Songs** The music player on an iPhone plays each of 10 songs. Once a song is played, it is not repeated until all the songs have been played. In how many different ways can the player play the 10 songs?
- 37. Stocks on the NYSE** Companies whose stocks are listed on the New York Stock Exchange (NYSE) have their company name represented by either one, two, or three letters (repetition of letters is allowed). What is the maximum number of companies that can be listed on the New York Stock Exchange?
- 38. Stocks on the NASDAQ** Companies whose stocks are listed on the NASDAQ stock exchange have their company name represented by either four or five letters (repetition of letters is allowed). What is the maximum number of companies that can be listed on the NASDAQ?
- 39. Garage Door Code** Outside a home there is a keypad that will open the garage if the correct four-digit code is entered.
- (a) How many codes are possible?
(b) What is the probability of entering the correct code on the first try, assuming that the owner doesn't remember the code?
- 40. Social Security Numbers** A Social Security number is used to identify each resident of the United States uniquely. The number is of the form $xxx-xx-xxxx$, where each x is a digit from 0 to 9.
- (a) How many Social Security numbers can be formed?
(b) What is the probability of correctly guessing the Social Security number of the president of the United States?
- 41. User Names** Suppose that a local area network requires eight letters for user names. Lower- and uppercase letters are considered the same. How many user names are possible for the local area network?
- 42. User Names** How many user names are possible in Problem 41 if the last character must be a digit?
- 43. Combination Locks** A combination lock has 50 numbers on it. To open it, you turn counterclockwise to a number, then rotate

clockwise to a second number, and then counterclockwise to the third number. Repetitions are allowed.

- (a) How many different lock combinations are there?
(b) What is the probability of guessing a lock combination on the first try?

- 44. Forming License Plate Numbers** How many different license plate numbers can be made by using one letter followed by five digits selected from the digits 0 through 9?

- NW 45. INDY 500** Suppose 40 cars start at the Indianapolis 500. In how many ways can the top 3 cars finish the race?

- 46. Betting on the Perfecta** In how many ways can the top 2 horses finish in a 10-horse race?

- 47. Forming a Committee** Four members from a 20-person committee are to be selected randomly to serve as chairperson, vice-chairperson, secretary, and treasurer. The first person selected is the chairperson; the second, the vice-chairperson; the third, the secretary; and the fourth, the treasurer. How many different leadership structures are possible?

- 48. Forming a Committee** Four members from a 50-person committee are to be selected randomly to serve as chairperson, vice-chairperson, secretary, and treasurer. The first person selected is the chairperson; the second, the vice-chairperson; the third, the secretary; and the fourth, the treasurer. How many different leadership structures are possible?

- 49. Lottery** A lottery exists where balls numbered 1–25 are placed in an urn. To win, you must match the four balls chosen in the correct order. How many possible outcomes are there for this game?

- 50. Forming a Committee** In the U.S. Senate, there are 21 members on the Committee on Banking, Housing, and Urban Affairs. Nine of these 21 members are selected to be on the Subcommittee on Economic Policy. How many different committee structures are possible for this subcommittee?

- NW 51. Simple Random Sample** How many different simple random samples of size 5 can be obtained from a population whose size is 50?

- 52. Simple Random Sample** How many different simple random samples of size 7 can be obtained from a population whose size is 100?

- 53. Children** A family has six children. If this family has exactly two boys, how many different birth and gender orders are possible?

- 54. Children** A family has eight children. If this family has exactly three boys, how many different birth and gender orders are possible?

- NW 55. DNA Sequences** (See Example 10.) How many distinguishable DNA sequences can be formed using three As, two Cs, two Gs, and three Ts?

- 56. DNA Sequences** (See Example 10.) How many distinguishable DNA sequences can be formed using one A, four Cs, three Gs, and four Ts?

- 57. Landscape Design** A golf-course architect has four linden trees, five white birch trees, and two bald cypress trees to plant in a row along a fairway. In how many ways can the landscaper plant the trees in a row, assuming that the trees are evenly spaced?

58. Starting Lineup A baseball team consists of three outfielders, four infielders, a pitcher, and a catcher. Assuming that the outfielders and infielders are indistinguishable, how many batting orders are possible?

59. Little Lotto In the Illinois Lottery game Little Lotto, an urn contains balls numbered 1–39. From this urn, 5 balls are chosen randomly, without replacement. For a \$1 bet, a player chooses one set of five numbers. To win, all five numbers must match those chosen from the urn. The order in which the balls are selected does not matter. What is the probability of winning Little Lotto with one ticket?

60. Mega Millions In Mega Millions, an urn contains balls numbered 1–56, and a second urn contains balls numbered 1–46. From the first urn, 5 balls are chosen randomly, without replacement and without regard to order. From the second urn, 1 ball is chosen randomly. For a \$1 bet, a player chooses one set of five numbers to match the balls selected from the first urn and one number to match the ball selected from the second urn. To win, all six numbers must match; that is, the player must match the first 5 balls selected from the first urn *and* the single ball selected from the second urn. What is the probability of winning the Mega Millions with a single ticket?

NW 61. Selecting a Jury The grade appeal process at a university requires that a jury be structured by selecting five individuals randomly from a pool of eight students and ten faculty.

- (a) What is the probability of selecting a jury of all students?
- (b) What is the probability of selecting a jury of all faculty?
- (c) What is the probability of selecting a jury of two students and three faculty?

62. Selecting a Committee Suppose that there are 55 Democrats and 45 Republicans in the U.S. Senate. A committee of seven senators is to be formed by selecting members of the Senate randomly.

- (a) What is the probability that the committee is composed of all Democrats?
- (b) What is the probability that the committee is composed of all Republicans?
- (c) What is the probability that the committee is composed of three Democrats and four Republicans?

63. Acceptance Sampling Suppose that a shipment of 120 electronic components contains 4 defective components. To determine whether the shipment should be accepted, a quality-control engineer randomly selects 4 of the components and tests them. If 1 or more of the components is defective, the shipment is rejected. What is the probability that the shipment is rejected?

64. In the Dark A box containing twelve 40-watt light bulbs and eighteen 60-watt light bulbs is stored in your basement. Unfortunately, the box is stored in the dark and you need two 60-watt bulbs. What is the probability of randomly selecting two 60-watt bulbs from the box?

65. Randomly Playing Songs Suppose a Spotify playlist you just created has 13 tracks. After listening to the playlist, you decide that you like 5 of the songs. The random feature on Spotify is set up to play each of the 13 songs once in a random order. Find the probability that among the first 4 songs played

- (a) you like 2 of them;

- (b) you like 3 of them;
- (c) you like all 4 of them.

66. Packaging Error Through a manufacturing error, three cans marked “regular soda” were accidentally filled with diet soda and placed into a 12-pack. Suppose that three cans are randomly selected from the 12-pack.

- (a) Determine the probability that exactly two contain diet soda.
- (b) Determine the probability that exactly one contains diet soda.
- (c) Determine the probability that all three contain diet soda.

67. Three of a Kind You are dealt 5 cards from a standard 52-card deck. Determine the probability of being dealt three of a kind (such as three aces or three kings) by answering the following questions:

- (a) How many ways can 5 cards be selected from a 52-card deck?
- (b) Each deck contains 4 twos, 4 threes, and so on. How many ways can three of the same card be selected from the deck?
- (c) The remaining 2 cards must be different from the 3 chosen and different from each other. For example, if we drew three kings, the 4th card cannot be a king. After selecting the three of a kind, there are 12 different ranks of card remaining in the deck that can be chosen. If we have three kings, then we can choose twos, threes, and so on. Of the 12 ranks remaining, we choose 2 of them and then select one of the 4 cards in each of the two chosen ranks. How many ways can we select the remaining 2 cards?
- (d) Use the General Multiplication Rule to compute the probability of obtaining three of a kind. That is, what is the probability of selecting three of a kind and two cards that are not like?

68. Two of a Kind Follow the outline presented in Problem 67 to determine the probability of being dealt exactly one pair.

69. Acceptance Sampling Suppose that you have just received a shipment of 20 modems. Although you don’t know this, 3 of the modems are defective. To determine whether you will accept the shipment, you randomly select 4 modems and test them. If all 4 modems work, you accept the shipment. Otherwise, the shipment is rejected. What is the probability of accepting the shipment?

70. Acceptance Sampling Suppose that you have just received a shipment of 100 televisions. Although you don’t know this, 6 are defective. To determine whether you will accept the shipment, you randomly select 5 televisions and test them. If all 5 televisions work, you accept the shipment; otherwise, the shipment is rejected. What is the probability of accepting the shipment?

71. ID Numbering The Federal Bureau of Investigation (FBI) maintains a records system that houses civil background checks and criminal histories in a database. Each file has its own identification number (ID). With its initial algorithm for generating IDs, the system will generate only 400 million unique IDs. Because of the number of individuals in the database, this algorithm was not sufficient—more numbers were needed. The new algorithm consists of eight characters. Each character can be a digit from 0 to 9 or one of 17 letters of the alphabet (letters

that can be confused with numbers, such as the number 1 and letter I, are excluded).

- (a) To the nearest billion, how many identification numbers are possible with this scheme?
- (b) The FBI does not allow the first character to be 0. To the nearest billion, how many identification numbers are possible?
- (c) The FBI does not allow any combinations of letters that spell out obscenities. So, there are 271 billion unique ID numbers that can be created by the new FBI algorithm. Based on your answer to part (b), how many combinations of obscenities are possible with the new FBI algorithm?

72. Password Policy According to the Sefton Council Password Policy (August 2007), the United Kingdom government recommends the use of “*Environ passwords with the following format: consonant, vowel, consonant, consonant, vowel, consonant, number, number (for example, pinray45).*”

- (a) Assuming passwords are not case sensitive, how many such passwords are possible (assume that there are 5 vowels and 21 consonants)?
- (b) How many passwords are possible if they are case sensitive?

5.6 Simulating Probability Experiments



Objective ① Use simulation to approximate probabilities

1 Use Simulation to Approximate Probabilities

Historical Note

Because simulations can be done physically, there is a long history of the practice. For example, Karl Pearson analyzed the outcomes of a large number of spins on roulette tables in Monte Carlo, France and learned that the roulette wheels in the casino may be out of balance. During the Second World War, researchers at the Los Alamos Scientific Laboratory needed to know how far neutrons travel through various materials. John von Neumann and Stanislas Ulam recommended that the problem be solved by modeling the experiment on a computer (because the question could not be solved using theoretical calculations). Because their work was secret, von Neumann called the method “Monte Carlo.” Ever since that time, this method of simulation has been called the *Monte Carlo Method*.

We introduced the idea of simulation in Section 5.1, where a computer was used to randomly flip a fair coin many times. This form of simulation was called a *random process* because the outcome of any particular flip of the coin could not be determined ahead of time, but over the course of many flips, the proportion of heads settled down to a specific value—namely, 0.5. We called this proportion the probability of observing a head. In doing these simulations, statistical applets were used to randomly generate outcomes to represent results from a random event (instead of physically flipping a coin or rolling a die).

However, applets are not the only technique that may be used for simulating probability experiments. Simulations can be tactile or random number generators from statistical spreadsheets may be used. Use of random-number generators to simulate outcomes from random events can be complicated and may require computer programming. The examples we present here serve as an introduction to the process and will give you a sense of the simulation methods that can be used to estimate probabilities.

EXAMPLE 1

Getting Out of Jail in Monopoly

Problem In the board game Monopoly, a player can get out of jail in one of three ways.

1. The player pays a \$50 fine.
2. The player uses a “Get Out of Jail” card.
3. The player rolls doubles.

If the player does not roll doubles after three rolls, the player must pay the \$50 fine. Use simulation to determine the probability that a player will not roll doubles after three consecutive rolls.

Approach Use a statistical spreadsheet to randomly generate integers from 1 to 6 in two columns. Each column will represent the outcome of the die (the number of pips displayed). Then, for each entry in a row, determine if both columns match in value (both are 1, both are 2, and so on). If the entries match, record “Yes” in the Doubles column and in the Outcome column enter “Success.” If the entries do not match, record “No” in the Doubles column. If there are three consecutive entries of “No,” then record “Fail” in the Outcome column. This will be done for 600 rolls.

Solution In the data set, each row represents a roll of the dice. See Figure 14 for a partial output of the simulated data.

Figure 14

Row	Die1	Die2	Doubles	Outcome
1	5	2	No	
2	3	4	No	
3	1	2	No	Fall
4	6	1	No	
5	4	3	No	
6	6	5	No	Fall
7	3	5	No	
8	2	6	No	
9	1	3	No	Fall
10	2	2	Yes	Success
11	6	6	Yes	Success
12	6	2	No	
13	6	2	No	
14	4	2	No	Fall

In the first roll, the first die was 5 and the second die was 2, so the player stays in jail. On the second roll, the first die was 3 and the second die was 4, so the player stays in jail. On the third roll, the first die was a 1 and the second die was a 2, so the player “Fails” (due to three consecutive rolls without a pair) and pays \$50. The next three rolls are not doubles, so we record “Fail.” Rolls seven through nine were not doubles, so we record “Fail”. On the tenth roll, the player got a pair of 2s, so the player got out of jail and we record “Success.” Continuing with this logic, the entries in the column “Outcome” are determined.

Now, determine the relative frequency of the data in the Outcome column. See Figure 15, which represents the relative frequencies of Fail and Success using StatCrunch. With 600 rolls of the dice, we were able to determine 237 Outcomes (the total number of Fail or Success).

Figure 15

Frequency table results for Outcome:	
Count = 237	
Outcome	Relative Frequency
Fail	0.62025316
Success	0.37974684

From the simulation, the approximate probability of rolling a pair within three rolls of the dice is 0.380.

Note: The exact probability assuming independence and using the Complement Rule is $1 - \left(\frac{5}{6}\right)^3 = 0.421$.

NW Now Work Problem 1

Now we look at another form of simulation. In Chapter 1 we discussed the role that randomness plays in data collection. When collecting data for an observational study, it is important that individuals are *randomly selected* to be in the study. This allows the results of the study to be extended to the population from which the individuals were randomly selected. When collecting data for a designed experiment, it is important that the individuals are *randomly assigned* to the various treatment groups in the study. This allows us to make statements of causation between the levels of treatment and the response variable in the study. In this section, we look at the role of randomness in randomly selecting individuals to be in a study. We will consider random assignment later in the course.

The randomness in selecting or assigning individuals leads to outcomes that are uncertain. For example, polling agencies (such as Gallup) routinely randomly select about 1000 individuals and pose a question to them such as “Do you believe the amount of income tax you pay to the federal government is fair?” Because the individuals are randomly selected, the number of individuals out of 1000 who believe the amount of federal income tax they pay is fair will likely be different for two different repetitions of the study. Simulation can be used to help describe these differences.

EXAMPLE 2

Random Selection—Qualitative Response

Problem: Unplugging refers to eliminating the use of social media, cell phones, and other technology. According to Harris Interactive, the proportion of adult Americans (aged 18 years or older) who attempt to “unplug” at least once a week is 0.45. There are approximately 241,000,000 Americans aged 18 years or older in the United States.

- Simulate obtaining a simple random sample of size 500 from the population. How many of the individuals sampled unplug? How many do not unplug? What proportion unplug at least once a week?
- Simulate obtaining a second simple random sample of size 500 from the population. How many of the individuals sampled unplug? How many do not unplug? What proportion unplug at least once a week? Why will the results of the first sample differ from those in the second sample?
- Now simulate obtaining at least 2000 more simple random samples of size 500 from the population. For each sample, determine the proportion of individuals sampled who unplug at least once a week. Draw a histogram of the 2000 proportions.
- Based on the simulation, what is the probability of obtaining a random sample where the proportion who unplug at least once a week is greater than 0.50 (even though the proportion of people in the population who unplug at least once a week is 0.45)? Would it be unusual to obtain a sample proportion greater than 0.5 from this population? Explain.

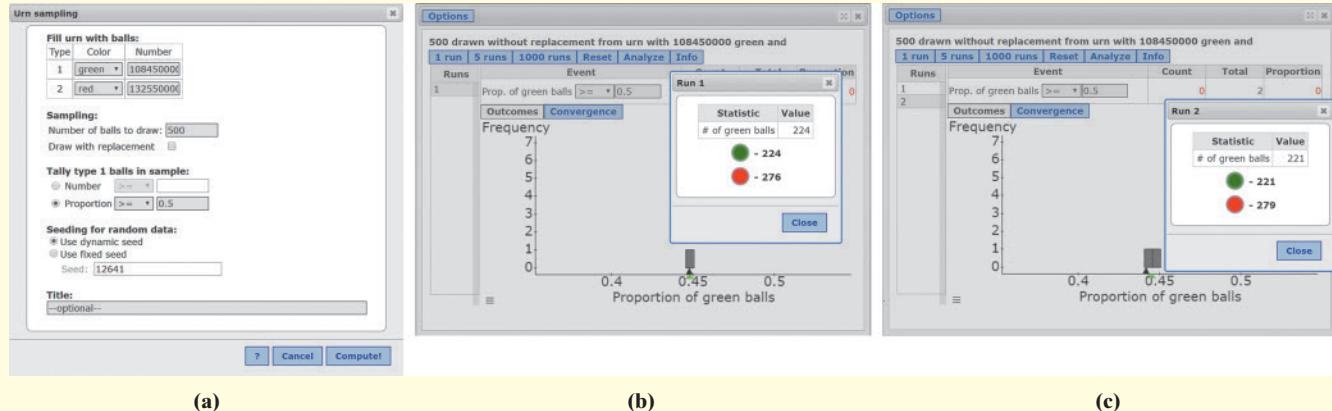
Approach:

- Assume the population proportion of adult Americans who unplug at least once a week is 0.45. Use this proportion to build a population in which $0.45(241,000,000) = 108,450,000$ individuals attempt to unplug at least once a week and the remainder ($241,000,000 - 108,450,000 = 132,550,000$) do not attempt to unplug at least once a week. In the StatCrunch Urn applet, let the 108,450,000 individuals who attempt to unplug be green balls; let the 132,550,000 individuals who do not unplug be red balls. Use the applet to simulate taking a simple random sample of size 500 from the population of adult Americans. For each individual selected, we are going to ask, “Do you unplug at least once a week?” If the ball selected is green, the answer to the question is “Yes”; if the ball selected is red, the answer to the question is “No.” Obtain the sample and results by clicking “1 run” in the applet.
- Obtain a second simple random sample by clicking “1 run” again.
- The applet may be used to obtain simple random samples 1000 at a time by clicking “1000 runs.” The StatCrunch Urn applet automatically draws a histogram of the results.
- The StatCrunch Urn applet will also determine the proportion of simulations where the proportion of individuals in the sample who unplug at least once a week is greater than 0.5.

Solution

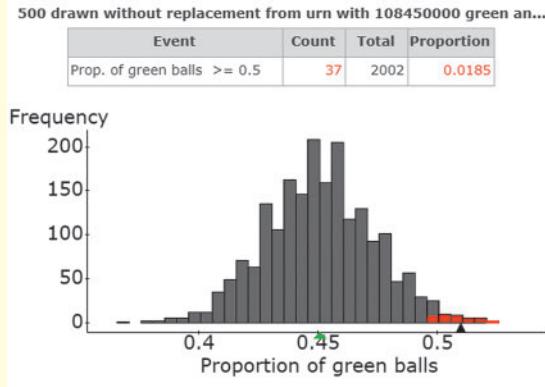
- In StatCrunch, select Applets > Simulation > Urn sampling. Fill in the dialogue window as shown in Figure 16(a). The green balls will be individuals who unplug, and the red balls will be individuals who do not unplug. Click Compute! to create the sampling urn. In the applet, click “1 run.” See Figure 16(b). The applet randomly selects 500 balls from the Urn and records the color. We obtained 224 green balls, which means 224 of the 500 individuals surveyed unplug at least once a week. The sample proportion of individuals who unplug at least once a week is $224/500 = 0.448$.
- Click “1 run” a second time. The results are shown in Figure 16(c). Notice we obtained 221 green balls, which means 221 of the 500 individuals surveyed unplug at least once a week. The sample proportion of individuals who unplug at least once a week is $221/500 = 0.442$. The results differ from those in part (a) because we “surveyed” 500 different individuals. So, it is the random sampling that leads to a different proportion of individuals who unplug (even though the proportion of individuals in the overall population who unplug has not changed).

Figure 16



- (c) Click “1000 runs” twice to obtain another 2000 simple random samples from the population. Figure 17 shows the proportion of the 2002 simple random samples that resulted in a proportion of individuals who unplug at least once a week.

Figure 17



- (d) Notice that 37 of the 2002, or $0.0185 = 1.85\%$, of the simple random samples resulted in a proportion of “unpluggers” greater than 0.5. Obtaining a simple random sample where the proportion of individuals who unplug is greater than 0.5 (when the population proportion is 0.45) would be unusual (because it is only expected to occur in about 2 of every 100 repetitions of this study).

Technology Step-by-Step

TI-83/84 Plus

Random Integers

Select MATH. Highlight PROB and select 5: randInt(. Type the lower limit, upper limit, and number of integers, n , to generate. Press $\text{STO} >$ and then 2ND 1 (to store the data in list L1). Press ENTER.

Excel

Random Integers

- Place the cursor in the cell in which you want to enter a random number. Type =RANDBETWEEN(lower limit, upper limit). For example, to generate a random integer between 1 and 6, type

$$= \text{RANDBETWEEN}(1,6)$$

- Select the bottom right corner of the cell with the random number and drag down to create the number of random integers desired.

Other Useful Excel Commands

- To find the sum of two columns (say A and B), type =A1 + B1 in cell C1. Drag the command down to create the sum of the two cells for all random observations.
Note: To hold a cell constant, use the \$ sign (for example, \$A\$1 holds the entry in cell A1 while using the drag feature).
- Many simulations require If/Else statements. In Excel, the syntax is =IF("logical test", value if TRUE, value if FALSE). For example, suppose we want to know if the entry in cell A1 is greater than the entry in cell B1, we would enter the following in cell C1: =IF(A1>B1, True, False)

StatCrunch**Random Integers**

1. Select Data > Simulate > Discrete Uniform.
2. Enter the number of rows and columns desired. Enter the Minimum and Maximum values of the parameters (such as 1 and 6 for a 6-sided die). Click Compute!.

Bernoulli Random Values

1. Select Data > Simulate > Bernoulli.
2. Enter the number of rows and columns desired. Enter the value of p , the proportion of the population with the characteristic. Click Compute!.

Computing Expressions in StatCrunch

1. Select Data > Compute > Expression.
2. Click Build.
3. Build the expression using the features available. For example, if column 1 is named "Die1" and column 2 is named "Die2", we could compute "Die1 + Die2". If we want to decide if Die1 is greater than Die2, we could compute "ifelse(Die1>Die2,True,False)".
4. Click Okay.
5. Click Compute!.



5.6 Assess Your Understanding

Applying the Concepts

NW **1. The Birthday Problem** An interesting problem from probability is the birthday problem. The problem deals with determining the likelihood that at least two people share the same birthday in a room of n individuals. The data file 5_6_1 located at www.pearsonhighered.com/sullivanstats represents simulated birthdates of 15 individuals 2000 times. Each birthdate was assigned a number from 1 (January 1) to 365 (December 31).

- (a) What is the birthdate of the randomly selected individual in row 1, column 1? Did the first simulation result in at least two people sharing the same birthdate?
- (b) Based on the simulation, what is the probability that at least two people will share the same birthdate?
- (c) Use the Birthday problem applet in StatCrunch to simulate randomly selecting 15 individuals. Repeat the simulation at least 2000 times and determine the proportion of samples that result in at least one common birthday.

DATA **2. Craps** Craps is a popular casino game. In its most basic form, two dice are initially thrown. If the outcome of the roll is 2, 3, or 12 (that is, the sum of the two die is 2, 3, or 12), the player loses his or her "pass line" bet. If the outcome of the roll is 7 or 11, the player wins his or her "pass line" bet. Any other number serves as the point. The player must roll that number again prior to rolling a seven. For example, if the point is eight, the player must roll an eight prior to a seven. The data at 5_6_2 located at www.pearsonhighered.com/sullivanstats represents outcomes from a simulated game of craps. Use the data to determine the probability of winning on a particular initial throw or point in craps using the column titled "Outcome."

DATA **3. The Price Is Right** In the gameshow *The Price Is Right*, two players get to spin a wheel with the values 05, 10, 15, ..., 95, 100 on it. In the game, Player 1 spins the wheel one time and gets to decide whether he/she would like to spin a second time in the hope that the sum of the two values gets closer to, but does not exceed, 100. Player 2 then spins the wheel; if the value shown is less than the score earned by Player 1, the player must spin again. If the second spin results in a sum less than that earned by Player 1 or greater than 100, Player 2 loses; if the second spin results in a sum greater than that earned by Player 1, but less than or equal to 100, Player 2 wins. If the two spins result in the same value as Player 1, the players tie. Suppose Player 1 ends up with 80 points. The data set 5_6_3 located at

www.pearsonhighered.com/sullivanstats represents simulated results. Use these simulated results to approximate the following probabilities.

- (a) Explain how to do a simulation to determine the probability that Player 2 will beat Player 1.
- (b) The column "Outcome_Spin1" represents the results of the first spin by Player 2. What is the probability Player 2 wins on the first spin?
- (c) What is the probability Player 2 ties on the first spin?
- (d) The results in the column "Strategy1" represents simulated results assuming the player does not accept the tie if Spin 1 results in 80. In the case of a tie after Spin 1, there is a second spin. In the case of a tie after Spin 2, there is a one-spin playoff. The results of any playoff are in the columns "Player1_Spin3" and "Player2_Spin3". Explain the sequence of events that led Player2 to win the game in Row 1. Explain the sequence of events that led Player2 to lose the game in Row 68.
- (e) What is the probability of Player 2 winning the game following Strategy 1?
- (f) The results in the column "Strategy2" represent simulated results assuming the player accepts the tie if Spin 1 results in 80. In the case of a tie after Spin 1, there is a one-spin playoff. The results of any playoff are in the columns "Player1_Spin3" and "Player2_Spin3". Explain the sequence of events that led Player2 to lose the game in Row 72.
- (g) What is the probability of Player 2 winning the game following Strategy 2?
- (h) Recommend a strategy for Player 2 to follow based on the results of parts (e) and (g).
- 4. Simulating Election Results** Suppose that polls suggest a candidate in a local school board election has 52% of the voters in favor of her as the candidate.
 - (a) Explain how you could use the integers from 1 to 100 to simulate votes.
 - (b) Use statistical software or a random table to simulate votes cast for the candidate assuming 500 ballots are cast. Repeat the simulation 20 times.
 - (c) Determine the proportion of votes cast for the candidate for each of the 20 simulations.
 - (d) What proportion of the simulations resulted in less than a majority (less than 50%) of the ballots cast for the candidate?

- (e) Repeat the simulation assuming 1000 ballots are cast for 20 simulations. What proportion of the simulations resulted in less than a majority of the ballots cast for the candidate? Explain what the result suggests.
- 5. Waiting for an Outcome** Suppose you would like to estimate the number of rolls it typically takes before observing a seven when rolling a pair of fair dice.
- Explain how you might use simulation to determine the typical number of rolls before observing a seven.
 - Use statistical software to simulate rolling a pair of dice and recording the sum. Determine the number of rolls needed before observing a seven. Repeat this simulation at least 20 times.
 - Approximate the number of rolls necessary before observing a seven by computing the sample mean number of rolls required from your simulations.
- 6. Online Dating** Recently, Pew Research reported that the proportion of adult Americans aged 18 or older who believe online dating is a good way to meet people is 0.59. There are approximately 241,000,000 Americans aged 18 years or older in the United States.
- Use the Urn applet in StatCrunch to simulate obtaining a simple random sample of size 500 from the population. How many of the individuals sampled believe online dating is a good way to meet people? How many do not believe online dating is a good way to meet people? What proportion believe online dating is a good way to meet people?
 - Simulate obtaining a second simple random sample of size 500 from the population. How many of the individuals sampled believe online dating is a good way to meet people? How many do not believe online dating is a good way to meet people? What proportion believe online dating is a good way to meet people? Why will the results of the first sample differ from those in the second sample?
 - Now simulate obtaining at least 2000 more simple random samples of size 500 from the population. Based on the

simulation, what is the probability of obtaining a random sample where the proportion of those who believe online dating is a good way to meet people is less than 0.55? Would it be unusual to obtain a sample proportion less than 0.55 from this population? Explain.

- 7. Moral Values** Recently, the Gallup Organization reported that the proportion of adult Americans aged 18 years or older who believe the state of moral values in the country is getting worse is 0.77. There are approximately 241,000,000 Americans aged 18 years or older in the United States.
- Use the Urn applet in StatCrunch to simulate obtaining a simple random sample of size 400 from the population. How many of the individuals sampled believe the state of moral values in the country is getting worse? How many do not believe the state of moral values in the country is getting worse? What proportion believe the state of moral values in the country is getting worse?
 - Simulate obtaining a second simple random sample of size 400 from the population. How many of the individuals sampled believe the state of moral values in the country is getting worse? How many do not believe the state of moral values in the country is getting worse? What proportion believe the state of moral values in the country is getting worse? Why will the results of the first sample differ from those in the second sample?
 - Now simulate obtaining at least 2000 more simple random samples of size 400 from the population. Based on the simulation, what is the probability of obtaining a random sample where the proportion of those who believe the state of moral values in the country is getting worse is greater than 0.8? Would it be unusual to obtain a sample proportion greater than 0.8 from this population? Explain.
 - Repeat parts (a) through (c) using a sample of size 800. Comment on the role sample size plays in the variability of outcomes. What law does this illustrate?

5.7 Putting It Together: Which Method Do I Use?

Objectives

- Determine the appropriate probability rule to use
- Determine the appropriate counting technique to use

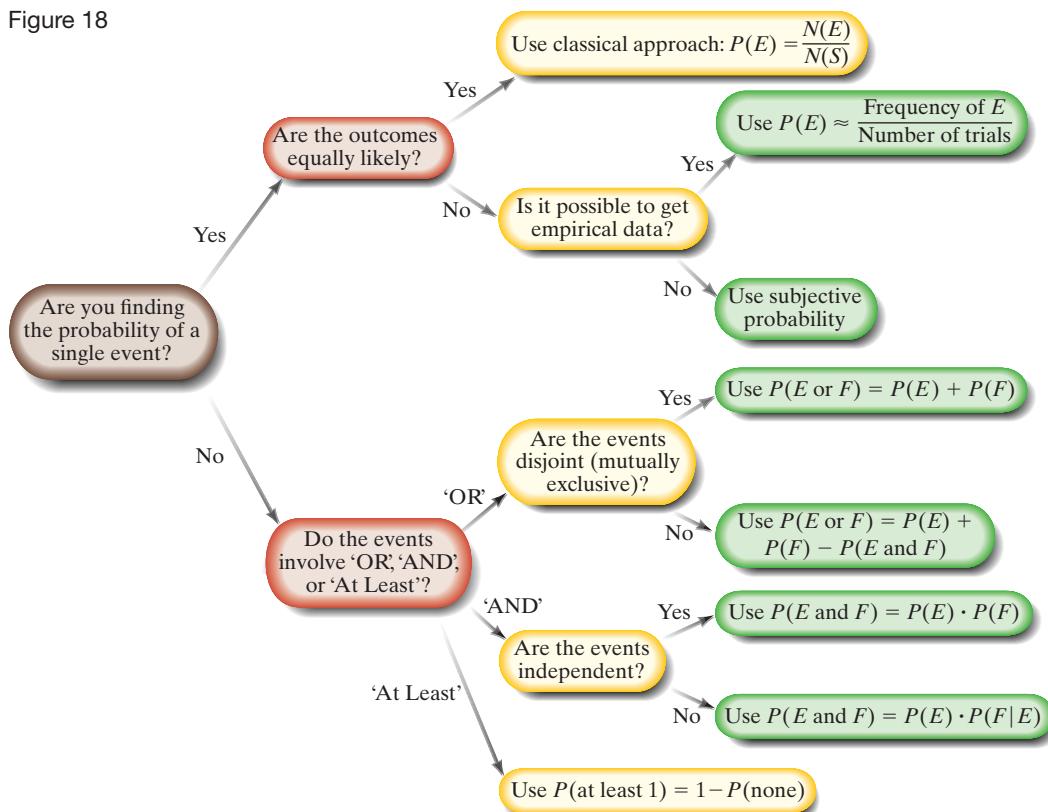


1 Determine the Appropriate Probability Rule to Use

This section will help you learn when to use a particular rule. To aid you, consider the flowchart in Figure 18 on the next page. While not all situations can be handled directly with the formulas provided, they can be combined and expanded to many more situations.

The first step is to determine whether we are finding the probability of a single event. If we are dealing with a single event, we must decide whether to use the classical method (equally likely outcomes), the empirical method (relative frequencies), or subjective probability. For experiments involving more than one event, we first decide which type of statement we have. For events involving “AND,” we must know if the events are independent. For events involving “OR,” we need to know if the events are disjoint (mutually exclusive).

Figure 18



EXAMPLE 1 Probability: Which Rule Do I Use?

Problem In the game show *Deal or No Deal?*, a contestant is presented with 26 suitcases that contain amounts ranging from \$0.01 to \$1,000,000. The contestant must pick an initial case that is set aside as the game progresses. The amounts are randomly distributed among the suitcases prior to the game as shown in Table 11. What is the probability that the contestant picks a case worth at least \$100,000?

Table 11

Prize	Number of Suitcases
\$0.01–\$100	8
\$200–\$1000	6
\$5,000–\$50,000	5
\$100,000–\$1,000,000	7

Approach Follow the flowchart in Figure 18.

Solution There is a single event, so we must decide among the empirical, classical, or subjective approaches to determine the probability. The probability experiment is selecting a suitcase. Each prize amount is randomly assigned to one of the 26 suitcases, so the outcomes are equally likely. Table 11 shows that seven cases contain at least \$100,000. Letting E = “worth at least \$100,000,” we compute $P(E)$ using the classical approach.

$$P(E) = \frac{N(E)}{N(S)} = \frac{7}{26} = 0.269$$

The probability the contestant selects a suitcase worth at least \$100,000 is 0.269. In 100 different games, we would expect about 27 games to result in a contestant choosing a suitcase worth at least \$100,000.

EXAMPLE 2 Probability: Which Rule Do I Use?

Problem According to a Harris poll, 14% of adult Americans have one or more tattoos, 50% have pierced ears, and 65% of those with one or more tattoos also have pierced ears. What is the probability that a randomly selected adult American has one or more tattoos and pierced ears?

Approach Follow the flowchart in Figure 18.

Solution We are finding the probability of an event involving ‘AND’. Letting T = “one or more tattoos” and E = “ears pierced,” we must find $P(T \text{ and } E)$. We need to determine if the two events, T and E , are independent. The problem statement tells us that $P(E) = 0.50$ and $P(E|T) = 0.65$. Because $P(E) \neq P(E|T)$, the two events are not independent. We find $P(T \text{ and } E)$ using the General Multiplication Rule.

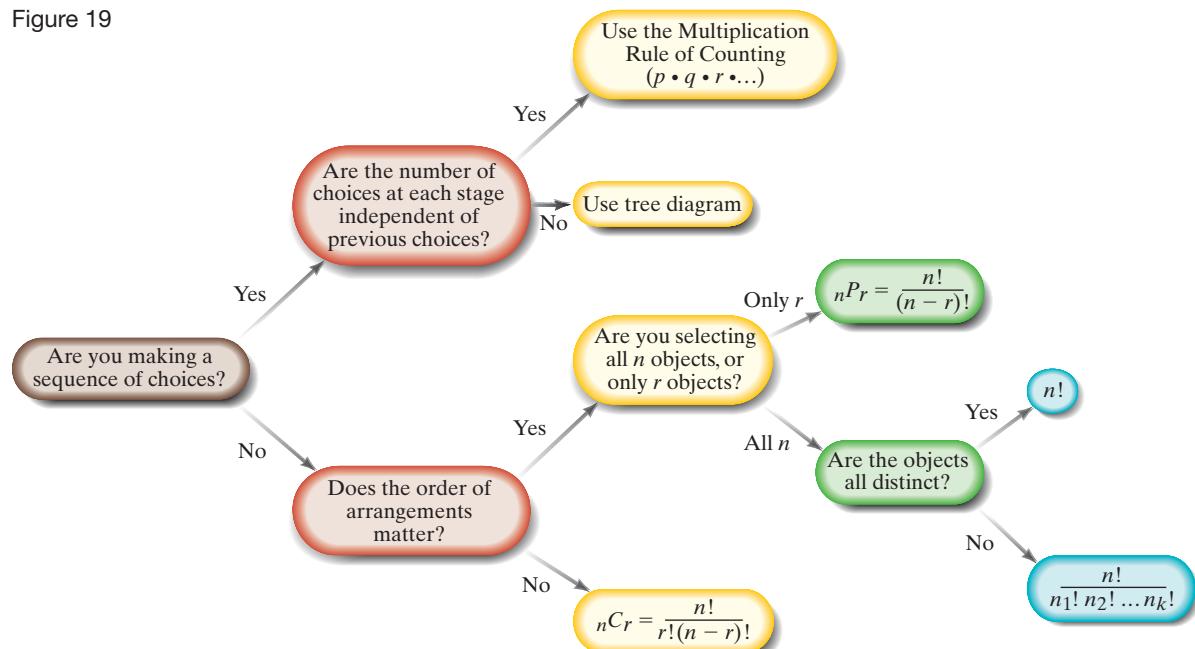
$$\begin{aligned}P(T \text{ and } E) &= P(T) \cdot P(E|T) \\&= (0.14)(0.65) \\&= 0.091\end{aligned}$$

So the probability of selecting an adult American at random who has one or more tattoos and pierced ears is 0.091. ●

② Determine the Appropriate Counting Technique to Use

To determine the appropriate counting technique to use, we need the ability to distinguish between a sequence of choices and an arrangement of items. We also need to determine whether order matters in the arrangements. See Figure 19. Keep in mind that one problem may require several counting rules.

Figure 19



We first must decide whether we have a sequence of choices or an arrangement of items. For a sequence of choices, we use the Multiplication Rule of Counting if the number of choices at each stage is independent of previous choices. If the number of choices at each stage is not independent of previous choices, we use a tree diagram. When determining the number of arrangements of items, we want to know whether the

order of selection matters. If order matters, we also want to know whether we are arranging all the items available or a subset of the items.

EXAMPLE 3 Counting: Which Technique Do I Use?

Problem The Hazelwood city council consists of five men and four women. How many different subcommittees can be formed that consist of three men and two women?

Approach Follow the flowchart in Figure 19.

Solution We need to find the number of subcommittees having three men and two women. So we consider a sequence of events: select the men, then select the women. Since the number of choices at each stage is independent of previous choices (the men chosen will not impact which women are chosen), we use the Multiplication Rule of Counting to obtain

$$N(\text{subcommittees}) = N(\text{ways to pick three men}) \cdot N(\text{ways to pick two women})$$

To select the three men, we must consider the number of arrangements of five men taken three at a time. Since the order of selection does not matter, we use the combination formula.

$$N(\text{ways to pick three men}) = {}_5C_3 = \frac{5!}{3! \cdot 2!} = 10$$

To select the two women, we must consider the number of arrangements of four women taken two at a time. Since the order of selection does not matter, we use the combination formula again.

$$N(\text{ways to pick two women}) = {}_4C_2 = \frac{4!}{2! \cdot 2!} = 6$$

Combining our results, we obtain $N(\text{subcommittees}) = 10 \cdot 6 = 60$. There are 60 possible subcommittees that contain three men and two women. 

EXAMPLE 4 Counting: Which Technique Do I Use?

Problem The Daytona 500, the season opening NASCAR event, has 43 drivers in the race. In how many different ways could the top four finishers (first, second, third, and fourth place) occur?

Approach Follow the flowchart in Figure 19.

Solution We need to find the number of ways to select the top four finishers. There are two different ways to solve this problem.

- View this as a sequence of choices, where the first choice is the first-place driver, the second choice is the second-place driver, and so on. There are 43 ways to pick the first driver, 42 ways to pick the second, 41 ways to pick the third, and 40 ways to pick the fourth. Use the Multiplication Rule of Counting. The number of ways the top four finishers can occur is

$$N(\text{top four}) = 43 \cdot 42 \cdot 41 \cdot 40 = 2,961,840$$

- Approach this problem as an arrangement of units. Because each race position is distinguishable, order matters. We are arranging the 43 drivers taken four at a time. Using our permutation formula, we get

$$N(\text{top four}) = {}_{43}P_4 = \frac{43!}{(43 - 4)!} = \frac{43!}{39!} = 43 \cdot 42 \cdot 41 \cdot 40 = 2,961,840$$

Again there are 2,961,840 different ways that the top four finishers could occur. 



5.7 Assess Your Understanding

Vocabulary and Skill Building

- What is the difference between a permutation and a combination?
- What method of assigning probabilities to a simple event uses relative frequencies?
- Which type of compound event is generally associated with multiplication? Which is generally associated with addition?
- Suppose that you roll a pair of dice 1000 times and get seven 350 times. Based on these results, what is the probability that the next roll results in seven?

For Problems 5 and 6, let the sample space be $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Suppose that the outcomes are equally likely. Compute the probability of the event:

- $E = \{1, 3, 5, 10\}$.
- $F = \text{"a number divisible by three."}$
- List all permutations of five objects a, b, c, d , and e taken three at a time without replacement.

In Problems 8 and 9, find the probability of the indicated event if $P(E) = 0.7$ and $P(F) = 0.2$.

- $P(E \text{ or } F)$ if E and F are mutually exclusive
- $P(E \text{ or } F)$ if $P(E \text{ and } F) = 0.15$

In Problems 10–12, evaluate each expression.

10. $\frac{6!2!}{4!}$ 11. ${}_7P_3$ 12. ${}_9C_4$

- Suppose that events E and F are independent, $P(E) = 0.8$, and $P(F) = 0.5$. What is $P(E \text{ and } F)$?
- Suppose that E and F are two events, $P(E \text{ and } F) = 0.4$, and $P(E) = 0.9$. Find $P(F|E)$.
- Suppose that E and F are two events, $P(E) = 0.9$, and $P(F|E) = 0.3$. Find $P(E \text{ and } F)$.
- List all combinations of five objects a, b, c, d , and e taken three at a time without replacement.

Applying the Concepts

- Soccer?** In a survey of 500 randomly selected Americans, it was determined that 22 play soccer. What is the probability that a randomly selected American plays soccer?

- Apartment Vacancy** A real estate agent conducted a survey of 200 landlords and asked how long their apartments remained vacant before a tenant was found. The results of the survey are shown in the table. The data are based on information obtained from the U.S. Census Bureau.

Duration of Vacancy	Frequency
Less than 1 month	42
1–2 months	38
2–4 months	45
4–6 months	30
6–12 months	24
1–2 years	13
2 years or more	8

- Construct a probability model for duration of vacancy.
- Is it unusual for an apartment to remain vacant for 2 years or more?

- Determine the probability that a randomly selected apartment is vacant for 1–4 months.

- Determine the probability that a randomly selected apartment is vacant for less than 2 years.

- Seating Arrangements** In how many ways can three men and three women be seated around a circular table (that seats six) assuming that women and men must alternate seats?

- Starting Lineups** Payton's futsal team consists of 10 girls, but only 5 can be on the field at any given time (four fielders and a goalie).

- How many starting lineups are possible if all players are selected without regard to position played?

- How many starting lineups are possible if either Payton or Jordyn must play goalie?

- Titanic Survivors** The following data represent the survival data for the ill-fated *Titanic* voyage by gender. The males are adult males and the females are adult females.

	Male	Female	Child	Total
Survived	338	316	57	711
Died	1352	109	52	1513
Total	1690	425	109	2224

Suppose a passenger is selected at random.

- What is the probability that the passenger survived?

- What is the probability that the passenger was female?

- What is the probability that the passenger was female or a child?

- What is the probability that the passenger was female and survived?

- What is the probability that the passenger was female or survived?

- If a female passenger is selected at random, what is the probability that she survived?

- If a child passenger is selected at random, what is the probability that the child survived?

- If a male passenger is selected at random, what is the probability that he survived?

- Do you think the adage "women and children first" was adhered to on the *Titanic*?

- Suppose two females are randomly selected. What is the probability both survived?

- Marijuana Use** According to the *Statistical Abstract of the United States*, about 17% of all 18- to 25-year-olds are current marijuana users.

- What is the probability that four randomly selected 18- to 25-year-olds are all marijuana users?

- What is the probability that among four randomly selected 18- to 25-year-olds at least one is a marijuana user?

23. SAT Reports Shawn is planning for college and takes the SAT test. While registering for the test, he is allowed to select three schools to which his scores will be sent at no cost. If there are 12 colleges he is considering, how many different ways could he fill out the score report form?

24. National Honor Society The distribution of National Honor Society members among the students at a local high school is shown in the table. A student's name is drawn at random.

Class	Total	National Honor Society
Senior	92	37
Junior	112	30
Sophomore	125	20
Freshman	120	0

- (a) What is the probability that the student is a junior?
 (b) What is the probability that the student is a senior, given that the student is in the National Honor Society?

25. Instant Winner In 2002, Valerie Wilson won \$1 million in a scratch-off game (Cool Million) from the New York lottery. Four years later, she won \$1 million in another scratch-off game (\$3,000,000 Jubilee), becoming the first person in New York state lottery history to win \$1 million or more in a scratch-off game twice. In the first game, she beat odds of 1 in 5.2 million to win. In the second, she beat odds of 1 in 705,600.

- (a) What is the probability that an individual would win \$1 million in both games if they bought one scratch-off ticket from each game?
 (b) What is the probability that an individual would win \$1 million twice in the \$3,000,000 Jubilee scratch-off game?

26. Text Twist In the game Text Twist, six letters are given and the player must form words of varying lengths using the letters provided. Suppose that the letters in a particular game are ENHSIC.

- (a) How many different arrangements are possible using all 6 letters?
 (b) How many different arrangements are possible using only 4 letters?
 (c) The solution to this game has three 6-letter words. To advance to the next round, the player needs at least one of the six-letter words. If the player simply guesses, what is the probability that he or she will get one of the six-letter words on their first guess of six letters?

27. Marriage and Education According to the U.S. Census Bureau, 20.2% of American women aged 25 years or older have a Bachelor's Degree; 16.5% of American women aged 25 years or older have never married; among American women aged 25 years or older who have never married, 22.8% have a Bachelor's Degree; and among American women aged 25 years or older who have a Bachelor's Degree, 18.6% have never married.

- (a) Are the events "have a Bachelor's Degree" and "never married" independent? Explain.

(b) Suppose an American woman aged 25 years or older is randomly selected, what is the probability she has a Bachelor's Degree and has never married? Interpret this probability.

28. Weather Forecast The weather forecast says there is a 10% chance of rain on Thursday. Jim wakes up on Thursday and sees overcast skies. Since it has rained for the past three days, he believes that the chance of rain is more likely 60% or higher. What method of probability assignment did Jim use?

29. Essay Test An essay test in European History has 12 questions. Students are required to answer 8 of the 12 questions. How many different sets of questions could be answered?

30. Exercise Routines Todd is putting together an exercise routine and feels that the sequence of exercises can affect his overall performance. He has 12 exercises to select from, but only has enough time to do 9. How many different exercise routines could he put together?

31. New Cars If the 2019 Hyundai Genesis has 2 engine types, 2 vehicle styles, 3 option packages, 8 exterior color choices, and 2 interior color choices, how many different Genesis's are possible?

32. Lingo In the gameshow *Lingo*, the team that correctly guesses a mystery word gets a chance to pull two Lingo balls from a bin. Balls in the bin are labeled with numbers corresponding to the numbers remaining on their Lingo board. There are also three prize balls and three red "stopper" balls in the bin. If a stopper ball is drawn first, the team loses their second draw. To form a Lingo, the team needs five numbers in a vertical, horizontal, or diagonal row. Consider the sample Lingo board below for a team that has just guessed a mystery word.

L	I	N	G	O
10			48	66
		34		74
		22	58	68
4	16		40	70
	26	52		64

- (a) What is the probability that the first ball selected is on the Lingo board?
 (b) What is the probability that the team draws a stopper ball on its first draw?
 (c) What is the probability that the team makes a Lingo on their first draw?
 (d) What is the probability that the team makes a Lingo on their second draw?



Chapter 5 Review

Summary

In this chapter, we took a break from the statistical process. The reason for this break is that probability forms the basis for statistical inference (which is the last step in the statistical process). Remember, inference takes information learned from a sample and generalizes it to a population along with a measure of reliability. Probability is used to measure the reliability in the results.

Probability is a measure of the likelihood of a random phenomenon or chance behavior. Because we are measuring a random phenomenon, there is short-term uncertainty. However, this short-term uncertainty gives rise to long-term predictability.

Probabilities are numbers between zero and one, inclusive. The closer a probability is to one, the more likely the event is to occur. If an event has probability zero, it is said to be impossible. Events with probability one are said to be certain.

We introduced four methods for computing probabilities: (1) the empirical method, (2) the classical method, (3) subjective probabilities, and (4) simulation. Empirical probabilities rely on the relative frequency with which an event happens. Classical probabilities require the outcomes in the experiment to be equally likely. We count the number of ways an event can happen and divide this by the number of possible outcomes of the experiment. Empirical probabilities require that an experiment be performed, whereas classical probability does not. Subjective probabilities are probabilities based on the opinion of the individual providing the probability. They are educated guesses about the likelihood of an event occurring, but still represent a legitimate way of assigning probabilities. Simulation is a technique used to re-create a random event. Simulations can be tactile or virtual (such as using a computer to pretend it is flipping a coin).

We are also interested in probabilities of multiple outcomes. For example, we might be interested in the

probability that either event E or event F happens. The Addition Rule is used to compute the probability of E or F . There are two versions of the Addition Rule. There is the Addition Rule for Disjoint Events. Two events are disjoint (or mutually exclusive) if they do not have any outcomes in common. That is, mutually exclusive events cannot happen at the same time. There is also the General Addition Rule, which does not require disjoint events.

The Multiplication Rule is used to compute the probability that both E and F occur. There are also two versions of the Multiplication Rule. There is the Multiplication Rule for Independent Events. Two events E and F are independent if knowing that one of the events occurs does not affect the probability of the other. There is also the General Multiplication Rule, which does not require independence.

Essentially, when you see the word “or” think Addition Rule; when you see the word “and” think Multiplication Rule.

The complement of an event E , denoted E^C , is all the outcomes in the sample space that are not in E .

We then introduced counting methods. The Multiplication Rule of Counting is used to count the number of ways a sequence of events can occur. Permutations are used to count the number of ways r items can be arranged from a set of n distinct items. We looked at two types of permutations—permutations without repetition (all items are distinct) and permutations with repetition (nondistinct items). Combinations are used to count the number of ways r items can be selected from a set of n distinct items without replacement and without regard to order. These counting techniques can be used to calculate probabilities using the classical method.

Finally, we introduced simulation, where we generated data to approximate probabilities of events or used an applet to approximate probabilities.

Vocabulary

Simulation (p. 227)

Random process (p. 227)

Probability (p. 228)

Outcome (p. 228)

Law of Large Numbers (p. 228)

Experiment (p. 229)

Sample space (p. 229)

Event (p. 229)

Probability model (p. 230)

Impossible event (p. 230)

Certainty (p. 230)

Unusual event (p. 230)

Equally likely outcomes (p. 232)

Tree diagram (p. 235)

Subjective probability (p. 236)

Disjoint (p. 242)

Mutually exclusive (p. 242)

Venn diagram (p. 242)

Contingency table (p. 246)

Two-way table (p. 246)

Row variable (p. 246)

Column variable (p. 246)

Cell (p. 246)

Complement (p. 247)

Independent (p. 253)

Dependent (p. 253)

Conditional probability (p. 260)

Factorial symbol (p. 272)

Permutation (p. 272)

Factorial notation (p. 273)

Combination (p. 274)

Formulas

Empirical Probability

$$P(E) \approx \frac{\text{frequency of } E}{\text{number of trials of experiment}}$$

Classical Probability

$$P(E) = \frac{\text{number of ways that } E \text{ can occur}}{\text{number of possible outcomes}} = \frac{N(E)}{N(S)}$$

Addition Rule for Disjoint Events

$$P(E \text{ or } F) = P(E) + P(F)$$

General Addition Rule

$$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$$

Probabilities of Complements

$$P(E^c) = 1 - P(E)$$

Multiplication Rule for Independent Events

$$P(E \text{ and } F) = P(E) \cdot P(F)$$

Multiplication Rule for n Independent Events

$$P(E_1 \text{ and } E_2 \text{ and } E_3 \cdots \text{ and } E_n) = P(E_1) \cdot P(E_2) \cdots \cdots P(E_n)$$

Conditional Probability Rule

$$P(F | E) = \frac{P(E \text{ and } F)}{P(E)} = \frac{N(E \text{ and } F)}{N(E)}$$

General Multiplication Rule

$$P(E \text{ and } F) = P(E) \cdot P(F|E)$$

Factorial Notation

$$n! = n \cdot (n-1) \cdot (n-2) \cdots \cdots 3 \cdot 2 \cdot 1$$

Combination

$${}_nC_r = \frac{n!}{r!(n-r)!}$$

Permutation

$${}_nP_r = \frac{n!}{(n-r)!}$$

Permutations with Nondistinct Items

$$\frac{n!}{n_1! \cdot n_2! \cdots \cdots n_k!}$$

Objectives

Section	You should be able to ...	Example(s)	Review Exercises
5.1	1 Understand random processes and the Law of Large Numbers (p. 227) 2 Apply the rules of probabilities (p. 230) 3 Compute and interpret probabilities using the empirical method (p. 231) 4 Compute and interpret probabilities using the classical method (p. 232) 5 Recognize and interpret subjective probabilities (p. 236)	p. 227 2 3, 4, 7(a) 5, 6, 7(b)	33 1, 13(d), 15 14(a), 15, 16(a) and (b), 17(a), 30 2–4, 13(a) and (d), 32(a) and (b) 28
5.2	1 Use the Addition Rule for Disjoint Events (p. 242) 2 Use the General Addition Rule (p. 245) 3 Compute the probability of an event using the Complement Rule (p. 247)	1 and 2 3 and 4 5 and 6	3, 4, 7, 13(b) and (c) 6, 16(d) 5, 14(b), 17(b)
5.3	1 Identify independent events (p. 253) 2 Use the Multiplication Rule for Independent Events (p. 254) 3 Compute at-least probabilities (p. 256)	1 2 and 3 4	9, 16(g), 32(f) 8, 14(c) and (d), 17(c) and (e), 18, 19 14(e), 17(d) and (f)
5.4	1 Compute conditional probabilities (p. 260) 2 Compute probabilities using the General Multiplication Rule (p. 262)	1 through 3 4 through 6	11, 16(f), 32(c) and (d) 10, 20, 29
5.5	1 Solve counting problems using the Multiplication Rule (p. 269) 2 Solve counting problems using permutations (p. 272) 3 Solve counting problems using combinations (p. 274) 4 Solve counting problems involving permutations with nondistinct items (p. 276) 5 Compute probabilities involving permutations and combinations (p. 278)	1 through 4 5 and 6 7 through 9 10 and 11 12 and 13	21 12(e) and (f), 22 12(c) and (d), 24 23 25, 26
5.6	1 Use simulation to approximate probabilities (p. 282)	1 and 2	27
5.7	1 Determine the appropriate probability rule to use (p. 287) 2 Determine the appropriate counting technique to use (p. 289)	1 and 2 3 and 4	13–20, 25, 26, 29(c) and (d), 30 21–25

Review Exercises

- 1. (a)** Which among the following numbers could be the probability of an event?

0, -0.01, 0.75, 0.41, 1.34

- (b)** Which among the following numbers could be the probability of an event?

$$\frac{2}{5}, \frac{1}{3}, -\frac{4}{7}, \frac{4}{3}, \frac{6}{7}$$

For Problems 2–5, let the sample space be

$S = \{ \text{red, green, blue, orange, yellow} \}$. Suppose that the outcomes are equally likely. Compute the probability of the event:

2. $E = \{\text{yellow}\}$
 3. $F = \{\text{green or orange}\}$
 4. $E = \{\text{red or blue or yellow}\}$
 5. Suppose that $E = \{\text{yellow}\}$. Compute the probability of E^c .
 6. Suppose that $P(E) = 0.76$, $P(F) = 0.45$, and $P(E \text{ and } F) = 0.32$. What is $P(E \text{ or } F)$?
 7. Suppose that $P(E) = 0.36$, $P(F) = 0.12$, and E and F are mutually exclusive. What is $P(E \text{ or } F)$?
 8. Suppose that events E and F are independent. In addition, $P(E) = 0.45$ and $P(F) = 0.2$. What is $P(E \text{ and } F)$?
 9. Suppose that $P(E) = 0.8$, $P(F) = 0.5$, and $P(E \text{ and } F) = 0.24$. Are events E and F independent? Why?
 10. Suppose that $P(E) = 0.59$ and $P(F|E) = 0.45$. What is $P(E \text{ and } F)$?
 11. Suppose that $P(E \text{ and } F) = 0.35$ and $P(F) = 0.7$. What is $P(E|F)$?

12. Determine the value of each of the following:

- (a)** $7!$ **(b)** $0!$
(c) ${}_9C_4$ **(d)** ${}_{10}C_3$
(e) ${}_9P_2$ **(f)** ${}_{12}P_4$

13. Roulette In the game of roulette, a wheel consists of 38 slots, numbered 0, 00, 1, 2, . . . , 36. (See the photo in Problem 27 from Section 5.1.) To play the game, a metal ball is spun around the wheel and allowed to fall into one of the numbered slots. The slots numbered 0 and 00 are green, the odd numbers are red, and the even numbers are black.

- (a) Determine the probability that the metal ball falls into a green slot. Interpret this probability.
 - (b) Determine the probability that the metal ball falls into a green or a red slot. Interpret this probability.
 - (c) Determine the probability that the metal ball falls into 00 or a red slot. Interpret this probability.
 - (d) Determine the probability that the metal ball falls into the number 31 and a black slot simultaneously. What term is used to describe this event?

14. Traffic Fatalities In 2016, there were 34,439 traffic fatalities in the United States. Of these, 9477 were alcohol related.

- (a) What is the probability that a randomly selected traffic fatality in 2016 was alcohol related?
 - (b) What is the probability that a randomly selected traffic fatality in 2016 was not alcohol related?
 - (c) What is the probability that two randomly selected traffic fatalities in 2016 were both alcohol related?
 - (d) What is the probability that neither of two randomly selected traffic fatalities in 2016 were alcohol related?
 - (e) What is the probability that of two randomly selected traffic fatalities in 2016 at least one was alcohol related?

- 15. Long Life?** In a poll conducted by Genworth Financial, a random sample of adults was asked, “What age would you like to live to?” The results of the survey are given in the table.

Age	Number
18–79	126
80–89	262
90–99	263
100 or older	388

- (a) Construct a probability model of the data.
 - (b) Is it unusual for an individual to want to live between 18 and 79 years?

- 16. Gestation Period versus Weight** The following data represent the birth weights (in grams) of babies born in 2017, along with the period of gestation.

Birth Weight (grams)	Period of Gestation			Total
	Preterm	Term	Postterm	
Less than 1000	25,258	167	24	25,449
1000–1999	79,593	10,623	1075	91,291
2000–2999	222,289	663,255	38,754	924,298
3000–3999	115,256	2,241,340	155,098	2,511,694
4000–4999	7171	267,556	23,810	298,537
At least 5000	201	3667	363	4231
Total	449,768	3,186,608	219,124	3,855,500

Source: National Vital Statistics Report

- (a) What is the probability that a randomly selected baby born in 2017 was postterm?
 - (b) What is the probability that a randomly selected baby born in 2017 weighed 3000–3999 grams?
 - (c) What is the probability that a randomly selected baby born in 2017 weighed 3000–3999 grams and was postterm?
 - (d) What is the probability that a randomly selected baby born in 2017 weighed 3000–3999 grams or was postterm?
 - (e) What is the probability that a randomly selected baby born in 2017 weighed less than 1000 grams and was postterm? Is this event impossible?
 - (f) What is the probability that a randomly selected baby born in 2017 weighed 3000–3999 grams, given the baby was postterm?
 - (g) Are the events “postterm baby” and “weighs 3000–3999 grams” independent? Why?

17. Who Do You Trust? According to the National Constitution Center, 18% of Americans trust organized religion.

 - (a) If an American is randomly selected, what is the probability he or she trusts organized religion?
 - (b) If an American is randomly selected, what is the probability he or she does not trust organized religion?
 - (c) In a random sample of three Americans, all three indicated that they trust organized religion. Is this result surprising?
 - (d) If three Americans are randomly selected, what is the probability that at least one does not trust organized religion?
 - (e) Would it be surprising if a random sample of five Americans resulted in none indicating they trust organized religion?
 - (f) If five Americans are randomly selected, what is the probability that at least one trusts organized religion?

18. Pick 3 For the Illinois Lottery's PICK 3 game, a player must match a sequence of three repeatable numbers, ranging from 0 to 9, in exact order (for example, 3–7–2). With a single ticket, what is the probability of matching the three winning numbers?

19. Pick 4 The Illinois Lottery's PICK 4 game is similar to PICK 3, except a player must match a sequence of four repeatable numbers, ranging from 0 to 9, in exact order (for example, 5–8–5–1). With a single ticket, what is the probability of matching the four winning numbers?

20. Drawing Cards Suppose that you draw 3 cards without replacement from a standard 52-card deck. What is the probability that all 3 cards are aces?

21. Forming License Plates A license plate is designed so that the first two characters are letters and the last four characters are digits (0 through 9). How many different license plates can be formed assuming that letters and numbers can be used more than once?

22. Choosing a Seat If four students enter a classroom that has 10 vacant seats, in how many ways can they be seated?

23. Arranging Flags How many different vertical arrangements are there of 10 flags if 4 are white, 3 are blue, 2 are green, and 1 is red?

24. Simple Random Sampling How many different simple random samples of size 8 can be obtained from a population whose size is 55?

25. Arizona's Pick 5 In one of Arizona's lotteries, balls are numbered 1–35. Five balls are selected randomly, without replacement. The order in which the balls are selected does not matter. To win, your numbers must match the five selected. Determine your probability of winning Arizona's Pick 5 with one ticket.

26. Packaging Error Because of a mistake in packaging, a case of 12 bottles of red wine contained five Merlot and seven Cabernet, each without labels. All the bottles look alike and have an equal probability of being chosen. Three bottles are randomly selected.

- (a) What is the probability that all three are Merlot?
- (b) What is the probability that exactly two are Merlot?
- (c) What is the probability that none is a Merlot?

27. Simulation Use a graphing calculator or statistical software to simulate the playing of the game of roulette, using an integer distribution with numbers 1 through 38. Repeat the simulation 100 times. Let the number 37 represent 0 and the number 38 represent 00. Use the results of the simulation to answer the following questions.

- (a) What is the probability that the ball lands in the slot marked 7?
- (b) What is the probability that the ball lands either in the slot marked 0 or in the one marked 00?

28. Explain what is meant by a subjective probability. List some examples of subjective probabilities.

29. Playing Five-Card Stud In the game of five-card stud, one card is dealt face down to each player and the remaining four cards are dealt face up. After two cards are dealt (one down and one up), the players bet. Players continue to bet after each additional card is dealt. Suppose three cards have been dealt to each of five players at the table. You currently have three clubs in your hand, so you will attempt to get a flush (all cards in the same suit). Of the cards dealt, there are two clubs showing in other player's hands.

- (a) How many clubs are in a standard 52-card deck?
- (b) How many cards remain in the deck or are not known by you? Of this amount, how many are clubs?
- (c) What is the probability that you get dealt a club on the next card?
- (d) What is the probability that you get dealt two clubs in a row?
- (e) Should you stay in the game?

30. Mark McGwire During the 1998 Major League Baseball season, Mark McGwire of the St. Louis Cardinals hit 70 home runs. Of the 70 home runs, 34 went to left field, 20 went to left center field, 13 went to center field, 3 went to right center field, and 0 went to right field. Source: Miklasz, B., et al. *Celebrating 70: Mark McGwire's Historic Season*, Sporting News Publishing Co., 1998, p. 179.

- (a) What is the probability that a randomly selected home run was hit to left field? Interpret this probability.
- (b) What is the probability that a randomly selected home run was hit to right field?
- (c) Was it impossible for Mark McGwire to hit a homer to right field?

31. Lottery Luck In 1996, a New York couple won \$2.5 million in the state lottery. Eleven years later, the couple won \$5 million in the state lottery using the same set of numbers. The odds of winning the New York lottery twice are roughly 1 in 16 trillion, described by a lottery spokesperson as "galactically astronomical." Although it is highly unlikely that an individual will win the lottery twice, it is not "galactically astronomical" that *someone* will win a lottery twice. Explain why this is the case.

32. Coffee Sales The following data represent the number of cases of coffee or filters sold by four sales reps in a recent sales competition.

Salesperson	Gourmet	Single Cup	Filters	Total
Connor	142	325	30	497
Paige	42	125	40	207
Bryce	9	100	10	119
Mallory	71	75	40	186
Total	264	625	120	1009

- (a) What is the probability that a randomly selected case was sold by Bryce? Is this unusual?
- (b) What is the probability that a randomly selected case was Gourmet?
- (c) What is the probability that a randomly selected Single-Cup case was sold by Mallory?
- (d) What is the probability that a randomly selected Gourmet case was sold by Bryce? Is this unusual?
- (e) What can be concluded from the results of parts (a) and (d)?
- (f) Are the events "Mallory" and "Filters" independent? Explain.
- (g) Are the events "Paige" and "Gourmet" mutually exclusive? Explain.
- (h) Patti and John each played 100 games of Solitaire on their smart phone. Patti won 7 of her games and estimated the probability of winning Solitaire as 0.07. John won 6 of his games and estimated the probability of winning Solitaire as 0.06. Why are their estimates different?
- (i) Would an empirical probability based on 100 games or 1000 games provide a better estimate of the probability of winning Solitaire? Why?



Chapter Test

1. Which among the following numbers could be the probability of an event? $0.23, 0, \frac{3}{2}, \frac{3}{4}, -1.32$

For Problems 2–4, let the sample space be $S = \{\text{Chris, Adam, Elaine, Brian, Jason}\}$. Suppose that the outcomes are equally likely.

2. Compute the probability of the event $E = \{\text{Jason}\}$.
3. Compute the probability of the event $E = \{\text{Chris or Elaine}\}$.
4. Suppose that $E = \{\text{Adam}\}$. Compute the probability of E^c .
5. Suppose that $P(E) = 0.37$ and $P(F) = 0.22$.
 - (a) Find $P(E \text{ or } F)$ if E and F are mutually exclusive.
 - (b) Find $P(E \text{ and } F)$ if E and F are independent.
6. Suppose that $P(E) = 0.15$, $P(F) = 0.45$, and $P(F|E) = 0.70$.
 - (a) What is $P(E \text{ and } F)$?
 - (b) What is $P(E \text{ or } F)$?
 - (c) What is $P(E|F)$?
 - (d) Are E and F independent?

7. Determine the value of each of the following:

- (a) $8!$
- (b) ${}_{12}C_6$
- (c) ${}_{14}P_8$

8. Craps is a dice game in which two fair dice are cast. If the roller shoots a 7 or 11 on the first roll, he or she wins. If the roller shoots a 2, 3, or 12 on the first roll, he or she loses.

- (a) Compute the probability that the shooter wins on the first roll. Interpret this probability.
- (b) Compute the probability that the shooter loses on the first roll. Interpret this probability.
9. According to Gallup, 26% of adult Americans believe their diet is very healthy.
 - (a) What is the probability that a randomly selected adult American believes his or her diet is very healthy? Interpret this probability.
 - (b) What is the probability that a randomly selected adult American does not believe his or her diet is very healthy?

10. The following probability model shows the distribution of the most-popular-selling Girl Scout Cookies®.

Cookie Type	Probability
Thin Mints	0.25
Samoas®/Caramel deLites™	0.19
Peanut Butter Patties®/Tagalongs™	0.13
Peanut Butter Sandwich/Do-si-dos™	0.11
Shortbread/Trefoils	0.09
Other varieties	0.23

Source: www.girlscouts.org

- (a) Verify that this is a probability model.
- (b) If a girl scout is selling cookies to people who randomly enter a shopping mall, what is the probability that the next box sold will be Peanut Butter Patties®/Tagalongs™ or Peanut Butter Sandwich/Do-si-dos™?

- (c) If a girl scout is selling cookies to people who randomly enter a shopping mall, what is the probability that the next box sold will be Thin Mints, Samoas®/Caramel deLites™, or Shortbread/Trefoils?

- (d) What is the probability that the next box sold will not be Thin Mints?

11. The following represent the results of a survey in which individuals were asked to disclose what they perceive to be the ideal number of children.

	0	1	2	3	4	5	6	Total
Female	5	2	87	61	28	3	2	188
Male	3	2	68	28	8	0	0	109
Total	8	4	155	89	36	3	2	297

Source: Sullivan Statistics Survey.

- (a) What is the probability an individual believes the ideal number of children is 2?

- (b) What is the probability an individual is female and believes the ideal number of children is 2?

- (c) What is the probability a randomly selected individual who took this survey is female or believes the ideal number of children is 2?

- (d) Among the females, what is the probability the individual believes the ideal number of children is 2?

- (e) If an individual who believes the ideal number of children is 4 is randomly selected, what is the probability the individual is male?

12. During the 2018 season, the Chicago Cubs won 58% of their games. Assuming that the outcomes of the baseball games are independent and that the percentage of wins this season will be the same as in 2018, answer the following questions:

- (a) What is the probability that the Cubs will win two games in a row?

- (b) What is the probability that the Cubs will win seven games in a row?

- (c) What is the probability that the Cubs will lose at least one of their next seven games?

13. You just received a shipment of 10 DVD players. One DVD player is defective. You will accept the shipment if two randomly selected DVD players work. What is the probability that you will accept the shipment?

14. In the game of Jumble, the letters of a word are scrambled. The player must form the correct word. In a recent game in a local newspaper, the Jumble “word” was LINCEY. How many different arrangements are there of the letters in this “word”?

15. The U.S. Senate Appropriations Committee has 29 members and a subcommittee is to be formed by randomly selecting 5 of its members. How many different committees could be formed?

16. In Pennsylvania’s Cash 5 lottery, balls are numbered 1 to 43. Five balls are selected randomly, without replacement. The order in which the balls are selected does not matter. To win, your numbers must match the five selected. Determine your probability of winning Pennsylvania’s Cash 5 with one ticket.

17. A local area network requires eight characters for a password. The first character must be a letter, but the remaining seven characters can be either a letter or a digit (0 through 9). Lower- and uppercase letters are considered the same. How many passwords are possible for the local area network?

18. A survey distributed at the 28th Lunar and Planetary Science Conference asked respondents to estimate the chance that there was life on Mars. The median response was a 57% chance of life on Mars. Which method of finding probabilities was used to obtain this result? Explain why.

19. How many distinguishable DNA sequences can be formed using two As, four Cs, four Gs, and five Ts?

20. A student is taking a 40-question multiple-choice test. Each question has five possible answers. Since the student did not study for the test, he guesses on all the questions. Letting 0 or 1 indicate a correct answer, use the following line from a table of random digits to simulate the probability that the student will guess a question correctly.

73634 79304 78871 25956 59109 30573 18513 61760

Making an Informed Decision

The Effects of Drinking and Driving

Everyone knows that alcohol impairs reaction time. For example, one study indicates that reaction time at a blood alcohol concentration of 0.08% doubles from 1.5 to 3 seconds. Your job is to compile some probabilities based on existing data to help convince people that it is a really bad idea to get behind the wheel of a car after consuming alcohol. It is also a bad idea to get into a car with an individual who has had a few drinks.

Go to the Fatality Analysis Reporting System Encyclopedia, which is online at the National Highway Traffic Safety Administration website (under Data). This encyclopedia contains records of all fatal car crashes for any given year. Click Run a Query Using the FARS Web-Based Encyclopedia. Click Query FARS data. You might consider reading the documentation provided to get a sense as to the type of data that can be obtained. Select a year you wish to analyze and click Submit.

You should see a table with variables that can be considered as they relate to fatal vehicle crashes. For example, if you want to determine the probability that a crash resulted from driver alcohol involvement, you could check the Driver

Alcohol Involvement box. Then click Submit. In the drop-down menu, you can fine-tune the data obtained. Click Univariate Tabulation. The next screen allows for further refinement. Click Submit, and you will have the data for the current year. Compile a few probabilities for univariate analysis.



Next, consider relations among two variables. For example, is there a relation between gender and driver alcohol involvement? Check both variables and click Submit. Again, you can fine-tune the data obtained. Once this is complete, select Cross Tabulation. Now choose the appropriate column and row variables to create a contingency table and click Submit. Develop a few conditional probabilities to see the relation between your two variables. Are they independent? If not, what does the dependency tell us?

Write a report, complete with the probabilities showing variable dependencies, that can be used to convince people of the dangers of mixing alcohol consumption with driving.

6

Discrete Probability Distributions

Outline

- 6.1 Discrete Random Variables
- 6.2 The Binomial Probability Distribution

Making an Informed Decision



A woman who was shopping in Los Angeles had her purse stolen by a young, blonde female who was wearing a ponytail. Because there were no eyewitnesses and no real evidence, the prosecution used probability to make its case against the defendant. Your job is to play the role of both the prosecution and defense attorney to make probabilistic arguments both for and against the defendant. See the Decisions project on page 331.

Putting It Together

Recall, the probability of an event is the long-term proportion with which the event is observed. That is, if we conduct an experiment 1000 times and observe an outcome 300 times, we estimate that the probability of the outcome is $300/1000 = 0.3$. The more times we conduct the experiment, the more accurate this empirical probability will be. This is the Law of Large Numbers. Counting techniques may be used to obtain theoretical probabilities if the outcomes in the experiment are equally likely. This is called classical probability.

A probability model lists the possible outcomes of a probability experiment and each outcome's probability. A probability model must satisfy the rules of probability. In particular, all probabilities must be between 0 and 1, inclusive, and the sum of the probabilities must equal 1.

In this chapter, we introduce probability models for **random variables**. A random variable is a numerical measure of the outcome to a probability experiment. So, rather than listing specific outcomes of a probability experiment, such as heads or tails, we might list the number of heads obtained in, say, three flips of a coin. In Section 6.1, we discuss random variables and describe the distribution of discrete random variables (shape, center, and spread). In Section 6.2, we discuss a specific discrete probability distribution: **the binomial probability distribution**.

6.1 Discrete Random Variables



Preparing for This Section Before getting started, review the following:

- Discrete versus continuous variables (Section 1.1, pp. 7–9)
- Relative frequency histograms for discrete data (Section 2.2, pp. 77–78)
- Mean (Section 3.1, pp. 108–110)
- Standard deviation (Section 3.2, pp. 123–128)
- Mean from grouped data (Section 3.3, pp. 139–140)
- Standard deviation from grouped data (Section 3.3, pp. 141–143)

Objectives

- ① Distinguish between discrete and continuous random variables
- ② Identify discrete probability distributions
- ③ Graph discrete probability distributions
- ④ Compute and interpret the mean of a discrete random variable
- ⑤ Interpret the mean of a discrete random variable as an expected value
- ⑥ Compute the standard deviation of a discrete random variable

1 Distinguish between Discrete and Continuous Random Variables

Consider a probability experiment in which we flip a coin two times. The outcomes of the experiment are $\{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$. Rather than being interested in a particular outcome, we might be interested in the number of heads (0 Heads, 1 Head, 2 Heads). If the outcome of a probability experiment is a numerical result, we say the outcome is a *random variable*.

Definition

A **random variable** is a numerical measure of the outcome of a probability experiment, so its value is determined by chance. Random variables are typically denoted using capital letters such as X .

So, in our coin-flipping example, if the random variable X represents the number of heads in two flips of a coin, the possible values of X are $x = 0, 1$, or 2 . Notice that we follow the practice of using a capital letter, such as X , to identify the random variable and a lowercase letter, x , to list the possible values of the random variable, or the sample space of the experiment.

As another example, consider an experiment that measures the time between arrivals of cars at a drive-through. The random variable T describes the time between arrivals, so the sample space of the experiment is $t > 0$.

There are two types of random variables, *discrete* and *continuous*.

Definitions

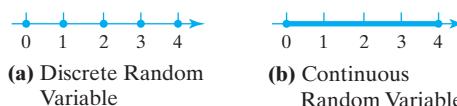
A **discrete random variable** has either a finite or countable number of values. The values of a discrete random variable can be plotted on a number line with space between each point. See Figure 1(a).

A **continuous random variable** has infinitely many values. The values of a continuous random variable can be plotted on a line in an uninterrupted fashion. See Figure 1(b).

IN OTHER WORDS

Discrete random variables typically result from counting ($0, 1, 2, 3$, and so on). Continuous random variables result from measurement.

Figure 1



EXAMPLE 1**Distinguishing between Discrete and Continuous Random Variables****CAUTION!**

Even though a radar gun may report the speed of a car as 37 miles per hour, it is actually any number greater than or equal to 36.5 mph and less than 37.5 mph. That is, $36.5 \leq s < 37.5$.

NW Now Work Problem 5

- (a) The number of As earned in a section of statistics with 15 students enrolled is a discrete random variable because its value results from counting. If the random variable X represents the number of As, then the possible values of X are $x = 0, 1, 2, \dots, 15$.
- (b) The number of cars that travel through a McDonald's drive-through in the next hour is a discrete random variable because its value results from counting. If the random variable X represents the number of cars, the possible values of X are $x = 0, 1, 2, \dots$.
- (c) The speed of the next car that passes a state trooper is a continuous random variable because speed is measured. If the random variable S represents the speed, the possible values of S are all positive real numbers; that is, $s > 0$.

In this chapter, we will concentrate on probabilities of discrete random variables. Chapter 7 discusses how to obtain probabilities for certain continuous random variables.

2 Identify Discrete Probability Distributions

Because the value of a random variable is determined by chance, we may assign probabilities to the possible values of the random variable.

Definition

The **probability distribution** of a discrete random variable X provides the possible values of the random variable and their corresponding probabilities. A probability distribution can be in the form of a table, graph, or mathematical formula.

EXAMPLE 2**A Discrete Probability Distribution****Table 1**

x	$P(x)$
0	0.01
1	0.10
2	0.38
3	0.51

Suppose we ask a basketball player to shoot three free throws. Let the random variable X represent the number of shots made, so $x = 0, 1, 2$, or 3. Table 1 shows a probability distribution for the random variable X .

We denote probabilities using the notation $P(x)$, where x is a specific value of the random variable. We read $P(x)$ as “the probability that the random variable X equals x .” For example, $P(3) = 0.51$ is read “the probability that the random variable X equals 3 is 0.51.”

Recall from Section 5.1 that probabilities must obey certain rules. Below are the rules for a discrete probability distribution using the notation just introduced.

Rules for a Discrete Probability Distribution

Let $P(x)$ denote the probability that the random variable X equals x ; then

1. $\sum P(x) = 1$
2. $0 \leq P(x) \leq 1$

IN OTHER WORDS

The first rule states that the sum of the probabilities must equal 1. The second rule states that each probability must be between 0 and 1, inclusive.

Table 1 from Example 2 is a discrete probability distribution because the sum of the probabilities equals 1 and each probability is between 0 and 1, inclusive.

EXAMPLE 3 Identifying Discrete Probability Distributions

Problem Which of the following is a discrete probability distribution?

(a)

x	$P(x)$
1	0.20
2	0.35
3	0.12
4	0.40
5	-0.07

(b)

x	$P(x)$
1	0.20
2	0.25
3	0.10
4	0.14
5	0.49

(c)

x	$P(x)$
1	0.20
2	0.25
3	0.10
4	0.14
5	0.31

Approach In a discrete probability distribution, the sum of the probabilities must equal 1, and all probabilities must be between 0 and 1, inclusive.

Solution

- (a) This is not a discrete probability distribution because $P(5) = -0.07$, which is less than 0.
- (b) This is not a discrete probability distribution because

$$\begin{aligned}\sum P(x) &= 0.20 + 0.25 + 0.10 + 0.14 + 0.49 \\ &= 1.18 \\ &\neq 1\end{aligned}$$

- (c) This is a discrete probability distribution because the sum of the probabilities equals 1, and each probability is between 0 and 1, inclusive.

NW Now Work Problem 9



Table 1 is an example of a discrete probability distribution in table form. We discuss discrete probability distributions using graphs next and using mathematical formulas in Section 6.2.

③ Graph Discrete Probability Distributions

In the graph of a discrete probability distribution, the horizontal axis represents the values of the discrete random variable and the vertical axis represents the corresponding probability of the discrete random variable. When graphing a discrete probability distribution, we want to emphasize that the data are discrete. Therefore, the graph of discrete probability distributions is drawn using vertical lines above each value of the random variable to a height that is the probability of the random variable.

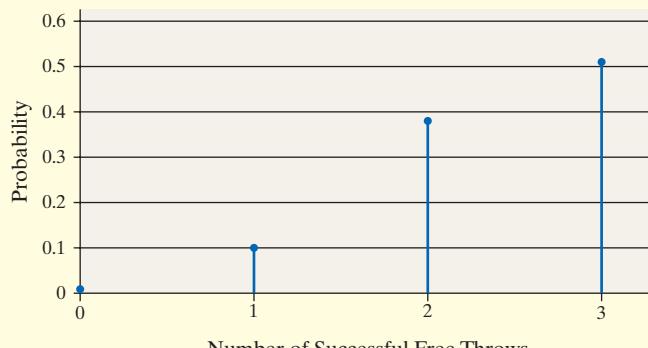
EXAMPLE 4 Graphing a Discrete Probability Distribution

Problem Graph the discrete probability distribution given in Table 1 from Example 2.

Approach In the graph of a discrete probability distribution, the horizontal axis represents the values of the discrete random variable and the vertical axis represents the corresponding probability of the discrete random variable. Draw the graph using vertical lines above each value of the random variable to a height that is the probability of the random variable.

Solution Figure 2 on the following page shows the graph of the distribution in Table 1.

Figure 2

Shooting Three Free Throws**NW Now Work Problems 17(a) and (b)**

Graphs of discrete probability distributions help determine the shape of the distribution. Recall that we describe distributions as skewed left, skewed right, or symmetric. The graph in Figure 2 is skewed left.

④ Compute and Interpret the Mean of a Discrete Random Variable

Remember, when describing the distribution of a variable, we describe its center, spread, and shape. We will use the mean to describe the center and use the standard deviation to describe the spread.

Let's see where the formula for computing the mean of a discrete random variable comes from. One semester I asked a small statistics class of 10 students to disclose the number of people living in their households. I obtained the following data:

$$2, 4, 6, 6, 4, 4, 2, 3, 5, 5$$

What is the mean number of people in the 10 households? We could find the mean by adding the observations and dividing by 10, but we will take a different approach. Letting the random variable X represent the number of people in the household, we obtain the probability distribution in Table 2.

Now compute the mean as follows:

$$\begin{aligned} \mu &= \frac{\sum x_i}{N} = \frac{2 + 4 + 6 + 6 + 4 + 4 + 2 + 3 + 5 + 5}{10} \\ &= \frac{\overbrace{2+2}^2 + \overbrace{3+1}^1 + \overbrace{4+4+4}^3 + \overbrace{5+5}^2 + \overbrace{6+6}^2}{10} \\ &= \frac{2 \cdot 2 + 3 \cdot 1 + 4 \cdot 3 + 5 \cdot 2 + 6 \cdot 2}{10} \\ &= 2 \cdot \frac{2}{10} + 3 \cdot \frac{1}{10} + 4 \cdot \frac{3}{10} + 5 \cdot \frac{2}{10} + 6 \cdot \frac{2}{10} \\ &= 2 \cdot P(2) + 3 \cdot P(3) + 4 \cdot P(4) + 5 \cdot P(5) + 6 \cdot P(6) \\ &= 2(0.2) + 3(0.1) + 4(0.3) + 5(0.2) + 6(0.2) \\ &= 4.1 \end{aligned}$$

We conclude that the mean of a discrete random variable is found by multiplying each possible value of the random variable by its corresponding probability and then adding these products.

Table 2

x	$P(x)$
2	$\frac{2}{10} = 0.2$
3	$\frac{1}{10} = 0.1$
4	$\frac{3}{10} = 0.3$
5	$\frac{2}{10} = 0.2$
6	$\frac{2}{10} = 0.2$

IN OTHER WORDS

To find the mean of a discrete random variable, multiply each value of the random variable by its probability. Then add these products.

The Mean of a Discrete Random Variable

The mean of a discrete random variable is given by the formula

$$\mu_X = \sum [x \cdot P(x)] \quad (1)$$

where x is the value of the random variable and $P(x)$ is the probability of observing the value x .

EXAMPLE 5**Computing the Mean of a Discrete Random Variable**

Problem Compute the mean of the discrete random variable given in Table 1 from Example 2.

Approach Find the mean of a discrete random variable by multiplying each value of the random variable by its probability and adding these products.

Solution Refer to Table 3. The first two columns represent the discrete probability distribution. The third column represents $x \cdot P(x)$.

Substitute into Formula (1) to find the mean number of free throws made.

$$\mu_X = \sum [x \cdot P(x)] = 0(0.01) + 1(0.10) + 2(0.38) + 3(0.51) = 2.39 \approx 2.4$$

Table 3

x	$P(x)$	$x \cdot P(x)$
0	0.01	$0 \cdot 0.01 = 0$
1	0.10	$1 \cdot 0.1 = 0.1$
2	0.38	$2 \cdot 0.38 = 0.76$
3	0.51	$3 \cdot 0.51 = 1.53$

We will follow the practice of rounding the mean and standard deviation to one more decimal place than the values of the random variable.

How to Interpret the Mean of a Discrete Random Variable

The mean of a discrete random variable can be thought of as the mean outcome of the probability experiment if we repeated the experiment many times. If we repeated the experiment in Example 5 of shooting three free throws many times, we would expect the mean number of free throws made to be around 2.4.

Interpretation of the Mean of a Discrete Random Variable

Suppose an experiment is repeated n independent times and the value of the random variable X is recorded. As the number of repetitions of the experiment increases, the mean value of the n trials will approach μ_X , the mean of the random variable X . In other words, let x_1 be the value of the random variable X after the first experiment, x_2 be the value of the random variable X after the second experiment, and so on. Then

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

The difference between \bar{x} and μ_X gets closer to 0 as n increases.

EXAMPLE 6**Interpretation of the Mean of a Discrete Random Variable**

Problem The basketball player from Example 2 is asked to shoot three free throws 100 times. Compute the mean number of free throws made.

Approach The player shoots three free throws and the number made is recorded. We repeat this experiment 99 more times and then compute the mean number of free throws made.

Solution Table 4 shows the results.

Table 4										
First experiment →	3	2	3	3	3	3	1	2	3	2
Second experiment →	2	3	3	1	2	2	2	2	2	3
Third experiment →	3	3	2	2	3	2	3	2	2	2
	3	3	2	3	2	3	3	2	3	1
	3	2	2	2	2	0	2	3	1	2
	3	3	2	3	2	3	2	1	3	2
	2	3	3	3	1	3	3	1	3	3
	3	2	2	1	3	2	2	2	3	2
	3	2	2	2	3	3	2	2	3	3
	2	3	2	1	2	3	3	2	3	3
	← Hundredth experiment									

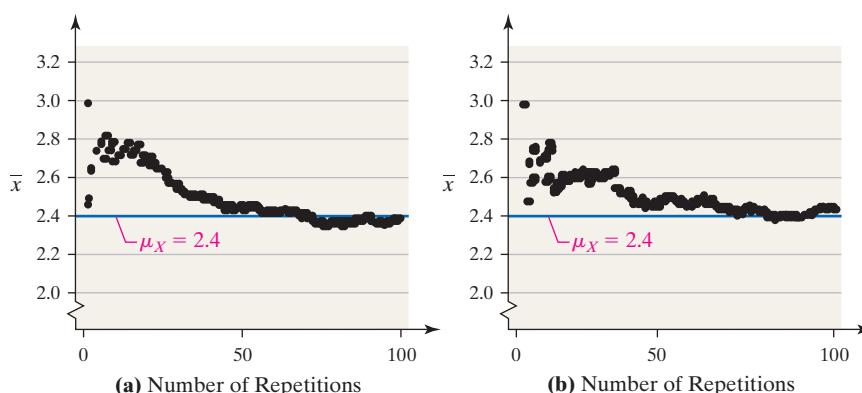
In the first experiment, the player made all three free throws. In the second experiment, the player made two out of three free throws. In the hundredth experiment, the player made all three free throws. The mean number of free throws made was

$$\bar{x} = \frac{3 + 2 + 3 + \dots + 3}{100} = 2.35$$

This is close to the theoretical mean of 2.4 (from Example 5). As the number of repetitions of the experiment increases, we expect \bar{x} to get even closer to 2.4. 

Figures 3(a) and 3(b) further illustrate the mean of a discrete random variable. Figure 3(a) shows the mean number of free throws made versus the number of repetitions of the experiment for the data in Table 4. Figure 3(b) shows the same information when the experiment of shooting three free throws is conducted a second time for 100 repetitions. In both plots the player starts “hot,” since the mean number of free throws made is above the theoretical level of 2.4. However, both graphs approach the theoretical mean of 2.4 as the number of repetitions of the experiment increases.

Figure 3



NW Now Work Problem 17(c)

5 Interpret the Mean of a Discrete Random Variable as an Expected Value

IN OTHER WORDS

The expected value of a discrete random variable is the mean of the discrete random variable.

Because the mean of a random variable represents what we would expect to happen in the long run, it is also called the **expected value**, $E(X)$. The interpretation of expected value is the same as the interpretation of the mean of a discrete random variable.

EXAMPLE 7 Finding the Expected Value

Historical Note

Christiaan Huygens was born on April 14, 1629, into an influential Dutch family. He studied Law and Mathematics at the University of Leiden from 1645 to 1647. From 1647 to 1649, he continued to study Law and Mathematics at the College of Orange at Breda. Among his many accomplishments, Huygens discovered the first moon of Saturn in 1655 and the shape of the rings of Saturn in 1656. While in Paris sharing his discoveries, he learned about probability through Fermat and Pascal. In 1657, Huygens published the first book on probability theory, in which he introduced the idea of expected value.



Table 5

x	$P(x)$
\$350 (survives)	0.998937
-\$249,650 (dies)	0.001063

NW Now Work Problem 25

Problem A term life insurance policy will pay a beneficiary a certain sum of money upon the death of the policyholder. These policies have premiums that must be paid annually. Suppose an 18-year-old male buys a \$250,000 1-year term life insurance policy for \$350. According to the *National Vital Statistics Report*, Vol. 58, No. 21, the probability that the male will survive the year is 0.998937. Compute the expected value of this policy to the insurance company.

Approach The experiment has two possible outcomes: survival or death. Let the random variable X represent the *payout* (money lost or gained), depending on survival or death of the insured. Assign probabilities to each payout and substitute these values into Formula (1).

Solution

Step 1 Because $P(\text{survives}) = 0.998937$, $P(\text{dies}) = 0.001063$. If the client survives the year, the insurance company makes \$350, or $x = \$350$. If the client dies during the year, the insurance company must pay \$250,000 to the client's beneficiary, but still keeps the \$350 premium, so $x = \$350 - \$250,000 = -\$249,650$. The value is negative because it is money paid by the insurance company. The probability distribution is listed in Table 5.

Step 2 The expected value of the policy (from the point of view of the insurance company) is

$$E(X) = \mu_X = \sum [x \cdot P(x)] = \$350(0.998937) + (-\$249,650)(0.001063) = \$84.25$$

Interpretation The company expects to make \$84.25 for each 18-year-old male client it insures. The \$84.25 profit of the insurance company is a long-term result. It does not make \$84.25 on each 18-year-old male it insures, but rather the average profit per 18-year-old male insured is \$84.25. Because this is a long-term result, the insurance "idea" will not work with only a few insured.

6 Compute the Standard Deviation of a Discrete Random Variable

We now introduce a method for computing the standard deviation of a discrete random variable.

Standard Deviation of a Discrete Random Variable

The standard deviation of a discrete random variable X is given by

$$\sigma_X = \sqrt{\sum [(x - \mu_X)^2 \cdot P(x)]} \quad (2a)$$

$$= \sqrt{\sum [x^2 \cdot P(x)] - \mu_X^2} \quad (2b)$$

where x is the value of the random variable, μ_X is the mean of the random variable, and $P(x)$ is the probability of observing a value of the random variable.

IN OTHER WORDS

The standard deviation of a discrete random variable is the square root of a weighted average of the squared deviations for which the weights are the probabilities.

EXAMPLE 8 Computing the Standard Deviation of a Discrete Random Variable

Problem Find the standard deviation of the discrete random variable given in Table 1 from Example 2.

Approach We will use Formula (2a) with the unrounded mean $\mu_X = 2.39$.

Solution Refer to Table 6. Columns 1 and 2 represent the discrete probability distribution. Column 3 represents $(x - \mu_X)^2 \cdot P(x)$. Find the sum of the entries in Column 3.

Table 6

x	$P(x)$	$(x - \mu_X)^2 \cdot P(x)$
0	0.01	$(0 - 2.39)^2 \cdot 0.01 = 0.057121$
1	0.10	$(1 - 2.39)^2 \cdot 0.10 = 0.19321$
2	0.38	$(2 - 2.39)^2 \cdot 0.38 = 0.057798$
3	0.51	$(3 - 2.39)^2 \cdot 0.51 = 0.189771$
$\sum [(x - \mu_X)^2 \cdot P(x)] = 0.4979$		

The standard deviation of the discrete random variable X is

$$\sigma_X = \sqrt{\sum [(x - \mu_X)^2 \cdot P(x)]} = \sqrt{0.4979} \\ \approx 0.7$$

NW Now Work Problem 17(d)

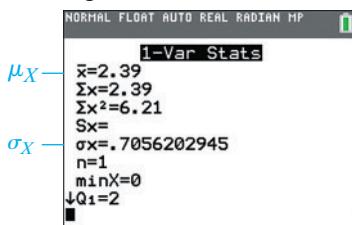
The variance of the discrete random variable is the value under the square root in the computation of the standard deviation. The variance of the discrete random variable in Example 8 is

$$\sigma_X^2 = 0.4979 \approx 0.5$$

The standard deviation of the discrete random variable X in the life insurance scenario from Example 7 is $\sigma_X = \$8146.6$. The standard deviation is large because there is a large range of possible outcomes (either a \$350 gain or \$249,650 loss).

EXAMPLE 9**Obtaining the Mean and Standard Deviation of a Discrete Random Variable Using Technology**

Figure 4



Problem Use statistical software or a graphing calculator to find the mean and the standard deviation of the random variable whose distribution is given in Table 1.

Approach We will use a TI-84 Plus CE graphing calculator to obtain the mean and standard deviation. The steps for determining the mean and standard deviation using a TI-83/84 Plus graphing calculator, StatCrunch, and Excel are given in the Technology Step-by-Step below.

Solution Figure 4 shows the results from a TI-84 Plus CE graphing calculator.

Note: The TI-84 does not find s_X when the sum of L_2 is 1.

Technology Step-by-Step**Finding the Mean and Standard Deviation of a Discrete Random Variable Using Technology****TI-83/84 Plus**

- Enter the values of the random variable in L1 and their corresponding probabilities in L2.
- Press STAT, highlight CALC, select 1: 1-VAR Stats, and press ENTER.
- Select L1 for List:; Select L2 for FreqList:; Highlight Calculate and press ENTER.

StatCrunch

- Enter the values of the random variable in column var1 and the corresponding probabilities in column var2. Name var1 x and var2 $P(x)$.
- Select **Stat**, **Calculators**, **Custom**. Select x for “Values in:”; Select $P(x)$ for “Weight in:” Click Compute!.

(continued)

Excel**Finding the Mean**

- Enter the values of the random variable in column A, and the corresponding probabilities in column B.
- Multiply each value from column A by its corresponding probability in column B. Place the product in column C. For example, in cell C1 enter “= A1*B1”. Copy the contents of cell C1 for the remaining values.
- Find the sum of the entries in column C using the sum command. For example, if there are entries in cells C1 through C10, enter “= sum(C1:C10)” into cell C11.

Finding the Standard Deviation

- Find the mean of the random variable.
- Subtract the mean from each value of the random variable in column A. Enter these values in column D.
- Square the entries in column D. Store the results in column E.
- Multiply the entries in column E by the probabilities in column B. Store the results in column F.
- Find the sum of the entries in column F. This represents the variance of the random variable.
- Find the square root of the sum of the entries in column F using the “= SQRT” command.



6.1 Assess Your Understanding

Vocabulary and Skill Building

- What is a random variable?
- What is the difference between a discrete random variable and a continuous random variable? Provide your own examples of each.
- What are the two requirements for a discrete probability distribution?
- In your own words, provide an interpretation of the mean (or expected value) of a discrete random variable.

In Problems 5–8, determine whether the random variable is discrete or continuous. In each case, state the possible values of the random variable.

- NW** 5. (a) The number of light bulbs that burn out in the next week in a room with 20 bulbs.
 (b) The time it takes to fly from New York City to Los Angeles.
 (c) The number of hits to a website in a day.
 (d) The amount of snow in Toronto during the winter.
6. (a) The time it takes for a light bulb to burn out.
 (b) The weight of a T-bone steak.
 (c) The number of free-throw attempts before the first shot is made.
 (d) In a random sample of 20 people, the number with type A blood.
7. (a) The amount of rain in Seattle during April.
 (b) The number of fish caught during a fishing tournament.
 (c) The number of customers arriving at a bank between noon and 1:00 P.M.
 (d) The time required to download a file from the Internet.
8. (a) The number of defects in a roll of carpet.
 (b) The distance a baseball travels in the air after being hit.
 (c) The number of points scored during a basketball game.
 (d) The square footage of a house.

In Problems 9–14, determine whether the distribution is a discrete probability distribution. If not, state why.

NW 9.

<i>x</i>	<i>P(x)</i>
0	0.2
1	0.2
2	0.2
3	0.2
4	0.2

10.

<i>x</i>	<i>P(x)</i>
0	0.1
1	0.5
2	0.05
3	0.25
4	0.1

11.

<i>x</i>	<i>P(x)</i>
10	0.1
20	0.23
30	0.22
40	0.6
50	-0.15

12.

<i>x</i>	<i>P(x)</i>
1	0
2	0
3	0
4	0
5	1

13.

<i>x</i>	<i>P(x)</i>
100	0.1
200	0.25
300	0.2
400	0.3
500	0.1

14.

<i>x</i>	<i>P(x)</i>
100	0.25
200	0.25
300	0.25
400	0.25
500	0.25

In Problems 15 and 16, determine the required value of the missing probability to make the distribution a discrete probability distribution.

15.

<i>x</i>	<i>P(x)</i>
3	0.4
4	?
5	0.1
6	0.2

16.

<i>x</i>	<i>P(x)</i>
0	0.30
1	0.15
2	?
3	0.20
4	0.15
5	0.05

Applying the Concepts

- NW 17. Televisions** In the Sullivan Statistics Survey I, individuals were asked to disclose the number of televisions in their household. In the following probability distribution, the random variable X represents the number of televisions in households.

Number of Televisions, x	$P(x)$
0	0
1	0.161
2	0.261
3	0.176
4	0.186
5	0.116
6	0.055
7	0.025
8	0.010
9	0.010

Source: Sullivan Statistics Survey I

- (a) Verify this is a discrete probability distribution.
- (b) Draw a graph of the probability distribution. Describe the shape of the distribution.
- (c) Determine and interpret the mean of the random variable X .
- (d) Determine the standard deviation of the random variable X .
- (e) What is the probability that a randomly selected household has three televisions?
- (f) What is the probability that a randomly selected household has three or four televisions?
- (g) What is the probability that a randomly selected household has no televisions? Would you consider this to be an impossible event?

- 18. Marriage** In the following probability distribution, the random variable X represents the number of marriages an individual aged 15 years or older has been involved in.

x	$P(x)$
0	0.272
1	0.575
2	0.121
3	0.027
4	0.004
5	0.001

Source: Based on data from the U.S. Census Bureau.

- (a) Verify that this is a discrete probability distribution.
- (b) Draw a graph of the probability distribution. Describe the shape of the distribution.
- (c) Compute and interpret the mean of the random variable X .
- (d) Compute the standard deviation of the random variable X .
- (e) What is the probability that a randomly selected individual 15 years or older was involved in two marriages?
- (f) What is the probability that a randomly selected individual 15 years or older was involved in at least two marriages?

- 19. Ichiro's Hit Parade** In the 2004 baseball season, Ichiro Suzuki of the Seattle Mariners set the record for the most hits in a season with a total of 262 hits. In the following probability distribution, the random variable X represents the number of hits Ichiro obtained in a game.

x	$P(x)$
0	0.1677
1	0.3354
2	0.2857
3	0.1491
4	0.0373
5	0.0248

Source: Chicago Tribune.

- (a) Verify that this is a discrete probability distribution.
- (b) Draw a graph of the probability distribution. Describe the shape of the distribution.
- (c) Compute and interpret the mean of the random variable X .
- (d) Compute the standard deviation of the random variable X .
- (e) What is the probability that in a randomly selected game Ichiro got 2 hits?
- (f) What is the probability that in a randomly selected game Ichiro got more than 1 hit?

- DATA 20. Waiting in Line** A Wendy's manager performed a study to determine a probability distribution for the number of people, X , waiting in line during lunch. The results were as follows:

x	$P(x)$	x	$P(x)$
0	0.011	7	0.098
1	0.035	8	0.063
2	0.089	9	0.035
3	0.150	10	0.019
4	0.186	11	0.004
5	0.172	12	0.006
6	0.132		

- (a) Verify that this is a discrete probability distribution.
- (b) Draw a graph of the probability distribution. Describe the shape of the distribution.
- (c) Compute and interpret the mean of the random variable X .
- (d) Compute the standard deviation of the random variable X .
- (e) What is the probability that eight people are waiting in line for lunch?
- (f) What is the probability that 10 or more people are waiting in line for lunch? Would this be unusual?

In Problems 21 and 22, (a) construct a discrete probability distribution for the random variable X [Hint: $P(x_i) = \frac{f_i}{N}$], (b) draw a graph of the probability distribution, (c) compute and interpret the mean of the random variable X , and (d) compute the standard deviation of the random variable X .

- 21. The World Series** The following data represent the number of games played in each World Series from 1923 to 2018.

x (games played)	Frequency
4	18
5	20
6	20
7	37

Source: Major League Baseball.

- 22. Ideal Number of Children** What is the ideal number of children to have in a family? The following data represent the ideal number of children for a random sample of 900 adult Americans.

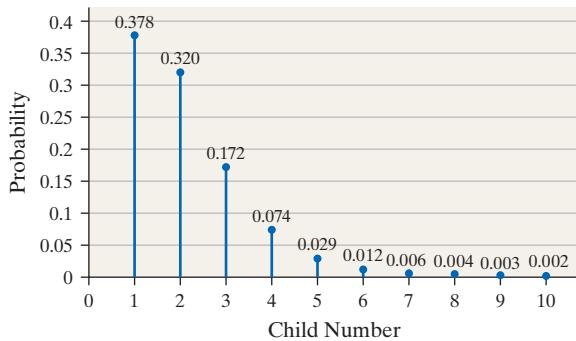
x (number of children)	Frequency
0	10
1	30
2	520
3	250
4	70
5	17
6	3

Source: Based on data from a Gallup poll.

- 23. Number of Births** The graph of the discrete probability distribution below represents the number of live births by a mother 50–54 years old who had a live birth in 2017.

Source: National Vital Statistics Report.

Number of Live Births,
50- to 54-Year-Old Mother

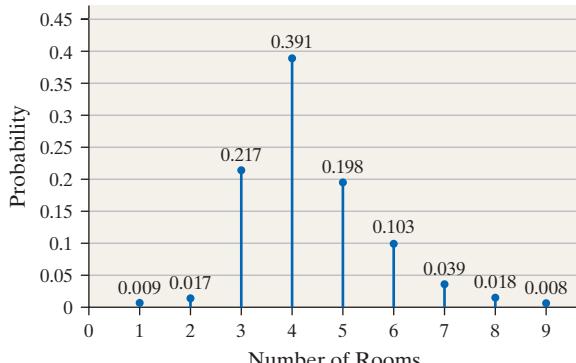


- (a) What is the probability that a randomly selected 50- to 54-year-old mother who had a live birth in 2017 has had her fourth live birth?
(b) What is the probability that a randomly selected 50- to 54-year-old mother who had a live birth in 2017 has had her fourth or fifth live birth?
(c) What is the probability that a randomly selected 50- to 54-year-old mother who had a live birth in 2017 has had her sixth or more live birth?
(d) If a 50- to 54-year-old mother who had a live birth in 2017 is randomly selected, how many live births would you expect the mother to have had?

- 24. Rental Units** The graph of the discrete probability distribution represents the number of rooms in rented housing units in 2017.

Source: U.S. Department of Commerce.

Number of Rooms in Rental Unit



- (a) What is the probability that a randomly selected rental unit has five rooms?
(b) What is the probability that a randomly selected rental unit has five or six rooms?
(c) What is the probability that a randomly selected rental unit has seven or more rooms?
(d) If a rental unit is randomly selected, how many rooms would you expect the unit to have?

NW 25. Life Insurance A life insurance company sells a \$250,000

1-year term life insurance policy to a 20-year-old female for \$200. According to the *National Vital Statistics Report*, 58(21), the probability that the female survives the year is 0.999544. Compute and interpret the expected value of this policy to the insurance company.

- 26. Life Insurance** A life insurance company sells a \$250,000 1-year term life insurance policy to a 20-year-old male for \$350. According to the *National Vital Statistics Report*, 58(21), the probability that the male survives the year is 0.998734. Compute and interpret the expected value of this policy to the insurance company.

- 27. Life Insurance Revisited** To the nearest dollar, what is the standard deviation of the value of the life insurance policy from Problem 25? Why is the value so high?

- 28. Life Insurance Revisited** To the nearest dollar, what is the standard deviation of the value of the life insurance policy from Problem 26? Why is the value so high?

- 29. Blackjack** Blackjack is a popular casino game in which a player is dealt two cards where the value of the card corresponds to the number on the card, face cards are worth ten, and aces are worth either one or eleven. The object is to get as close to 21 as possible without going over and have cards whose value exceeds that of the dealer. A blackjack is an ace and a ten in two cards. It pays 1.5 times the bet. The dealer plays last and must draw a card with sixteen and hold with seventeen or more. The following distribution shows the winnings and probability for a \$20 bet. In cases where the dealer and player have the same value, there is a tie (called a “push”). Source: “Examining a Gambler’s Claims: Probabilistic Fact-Checking and Don Johnson’s Extraordinary Winning Streak” by W.J. Hurley, Jack Brimberg, and Richard Kohar. *Chance* Vol. 27, 2014.

Winnings	Probability
0	0.0982
\$30	0.0483
\$20	0.389275
-\$20	0.464225

- (a) Compute and interpret the expected value of the game from the player’s point of view.
(b) Suppose over the course of one hour, a player can expect to be dealt about 40 hands. How much should a player expect to win or lose over the course of three hours?

- 30. Investment** An investment counselor calls with a hot stock tip. He believes that if the economy remains strong, the investment will result in a profit of \$50,000. If the economy grows at a moderate pace, the investment will result in a profit of \$10,000. However, if the economy goes into recession, the investment will result in a loss of \$50,000. You contact an economist who believes there is a 20% probability the economy will remain strong, a 70% probability the economy will grow at a moderate pace, and a 10% probability the economy will slip into recession. What is the expected profit from this investment?

31. Roulette In the game of roulette, a player can place a \$5 bet on the number 17 and have a $\frac{1}{38}$ probability of winning. If the metal ball lands on 17, the player wins \$175. Otherwise, the casino takes the player's \$5. What is the expected value of the game to the player? If you played the game 1000 times, how much would you expect to lose?

32. Video Poker The following table shows the net winnings from a \$1 bet in a video poker game.

Outcome	Probability	Profit
Royal Flush	0.0000023	\$799
Straight Flush	0.000142	\$49
Four of a Kind	0.00225	\$24
Full House	0.01098	\$8
Flush	0.01572	\$4
Straight	0.01842	\$3
Three of a Kind	0.06883	\$2
Two Pair	0.11960	\$1
Jacks or Better	0.18326	\$0
Less than Jacks or Better	0.58076	-\$1

Source: The Wizard of Odds.

- (a) Compute and interpret the expected net winnings from the player's point of view. Round your answer to three decimal places (nearest tenth of a penny).
- (b) Suppose over the course of one hour a player can expect to play 90 games. How much should the player expect to win or lose over the course of one hour?
- (c) What is the standard deviation of net winnings? What does this result suggest?

33. Powerball Powerball is a multistate lottery. The following probability distribution represents the cash prizes of Powerball with their corresponding probabilities.

x (cash prize, \$)	P(x)
Grand prize	0.00000000684
200,000	0.00000028
10,000	0.000001711
100	0.000153996
7	0.004778961
4	0.007881463
3	0.01450116
0	0.9726824222

Source: www.powerball.com

- (a) If the grand prize is \$15,000,000, find and interpret the expected cash prize. If a ticket costs \$1, what is your expected profit from one ticket?
- (b) If the grand prize is \$15,000,000, to the nearest dollar, what is the standard deviation of the cash prize? What does this value suggest?

- (c) To the nearest million, how much should the grand prize be so that you can expect a profit? Assume nobody else wins so that you do not have to share the grand prize.

- (d) Does the size of the grand prize affect your chance of winning? Explain.

34. SAT Test Penalty Some standardized tests, such as the SAT test, incorporate a penalty for wrong answers. For example, a multiple-choice question with five possible answers will have 1 point awarded for a correct answer and $\frac{1}{4}$ point deducted for an incorrect answer. Questions left blank are worth 0 points.

- (a) Find the expected number of points received for a multiple-choice question with five possible answers when a student just guesses.

- (b) Explain why there is a deduction for wrong answers.

35. The Extra Point After scoring a touchdown in football, the scoring team is entitled to either earn 1 extra point (by kicking the ball through goal posts) or 2 extra points (by advancing the ball past the goal line). Up until the 2016–2017 football season, teams who attempted to earn 1 extra point had a success rate of 0.993; teams who attempted to earn 2 extra points had a success rate of 0.480.

- (a) Use the concept of expected value to explain why teams would typically go for 1 extra point.

- (b) Use the concept of expected value to explain the justification in moving the snap for an extra point back to the 15-yard line after scoring a touchdown in professional football. **Hint:** The success rate of kicks where the snap is at the 15-yard line is 0.959.

36. Simulation Use the probability distribution from Problem 20 and a DISCRETE command for some statistical software to simulate 100 repetitions of the experiment. Approximate the mean and standard deviation of the random variable X based on the simulation. Repeat the simulation by performing 500 repetitions of the experiment. Approximate the mean and standard deviation of the random variable. Compare your results to the theoretical mean and standard deviation. What property is being illustrated?



37. Putting It Together: Sullivan Statistics Survey I One question from the Sullivan Statistics Survey I was "How many credit cards do you currently have?" This question was asked only of those individuals who have at least one credit card. Go to www.pearsonhighered.com/sullivanstats to obtain the survey results. Answer the following questions based on the results of the survey.

- (a) Determine the mean number of credit cards based on the raw data.

- (b) Determine the standard deviation number of credit cards based on the raw data.

- (c) Determine a probability distribution for the random variable, X , the number of credit cards issued to an individual.

- (d) Draw a graph of the discrete probability distribution for the random variable X . Describe the shape of the distribution.

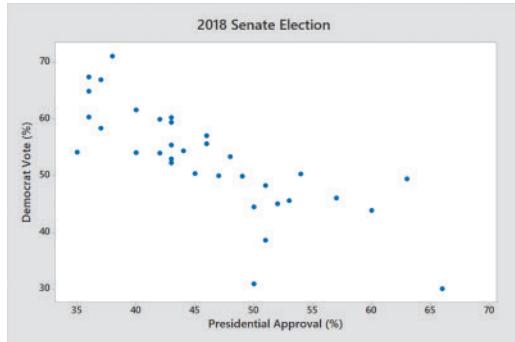
- (e) Determine the mean and standard deviation number of credit cards from the probability distribution found in part (c).

- (f) Determine the probability of randomly selecting an individual whose number of credit cards is more than two standard deviations from the mean. Is this result unusual?

- (g) Determine the probability of randomly selecting two individuals who are issued exactly two credit cards. [Hint: Are the events independent?] Interpret this result.

Retain Your Knowledge

38. 2018 Senate Election A major theme during the 2018 election in both the House and Senate was the popularity of President Trump (a Republican). The scatter diagram below shows the presidential approval rating (percent of registered voters who approve of the job the president is doing) versus the percentage of votes received by the Democratic senatorial candidate in the 2018 election.



- (a) What type of relation appears to exist between presidential approval ratings and percentage of Democrat votes?

- (b) There are 34 observations in the data set. The correlation between presidential approval and percent Republican is -0.781 . Does this suggest a linear relation exists between presidential approval and percent Republican? **Hint:** Use $n = 30$ from Table II.
- (c) The least-squares regression line treating presidential approval, x , as the explanatory variable and percent Democrat as the response variable, y , is $\hat{y} = -0.9147x + 94.9933$. Interpret the slope. What does this say about candidates of the same party as the president in Senate elections?
- (d) Does it make sense to interpret the intercept? Why or why not?
- (e) If President Trump's approval rating was 50%, what would you expect the percent of votes for the Democrat candidate to be?
- (f) What would President Trump's approval rating need to be in order for the Democrat candidate to expect to receive 50% of the vote? Round your answer to the nearest tenth of a percent.
- (g) In Wisconsin, Tammy Baldwin (a Democrat) won the election with 55.4% of the vote; President Trump's approval rating was 43%. Did Baldwin get a higher percent of the votes than would be expected for this presidential approval rating?

6.2 The Binomial Probability Distribution



Preparing for This Section Before getting started, review the following:

- Empirical Rule (Section 3.2, pp. 129–131)
- Addition Rule for Disjoint Events (Section 5.2, pp. 242–244)
- Complement Rule (Section 5.2, pp. 247–249)
- Independence (Section 5.3, pp. 253–254)
- Multiplication Rule for Independent Events (Section 5.3, pp. 254–256)
- Combinations (Section 5.5, pp. 274–276)

Objectives

- ① Determine whether a probability experiment is a binomial experiment
- ② Compute probabilities of binomial experiments
- ③ Compute the mean and standard deviation of a binomial random variable
- ④ Graph a binomial probability distribution

1 Determine Whether a Probability Experiment Is a Binomial Experiment

In Section 6.1, we stated that probability distributions could be presented using tables, graphs, or mathematical formulas. In this section, we introduce a specific type of discrete probability distribution that can be presented using a formula, the *binomial probability distribution*.

The **binomial probability distribution** is a discrete probability distribution that describes probabilities for experiments in which there are two mutually exclusive (disjoint) outcomes. These two outcomes are generally referred to as *success* (such as making a free throw) and *failure* (such as missing a free throw). Experiments in which only two outcomes are possible are referred to as *binomial experiments*, provided that certain criteria are met.

IN OTHER WORDS

The prefix *bi* means “two.” This should help remind you that binomial experiments deal with situations in which there are only two outcomes: success or failure.

Criteria for a Binomial Probability Experiment

An experiment is said to be a **binomial experiment** if

1. The experiment is performed a fixed number of times. Each repetition of the experiment is called a **trial**.
2. The trials are independent. This means that the outcome of one trial will not affect the outcome of the other trials.
3. For each trial, there are two mutually exclusive (disjoint) outcomes: success or failure.
4. The probability of success is the same for each trial of the experiment.

Let the random variable X be the number of successes in n trials of a binomial experiment. Then X is called a **binomial random variable**. Before introducing the method for computing binomial probabilities, it is worthwhile to introduce some notation.

Notation Used in the Binomial Probability Distribution

- There are n independent trials of the experiment.
- Let p denote the probability of success for each trial so that $1 - p$ is the probability of failure for each trial.
- Let X denote the number of successes in n independent trials of the experiment. So $0 \leq x \leq n$.

EXAMPLE 1**Identifying Binomial Experiments****Historical Note**

Jacob Bernoulli was born on December 27, 1654, in Basel, Switzerland. He studied philosophy and theology at the urging of his parents. (He resented this.) In 1671, he graduated from the University of Basel with a master's degree in philosophy. In 1676, he received a licentiate in theology. After earning his philosophy degree, Bernoulli traveled to Geneva to tutor. From there, he went to France to study with the great mathematicians of the time. One of Bernoulli's greatest works is *Ars Conjectandi*, published 8 years after his death. In this publication, Bernoulli proved the binomial probability formula. To this day, each trial in a binomial probability experiment is called a *Bernoulli trial*.



Problem Determine which of the following probability experiments qualify as binomial experiments. For those that are binomial experiments, identify the number of trials, probability of success, probability of failure, and possible values of the random variable X .

- (a) An experiment in which a basketball player who historically makes 80% of his free throws is asked to shoot three free throws, and the number of free throws made is recorded.
- (b) According to a recent Harris Poll, 28% of Americans state that chocolate is their favorite flavor of ice cream. Suppose a simple random sample of size 10 is obtained and the number of Americans who choose chocolate as their favorite ice cream flavor is recorded.
- (c) A probability experiment in which three cards are drawn from a deck without replacement and the number of aces is recorded.

Approach Determine whether the four conditions for a binomial experiment are satisfied.

1. The experiment is performed a fixed number of times.
2. The trials are independent.
3. There are only two possible outcomes of the experiment.
4. The probability of success for each trial is constant.

Solution

- (a) This is a binomial experiment because

1. There are $n = 3$ trials.
2. The trials are independent.
3. There are two possible outcomes: make or miss.
4. The probability of success (make) is 0.8 and the probability of failure (miss) is 0.2. The probabilities are the same for each trial.

The random variable X is the number of free throws made with $x = 0, 1, 2$, or 3 .

(continued)

- (b)** This is a binomial experiment because
1. There are 10 trials (the 10 randomly selected people).
 2. The trials are independent.*
 3. There are two possible outcomes: finding an American who chooses chocolate as his or her favorite ice cream, or not.
 4. The probability of success is 0.28 and the probability of failure is $1 - 0.28 = 0.72$. The random variable X is the number of people who choose chocolate as their favorite ice cream with $x = 0, 1, 2, 3, \dots, 10$.
- (c)** This is not a binomial experiment because the trials are not independent. The probability of an ace on the first trial is $\frac{4}{52} \approx 0.077$. Because we are sampling without replacement, if an ace is selected on the first trial, the probability of an ace on the second trial is $\frac{3}{51} \approx 0.059$. If an ace is not selected on the first trial, the probability of an ace on the second trial is $\frac{4}{51} \approx 0.078$.

NW Now Work Problem 9**CAUTION!**

The probability of success, p , is always associated with the random variable X , the number of successes. So if X represents the number of 18-year-olds involved in an accident, then p represents the probability of an 18-year-old being involved in an accident.

Note that the word *success* does not necessarily imply something positive. Success means that an outcome has occurred that corresponds with p , the probability of success. For example, a probability experiment might be to randomly select ten 18-year-old male drivers. If X denotes the number who have been involved in an accident within the last year, a success would mean obtaining an 18-year-old male who was involved in an accident. This outcome is not positive, but it is a success as far as the experiment goes.

② Compute Probabilities of Binomial Experiments

Now we will compute probabilities for a binomial random variable X . We present three methods for obtaining binomial probabilities: (1) the binomial probability distribution formula, (2) a table of binomial probabilities, and (3) technology. We develop the binomial probability formula in Example 2.

EXAMPLE 2**Constructing a Binomial Probability Distribution**

Problem According to the American Red Cross, 7% of people in the United States have blood type O-negative. A simple random sample of size 4 is obtained, and the number of people X with blood type O-negative is recorded. Construct a probability distribution for the random variable X .

Approach This is a binomial experiment with $n = 4$ trials. We define a success as selecting an individual with blood type O-negative. The probability of success, p , is 0.07, and X is the random variable representing the number of successes with $x = 0, 1, 2, 3$, or 4.

Step 1 Construct a tree diagram listing the various outcomes of the experiment by listing each outcome as *S* (success) or *F* (failure).

Step 2 Compute the probabilities for each value of the random variable X .

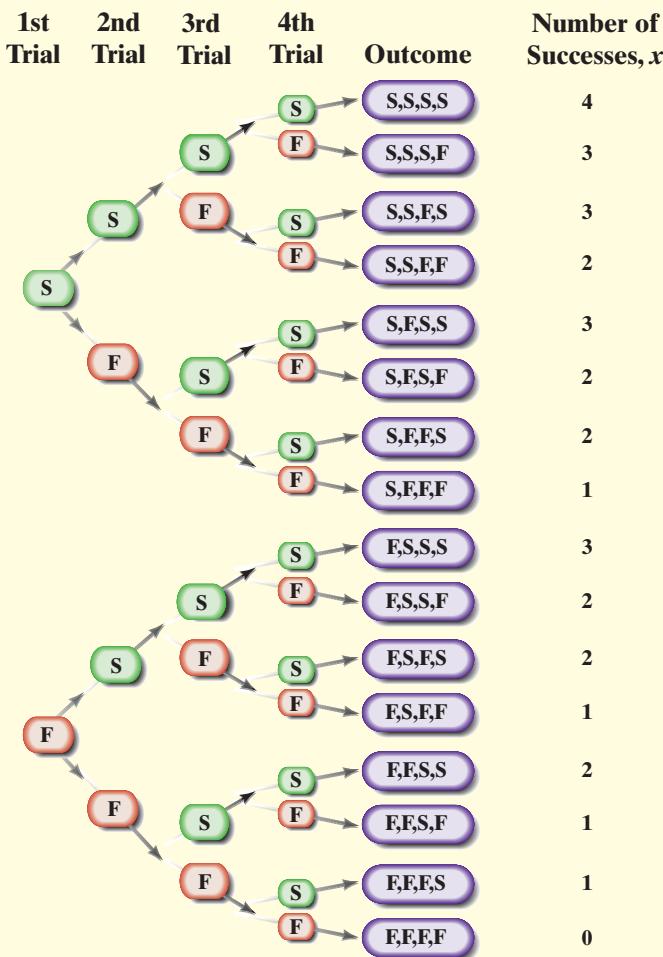
Step 3 Construct the probability distribution.

Solution

Step 1 The tree diagram in Figure 5 on the following page lists the 16 possible outcomes of the experiment.

*In sampling from large populations without replacement, the trials are assumed to be independent, provided that the sample size is small in relation to the size of the population. As a rule of thumb, if the sample size is less than 5% of the population size, the trials are assumed to be independent, although they are technically dependent. See Example 6 in Section 5.4.

Figure 5



Step 2 Now compute the probability for each possible value of the random variable X .

$$\begin{aligned}
 P(0) &= P(FFFF) = P(F) \cdot P(F) \cdot P(F) \cdot P(F) && \text{Multiplication Rule for Independent Events} \\
 &= (0.93)(0.93)(0.93)(0.93) \\
 &= (0.93)^4 \\
 &= 0.74805
 \end{aligned}$$

$$\begin{aligned}
 P(1) &= P(SFFF \text{ or } FSFF \text{ or } FFSF \text{ or } FFFS) && \text{Addition Rule for Disjoint Events} \\
 &= P(SFFF) + P(FSFF) + P(FFSF) + P(FFFS) \\
 &= (0.07)^1(0.93)^3 + (0.07)^1(0.93)^3 + (0.07)^1(0.93)^3 + (0.07)^1(0.93)^3 \\
 &= 4(0.07)^1(0.93)^3 && \text{Multiplication Rule for Independent Events} \\
 &= 0.22522
 \end{aligned}$$

$$\begin{aligned}
 P(2) &= P(SSFF \text{ or } SFSF \text{ or } SFFS \text{ or } FSSF \text{ or } FSFS \text{ or } FFSS) \\
 &= P(SSFF) + P(SFSF) + P(SFFS) + P(FSSF) + P(FSFS) + P(FFSS) \\
 &= (0.07)^2(0.93)^2 + (0.07)^2(0.93)^2 + (0.07)^2(0.93)^2 + (0.07)^2(0.93)^2 \\
 &\quad + (0.07)^2(0.93)^2 + (0.07)^2(0.93)^2 \\
 &= 6(0.07)^2(0.93)^2 \\
 &= 0.02543
 \end{aligned}$$

Compute $P(3)$ and $P(4)$ similarly and obtain $P(3) = 0.00128$ and $P(4) = 0.00002$. You are encouraged to verify these probabilities.

Step 3 We present these results in the probability distribution in Table 8.

Table 8

x	$P(x)$
0	0.74805
1	0.22522
2	0.02543
3	0.00128
4	0.00002

Notice some interesting results in Example 2. Consider the probability of obtaining $x = 1$ success:

$$P(1) = 4(0.07)^1(0.93)^3$$

4 is the number of ways we obtain 1 success in 4 trials of the experiment. Here, it is ${}_4C_1$. 0.07 is the probability of success and the exponent 1 is the number of successes. 0.93 is the probability of failure and the exponent 3 is the number of failures.

The coefficient 4 is the number of ways of obtaining one success in four trials. In general, the coefficient is ${}_nC_x$, the number of ways of obtaining x successes in n trials. The second factor in the formula, $(0.07)^1$, is the probability of success, p , raised to the number of successes, x . The third factor in the formula, $(0.93)^3$, is the probability of failure, $1 - p$, raised to the number of failures, $n - x$. The following *binomial probability distribution function (pdf)* formula holds for all binomial experiments.

CAUTION!

Before using the binomial probability distribution function, be sure the requirements for a binomial experiment are satisfied.

Binomial Probability Distribution Function

The probability of obtaining x successes in n independent trials of a binomial experiment is given by

$$P(x) = {}_nC_x p^x (1 - p)^{n-x} \quad x = 0, 1, 2, \dots, n \quad (1)$$

where p is the probability of success.

When reading probability problems, pay special attention to key phrases that translate into mathematical symbols. Table 9 lists various phrases and their corresponding mathematical equivalent.

Table 9

Phrase	Math Symbol
“at least” or “no less than” or “greater than or equal to”	\geq
“more than” or “greater than”	$>$
“fewer than” or “less than”	$<$
“no more than” or “at most” or “less than or equal to”	\leq
“exactly” or “equals” or “is”	$=$

EXAMPLE 3 Using the Binomial Probability Distribution Function

Problem According to CTIA, 72% of all adult Americans would rather give up chocolate than their cell phone. In a random sample of 10 adult Americans, what is the probability that

- (a) exactly 8 would rather give up chocolate?
- (b) fewer than 3 would rather give up chocolate?
- (c) at least 3 would rather give up chocolate?
- (d) the number of adult Americans who would rather give up chocolate is between 5 and 7, inclusive?

Approach This is a binomial experiment with $n = 10$ independent trials. We define a success as selecting an adult American who would rather give up chocolate. The probability of success, p , is 0.72. The possible values of the random variable X are $x = 0, 1, 2, \dots, 10$. We use Formula (1) to compute the probabilities.

Solution

$$\begin{aligned}
 \text{(a)} \quad P(8) &= {}_{10}C_8(0.72)^8(1 - 0.72)^{10-8} & n = 10, x = 8, p = 0.72 \\
 &= \frac{10!}{8!(10-8)!}(0.72)^8(0.28)^2 & {}_nC_x = \frac{n!}{x!(n-x)!} \\
 &= 45(0.72)^8(0.28)^2 \\
 &= 0.2548
 \end{aligned}$$

Interpretation The probability of getting exactly 8 adult Americans out of 10 who would rather give up chocolate is 0.2548. In 100 trials of this study (that is, if we surveyed 10 adult Americans 100 different times), we would expect about 25 trials to result in 8 adult Americans who would rather give up chocolate than their cell phone.

(b) The phrase *fewer than* means “less than.” The values of the random variable X less than 3 are $x = 0, 1$, or 2 .

$$\begin{aligned}
 P(X < 3) &= P(0 \text{ or } 1 \text{ or } 2) \\
 &= P(0) + P(1) + P(2) \quad \text{Addition Rule for Disjoint Events} \\
 &= {}_{10}C_0(0.72)^0(1 - 0.72)^{10-0} + {}_{10}C_1(0.72)^1(1 - 0.72)^{10-1} \\
 &\quad + {}_{10}C_2(0.72)^2(1 - 0.72)^{10-2} \\
 &= 0.000003 + 0.000076 + 0.000881 \\
 &= 0.00096
 \end{aligned}$$

NOTE

$$\begin{aligned}
 {}_{10}C_0(0.72)^0(1 - 0.72)^{10} &= 0.28^{10} = 2.96 \times 10^{-6}. \text{ And,} \\
 2.96 \times 10^{-6} &= 0.00000296
 \end{aligned}$$

NOTE

$$\begin{aligned}
 0.000096 &\approx 0.0001 \text{ and} \\
 0.0001 &= \frac{1}{10,000}.
 \end{aligned}$$

Interpretation The probability that, in a random sample of 10 adult Americans, fewer than three would rather give up chocolate than their cell phone is 0.00096. In 10,000 trials of this study, we would expect about 1 trial to result in fewer than 3 adult Americans who would rather give up chocolate.

(c) The values of the random variable X that are at least 3 are $x = 3, 4, 5, \dots, 10$. Rather than compute $P(X \geq 3)$ directly by computing $P(3) + P(4) + \dots + P(10)$, we can use the Complement Rule.

$$P(X \geq 3) = 1 - P(X < 3) = 1 - 0.00096 = 0.9990$$

Interpretation The probability that, in a random sample of 10 adult Americans, at least 3 would rather give up chocolate than their cell phone is 0.9990. In 1000 trials of this study, we would expect about 999 trials to result in at least 3 stating they would rather give up chocolate.

(d) The word *inclusive* means “including,” so we want to determine the probability that 5, 6, or 7 adult Americans would rather give up chocolate.

$$\begin{aligned}
 P(5 \leq X \leq 7) &= P(5 \text{ or } 6 \text{ or } 7) \\
 &= P(5) + P(6) + P(7) \quad \text{Addition Rule for Disjoint Events} \\
 &= {}_{10}C_5(0.72)^5(1 - 0.72)^{10-5} + {}_{10}C_6(0.72)^6(1 - 0.72)^{10-6} \\
 &\quad + {}_{10}C_7(0.72)^7(1 - 0.72)^{10-7} \\
 &= 0.0839 + 0.1798 + 0.2642 \\
 &= 0.5279
 \end{aligned}$$

Interpretation The probability that the number of adult Americans who would rather give up chocolate is between 5 and 7, inclusive, is 0.5279. In 100 trials of this study, we would expect about 53 trials to result in 5 to 7 adult Americans who would rather give up chocolate.

Obtaining Binomial Probabilities from Tables

Another method for obtaining probabilities is the binomial probability table. Table III in Appendix A gives probabilities for a binomial random variable X taking on a specific value, such as $P(10)$, for select values of n and p . Table IV in Appendix A gives cumulative probabilities of a binomial random variable X . This means that Table IV gives “less than or equal to” binomial probabilities, such as $P(X \leq 6)$. We illustrate how to use Tables III and IV in Example 4.

EXAMPLE 4 Computing Binomial Probabilities Using the Binomial Table

Problem According to the Gallup Organization, 65% of adult Americans are in favor of the death penalty for individuals convicted of murder. In a random sample of 15 adult Americans, what is the probability that

- (a) exactly 10 favor the death penalty?
- (b) no more than 6 favor the death penalty?

Approach We use Tables III and IV in Appendix A to obtain the probabilities.

Solution

- (a) We have $n = 15$, $p = 0.65$, and $x = 10$. In Table III, Appendix A, go to the section that contains $n = 15$ and the column that contains $p = 0.65$. The value at which the $x = 10$ row intersects the $p = 0.65$ column is the probability we seek. See Figure 6. So $P(10) = 0.2123$.

Figure 6

n	x	p														
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70
15	0	0.8601	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
	1	0.1303	0.3658	0.3432	0.2312	0.1319	0.0668	0.0305	0.0126	0.0047	0.0016	0.0005	0.0001	0.0000+	0.0000+	0.0000+
	2	0.0092	0.1348	0.2669	0.2856	0.2309	0.1559	0.0916	0.0476	0.0219	0.0090	0.0032	0.0010	0.0003	0.0001	0.0000+
	3	0.0004	0.0307	0.1285	0.2184	0.2501	0.2252	0.1700	0.1110	0.0634	0.0318	0.0139	0.0052	0.0016	0.0004	0.0001
	4	0.0000+	0.0049	0.0428	0.1156	0.1876	0.2252	0.2186	0.1792	0.1268	0.0780	0.0417	0.0191	0.0074	0.0024	0.0006
	5	0.0000+	0.0006	0.0105	0.0449	0.1032	0.1651	0.2061	0.2123	0.1859	0.1404	0.0916	0.0515	0.0245	0.0096	0.0030
	6	0.0000+	0.0000+	0.0019	0.0132	0.0430	0.0917	0.1472	0.1906	0.2066	0.1914	0.1527	0.1048	0.0612	0.0298	0.0116
	7	0.0000+	0.0000+	0.0003	0.0030	0.0138	0.0393	0.0811	0.1319	0.1771	0.2013	0.1964	0.1647	0.1181	0.0710	0.0348
	8	0.0000+	0.0000+	0.0000+	0.0005	0.0035	0.0131	0.0348	0.0710	0.1181	0.1647	0.1964	0.2013	0.1771	0.1319	0.0811
	9	0.0000+	0.0000+	0.0000+	0.0001	0.0007	0.0034	0.0116	0.0298	0.0612	0.1048	0.1527	0.1914	0.2066	0.1906	0.1472
	10	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0007	0.0030	0.0096	0.0245	0.0515	0.0916	0.1404	0.1859	0.2123	0.2061
	11	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0006	0.0024	0.0074	0.0191	0.0417	0.0780	0.1268	0.1792	0.2186

Interpretation In 100 trials of this study (randomly selecting 15 adult Americans), we expect about 21 trials to result in exactly 10 adult Americans who favor the death penalty for individuals convicted of murder.

- (b) The phrase *no more than* means “less than or equal to.” To compute $P(X \leq 6)$, use the cumulative binomial table, Table IV in Appendix A, which lists binomial probabilities less than or equal to a specified value. We have $n = 15$, $p = 0.65$, so go to the section that contains $n = 15$ and the column that contains $p = 0.65$. The value at which the $x = 6$ row intersects the $p = 0.65$ column represents $P(X \leq 6)$. See Figure 7 on the next page. So $P(X \leq 6) = 0.0422$.

Figure 7

n	x	p														
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70
15	0	0.8601	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
	1	0.9904	0.8290	0.5490	0.3186	0.1671	0.0802	0.0353	0.0142	0.0052	0.0017	0.0005	0.0001	0.0000+	0.0000	0.0000
	2	0.9996	0.9638	0.8159	0.6042	0.3980	0.2361	0.1268	0.0617	0.0271	0.0107	0.0037	0.0011	0.0003	0.0001	0.0000+
	3	1.0000-	0.9945	0.9444	0.8227	0.6482	0.4613	0.2969	0.1727	0.0905	0.0424	0.0176	0.0063	0.0019	0.0005	0.0001
	4	1.0000-	0.9994	0.9873	0.9383	0.8358	0.6865	0.5155	0.3519	0.2173	0.1204	0.0592	0.0255	0.0093	0.0028	0.0007
	5	1.0000-	0.9999	0.9978	0.9832	0.9389	0.8518	0.7216	0.5643	0.4032	0.2608	0.1509	0.0769	0.0338	0.0124	0.0037
	6	1.0000-	1.0000-	0.9997	0.9964	0.9819	0.9434	0.8689	0.7548	0.6098	0.4522	0.3036	0.1818	0.0950	0.0422	0.0152
	7	1.0000-	1.0000-	1.0000-	0.9994	0.9958	0.9827	0.9500	0.8868	0.7869	0.6535	0.5000	0.3465	0.2131	0.1132	0.0500

Interpretation In 100 different trials of this study, we expect about 4 trials to result in no more than 6 adult Americans who favor the death penalty for individuals convicted of murder.



Obtaining Binomial Probabilities Using Technology

Statistical software and graphing calculators also have the ability to compute binomial probabilities.

EXAMPLE 5

Obtaining Binomial Probabilities Using Technology

Problem According to the Gallup Organization, 65% of adult Americans are in favor of the death penalty for individuals convicted of murder. In a random sample of 15 adult Americans, what is the probability that

- (a) exactly 10 favor the death penalty?
- (b) no more than 6 favor the death penalty?

Approach Statistical software or graphing calculators with advanced statistical features have the ability to determine binomial probabilities. The steps for determining binomial probabilities using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch can be found in the Technology Step-by-Step on pages 323–324.

Solution We will use StatCrunch to determine the probability for part (a) and a TI-84 Plus CE to determine the probability for part (b).

- (a) Using StatCrunch's binomial calculator, we obtain the results in Figure 8(a). So $P(10) = 0.2123$.

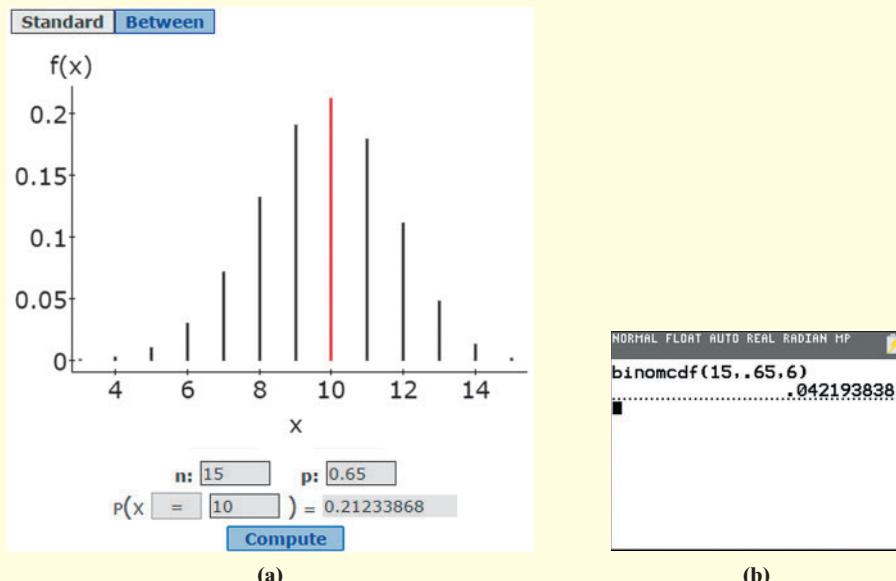
Interpretation In 100 trials of this study (randomly selecting 15 adult Americans), we expect about 21 trials to result in exactly 10 adult Americans who favor the death penalty for individuals convicted of murder.

- (b) The phrase *no more than* means “less than or equal to.” To compute $P(X \leq 6)$, we use the **cumulative distribution function (cdf)**, which computes probabilities less than or equal to a specified value. Using a TI-84 Plus CE graphing calculator to find $P(X \leq 6)$ with $n = 15$ and $p = 0.65$, we find $P(X \leq 6) = 0.0422$. See Figure 8(b).

Interpretation In 100 trials of this study, we expect about 4 trials to result in no more than 6 adult Americans who favor the death penalty for individuals convicted of murder.

(continued)

Figure 8

**NW** Now Work Problem 35

③ Compute the Mean and Standard Deviation of a Binomial Random Variable

We discussed finding the mean (or expected value) and standard deviation of a discrete random variable in Section 6.1. These formulas can be used to find the mean and standard deviation of a binomial random variable, but a simpler method exists.

IN OTHER WORDS

The mean of a binomial random variable equals the product of the number of trials of the experiment and the probability of success. It can be interpreted as the expected number of successes in n trials of the experiment.

Mean (or Expected Value) and Standard Deviation of a Binomial Random Variable

A binomial experiment with n independent trials and probability of success p has a mean and standard deviation given by the formulas

$$\mu_X = np \quad \text{and} \quad \sigma_X = \sqrt{np(1-p)} \quad (2)$$

EXAMPLE 6**Finding the Mean and Standard Deviation of a Binomial Random Variable**

Problem According to CTIA, 72% of adult Americans would rather give up chocolate than their cell phone. In a simple random sample of 300 households, determine the mean and standard deviation number who would rather give up chocolate than their cell phone.

Approach This is a binomial experiment with $n = 300$ and $p = 0.72$. Use Formula (2) to find the mean and standard deviation, respectively.

Solution

$$\mu_X = np = 300(0.72) = 216$$

and

$$\sigma_X = \sqrt{np(1-p)} = \sqrt{300(0.72)(1-0.72)} = \sqrt{60.48} = 7.8$$

Interpretation We expect that, in a random sample of 300 adult Americans, 216 would rather give up chocolate than their cell phone.

NW Now Work Problems 29(a)–(c)

④ Graph a Binomial Probability Distribution

To graph a binomial probability distribution, first find the probabilities for each possible value of the random variable. Then follow the same approach that was used to graph discrete probability distributions.

EXAMPLE 7

Graphing Binomial Probability Distributions

Problem

- Graph the binomial probability distribution with $n = 10$ and $p = 0.2$. Comment on the shape of the distribution.
- Graph the binomial probability distribution with $n = 10$ and $p = 0.5$. Comment on the shape of the distribution.
- Graph the binomial probability distribution with $n = 10$ and $p = 0.8$. Comment on the shape of the distribution.

Approach To graph a binomial probability distribution, first obtain the probability distribution and then graph the distribution using the approach in Section 6.1.

Solution

- Table 10 shows the probability distribution with $n = 10$ and $p = 0.2$. Although $P(9)$ is 0.000004096, it is written as 0.0000 to four significant digits. The same idea applies to $P(10)$. Figure 9 shows the graph of the distribution. The distribution is skewed right.

Figure 9

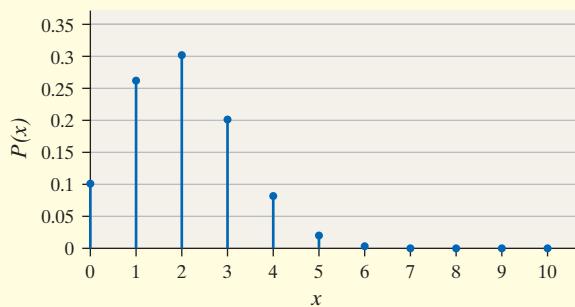


Table 10

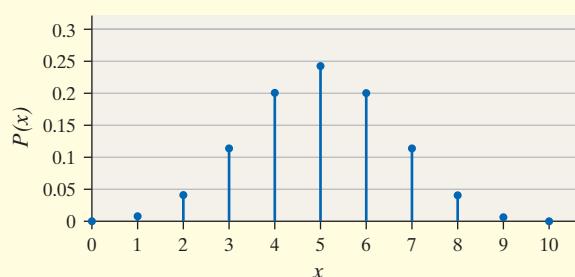
x	P(x)
0	0.1074
1	0.2684
2	0.3020
3	0.2013
4	0.0881
5	0.0264
6	0.0055
7	0.0008
8	0.0001
9	0.0000
10	0.0000

Table 11

x	P(x)
0	0.0010
1	0.0098
2	0.0439
3	0.1172
4	0.2051
5	0.2461
6	0.2051
7	0.1172
8	0.0439
9	0.0098
10	0.0010

- Table 11 shows the probability distribution with $n = 10$ and $p = 0.5$. Figure 10 shows the graph of the distribution. The distribution is symmetric and approximately bell shaped.

Figure 10



(continued)

Table 12

x	P(x)
0	0.0000
1	0.0000
2	0.0001
3	0.0008
4	0.0055
5	0.0264
6	0.0881
7	0.2013
8	0.3020
9	0.2684
10	0.1074

NW Now Work Problem 29(d)

- (c) Table 12 shows the probability distribution with $n = 10$ and $p = 0.8$. Figure 11 shows the graph of the distribution. The distribution is skewed left.

Figure 11

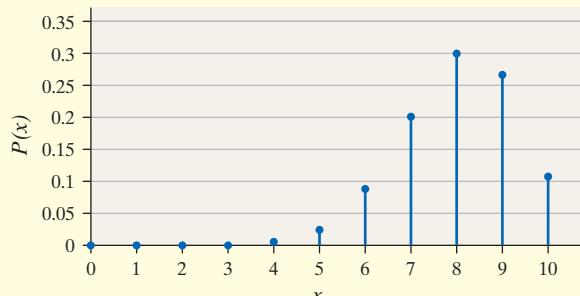


Figure 12

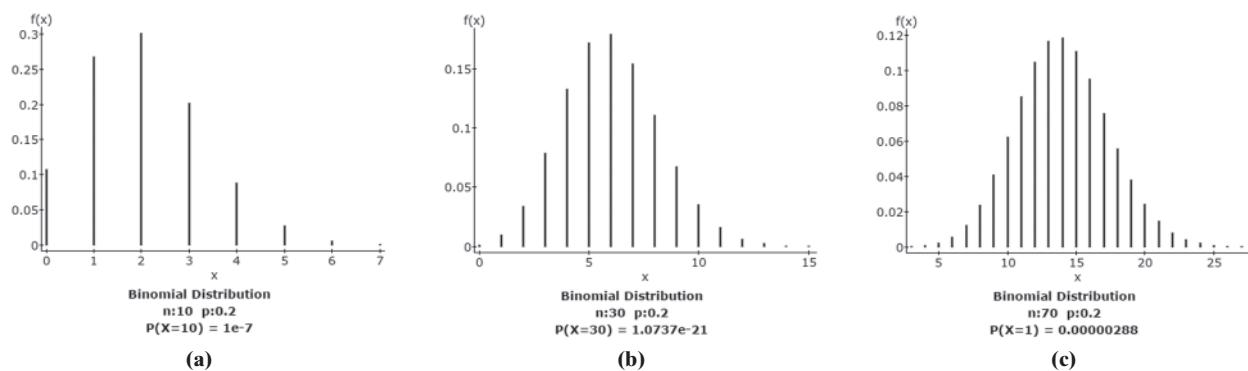


Figure 12(a) is skewed right, Figure 12(b) is slightly skewed right, and Figure 12(c) appears bell shaped. Our conclusion follows:

For a fixed p , as the number of trials n in a binomial experiment increases, the probability distribution of the random variable X becomes bell shaped. As a rule of thumb, if $np(1 - p) \geq 10$,* the probability distribution will be approximately bell shaped.

IN OTHER WORDS

Provided that $np(1 - p) \geq 10$, the interval $\mu - 2\sigma$ to $\mu + 2\sigma$ represents the “usual” observations. Observations outside this interval may be considered unusual.

This result allows us to use the Empirical Rule to identify unusual observations in a binomial experiment. Recall the Empirical Rule states that in a bell-shaped distribution about 95% of all observations lie within two standard deviations of the mean. That is, about 95% of the observations lie between $\mu - 2\sigma$ and $\mu + 2\sigma$. Any observation that lies outside this interval may be considered unusual because the observation occurs less than 5% of the time.

*P. P. Ramsey and P. H. Ramsey, “Evaluating the Normal Approximation to the Binomial Test,” *Journal of Educational Statistics* 13 (1998): 173–182.

EXAMPLE 8**Using the Mean, Standard Deviation, and Empirical Rule to Check for Unusual Results in a Binomial Experiment**

Problem According to CTIA, 72% of adult Americans would rather give up chocolate than their cell phone. In a simple random sample of 300 adult Americans, 230 indicated they would rather give up chocolate than their cell phone. Is this result unusual?

Approach Because $np(1 - p) = 300(0.72)(1 - 0.72) = 60.48 \geq 10$, the binomial probability distribution is approximately bell shaped. Therefore, we can use the Empirical Rule: If the observation is less than $\mu - 2\sigma$ or greater than $\mu + 2\sigma$, it is unusual.

Solution From Example 6, we have $\mu = 216$ and $\sigma = 7.8$.

$$\mu - 2\sigma = 216 - 2(7.8) = 216 - 15.6 = 200.4$$

and

$$\mu + 2\sigma = 216 + 2(7.8) = 216 + 15.6 = 231.6$$

Interpretation Since any value less than 200.4 or greater than 231.6 is unusual, 230 is not an unusual result.

NW Now Work Problem 43

**Technology Step-by-Step****Computing Binomial Probabilities via Technology****TI-83/84 Plus****Computing $P(x)$**

1. Press 2nd VARS to access the probability distribution menu.
2. Highlight binompdf(and hit ENTER.
3. Enter the number of trials n , the probability of success p , and the number of successes x . Highlight Paste and hit ENTER. For example, with $n = 15$, $p = 0.3$, and $x = 8$, you should see the following on the HOME screen

`binompdf(15,0.3,8)`

Then hit ENTER.

Computing $P(X \leq x)$

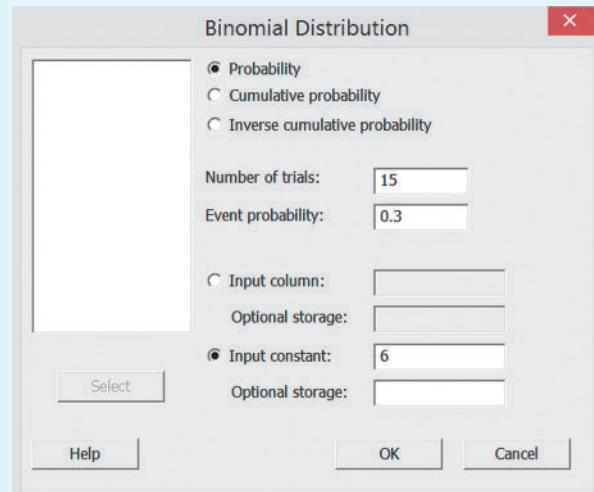
1. Press 2nd VARS to access the probability distribution menu.
2. Highlight binomcdf(and hit ENTER.
3. Enter the number of trials n , the probability of success p , and the number of successes x . Highlight Paste and hit ENTER. For example, with $n = 15$, $p = 0.3$, and $x \leq 8$, you should see the following on the HOME screen

`binomcdf(15,0.3,8)`

Then hit ENTER.

Minitab**Computing $P(x)$**

1. Select the CALC menu, highlight **Probability Distributions**, then highlight **Binomial . . .**.
2. With $n = 15$, $p = 0.3$, and $x = 6$, fill in the window as shown. Click OK.

**Computing $P(X \leq x)$**

Follow the same steps as those for computing $P(x)$. In the window above, select **Cumulative probability** instead of **Probability**.

Excel**Computing $P(x)$**

1. Click the Formulas tab. Select More Functions. Highlight Statistical in the Function category menu. Highlight BINOM.DIST in the Function name menu.
2. Fill in the window with the appropriate values. Type FALSE in the cumulative cell. Click OK.

(continued)

Computing $P(X \leq x)$

Follow the same steps as those presented for computing $P(x)$. In the BINOM.DIST window, type TRUE in the cumulative cell.

StatCrunch

1. Select **Stat**, highlight **Calculators**, select **Binomial**.
2. Enter the number of trials, n , and probability of success, p . If you want to compute $P(X < x)$, $P(X \leq x)$,

$P(X > x)$, or $P(X \geq x)$, highlight the Standard tab. In the pull-down menu, decide if you wish to compute $P(X \leq x)$, $P(X < x)$, and so on. Finally, enter the value of x . Click Compute.

If you want to compute $P(a \leq X \leq b)$, highlight the Between tab. Enter the values of a and b . Click Compute.



6.2 Assess Your Understanding

Vocabulary and Skill Building

1. State the criteria for a binomial probability experiment.
2. (a) A binomial experiment is performed a fixed number of times. Each repetition of the experiment is called a _____.
(b) For each repetition of a binomial experiment, there are two mutually exclusive outcomes: _____ or _____.
3. *True or False:* In the binomial probability distribution function, ${}_n C_x$ represents the number of ways of obtaining x successes in n trials.
4. The phrase “no more than” is represented by the math symbol _____.
5. The expected number of successes in a binomial experiment with n trials and probability of success p is _____.
6. As a rule of thumb, if _____, the probability distribution of a binomial random variable X is approximately bell shaped.

In Problems 7–16, determine which of the following probability experiments represents a binomial experiment. If the probability experiment is not a binomial experiment, state why.

7. A random sample of 15 college seniors is obtained, and the individuals selected are asked to state their ages.
8. A random sample of 30 cars in a used car lot is obtained, and their mileages recorded.

NW 9. An experimental drug is administered to 100 randomly selected individuals, with the number of individuals responding favorably recorded.

10. A poll of 1200 registered voters is conducted in which the respondents are asked whether they believe Congress should reform Social Security.

11. Three cards are selected from a standard 52-card deck without replacement. The number of aces selected is recorded.

12. Three cards are selected from a standard 52-card deck with replacement. The number of kings selected is recorded.

13. A basketball player who makes 80% of her free throws is asked to shoot free throws until she misses. The number of free-throw attempts is recorded.

14. A baseball player who reaches base safely 30% of the time is allowed to bat until he reaches base safely for the third time. The number of at-bats required is recorded.

15. One hundred randomly selected U.S. parents with at least one child under the age of 18 are surveyed and asked if they

have ever spanked their child. The number of parents who have spanked their child is recorded.

16. In a town with 400 citizens, 100 randomly selected citizens are asked to identify their religion. The number who identify with a Christian religion is recorded.

In Problems 17–28, a binomial probability experiment is conducted with the given parameters. Compute the probability of x successes in the n independent trials of the experiment.

17. $n = 10, p = 0.4, x = 3$
18. $n = 15, p = 0.85, x = 12$
19. $n = 40, p = 0.99, x = 38$
20. $n = 50, p = 0.02, x = 3$
21. $n = 8, p = 0.35, x = 3$
22. $n = 20, p = 0.6, x = 17$
23. $n = 9, p = 0.2, x \leq 3$
24. $n = 10, p = 0.65, x < 5$
25. $n = 7, p = 0.5, x > 3$
26. $n = 20, p = 0.7, x \geq 12$
27. $n = 12, p = 0.35, x \leq 4$
28. $n = 11, p = 0.75, x \geq 8$

In Problems 29–34, (a) construct a binomial probability distribution with the given parameters; (b) compute the mean and standard deviation of the random variable using the methods of Section 6.1; (c) compute the mean and standard deviation, using the methods of this section; and (d) draw a graph of the probability distribution and comment on its shape.

- NW** 29. $n = 6, p = 0.3$
30. $n = 8, p = 0.5$
 31. $n = 9, p = 0.75$
 32. $n = 10, p = 0.2$
 33. $n = 10, p = 0.5$
 34. $n = 9, p = 0.8$

Applying the Concepts

NW 35. **On-Time Flights** According to flightstats.com, American Airlines flights from Dallas to Chicago are on time 80% of the time. Suppose 15 flights are randomly selected, and the number of on-time flights is recorded.

- (a) Explain why this is a binomial experiment.
- (b) Determine the values of n and p .

- (c) Find and interpret the probability that exactly 10 flights are on time.
- (d) Find and interpret the probability that fewer than 10 flights are on time.
- (e) Find and interpret the probability that at least 10 flights are on time.
- (f) Find and interpret the probability that between 8 and 10 flights, inclusive, are on time.

36. Morality In a recent poll, the Gallup Organization found that 45% of adult Americans believe that the overall state of moral values in the United States is poor. Suppose a survey of a random sample of 25 adult Americans is conducted in which they are asked to disclose their feelings on the overall state of moral values in the United States.

- (a) Find and interpret the probability that exactly 15 of those surveyed feel the state of morals is poor.
- (b) Find and interpret the probability that no more than 10 of those surveyed feel the state of morals is poor.
- (c) Find and interpret the probability that more than 16 of those surveyed feel the state of morals is poor.
- (d) Find and interpret the probability that 13 or 14 believe the state of morals is poor.
- (e) Would it be unusual to find 20 or more adult Americans who believe the overall state of moral values is poor in the United States? Why?

37. Toilet Flushing In the Healthy Handwashing Survey conducted by Bradley Corporation, it was found that 64% of adult Americans operate the flusher of toilets in public restrooms with their foot. Suppose a random sample of $n = 20$ adult Americans is obtained and the number x who flush public toilets with their foot is recorded.

- (a) Explain why this is a binomial experiment.
- (b) Find and interpret the probability that exactly 12 flush public toilets with their foot.
- (c) Find and interpret the probability that at least 16 flush public toilets with their foot.
- (d) Find and interpret the probability that between 9 and 11, inclusive, flush public toilets with their foot.
- (e) Would it be unusual to find more than 17 who flush public toilets with their foot? Why?

38. Allergy Sufferers Clarinex-D is a medication whose purpose is to reduce the symptoms associated with a variety of allergies. In clinical trials of Clarinex-D, 5% of the patients in the study experienced insomnia as a side effect. A random sample of 20 Clarinex-D users is obtained, and the number of patients who experienced insomnia is recorded.

- (a) Find the probability that exactly 3 experienced insomnia as a side effect.
- (b) Find the probability that 3 or fewer experienced insomnia as a side effect.
- (c) Find the probability that between 1 and 4 patients inclusive, experienced insomnia as a side effect.
- (d) Would it be unusual to find 4 or more patients who experienced insomnia as a side effect? Why?

39. Sneeze According to a study done by Nick Wilson of Otago University Wellington, the probability a randomly selected individual will not cover his or her mouth when sneezing is 0.267. Suppose you sit on a bench in a mall and observe people's habits as they sneeze.

- (a) What is the probability that among 10 randomly observed individuals exactly 4 do not cover their mouth when sneezing?

- (b) What is the probability that among 10 randomly observed individuals fewer than 3 do not cover their mouth?
- (c) Would you be surprised if, after observing 10 individuals, fewer than half covered their mouth when sneezing? Why?

40. Sneeze Revisited According to a study done by Nick Wilson of Otago University Wellington, the probability a randomly selected individual will cover his or her mouth with a tissue, handkerchief, or elbow (the method recommended by public health officials) when sneezing is 0.047. Suppose you sit on a bench in a mall and observe people's habits as they sneeze.

- (a) What is the probability that among 15 randomly observed sneezing individuals exactly 2 cover their mouth with a tissue, handkerchief, or elbow?
- (b) What is the probability that among 15 randomly observed sneezing individuals fewer than 3 cover their mouth with a tissue, handkerchief, or elbow?
- (c) Would you be surprised if, after observing 15 sneezing individuals, more than 4 covered their mouth with a tissue, handkerchief, or elbow?

41. Jury Selection Twelve jurors are randomly selected from a population of 3 million residents. Of these 3 million residents, it is known that 45% are Hispanic. Of the 12 jurors selected, 2 are Hispanic.

- (a) What proportion of the jury described is Hispanic?
- (b) If 12 jurors are randomly selected from a population that is 45% Hispanic, what is the probability that 2 or fewer jurors will be Hispanic?
- (c) If you were the lawyer of a Hispanic defendant, what might you argue?

42. Sullivan Survey: Car Color According to paint manufacturer DuPont, 6% of all cars worldwide are red. In the Sullivan Statistics Survey, of 175 respondents, 17, or 9.7%, indicated the color of their car is red. Determine if the results of the Sullivan Survey contradict those of DuPont by computing $P(X \geq 17)$, where X is a binomial random variable with $n = 175$ and $p = 0.06$. Explain what the probability represents.

NW 43. On-Time Flights According to flightstats.com, American Airlines flights from Dallas to Chicago are on time 80% of the time. Suppose 100 flights are randomly selected.

- (a) Compute the mean and standard deviation of the random variable X , the number of on-time flights in 100 trials of the probability experiment.
- (b) Interpret the mean.
- (c) Would it be unusual to observe 75 on-time flights in a random sample of 100 flights from Dallas to Chicago? Why?

44. Morality In a recent poll, the Gallup Organization found that 45% of adult Americans believe that the overall state of moral values in the United States is poor.

- (a) Compute the mean and standard deviation of the random variable X , the number of adults who believe that the overall state of moral values in the United States is poor based on a random sample of 500 adult Americans.
- (b) Interpret the mean.
- (c) Would it be unusual to identify 240 adult Americans who believe that the overall state of moral values in the United States is poor based on a random sample of 500 adult Americans? Why?

45. Toilet Flushing In the Healthy Handwashing Survey conducted by Bradley Corporation, it was found that 64% of adult Americans operate the flusher of toilets in public restrooms with their foot.

- (a) If 500 adult Americans are randomly selected, how many would we expect to flush toilets in public restrooms with their foot?
- (b) Would it be unusual to observe 280 adult Americans who flush toilets in public restrooms with their foot?
- 46. Allergy Sufferers** Clarinex-D is a medication whose purpose is to reduce the symptoms associated with a variety of allergies. In clinical trials of Clarinex-D, 5% of the patients in the study experienced insomnia as a side effect.
- (a) If 240 users of Clarinex-D are randomly selected, how many would we expect to experience insomnia as a side effect?
- (b) Would it be unusual to observe 20 patients experiencing insomnia as a side effect in 240 trials of the probability experiment? Why?
- 47. Spanking** In March 1995, The Harris Poll reported that 80% of parents spank their children. Suppose a recent poll of 1030 adult Americans with children finds that 781 indicated that they spank their children. If we assume parents' attitude toward spanking has not changed since 1995, how many of 1030 parents surveyed would we expect to spank? Do the results of the survey suggest that parents' attitude toward spanking may have changed since 1995? Why?
- 48. Government Solutions?** In May, 2000, the Gallup Organization reported that 11% of adult Americans had a great deal of trust and confidence in the federal government handling domestic issues. Suppose a survey of a random sample of 1100 adult Americans finds that 84 have a great deal of trust and confidence in the federal government handling domestic issues. Would these results be considered unusual? Why?
- 49. Racial Profiling in New York City** The following excerpt is from the *Racial Profiling Data Collection Resource Center*.
- In 2006, the New York City Police Department stopped a half-million pedestrians for suspected criminal involvement. Raw statistics for these encounters suggest large racial disparities—89 percent of the stops involved nonwhites. Do these statistics point to racial bias in police officers' decisions to stop particular pedestrians? Do they indicate that officers are particularly intrusive when stopping nonwhites?*
- Source: Published by Northeastern University.
- Write a report that answers the questions posed using the fact that 44% of New York City residents were classified as white in 2006. In your report, cite some shortcomings in using the proportion of white residents in the city to formulate likelihoods.
- 50. Simulating Election Results** In an election for President of the United States, a pollster wishes to predict the winner of the popular vote. She will do this by taking a random sample of 1000 voters and ask each to disclose who he/she will vote for. Assume the election is a dichotomous choice between a Republican candidate and Democrat candidate and that the proportion of voters in the population who will vote for the Republican candidate is 0.48 (meaning the Democrat would win the election).
- (a) Explain why randomly selecting 1000 voters from the population of registered voters and asking each to disclose the candidate he/she supports is a binomial experiment. What are the values of n , the number of trials, and p , the probability of success?
- (b) Simulate obtaining a random sample of 1000 voters and record the number who disclose the Republican as their choice for President. Repeat this simulation 2000 times.
- (c) For each simulation determine the proportion of voters who selected the Republican candidate.

- (d) What proportion of the 2000 simulations resulted in a poll that suggests the Republican would win the election?
- (e) Repeat this simulation for a random sample of 1500 voters. Explain why the results differ.
- 51. Overbooking Flights** Historically, the probability that a passenger will miss a flight is 0.0995. *Source: Passenger-Based Predictive Modeling of Airline No-show Rates* by Richard D. Lawrence, Se June Hong, and Jacques Cherrier. Airlines do not like flights with empty seats, but it is also not desirable to have overbooked flights because passengers must be "bumped" from the flight. The Lockheed L49 Constellation has a seating capacity of 54 passengers.
- (a) If 56 tickets are sold, what is the probability 55 or 56 passengers show up for the flight resulting in an overbooked flight?
- (b) Suppose 60 tickets are sold; what is the probability a passenger will have to be "bumped"?
- (c) For a plane with seating capacity of 250 passengers, how many tickets may be sold to keep the probability of a passenger being "bumped" below 1%?
- 52. Athletics Participation** According to the High School Athletics Participation Survey, approximately 55% of students enrolled in high schools participate in athletic programs. You are performing a study of high school students and would like at least 11 students in the study to be participating in athletics. *Source: National Federation of State High School Associations.*
- (a) How many high school students do you expect to have to randomly select?
- (b) How many high school students do you have to select to have a 99% probability that the sample contains at least 12 who participate in athletics?
- 53. Geometric Probability Distribution** A probability distribution for the random variable X , the number of trials until a success is observed, is called the **geometric probability distribution**. It has the same criteria as the binomial distribution (see page 313), except that the number of trials is not fixed. Its probability distribution function (pdf) is
- $$P(x) = p(1 - p)^{x-1}, \quad x = 1, 2, 3, \dots$$
- where p is the probability of success.
- (a) What is the probability that a basketball player misses his first two field goal attempts and makes the third? Over his career, the player made 52.4% of his field goal attempts. That is, find $P(3)$.
- (b) Construct a probability distribution for the random variable X , the number of field goal attempts of he until he makes a field goal. Construct the distribution for $x = 1, 2, 3, \dots, 10$. The probabilities are small for $x > 10$.
- (c) Compute the mean of the distribution, using the formula presented in Section 6.1.
- (d) Compare the mean obtained in part (c) with the value $\frac{1}{p}$. Conclude that the mean of a geometric probability distribution is $\mu_X = \frac{1}{p}$. How many field goals do we expect the player to attempt before we observe a made field goal?
- 54. Negative Binomial Probability Distribution** The **negative binomial probability distribution** can be used to compute the probability of the random variable X , the number of trials necessary to observe r successes of a binomial experiment. The probability distribution function is given by
- $$P(x) = {}_{(x-1)}C_{r-1} p^r (1 - p)^{x-r}$$
- $$x = r, r + 1, r + 2, \dots$$

Consider a roulette wheel. Remember, a roulette wheel has 2 green slots, 18 red slots, and 18 black slots.

- (a) What is the probability that it will take $x = 1$ trial before observing $r = 1$ green?
- (b) What is the probability that it will take $x = 20$ trials before observing $r = 2$ greens?
- (c) What is the probability that it will take $x = 30$ trials before observing $r = 3$ greens?
- (d) The expected number of trials before observing r successes is $\frac{r}{p}$. What is the expected number of trials before observing 3 greens?

55. Putting It Together: A Drug Study Toxic Epidermal Necrolysis (TEN), also known as Lyell's syndrome, is a life-threatening disease characterized by blisters that cover the skin and extensive peeling off of skin. In a double-blind, randomized, placebo-controlled study of the effects of a drug called thalidomide on TEN, it was found that 10 of the 12 patients in the thalidomide group died compared with 3 of 10 in the placebo group. *Source: Randomized Comparison of Thalidomide Versus Placebo in Toxic Epidermal Necrolysis*, Wolkenstein, Pierre et al. *The Lancet*, 352 (9140): 1586–1589.

- (a) What type of experimental design is this?
- (b) What is the response variable in the study? Is it qualitative or quantitative?
- (c) How many subjects were in the study?
- (d) What does it mean for the study to be “double-blind”?
- (e) What does it mean for the study to be “randomized”?
- (f) Using the methods of this section, explain why the study was stopped. What does this result suggest?

56. Putting It Together: Beating the Stock Market One measure of successful investing is being able to “beat the market.” To beat the market in any given year, an investor must earn a rate of return greater than the rate of return of some market basket of stocks, such as the Dow Jones Industrial Average (DJIA) or Standard and Poor’s 500 (S&P 500). Suppose in any given year, there is a probability of 0.5 that a particular investment advisor beats the market for his/her clients.

- (a) If there are 5000 investment advisors across the country, how many would be expected to beat the market in any given year?
- (b) Assume beating the market in one year is independent of beating the market in any other year. What is the probability

that a randomly selected investment advisor beats the market in two consecutive years? Based on this result, how many of 5000 investment advisors would be expected to beat the market for two consecutive years?

- (c) Assume beating the market in one year is independent of beating the market in any other year. What is the probability that a randomly selected investment advisor beats the market in five consecutive years? Based on this result, how many of 5000 investment advisors would be expected to beat the market for five consecutive years?
- (d) Assume beating the market in one year is independent of beating the market in any other year. What is the probability that a randomly selected investment advisor beats the market in ten consecutive years? Based on this result, how many of 5000 investment advisors would be expected to beat the market for ten consecutive years?
- (e) Assume a randomly selected investment advisor can beat the market with probability 0.5 and investment results from year to year are independent. Suppose we randomly select 5000 investment advisors and determine the number x who have beaten the market the past ten years. Explain why this is a binomial experiment (assuming there are tens of thousands of investment advisors in the population) and clearly state what a success represents.
- (f) Use the results of part (e) to determine the probability of identifying at least six investment advisors who will beat the market for ten consecutive years. Interpret this result. Is it unusual to identify at least six investment advisors who consistently beats the market even though his/her underlying ability to beat the market is 0.5? Explain.

Explaining the Concepts

- 57. Explain what “success” means in a binomial probability experiment.
- 58. Explain how the value of n , the number of trials in a binomial experiment, affects the shape of the distribution of a binomial random variable.
- 59. Explain how the value of p , the probability of success, affects the shape of the distribution of a binomial random variable.
- 60. When can the Empirical Rule be used to identify unusual results in a binomial experiment? Why can the Empirical Rule be used to identify results in a binomial experiment?



Chapter 6 Review

Summary

In this chapter, we introduced probability models for random variables. A random variable represents the numerical measurement of the outcome from a probability experiment. Discrete random variables have either a finite or a countable number of outcomes. The values of a discrete random variable can be plotted on a number line with space between each point. The term *countable* means that the values result from counting. Continuous random variables have infinitely many values and typically result from measurement. The values of a continuous random

variable can be plotted on a line in an uninterrupted fashion.

We also looked at the probability distribution of discrete random variables. Discrete probability distributions must satisfy the following two criteria: (1) All probabilities must be between 0 and 1, inclusive, and (2) the sum of all probabilities must equal 1.

Discrete probability distributions can be presented by a table, graph, or mathematical formula.

The mean and standard deviation of a random variable describe the center and spread of the distribution. The mean of a random variable is also called its expected value.

A probability experiment is considered a binomial experiment if

1. The experiment is performed a fixed number of times, n . Each repetition of the experiment is called a trial.
2. The trials are independent. This means that the outcome of one trial will not affect the outcome of the other trials.

3. For each trial, there are two mutually exclusive (disjoint) outcomes: success or failure.
4. The probability of success is the same for each trial of the experiment. We denote the probability of success as p .

Binomial probabilities were obtained by hand, a table, or via technology using its probability distribution function.

The mean and standard deviation of a binomial random variable may be used to identify unusual results in a binomial experiment using the Empirical Rule provided $np(1 - p) \geq 10$.

Vocabulary

Random variable (p. 300)
 Discrete random variable (p. 300)
 Continuous random variable (p. 300)
 Probability distribution (p. 301)

Expected value (p. 305)
 Binomial probability distribution (p. 312)
 Binomial experiment (p. 313)
 Trial (p. 313)

Binomial random variable (p. 313)
 Cumulative distribution function (p. 319)

Formulas

Mean (or Expected Value) of a Discrete Random Variable

$$\mu_X = E(X) = \sum [x \cdot P(x)]$$

Standard Deviation of a Discrete Random Variable

$$\begin{aligned}\sigma_X &= \sqrt{\sum [(x - \mu_X)^2 \cdot P(x)]} \\ &= \sqrt{\sum [x^2 \cdot P(x)] - \mu_X^2}\end{aligned}$$

Binomial Probability Distribution Function

$$P(x) = {}_n C_x p^x (1 - p)^{n-x} \quad x = 0, 1, 2, \dots, n$$

Mean of a Binomial Random Variable

$$\mu_X = np$$

Standard Deviation of a Binomial Random Variable

$$\sigma_X = \sqrt{np(1 - p)}$$

Objectives

Section	You should be able to . . .	Example(s)	Review Exercises
6.1	1 Distinguish between discrete and continuous random variables (p. 300) 2 Identify discrete probability distributions (p. 301) 3 Graph discrete probability distributions (p. 302) 4 Compute and interpret the mean of a discrete random variable (p. 303) 5 Interpret the mean of a discrete random variable as an expected value (p. 305) 6 Compute the standard deviation of a discrete random variable (p. 306)	1 2 and 3 4 5, 6, and 9 7 8 and 9	1 2, 3(a) 3(b) 3(c) 4 3(d)
6.2	1 Determine whether a probability experiment is a binomial experiment (p. 312) 2 Compute probabilities of binomial experiments (p. 314) 3 Compute the mean and standard deviation of a binomial random variable (p. 320) 4 Graph a binomial probability distribution (p. 321)	1 2–5 6 7	5 6(a)–(e), 7(a)–(d), 8(a), 11 6(f), 7(e), 8(b) 8(c)

Review Exercises

1. Determine whether the random variable is discrete or continuous. In each case, state the possible values of the random variable.
- The number of students in a randomly selected elementary school classroom
 - The amount of snow that falls in Minneapolis during the winter season
 - The flight time accumulated by a randomly selected Air Force fighter pilot
 - The number of points scored by the Miami Heat in a randomly selected basketball game
2. Determine whether the distribution is a discrete probability distribution. If not, state why.

(a)	x	P(x)
0	0.34	
1	0.21	
2	0.13	
3	0.04	
4	0.01	

(b)	x	P(x)
0	0.40	
1	0.31	
2	0.23	
3	0.04	
4	0.02	

3. **Stanley Cup** The Stanley Cup is a best-of-seven series to determine the champion of the National Hockey League. The following data represent the number of games played, X , in the Stanley Cup before a champion was determined from 1939 to 2019.

Note: There was no champion in 2005. The season was cancelled because of a labor dispute.

x	Frequency
4	20
5	19
6	24
7	17

Source: *Information Please Almanac*.

- Construct a probability model for the random variable X , the number of games in the Stanley Cup.
- Graph the discrete probability distribution.
- Compute and interpret the mean of the random variable X .
- Compute the standard deviation of the random variable X .

4. **Expected Value of Three-Card Poker** A popular casino table game is three-card poker. One aspect of the game is the “pair plus” bet in which a player is paid a dollar amount for any hand of a pair or better, regardless of the hand the dealer has. The table shows the profit and probability of various hands of a player playing the \$5 pair plus bet.

Outcome	Profit (\$)	Probability
Straight flush	200	12/5525
Three of a kind	150	1/425
Straight	30	36/1105
Flush	20	274/5525
Pair	5	72/425
Other	-5	822/1105

Source: <http://wizardofodds.com/threecardpoker>

- What is the expected profit when playing the \$5 pair plus bet in three card poker?
 - If you play the game for 4 hours with an average of 35 hands per hour, how much would you expect to lose?
 - What is the standard deviation of profit when playing the \$5 pair plus bet in three card poker? What does this value suggest?
5. Determine whether the probability experiment represents a binomial experiment. If not, explain why.
- According to the *Chronicle of Higher Education*, the probability that a randomly selected incoming freshman will graduate from college within 6 years is 0.54. Suppose that 10 incoming freshmen are randomly selected. After 6 years, each student is asked whether he or she graduated.
 - An experiment is conducted in which a single die is cast until a 3 comes up. The number of throws required is recorded.
6. **Emergency Room Visits** The probability that a randomly selected patient who visits the emergency room (ER) will die within 1 year of the visit is 0.05.
- Source:* SuperFreakonomics.
- What is the probability that exactly 1 of 10 randomly selected visitors to the ER will die within 1 year? Interpret this result.
 - What is the probability that fewer than 2 of 25 randomly selected visitors to the ER will die within 1 year? Interpret this result.
 - What is the probability that at least 2 of 25 randomly selected visitors to the ER will die within 1 year? Interpret this result.
 - What is the probability that at least 8 of 10 randomly selected visitors to the ER will *not* die within 1 year?
 - Would it be unusual if more than 3 of 30 randomly selected visitors to the ER died within 1 year? Why?
 - In a random sample of 1000 visitors to the ER, how many visitors are expected to die within the next year? What is the standard deviation number of deaths?
 - At a particular emergency room, a researcher obtains a random sample of 800 visitors and finds that after 1 year 51 of them have died. Do you think this particular emergency room should be investigated to see if something unusual is occurring?
7. **Driving Age** According to a Gallup poll, 60% of U.S. women 18 years old or older stated that the minimum driving age should be 18. In a random sample of 15 U.S. women 18 years old or older, find the probability that:
- Exactly 10 believe that the minimum driving age should be 18.
 - Fewer than 5 believe that the minimum driving age should be 18.
 - At least 5 believe that the minimum driving age should be 18.
 - Between 7 and 12, inclusive, believe that the minimum driving age should be 18.
 - In a random sample of 200 U.S. women 18 years old or older, what is the expected number who believe that the minimum driving age should be 18? What is the standard deviation?

- (f) If a random sample of 200 U.S. women 18 years old or older resulted in 110 who believe that the minimum driving age should be 18, would this be unusual? Why?
8. Consider a binomial probability distribution with parameters $n = 8$ and $p = 0.75$.
- Construct a binomial probability distribution with these parameters.
 - Compute the mean and standard deviation of the distribution.
 - Graph the discrete probability distribution. Comment on the shape of the distribution.
9. State the condition required to use the Empirical Rule to check for unusual observations in a binomial experiment.
10. In sampling from finite populations without replacement, the assumption of independence required for a binomial

experiment is violated. Under what circumstances can we sample without replacement and still use the binomial probability formula to approximate probabilities?

- 11. Self-Injury** According to the article “Self-injurious Behaviors in a College Population,” 17% of undergraduate or graduate students have had at least one incidence of self-injurious behavior. The researchers conducted a survey of 40 college students who reported a history of emotional abuse and found that 12 of them have had at least one incidence of self-injurious behavior. What do the results of this survey tell you about college students who report a history of emotional abuse?

Source: Janis Whitlock, John Eckenrode, and Daniel Silverman. “Self-injurious Behaviors in a College Population,” *Pediatrics* 117: 1939–1948.



Chapter Test

1. Determine whether the random variable is discrete or continuous. In each case, state the possible values of the random variable.
- The number of days with measurable rainfall in Honolulu, Hawaii, during a year
 - The miles per gallon of gasoline obtained by a randomly selected Toyota Prius
 - The number of golf balls hit into the ocean on the famous 18th hole at Pebble Beach on a randomly selected Sunday
 - The weight (in grams) of a randomly selected robin’s egg
2. Determine whether the distribution is a discrete probability distribution. If not, state why.

(a)	x	P(x)
0	0.324	
1	0.121	
2	0.247	
3	0.206	
4	0.102	

(b)	x	P(x)
0	0.34	
1	0.28	
2	0.26	
3	0.23	
4	-0.11	

3. At the Wimbledon Tennis Championship, to win a match in men’s singles a player must win the best of five sets. The following data represent the number of sets played, X , in the men’s singles final match for the years 1968 to 2019.

x	Frequency
3	23
4	14
5	15

Source: www.wimbledon.org

- Construct a probability model for the random variable X , the number of sets played in the Wimbledon men’s singles final match.
- Draw a graph of the discrete probability distribution.
- Compute and interpret the mean of the random variable X .

- Compute the standard deviation of the random variable X .
- A life insurance company sells a \$100,000 one-year term life insurance policy to a 35-year-old male for \$200. According to the *National Vital Statistics Report*, 56(9), the probability the male survives the year is 0.998725. Compute and interpret the expected value of this policy to the life insurance company.
- State the criteria that must be met for an experiment to be a binomial experiment.
- Determine whether the probability experiment represents a binomial experiment. If not, explain why.
- An urn contains 20 colored golf balls: 8 white, 6 red, 4 blue, and 2 yellow. A child is allowed to draw balls until he gets a yellow one. The number of draws required is recorded.
- According to the *Uniform Crime Report*, 2017, 17.6% of property crimes committed in the United States were cleared by arrest or exceptional means. Twenty-five property crimes from 2017 are randomly selected and the number that was cleared is recorded.
- According to a study conducted by CESI Debt Solutions, 80% of married people hide purchases from their mates. In a random sample of 20 married people, find and interpret:
 - The probability exactly 15 hide purchases from their mates.
 - The probability at least 19 hide purchases from their mates.
 - The probability fewer than 19 hide purchases from their mates.
 - The probability between 15 and 17, inclusive, hide purchases from their mates.
- Suppose the adult American population is equally split in their belief that the amount of tax (federal, state, property, sales, and so on) they pay is too high.
 - How many people would we expect to say they pay too much tax if we surveyed 1200 randomly selected adult Americans?

- (b) Explain why we can use the Empirical Rule with the idea of unusual events (events that occur with relative frequency less than 0.05) to identify any unusual results in a survey of 1200 adult Americans.
- (c) If a survey of 1200 adult Americans results in 640 stating they feel the amount of tax they pay is too high, would these results contradict the belief that adult Americans are equally split in their belief that the amount of tax they pay is too high? Why?

9. Consider a binomial probability distribution with parameters $n = 5$ and $p = 0.2$.
- (a) Construct a binomial probability distribution with these parameters.
- (b) Compute the mean and standard deviation of the distribution.
- (c) Graph the discrete probability distribution. Comment on the shape of the distribution.

Making an Informed Decision

Should We Convict?

A woman who was shopping in Los Angeles had her purse stolen by a young, blonde female who was wearing a ponytail. The blonde female got into a yellow car that was driven by a black male who had a mustache and a beard. The police located a blonde female named Janet Collins who wore her hair in a ponytail and had a friend who was a black male who had a mustache and beard and also drove a yellow car. The police arrested the two subjects.

Because there were no eyewitnesses and no real evidence, the prosecution used probability to make its case against the defendants. The probabilities below were presented by the prosecution for the known characteristics of the thieves.

Characteristic	Probability
Yellow car	$\frac{1}{10}$
Man with a mustache	$\frac{1}{4}$
Woman with a ponytail	$\frac{1}{10}$
Woman with blonde hair	$\frac{1}{3}$
Black man with beard	$\frac{1}{10}$
Interracial couple in car	$\frac{1}{1000}$

(a) Assuming that the characteristics listed are independent of each other, what is the probability that a randomly selected couple has all these characteristics? That is, what is P ("yellow car" and "man with a mustache" and ... and "interracial couple in a car")?

(b) Would you convict the defendants based on this probability? Why?

(c) Now let n represent the number of couples in the Los Angeles area who could have committed the crime. Let p represent the probability that a randomly selected couple has all six characteristics listed. Let the random variable X represent the number of couples who have all the characteristics listed in the table. Assuming that the random variable X follows the binomial probability function, we have

$$P(x) = {}_nC_x \cdot p^x(1-p)^{n-x} \quad x = 0, 1, 2, \dots, n$$

Assuming that there are $n = 1,000,000$ couples in the Los Angeles area, what is the probability that more than one of them has the characteristics listed in the table? Does this result cause you to change your mind regarding the defendants' guilt?

(d) Now let's look at this case from a different point of view. Compute the probability that more than one couple has the characteristics described, given that at least one couple has the characteristics.

$$\begin{aligned} P(X > 1 | X \geq 1) &= \frac{P(X > 1 \text{ and } X \geq 1)}{P(X \geq 1)} \\ &= \frac{P(X > 1)}{P(X \geq 1)} \end{aligned}$$

Compute this probability, assuming that $n = 1,000,000$. Compute this probability again, but this time assume that $n = 2,000,000$. Do you think that the couple should be convicted "beyond all reasonable doubt"? Why?





The Normal Probability Distribution

Outline

- 7.1** Properties of the Normal Distribution
- 7.2** Applications of the Normal Distribution
- 7.3** Assessing Normality
- 7.4** The Normal Approximation to the Binomial Probability Distribution

Making an Informed Decision



You are interested in modeling the behavior of stocks. In particular, you want to build a model that describes the rate of return on a basket of stocks, such as large capitalization companies. To build this model, you must identify and use historical rates of return on a basket of stocks. Then your model can be used to identify high-performing companies that might be worthy of your investment. See the Decisions project on page 368.

Putting It Together

In Chapter 6, we discussed discrete random variables. In particular, we learned how to find probabilities for binomial random variables using the binomial probability distribution function. This probability distribution function is a formula that allows us to determine probabilities. Recall, we mentioned that probability models can be in the form of a table, a formula, or a graph.

In this chapter, we discuss how to determine probabilities for continuous random variables. The approach used in determining probabilities for continuous random variables differs from that of discrete random variables. For discrete random variables, we use a formula to determine probabilities. For continuous random variables, we use a graph that models the random variable to determine probabilities. In particular, we discuss distributions for two continuous random variables: the ***uniform distribution*** and the ***normal distribution***. Most of the discussion will focus on the normal distribution, which has many applications.

7.1 Properties of the Normal Distribution



Preparing for This Section Before getting started, review the following:

- Continuous variable (Section 1.1, p. 8)
- The Empirical Rule (Section 3.2, pp. 129–131)
- Rules for a discrete probability distribution (Section 6.1, p. 301)

Objectives ① Use the uniform probability distribution

② Graph a normal curve

③ State the properties of the normal curve

④ Explain the role of area in the normal density function

① Use the Uniform Probability Distribution

First, a uniform distribution will be discussed to see the relation between area and probability.

EXAMPLE 1 The Uniform Distribution

Imagine that a friend of yours is always late. Let the random variable X represent the time from when you are supposed to meet your friend until he shows up. Suppose your friend could be on time ($x = 0$) or up to 30 minutes late ($x = 30$), with all equal intervals of time between $x = 0$ and $x = 30$ being equally likely. For example, your friend is just as likely to be 3–4 minutes late as he is to be 25–26 minutes late. The random variable X can be any value in the interval from 0 to 30, that is, $0 \leq x \leq 30$. Because any two intervals of equal length between 0 and 30, inclusive, are equally likely, the random variable X is said to follow a **uniform probability distribution**.

When computing probabilities for discrete random variables, the value of the random variable is usually substituted into a formula.

Things are not as easy for continuous random variables. Because an infinite number of outcomes are possible for continuous random variables, the probability of observing one *particular* value is zero. For example, the probability that your friend is exactly 12.9438823 minutes late is zero. This result is based on the fact that classical probability is found by dividing the number of ways an event can occur by the total number of possibilities: there is one way to observe 12.9438823, and there are an infinite number of possible values between 0 and 30. To resolve this problem, probabilities of continuous random variables are computed over an *interval* of values. For example, we might compute the probability that your friend is between 10 and 15 minutes late. To find probabilities for continuous random variables, we use a *probability density function*.

Definition

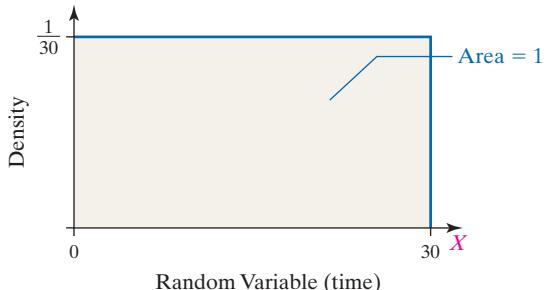
A **probability density function (pdf)** is an equation used to compute probabilities of continuous random variables. It must satisfy the following two properties:

1. The total area under the graph of the equation over all possible values of the random variable must equal 1.
2. The height of the graph of the equation must be greater than or equal to 0 for all possible values of the random variable.

Property 1 is similar to the rule for discrete probability distributions that stated the sum of the probabilities must add up to 1. Property 2 is similar to the rule that stated all probabilities must be greater than or equal to 0.

Figure 1 illustrates these properties for Example 1. Since any value of the random variable between 0 and 30 is equally likely, the graph of the probability density function is a rectangle. Because the random variable is any number between 0 and 30 inclusive, the width of the rectangle is 30. The area under the graph of the probability density function must equal 1, and the area of a rectangle equals height times width, so the height of the rectangle must be $\frac{1}{30}$.

Figure 1

**IN OTHER WORDS**

To find probabilities for continuous random variables, we do not use probability distribution functions (as we did for discrete random variables). Instead, we use probability density functions. The word **density** is used because it refers to the number of individuals per unit of area.

A pressing question remains: How are density functions used to find probabilities of continuous random variables?

The area under the graph of a density function over an interval represents the probability of observing a value of the random variable in that interval.

The following example illustrates this statement.

EXAMPLE 2**Area as a Probability**

Problem Refer to the situation in Example 1.

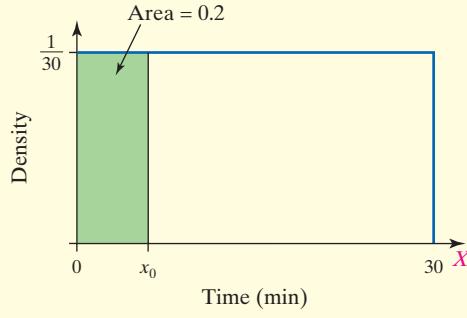
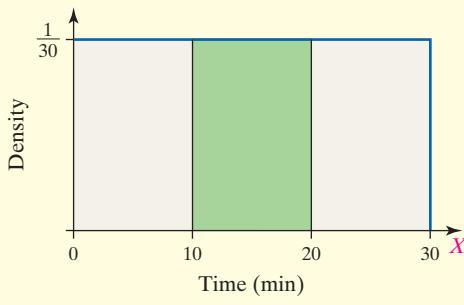
- What is the probability your friend will be between 10 and 20 minutes late?
- It is 10 A.M. There is a 20% probability your friend will arrive within the next _____ minutes.

Approach Use the graph of the density function in Figure 1 to find the solutions.

Solution

- We want to find the shaded area in Figure 2(a). The width of the shaded rectangle is 10 and its height is $\frac{1}{30}$. The area between 10 and 20 is $10\left(\frac{1}{30}\right) = \frac{1}{3}$. The probability your friend is between 10 and 20 minutes late is $\frac{1}{3}$.
- The area of the shaded region in Figure 2(b) is 0.2. Here we need to determine the width of the rectangle so that its area is 0.2. Solve $x_0 \cdot \frac{1}{30} = 0.2$ and find $x_0 = 30(0.2) = 6$. There is a 20% probability your friend will arrive within the next 6 minutes, or by 10:06 A.M.

Figure 2

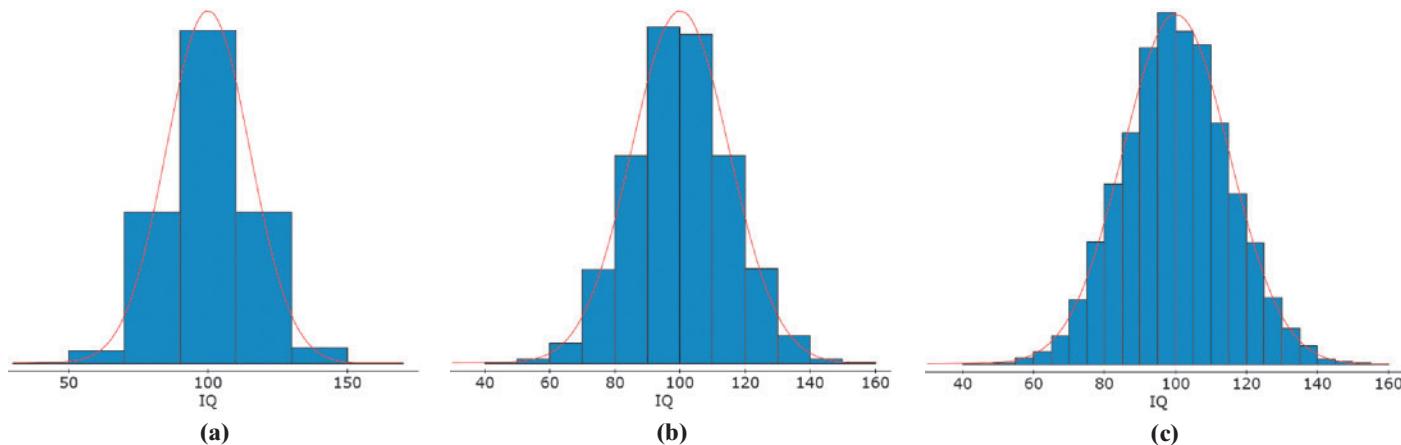


The uniform density function was introduced so we could associate probability with area. We are now better prepared to discuss the most frequently used continuous distribution, the *normal distribution*.

② Graph a Normal Curve

A rectangle is used to find the probability of observing an interval of numbers (such as 10–20 minutes after 10 A.M.) for a uniform random variable. However, not all continuous random variables follow a uniform distribution. For example, continuous random variables such as IQ scores and birth weights of babies have distributions that are symmetric and bell shaped. Consider the histograms in Figure 3, which represent the IQ scores of 10,000 randomly selected adults. Notice that as the class width of the histogram decreases, the histogram becomes closely approximated by the smooth red curve. For this reason, we can use the red curve to *model* the probability distribution of this continuous random variable.

Figure 3



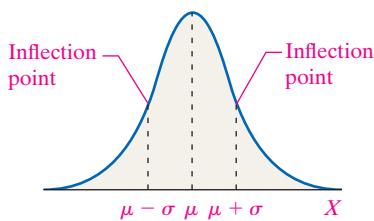
In mathematics, a **model** is an equation, table, or graph used to describe reality. The red curve in Figure 3 is a model called the **normal curve**, which is used to describe continuous random variables that are said to be *normally distributed*.

Definitions

A continuous random variable is **normally distributed**, or has a **normal probability distribution**, if its relative frequency histogram has the shape of a normal curve.

Figure 4* shows a normal curve, demonstrating the roles that the mean μ and standard deviation σ play in drawing the curve. The mode represents the “high point” of the graph of any distribution. The median represents the point where 50% of the area under the distribution is to the left and 50% is to the right. The mean represents the

Figure 4



*The vertical scale on the graph, which indicates **density**, is purposely omitted. The vertical scale, while important, will not play a role in any of the computations using this curve.

balancing point of the graph of the distribution (see Figure 2 on page 110 in Section 3.1). For symmetric distributions with a single peak, such as the normal distribution, the mean = median = mode. Because of this, the mean, μ , corresponds to the high point of the graph of the distribution.

The points at $x = \mu - \sigma$ and $x = \mu + \sigma$ are the **inflection points** on the normal curve, the points on the curve where the curvature of the graph changes. To the left of $x = \mu - \sigma$ and to the right of $x = \mu + \sigma$, the curve is drawn upward (↑ or ↗). Between $x = \mu - \sigma$ and $x = \mu + \sigma$, the curve is drawn downward (↓ or ↘).

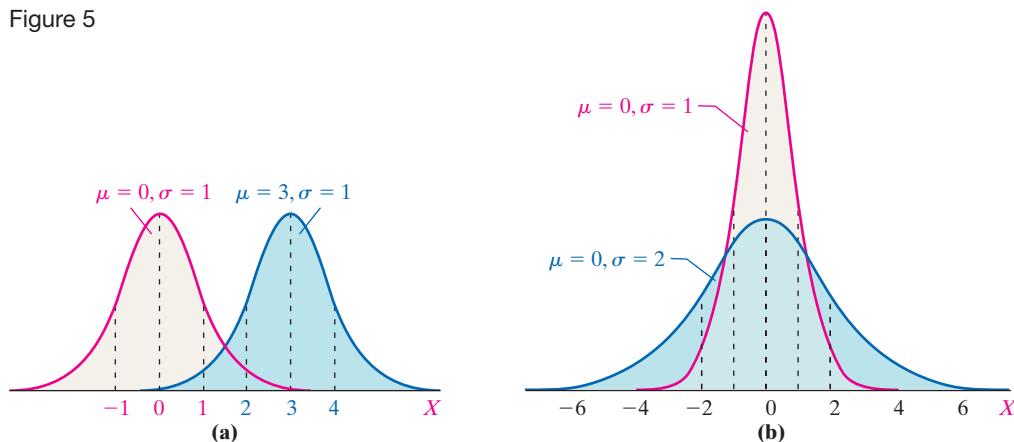
Figure 5 shows how changes in μ and σ change the position or shape of a normal curve. In Figure 5(a), one density curve has $\mu = 0, \sigma = 1$, and the other has $\mu = 3, \sigma = 1$. We can see that increasing the mean from 0 to 3 caused the graph to shift three units to the right but maintained its shape. In Figure 5(b), one density curve has $\mu = 0, \sigma = 1$, and the other has $\mu = 0, \sigma = 2$. We can see that increasing the standard deviation from 1 to 2 caused the graph to become flatter and more spread out but maintained its location of center.

Historical Note



Karl Pearson (of correlation fame) coined the phrase *normal curve*. He did not do this to imply that a distribution that is not normal is *abnormal*. Rather, Pearson wanted to avoid giving the name of the distribution a proper name, such as Gaussian (as in Carl Friedrich Gauss, who is incorrectly credited with the discovery of the normal curve). See the Historical Note on Abraham DeMoivre on the next page.

Figure 5



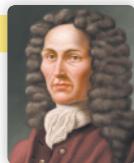
NW Now Work Problem 23

③ State the Properties of the Normal Curve

The normal probability density function satisfies all the requirements of probability distributions. We list the properties of the normal density curve next.

Properties of the Normal Density Curve

1. The normal curve is symmetric about its mean, μ .
2. Because mean = median = mode, the normal curve has a single peak and the highest point occurs at $x = \mu$.
3. The normal curve has inflection points at $\mu - \sigma$ and $\mu + \sigma$.
4. The area under the normal curve is 1.
5. The area under the normal curve to the right of μ equals the area under the curve to the left of μ , which equals $\frac{1}{2}$.
6. As x increases without bound (gets larger and larger), the graph approaches, but never reaches, the horizontal axis. As x decreases without bound (gets more and more negative), the graph approaches, but never reaches, the horizontal axis.

Historical Note

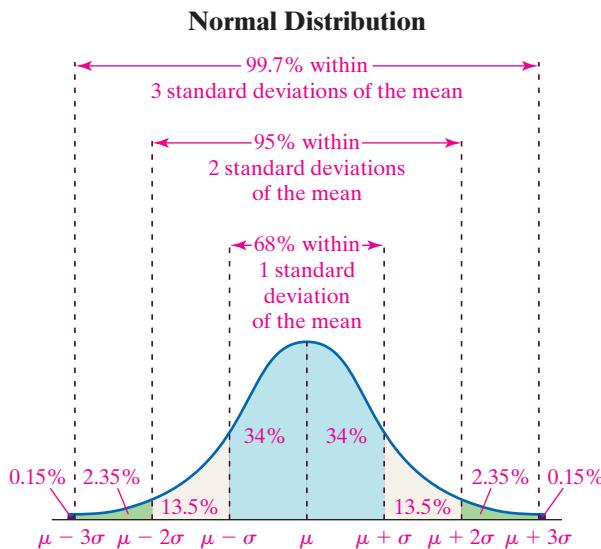
Abraham de Moivre was born in France on May 26, 1667. He is known as a great contributor to the areas of probability and trigonometry. In 1685, he moved to England. De Moivre was elected a fellow of the Royal Society in 1697. He was part of the commission to settle the dispute between Newton and Leibniz regarding who was the discoverer of calculus. He published *The Doctrine of Chance* in 1718. In 1733, he developed the equation that describes the normal curve. Unfortunately, de Moivre had a difficult time being accepted in English society and was able to make only a meager living tutoring mathematics. An interesting piece of information regarding de Moivre: he correctly predicted the day of his death, November 27, 1754.

7. The Empirical Rule:

- Approximately 68% of the area under the normal curve is between $x = \mu - \sigma$ and $x = \mu + \sigma$;
- approximately 95% of the area is between $x = \mu - 2\sigma$ and $x = \mu + 2\sigma$;
- approximately 99.7% of the area is between $x = \mu - 3\sigma$ and $x = \mu + 3\sigma$.

See Figure 6.

Figure 6

**4 Explain the Role of Area in the Normal Density Function**

Let's look at an example of a normally distributed random variable.

EXAMPLE 3**A Normal Random Variable****Table 1**

Height (inches)	Relative Frequency
29.0–29.9	0.005
30.0–30.9	0.005
31.0–31.9	0.005
32.0–32.9	0.025
33.0–33.9	0.02
34.0–34.9	0.055
35.0–35.9	0.075
36.0–36.9	0.09
37.0–37.9	0.115
38.0–38.9	0.15
39.0–39.9	0.12
40.0–40.9	0.11
41.0–41.9	0.07
42.0–42.9	0.06
43.0–43.9	0.035
44.0–44.9	0.025
45.0–45.9	0.025
46.0–46.9	0.005
47.0–47.9	0.005

Problem The relative frequency distribution given in Table 1 represents the heights of a pediatrician's three-year-old female patients. The raw data indicate that the mean height of the patients is $\mu = 38.72$ inches with standard deviation $\sigma = 3.17$ inches.

- Draw a relative frequency histogram of the data. Comment on the shape of the distribution.
- Draw a normal curve with $\mu = 38.72$ inches and $\sigma = 3.17$ inches on the relative frequency histogram. Compare the area of the rectangle for heights between 40 and 40.9 inches to the area under the normal curve for heights between 40 and 40.9 inches.

Approach

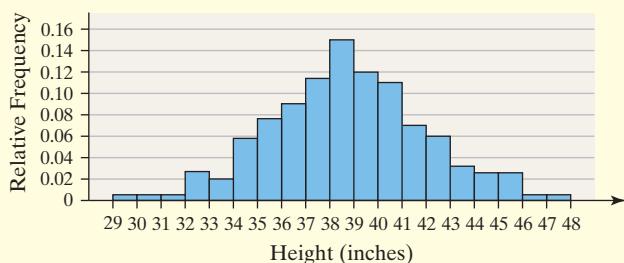
- Draw the relative frequency histogram. If the histogram is shaped like Figure 4, then height is approximately normal. We say “approximately normal,” rather than “normal,” because the normal curve is an idealized description of the data, and data rarely follow the curve exactly.
- Draw the normal curve on the histogram with the high point at μ and the inflection points at $\mu - \sigma$ and $\mu + \sigma$. Shade the rectangle corresponding to heights between 40 and 40.9 inches.

Solution

- Figure 7 shows the relative frequency histogram, which is symmetric and bell shaped.

(continued)

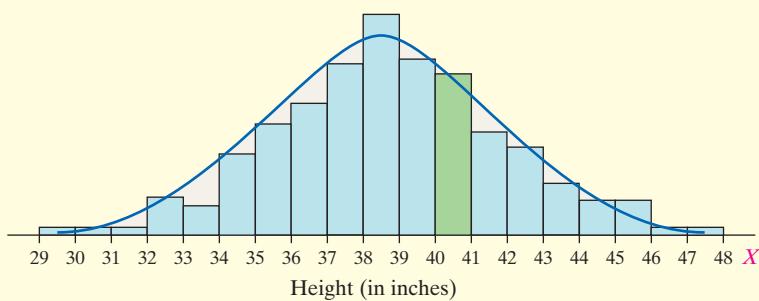
Figure 7

Height of Three-Year-Old Females

- (b) In Figure 8, the normal curve with $\mu = 38.72$ and $\sigma = 3.17$ is superimposed on the relative frequency histogram. The normal curve describes the data fairly well. We conclude that the heights of three-year-old females are approximately normal with $\mu = 38.72$ inches and $\sigma = 3.17$ inches.

Figure 8 also shows the rectangle whose area represents the proportion of three-year-old females between 40 and 40.9 inches. Notice that the area of this shaded region is very close to the area under the normal curve for the same region, so we can use the area under the normal curve to approximate the proportion of three-year-old females with heights between 40 and 40.9 inches!

Figure 8

Height of Three-Year-Old Females

The equation (or model) used to determine the probability of a continuous random variable is called a **probability density function** (or **pdf**). The **normal probability density function** is given by

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where μ is the mean and σ is the standard deviation of the normal random variable. Do not feel threatened by this equation, because we will not be using it in this text. Instead, we will use the normal distribution in graphical form by drawing the normal curve.

We now summarize the role area plays in the normal curve.

Area under a Normal Curve

Suppose that a random variable X is normally distributed with mean μ and standard deviation σ . The area under the normal curve for any interval of values of the random variable X represents either

- the proportion of the population with the characteristic described by the interval of values or
- the probability that a randomly selected individual from the population will have the characteristic described by the interval of values.

EXAMPLE 4**Interpreting the Area under a Normal Curve****Historical Note**

The normal probability distribution is often referred to as the Gaussian distribution in honor of Carl Gauss, the individual thought to have discovered the idea. However, it was actually Abraham de Moivre who first wrote down the equation of the normal distribution. Gauss was born in Brunswick, Germany, on April 30, 1777. Mathematical prowess was evident early in Gauss's life. At age 8 he was able to instantly add the first 100 integers. In 1799, Gauss earned his doctorate. The subject of his dissertation was the Fundamental Theorem of Algebra. In 1809, Gauss published a book on the mathematics of planetary orbits. In this book, he further developed the theory of least-squares regression by analyzing the errors. The analysis of these errors led to the discovery that errors follow a normal distribution. Gauss was considered to be "glacially cold" as a person and had troubled relationships with his family. Gauss died on February 23, 1855.

Problem The serum total cholesterol for males 20–29 years old is approximately normally distributed with mean $\mu = 180$ mg/dL and $\sigma = 36.2$ mg/dL, based on data obtained from the National Health and Nutrition Examination Survey.

- Draw a normal curve with the parameters labeled.
- An individual with total cholesterol greater than 200 mg/dL is considered to have high cholesterol. Shade the region under the normal curve to the right of $x = 200$.
- Suppose that the area under the normal curve to the right of $x = 200$ is 0.2903. (You will learn how to find this area in Section 7.2.) Provide two interpretations of this result.

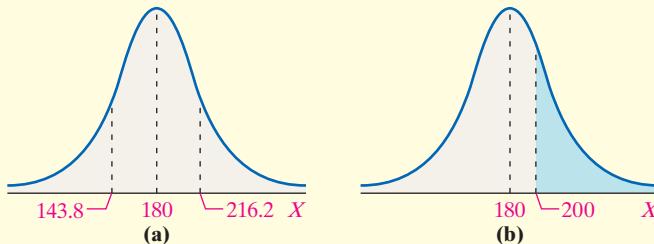
Approach

- Draw the normal curve with the mean $\mu = 180$ mg/dL labeled at the high point and the inflection points at $\mu - \sigma = 180 - 36.2 = 143.8$ and $\mu + \sigma = 180 + 36.2 = 216.2$.
- Shade the region under the normal curve to the right of $x = 200$.
- The two interpretations of the area under a normal curve are (1) a proportion and (2) a probability.

Solution

- Figure 9(a) shows the graph of the normal curve.
- Figure 9(b) shows the region under the normal curve to the right of $x = 200$ shaded.

Figure 9



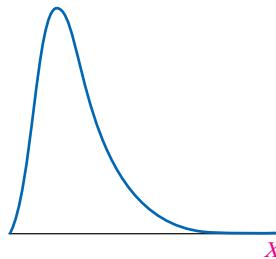
- The two interpretations for the area of this shaded region are (1) the proportion of 20- to 29-year-old males that have high cholesterol is 0.2903 and (2) the probability that a randomly selected 20- to 29-year-old male has high cholesterol is 0.2903.

NW Now Work Problems 29 and 33**7.1 Assess Your Understanding****Vocabulary and Skill Building**

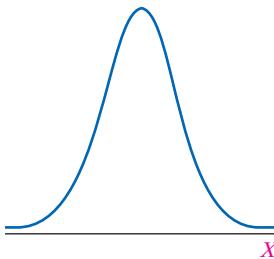
- What are the two properties that a probability density function must satisfy?
- A _____ is an equation, table, or graph used to describe reality.
- What are the properties of the normal density curve?
- The area under a normal curve can be interpreted as a _____ or _____.

For Problems 5–10, determine whether the graph can represent a normal curve. If it cannot, explain why.

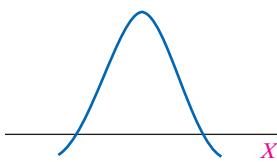
5.



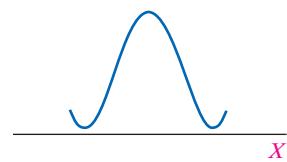
6.



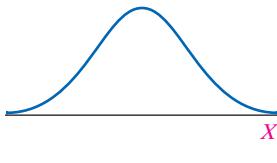
7.



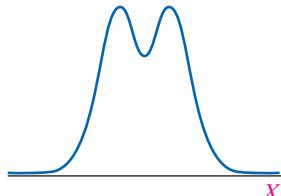
8.



9.



10.



Problems 11–14 use the information presented in Examples 1 and 2.

- NW** 11. (a) Find the probability that your friend is between 5 and 10 minutes late.

(b) It is 10 A.M. There is a 40% probability your friend will arrive within the next ___ minutes.

12. (a) Find the probability that your friend is between 15 and 25 minutes late.

(b) It is 10 A.M. There is a 90% probability your friend will arrive within the next ___ minutes.

13. Find the probability that your friend is at least 20 minutes late.

14. Find the probability that your friend is no more than 5 minutes late.

15. **Uniform Distribution** The random-number generator on calculators randomly generates a number between 0 and 1. The random variable X , the number generated, follows a uniform probability distribution.

- (a) Draw the graph of the uniform density function.
 (b) What is the probability of generating a number between 0 and 0.2?
 (c) What is the probability of generating a number between 0.25 and 0.6?
 (d) What is the probability of generating a number greater than 0.95?
 (e) Use your calculator or statistical software to randomly generate 200 numbers between 0 and 1. What proportion of the numbers are between 0 and 0.2? Compare the result with part (b).

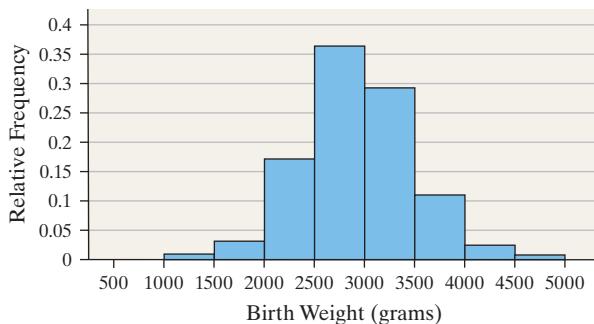
16. **Uniform Distribution** The reaction time X (in minutes) of a certain chemical process follows a uniform probability distribution with $5 \leq X \leq 10$.

- (a) Draw the graph of the density curve.
 (b) What is the probability that the reaction time is between 6 and 8 minutes?
 (c) What is the probability that the reaction time is between 5 and 8 minutes?
 (d) What is the probability that the reaction time is less than 6 minutes?

In Problems 17–20, determine whether or not the histogram indicates that a normal distribution could be used as a model for the variable.

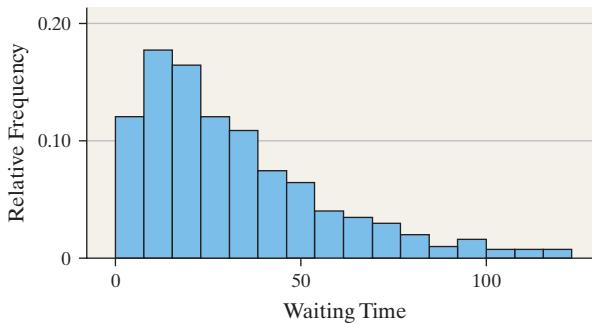
17. **Birth Weights** The relative frequency histogram represents the birth weights (in grams) of babies whose term was 36 weeks.

Birth Weights of Babies Whose Term Was 36 Weeks



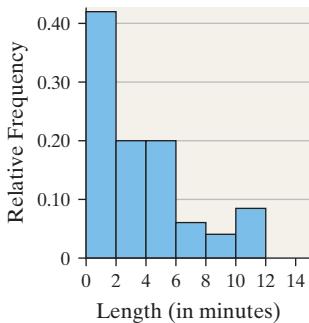
18. **Waiting in Line** The relative frequency histogram represents the waiting times (in minutes) to ride the American Eagle Roller Coaster for 2000 randomly selected people on a Saturday afternoon in the summer.

Waiting Time for the American Eagle Roller Coaster

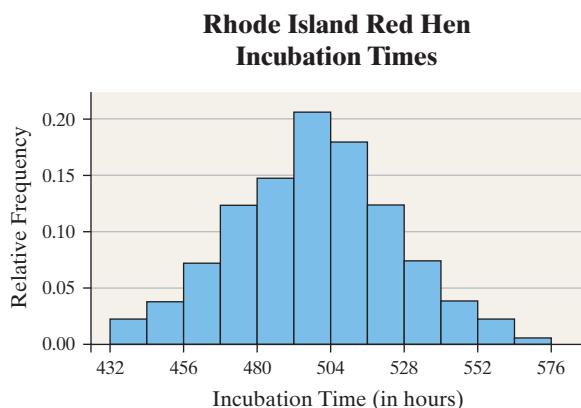


19. **Length of Phone Calls** The relative frequency histogram represents the length of phone calls on my wife's cell phone during the month of September.

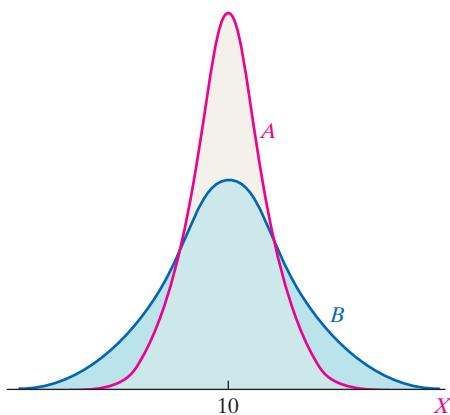
Length of Phone Calls



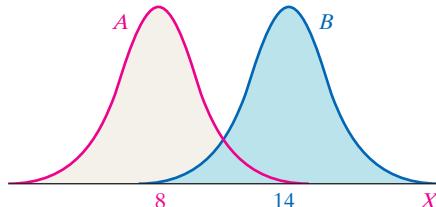
20. **Incubation Times** The relative frequency histogram on the following page represents the incubation times of a random sample of Rhode Island Red hens' eggs.



21. One graph in the figure below represents a normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 3$. The other graph represents a normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 2$. Determine which graph is which and explain how you know.

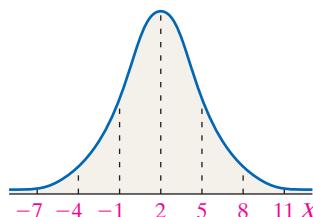


22. One graph in the figure below represents a normal distribution with mean $\mu = 8$ and standard deviation $\sigma = 2$. The other graph represents a normal distribution with mean $\mu = 14$ and standard deviation $\sigma = 2$. Determine which graph is which and explain how you know.

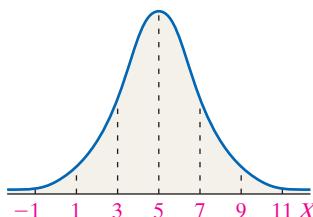


In Problems 23–26, the graph of a normal curve is given. Use the graph to identify the values of μ and σ .

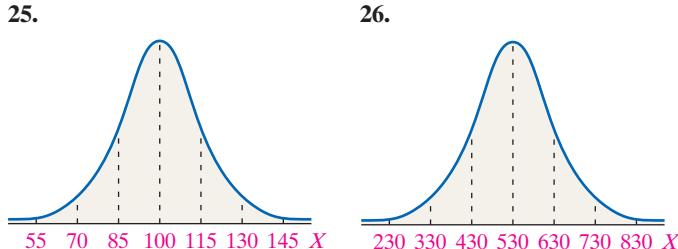
NW 23.



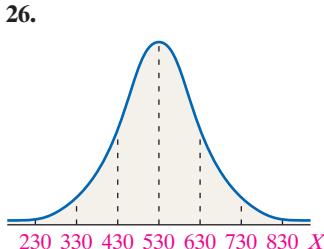
24.



25.



26.



In Problems 27 and 28, draw a normal curve and label the mean and inflection points.

27. $\mu = 30$ and $\sigma = 10$

28. $\mu = 50$ and $\sigma = 5$

Applying the Concepts

NW 29. You Explain It! Upper Arm Length The length of a 4-year-old male's upper arm is approximately normally distributed with mean $\mu = 21.8$ cm and standard deviation $\sigma = 2.0$ cm. *Source: Vital and Health Statistics, U.S. Department of Health and Human Services.*

- (a) Draw a normal curve with the parameters labeled.
- (b) Shade the region that represents the proportion of 4-year-old males whose upper arm length is less than 20 cm.
- (c) Suppose the area under the normal curve to the left of $x = 20$ cm is 0.1841. Provide two interpretations of this result.

30. You Explain It! Upper Leg Length The length of a 10-year-old female's upper leg is approximately normally distributed with mean $\mu = 33.8$ cm and standard deviation $\sigma = 2.1$ cm. *Source: Vital and Health Statistics, U.S. Department of Health and Human Services.*

- (a) Draw a normal curve with the parameters labeled.
- (b) Shade the region that represents the proportion of 10-year-old females whose upper leg length is more than 37 cm.
- (c) Suppose the area under the normal curve to the right of $x = 37$ cm is 0.0638. Provide two interpretations of this result.

31. You Explain It! Birth Weights The birth weights of full-term babies are normally distributed with mean $\mu = 3400$ grams and $\sigma = 505$ grams.

Source: Based on data obtained from the National Vital Statistics Report, Vol. 48, No. 3.

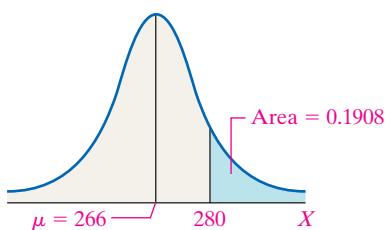
- (a) Draw a normal curve with the parameters labeled.
- (b) Shade the region that represents the proportion of full-term babies who weigh more than 4410 grams.
- (c) Suppose the area under the normal curve to the right of $x = 4410$ grams is 0.0228. Provide two interpretations of this result.

32. You Explain It! Height of 10-Year-Old Males The heights of 10-year-old males are normally distributed with mean $\mu = 55.9$ inches and $\sigma = 5.7$ inches.

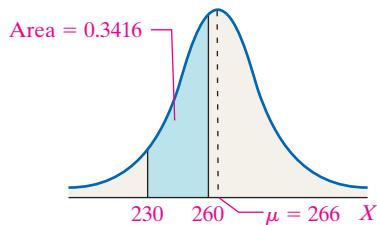
- (a) Draw a normal curve with the parameters labeled.
- (b) Shade the region that represents the proportion of 10-year-old males who are less than 46.5 inches tall.
- (c) Suppose the area under the normal curve to the left of $x = 46.5$ inches is 0.0496. Provide two interpretations of this result.

NW 33. You Explain It! Gestation Period The lengths of human pregnancies are normally distributed with $\mu = 266$ days and $\sigma = 16$ days.

- (a) The figure represents the normal curve with $\mu = 266$ days and $\sigma = 16$ days. The area to the right of $x = 280$ days is 0.1908. Provide two interpretations of this area.

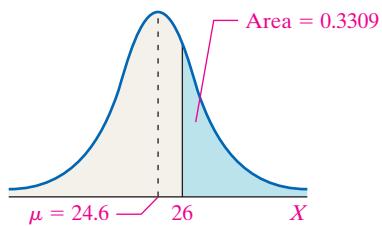


- (b) The figure represents the normal curve with $\mu = 266$ days and $\sigma = 16$ days. The area between $x = 230$ days and $x = 260$ days is 0.3416. Provide two interpretations of this area.

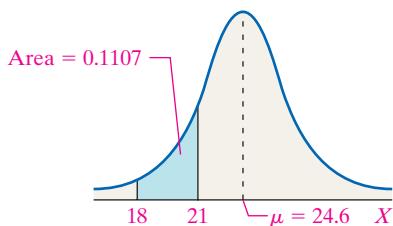


- 34. You Explain It! Miles per Gallon** Elena conducts an experiment in which she fills up the gas tank on her Toyota Camry 40 times and records the miles per gallon for each fill-up. A histogram of the miles per gallon (mpg) indicates that a normal distribution with a mean of 24.6 mpg and a standard deviation of 3.2 mpg could be used to model the gas mileage for her car.

- (a) The figure represents the normal curve with $\mu = 24.6$ miles per gallon and $\sigma = 3.2$ miles per gallon. The area under the curve to the right of $x = 26$ mpg is 0.3309. Provide two interpretations of this area.



- (b) The figure below represents the normal curve with $\mu = 24.6$ miles per gallon and $\sigma = 3.2$ miles per gallon. The area under the curve between $x = 18$ mpg and $x = 21$ mpg is 0.1107. Provide two interpretations of this area.



Retain Your Knowledge

- 37. Cardiac Arrest** Researchers conducted a prospective cohort study in which male patients who had an out-of-hospital cardiac arrest were submitted to therapeutic hypothermia (intravenous infusion of cold saline followed by surface cooling with the goal of maintaining body temperature of 33 degrees Celsius for 24 hours). Note that normal body temperature is 37 degrees Celsius. The survival status, length of stay in the intensive care unit (ICU), and time spent on a ventilator were measured. Each of these variables was compared to a historical cohort of patients who were treated prior to the availability of therapeutic

- 35. Hitting with a Pitching Wedge** In the game of golf, distance control is just as important as how far a player hits the ball. Michael went to the driving range with his range finder and hit 75 golf balls with his pitching wedge and measured the distance each ball traveled (in yards). He obtained the following data:

100	97	101	101	103	100	99	100	100
104	100	101	98	100	99	99	97	101
104	99	101	101	101	100	96	99	99
98	94	98	107	98	100	98	103	100
98	94	104	104	98	101	99	97	103
102	101	101	100	95	104	99	102	95
99	102	103	97	101	102	96	102	99
96	108	103	100	95	101	103	105	100
94	99	95						

- (a) Use statistical software to construct a relative frequency histogram. Comment on the shape of the distribution. Draw a normal density curve on the relative frequency histogram.
- (b) Do you think the normal density curve accurately describes the distance Michael hits with a pitching wedge? Why?

- 36. Heights of Five-Year-Old Females** The following data represent the heights (in inches) of 80 randomly selected five-year-old females.

44.5	42.4	42.2	46.2	45.7	44.8	43.3	39.5
45.4	43.0	43.4	44.7	38.6	41.6	50.2	46.9
39.6	44.7	36.5	42.7	40.6	47.5	48.4	37.5
45.5	43.3	41.2	40.5	44.4	42.6	42.0	40.3
42.0	42.2	38.5	43.6	40.6	45.0	40.7	36.3
44.5	37.6	42.2	40.3	48.5	41.6	41.7	38.9
39.5	43.6	41.3	38.8	41.9	40.3	42.1	41.9
42.3	44.6	40.5	37.4	44.5	40.7	38.2	42.6
44.0	35.9	43.7	48.1	38.7	46.0	43.4	44.6
37.7	34.6	42.4	42.7	47.0	42.8	39.9	42.3

- (a) Use statistical software to construct a relative frequency histogram. Comment on the shape of the distribution. Draw a normal density curve on the relative frequency histogram.
- (b) Do you think the normal density curve accurately describes the heights of five-year-old females? Why?

hypothermia. Of the 52 hypothermia patients, 37 survived; of the 74 patients in the control group, 43 survived. The median length of stay among survivors for the hypothermia patients was 14 days versus 21 days for the control group. The time on the ventilator among survivors for the hypothermia group was 219 hours versus 328 hours for the control group.

Source: Storem, Christian, et al. "Mild Therapeutic Hypothermia Shortens Intensive Care Unit Stay of Survivors After Out-of-Hospital Cardiac Arrest Compared to Historical Controls." *Critical Care* 2008, 12:R78 BioMed Central.

- (a) What does it mean to say this is a prospective cohort study?
- (b) What is the explanatory variable in the study? Is it qualitative or quantitative?
- (c) What are the three response variables in the study? For each, state whether the variable is qualitative or quantitative.
- (d) Are the reported times on the ventilator statistics or parameters? Explain.

- (e) To what population does this study apply?
- (f) Based on the results of this study, what is the probability a randomly selected male who has an out-of-hospital cardiac arrest and submits to therapeutic hypothermia will survive? What about those who do not submit to therapeutic hypothermia?

7.2 Applications of the Normal Distribution



Preparing for This Section Before getting started, review the following:

- z -scores (Section 3.4, pp. 146–147)
- Percentiles (Section 3.4, pp. 147–149)
- Complement Rule (Section 5.2, p. 247)

Objectives

- ① Find and interpret the area under a normal curve
- ② Find the value of a normal random variable

If X is a normally distributed random variable, the area under the normal curve represents the proportion of a population with a certain characteristic, or the probability that a randomly selected individual from the population has the characteristic.

The question then is, “How do I find the area under the normal curve?” We have two options—by-hand calculations with the aid of a table, or technology.

① Find and Interpret the Area under a Normal Curve

We use z -scores to help find the area under a normal curve by hand. Recall, the z -score allows us to transform a random variable X with mean μ and standard deviation σ into a random variable Z with mean 0 and standard deviation 1.

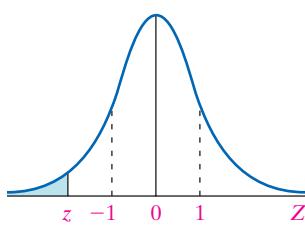
Standardizing a Normal Random Variable

Suppose that the random variable X is normally distributed with mean μ and standard deviation σ . Then the random variable

$$Z = \frac{X - \mu}{\sigma}$$

is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$. The random variable Z is said to have the **standard normal distribution**.

Figure 10



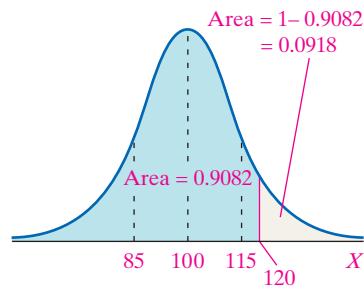
This result is powerful! If a normal random variable X has mean different from 0 or a standard deviation different from 1, X can be transformed into a **standard normal random variable** Z whose mean is 0 and standard deviation is 1. Then Table V (found on the inside back cover of the text and in Appendix A) may be used to find the area to the left of a specified z -score, z , as shown in Figure 10, which is also the area to the left of the value of x in the distribution of X . The graph in Figure 10 is called the **standard normal curve**.

For example, IQ scores can be modeled by a normal distribution with $\mu = 100$ and $\sigma = 15$. An individual whose IQ is 120, is $z = \frac{x - \mu}{\sigma} = \frac{120 - 100}{15} = 1.33$ standard deviations above the mean (recall, we round z -scores to two decimal places). Look in Table V and find the area under the standard normal curve to the left of $z = 1.33$ is 0.9082. See Figure 11. Therefore, the area under the normal curve to the left of $x = 120$ is 0.9082 as shown in Figure 12.

Figure 11

Standard Normal Distribution						
<i>z</i>	.00	.01	.02	.03	.04	.05
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265
1.5	0.9332	0.9355	0.9377	0.9397	0.9412	0.9427

Figure 12



To find the area to the right of the value of a random variable, use the Complement Rule and determine one minus the area to the left. For example, to find the area under the normal curve with mean $\mu = 100$ and standard deviation $\sigma = 15$ to the right of $x = 120$, compute

$$\begin{aligned} \text{Area} &= 1 - 0.9082 \\ &= 0.0918 \end{aligned}$$

as shown in Figure 12.

EXAMPLE 1 Finding Area under a Normal Curve

Problem A pediatrician obtains the heights of her three-year-old female patients. The heights are approximately normally distributed, with mean 38.72 inches and standard deviation 3.17 inches. Use the normal model to determine the proportion of the three-year-old females that have a height less than 35 inches.

By-Hand Approach

Step 1 Draw a normal curve and shade the desired area.

Step 2 Convert the value of x to a z -score using

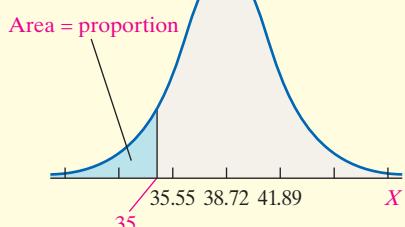
$$z = \frac{x - \mu}{\sigma}$$

Step 3 Use Table V to find the area to the left of the z -score found in Step 2.

By-Hand Solution

Step 1 Figure 13 shows the normal curve with the area to the left of 35 shaded.

Figure 13



Step 2 Convert $x = 35$ to a z -score.

$$z = \frac{x - \mu}{\sigma} = \frac{35 - 38.72}{3.17} = -1.17$$

Technology Approach

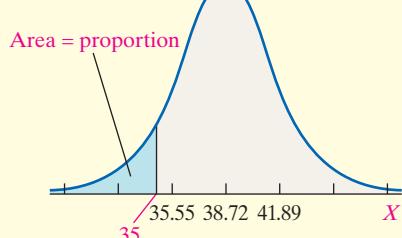
Step 1 Draw a normal curve and shade the desired area.

Step 2 Use a statistical spreadsheet or calculator with advanced statistical features to find the area. The steps for determining the area under any normal curve using the TI-83/84 Plus graphing calculator, Minitab, Excel, and StatCrunch are found in the Technology Step-by-Step on pages 350–351.

Technology Solution

Step 1 Figure 15 shows the normal curve with the area to the left of 35 shaded.

Figure 15



Step 3 Look up $z = -1.17$ in Table V and find the entry. The area to the left of $z = -1.17$ is 0.1210. See Figure 14. Therefore, the area to the left of $x = 35$ is 0.1210.

Figure 14

z	.00	.01	.02	.03	.04	.05	.06	.07	.08
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0080	0.0079	0.0078	0.0078	0.0078	0.0077	0.0075	0.0074	0.0073
-3.0	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838
-2.9	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003
-2.8	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190
-2.7	0.1587	0.1562	0.1530	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401

The normal model indicates that the proportion of the pediatrician's three-year-old females who are less than 35 inches tall is 0.1210.

Step 2 Figure 16 shows the results from Minitab. The area under the normal curve to the left of 35 is 0.1203.

Figure 16

Cumulative Distribution Function

Normal with mean = 38.72 and standard deviation = 3.17
 $x \quad P(X \leq x)$
 35 0.120297

The normal model indicates that the proportion of the pediatrician's three-year-old females who are less than 35 inches tall is 0.1203.



CAUTION!

Notice the by-hand solution and technology solution in Example 1 differ. The difference exists because we rounded the z-score in Step 2 of the by-hand solution, which leads to rounding error.

Table 2

Height (inches)	Relative Frequency
29.0–29.9	0.005
30.0–30.9	0.005
31.0–31.9	0.005
32.0–32.9	0.025
33.0–33.9	0.02
34.0–34.9	0.055
35.0–35.9	0.075
36.0–36.9	0.09
37.0–37.9	0.115
38.0–38.9	0.15
39.0–39.9	0.12
40.0–40.9	0.11
:	:
47.0–47.9	0.005

According to the results of Example 1, the proportion of three-year-old females who are shorter than 35 inches is approximately 0.12. If the normal curve is a good model for determining proportions (or probabilities), then about 12% of the three-year-olds in Table 1 (from Section 7.1) should be shorter than 35 inches. For convenience, part of Table 1 is repeated in Table 2.

The relative frequency distribution in Table 2 shows that $0.005 + 0.005 + 0.005 + 0.025 + 0.02 + 0.055 = 0.115 = 11.5\%$ of the three-year-old females are less than 35 inches tall. The results based on the normal curve are close to the actual results. The normal curve accurately models the heights.

If we wanted to know the proportion of three-year-old females whose height is greater than 35 inches, use the Complement Rule and find the proportion is $1 - 0.1210 = 0.879$ (using the “by-hand” computation).

Because the area under the normal curve represents a proportion, we can also use the area to find percentile ranks of scores. Recall that the k th percentile divides the lower $k\%$ of a data set from the upper $(100 - k)\%$. In Example 1, 12% of the females have a height less than 35 inches, and 88% of the females have a height greater than 35 inches, so a child whose height is 35 inches is at the 12th percentile.

EXAMPLE 2

Finding the Probability of a Normal Random Variable

Problem For the pediatrician presented in Example 1, find the probability that a randomly selected three-year-old girl is between 35 and 40 inches tall, inclusive. That is, find $P(35 \leq X \leq 40)$.

By-Hand Approach

Step 1 Draw a normal curve and shade the desired area.

Step 2 Convert the values of x to z -scores using

$$z = \frac{x - \mu}{\sigma}$$

Step 3 Use Table V to find the area to the left of each z -score found in Step 2. Use this result to find the area between the z -scores.

Technology Approach

Step 1 Draw a normal curve and shade the desired area.

Step 2 Use a statistical spreadsheet or calculator with advanced statistical features to find the area. The steps for determining the area under any normal curve using the TI-83/84 Plus graphing calculator, Minitab, Excel, and StatCrunch are found in the Technology Step-by-Step on pages 350–351.

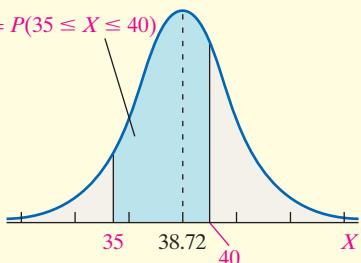
(continued)

By-Hand Solution

Step 1 Figure 17 shows the normal curve with the area between 35 and 40 shaded.

Figure 17

Area = $P(35 \leq X \leq 40)$



Step 2 Convert $x_1 = 35$ and $x_2 = 40$ to z -scores.

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{35 - 38.72}{3.17} = -1.17$$

$$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{40 - 38.72}{3.17} = 0.40$$

Step 3 Table V shows that the area to the left of $z_2 = 0.4$ (or $x_2 = 40$) is 0.6554 and the area to the left of $z_1 = -1.17$ (or $x_1 = 35$) is 0.1210, so the area between $z_1 = -1.17$ and $z_2 = 0.40$ is $0.6554 - 0.1210 = 0.5344$. The probability a randomly selected three-year-old female is between 35 and 40 inches tall is 0.5344. That is, $P(35 \leq X \leq 40) = P(-1.17 \leq Z \leq 0.40) = 0.5344$.

Interpretation If we randomly selected 100 three-year-old females, we would expect about 53 or 54 of them to be between 35 and 40 inches tall.

NW Now Work Problem 39

According to the relative frequency distribution in Table 2, the proportion of three-year-old females with heights between 35 and 40 inches is $0.075 + 0.09 + 0.115 + 0.15 + 0.12 = 0.55$. This is very close to the probability found in Example 2.

The methods for obtaining the area under a normal curve are summarized in Table 3.

Table 3

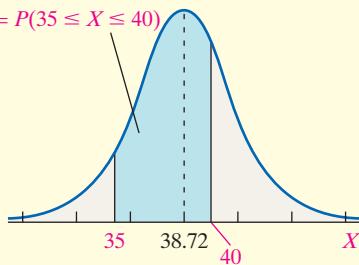
Problem	Approach	Solution
Find the area to the left of x .	Shade the area to the left of x .	<ul style="list-style-type: none"> Convert the value of x to a z-score. Use Table V to find the row and column that correspond to z. The area to the left of x is the value where the row and column intersect. <p style="text-align: center;">or</p> <ul style="list-style-type: none"> Use technology to find the area.
Find the area to the right of x .	Shade the area to the right of x .	<ul style="list-style-type: none"> Convert the value of x to a z-score. Use Table V to find the area to the left of z (which is also the area to the left of x). The area to the right of z (also x) is 1 minus the area to the left of z. <p style="text-align: center;">or</p> <ul style="list-style-type: none"> Use technology to find the area.

Technology Solution

Step 1 Figure 18 shows the normal curve with the area between 35 and 40 shaded.

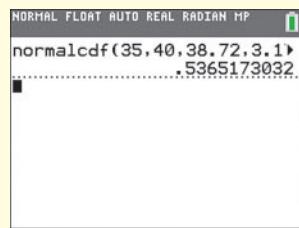
Figure 18

Area = $P(35 \leq X \leq 40)$



Step 2 Figure 19 shows the results from a TI-84 Plus CE graphing calculator.

Figure 19



The area between $x = 35$ and $x = 40$ is 0.5365. The probability a randomly selected three-year-old female is between 35 and 40 inches tall is 0.5365. That is, $P(35 \leq X \leq 40) = 0.5365$.

Table 3 (Continued)

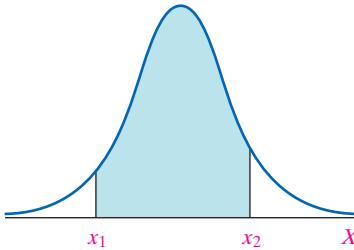
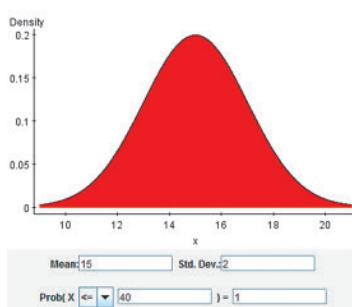
Problem	Approach	Solution
Find the area between x_1 and x_2 .	Shade the area between x_1 and x_2 .	<ul style="list-style-type: none"> Convert the values of x to z-scores. Use Table V to find the area to the left of z_1 and the area to the left of z_2. The area between x_1 and x_2 is the area between z_1 and z_2, which is $(\text{area to the left of } z_2) - (\text{area to the left of } z_1)$. or Use technology to find the area. 

Figure 20



Some Cautionary Thoughts

The normal curve extends indefinitely in both directions. For this reason, there is no range of values of a normal random variable for which the area under the curve is 1. For example, if asked to find the area under a normal curve to the left of $x = 40$ with $\mu = 15$ and $\sigma = 2$, StatCrunch (as well as other software and calculators) will state the area is 1, because it can only compute a limited number of decimal places. See Figure 20. However, the area under the curve to the left of $x = 40$ is not 1; it is some value slightly less than 1. So we will follow the practice of reporting such areas as >0.9999 . Similarly, if software reports an area of 0, we will report the area as <0.0001 .

When finding area under the normal curve by hand using Table V, we will report any area to the left of $z = -3.49$ (the smallest value of z in the table) or to the right of $z = 3.49$ (the largest value of z in the table) as <0.0001 . Any area under the normal curve to the left of $z = 3.49$ or to the right of $z = -3.49$ is reported as >0.9999 .

② Find the Value of a Normal Random Variable

Often, we do not want to find the proportion, probability, or percentile given a value of a normal random variable. Rather, we want to find the value of a normal random variable that corresponds to a certain proportion, probability, or percentile. For example, we might want to know the height of a three-year-old girl who is at the 20th percentile. Or we might want to know the scores on a standardized exam that separate the middle 90% of scores from the bottom and top 5%.

EXAMPLE 3 Finding the Value of a Normal Random Variable

Problem The heights of a pediatrician's three-year-old female patients are approximately normally distributed, with mean 38.72 inches and standard deviation 3.17 inches. Find the height of a three-year-old female at the 20th percentile.

By-Hand Approach

Step 1 Draw a normal curve and shade the desired area.

Step 2 Use Table V to find the z -score that corresponds to the shaded area.

Step 3 Obtain the normal value using the formula $x = \mu + z\sigma$.*

Technology Approach

Step 1 Draw a normal curve and shade the desired area.

Step 2 Use a statistical spreadsheet or calculator with advanced statistical features to find the score. The steps for determining the value of a normal random variable, given an area, using the TI-83/84 Plus graphing

(continued)

*The formula provided in Step 3 of the by-hand approach is the formula for computing a z -score, solved for x .

$$z = \frac{x - \mu}{\sigma} \quad \text{Formula for standardizing a value, } x, \text{ for a random variable } X$$

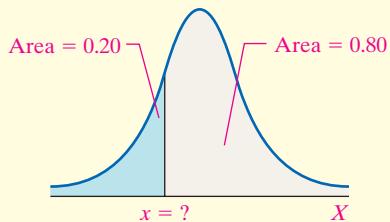
$$z\sigma = x - \mu \quad \text{Multiply both sides by } \sigma.$$

$$x = \mu + z\sigma \quad \text{Add } \mu \text{ to both sides.}$$

By-Hand Solution

Step 1 Figure 21 shows the normal curve with the unknown value of x at the 20th percentile, which separates the bottom 20% of the distribution from the top 80%.

Figure 21



Step 2 We want to find the z -score such that the area to the left of the z -score is 0.20. Refer to Table V and look in the body of the table for the area closest to 0.20. The area closest to 0.20 is 0.2005, which corresponds to a z -score of -0.84 . See Figure 22.

Figure 22

Standard Normal Distribution					
z	.00	.01	.02	.03	.04
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006
-3.1	0.0009	0.0009	0.0008	0.0008	0.0008
-3.0	0.0011	0.0011	0.0010	0.0010	0.0010
-2.9	0.0013	0.0013	0.0012	0.0012	0.0012
-2.8	0.0015	0.0015	0.0014	0.0014	0.0014
-2.7	0.0017	0.0017	0.0016	0.0016	0.0016
-2.6	0.0019	0.0019	0.0018	0.0018	0.0018
-2.5	0.0021	0.0021	0.0020	0.0020	0.0020
-2.4	0.0023	0.0023	0.0022	0.0022	0.0022
-2.3	0.0025	0.0025	0.0024	0.0024	0.0024
-2.2	0.0027	0.0027	0.0026	0.0026	0.0026
-2.1	0.0029	0.0029	0.0028	0.0028	0.0028
-2.0	0.0031	0.0031	0.0030	0.0030	0.0030
-1.9	0.0033	0.0033	0.0032	0.0032	0.0032
-1.8	0.0035	0.0035	0.0034	0.0034	0.0034
-1.7	0.0037	0.0037	0.0036	0.0036	0.0036
-1.6	0.0039	0.0039	0.0038	0.0038	0.0038
-1.5	0.0041	0.0041	0.0040	0.0040	0.0040
-1.4	0.0043	0.0043	0.0042	0.0042	0.0042
-1.3	0.0045	0.0045	0.0044	0.0044	0.0044
-1.2	0.0047	0.0047	0.0046	0.0046	0.0046
-1.1	0.0049	0.0049	0.0048	0.0048	0.0048
-1.0	0.0051	0.0051	0.0050	0.0050	0.0050
-0.9	0.0053	0.0053	0.0052	0.0052	0.0052
-0.8	0.0055	0.0055	0.0054	0.0054	0.0054
-0.7	0.0057	0.0057	0.0056	0.0056	0.0056
-0.6	0.0059	0.0059	0.0058	0.0058	0.0058
-0.5	0.0061	0.0061	0.0060	0.0060	0.0060
-0.4	0.0063	0.0063	0.0062	0.0062	0.0062
-0.3	0.0065	0.0065	0.0064	0.0064	0.0064
-0.2	0.0067	0.0067	0.0066	0.0066	0.0066
-0.1	0.0069	0.0069	0.0068	0.0068	0.0068
0.0	0.0071	0.0071	0.0070	0.0070	0.0070
0.1	0.0072	0.0072	0.0071	0.0071	0.0071
0.2	0.0073	0.0073	0.0072	0.0072	0.0072
0.3	0.0074	0.0074	0.0073	0.0073	0.0073
0.4	0.0075	0.0075	0.0074	0.0074	0.0074
0.5	0.0076	0.0076	0.0075	0.0075	0.0075
0.6	0.0077	0.0077	0.0076	0.0076	0.0076
0.7	0.0078	0.0078	0.0077	0.0077	0.0077
0.8	0.0079	0.0079	0.0078	0.0078	0.0078
0.9	0.0080	0.0080	0.0079	0.0079	0.0079
1.0	0.0081	0.0081	0.0080	0.0080	0.0080
1.1	0.0082	0.0082	0.0081	0.0081	0.0081
1.2	0.0083	0.0083	0.0082	0.0082	0.0082
1.3	0.0084	0.0084	0.0083	0.0083	0.0083
1.4	0.0085	0.0085	0.0084	0.0084	0.0084
1.5	0.0086	0.0086	0.0085	0.0085	0.0085
1.6	0.0087	0.0087	0.0086	0.0086	0.0086
1.7	0.0088	0.0088	0.0087	0.0087	0.0087
1.8	0.0089	0.0089	0.0088	0.0088	0.0088
1.9	0.0090	0.0090	0.0089	0.0089	0.0089
2.0	0.0091	0.0091	0.0090	0.0090	0.0090
2.1	0.0092	0.0092	0.0091	0.0091	0.0091
2.2	0.0093	0.0093	0.0092	0.0092	0.0092
2.3	0.0094	0.0094	0.0093	0.0093	0.0093
2.4	0.0095	0.0095	0.0094	0.0094	0.0094
2.5	0.0096	0.0096	0.0095	0.0095	0.0095
2.6	0.0097	0.0097	0.0096	0.0096	0.0096
2.7	0.0098	0.0098	0.0097	0.0097	0.0097
2.8	0.0099	0.0099	0.0098	0.0098	0.0098
2.9	0.0100	0.0100	0.0099	0.0099	0.0099
3.0	0.0100	0.0100	0.0099	0.0099	0.0099
3.1	0.0100	0.0100	0.0099	0.0099	0.0099
3.2	0.0100	0.0100	0.0099	0.0099	0.0099
3.3	0.0100	0.0100	0.0099	0.0099	0.0099
3.4	0.0100	0.0100	0.0099	0.0099	0.0099

Step 3 The height of a three-year-old female at the 20th percentile is

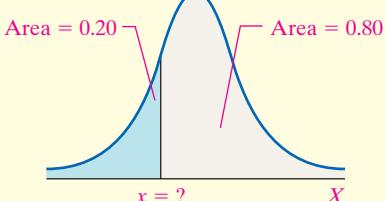
$$\begin{aligned}x &= \mu + z\sigma \\&= 38.72 + (-0.84)(3.17) \\&= 36.1 \text{ inches}\end{aligned}$$

calculator, Minitab, Excel, and StatCrunch are found in the Technology Step-by-Step on pages 350–351.

Technology Solution

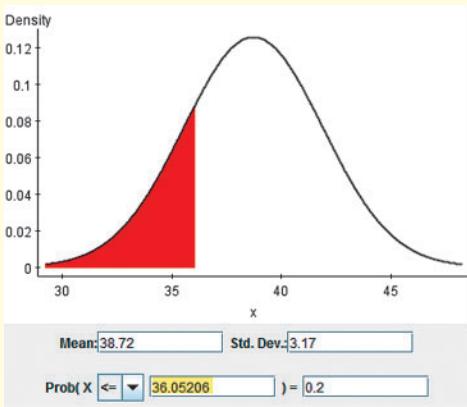
Step 1 Figure 23 shows the normal curve with the unknown value of x at the 20th percentile, which separates the bottom 20% of the distribution from the top 80%.

Figure 23



Step 2 Figure 24 shows the results obtained from StatCrunch. The height of a three-year-old female at the 20th percentile is 36.1 inches.

Figure 24

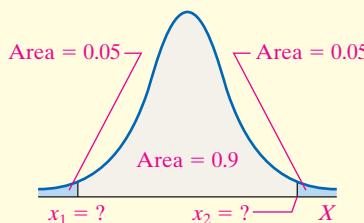
**NW Now Work Problem 47(a)****EXAMPLE 4 Finding the Value of a Normal Random Variable**

Problem The scores earned on the mathematics portion of the SAT, a college entrance exam, are approximately normally distributed with mean 516 and standard deviation 116. What scores separate the middle 90% of test takers from the bottom and top 5%? In other words, find the 5th and 95th percentiles. *Source:* The College Board.

By-Hand Solution

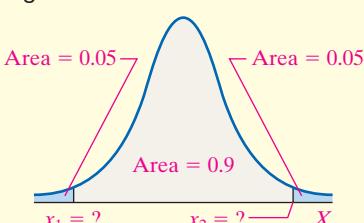
Step 1 Figure 25 shows the normal curve with the unknown values of x separating the bottom and top 5% of the distribution from the middle 90%.

Figure 25

**Technology Solution**

Step 1 Figure 27 shows the normal curve with the unknown values of x separating the bottom and top 5% of the distribution from the middle 90%.

Figure 27



Step 2 First, find the z -score that corresponds to an area of 0.05 to the left. In Table V, look in the body of the table and find that 0.0495 and 0.0505 are equally close to 0.05. See Figure 26. We agree to take the mean of the two z -scores corresponding to the areas. The z -score corresponding to an area of 0.0495 is -1.65 , and the z -score corresponding to an area of 0.0505 is -1.64 . The approximate z -score corresponding to an area of 0.05 to the left is $z_1 = -1.645$.

Figure 26

Standard Normal Distribution						
z	.00	.01	.02	.03	.04	.05
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0
-3.2	0.0006	0.0006	0.0006	0.0006	0.0006	0
-3.1	0.0008	0.0008	0.0008	0.0008	0.0008	0
-3.0	0.0011	0.0011	0.0011	0.0011	0.0011	0
-2.9	0.0015	0.0015	0.0015	0.0015	0.0015	0
-2.8	0.0021	0.0021	0.0021	0.0021	0.0021	0
-2.7	0.0030	0.0030	0.0030	0.0030	0.0030	0
-2.6	0.0042	0.0042	0.0042	0.0042	0.0042	0
-2.5	0.0057	0.0057	0.0057	0.0057	0.0057	0
-2.4	0.0075	0.0075	0.0075	0.0075	0.0075	0
-2.3	0.0096	0.0096	0.0096	0.0096	0.0096	0
-2.2	0.0121	0.0121	0.0121	0.0121	0.0121	0
-2.1	0.0151	0.0151	0.0151	0.0151	0.0151	0
-2.0	0.0185	0.0185	0.0185	0.0185	0.0185	0
-1.9	0.0223	0.0223	0.0223	0.0223	0.0223	0
-1.8	0.0263	0.0263	0.0263	0.0263	0.0263	0
-1.7	0.0303	0.0303	0.0303	0.0303	0.0303	0
-1.6	0.0343	0.0343	0.0343	0.0343	0.0343	0
-1.5	0.0382	0.0382	0.0382	0.0382	0.0382	0
-1.4	0.0420	0.0420	0.0420	0.0420	0.0420	0
-1.3	0.0457	0.0457	0.0457	0.0457	0.0457	0
-1.2	0.0493	0.0493	0.0493	0.0493	0.0493	0
-1.1	0.0528	0.0528	0.0528	0.0528	0.0528	0
-1.0	0.0561	0.0561	0.0561	0.0561	0.0561	0
-0.9	0.0591	0.0591	0.0591	0.0591	0.0591	0
-0.8	0.0618	0.0618	0.0618	0.0618	0.0618	0
-0.7	0.0643	0.0643	0.0643	0.0643	0.0643	0
-0.6	0.0668	0.0668	0.0668	0.0668	0.0668	0
-0.5	0.0688	0.0688	0.0688	0.0688	0.0688	0
-0.4	0.0702	0.0702	0.0702	0.0702	0.0702	0
-0.3	0.0718	0.0718	0.0718	0.0718	0.0718	0
-0.2	0.0730	0.0730	0.0730	0.0730	0.0730	0
-0.1	0.0739	0.0739	0.0739	0.0739	0.0739	0
0.0	0.0744	0.0744	0.0744	0.0744	0.0744	0
0.1	0.0744	0.0744	0.0744	0.0744	0.0744	0
0.2	0.0739	0.0739	0.0739	0.0739	0.0739	0
0.3	0.0730	0.0730	0.0730	0.0730	0.0730	0
0.4	0.0718	0.0718	0.0718	0.0718	0.0718	0
0.5	0.0702	0.0702	0.0702	0.0702	0.0702	0
0.6	0.0688	0.0688	0.0688	0.0688	0.0688	0
0.7	0.0668	0.0668	0.0668	0.0668	0.0668	0
0.8	0.0643	0.0643	0.0643	0.0643	0.0643	0
0.9	0.0618	0.0618	0.0618	0.0618	0.0618	0
1.0	0.0591	0.0591	0.0591	0.0591	0.0591	0
1.1	0.0561	0.0561	0.0561	0.0561	0.0561	0
1.2	0.0528	0.0528	0.0528	0.0528	0.0528	0
1.3	0.0493	0.0493	0.0493	0.0493	0.0493	0
1.4	0.0457	0.0457	0.0457	0.0457	0.0457	0
1.5	0.0420	0.0420	0.0420	0.0420	0.0420	0
1.6	0.0382	0.0382	0.0382	0.0382	0.0382	0
1.7	0.0343	0.0343	0.0343	0.0343	0.0343	0
1.8	0.0303	0.0303	0.0303	0.0303	0.0303	0
1.9	0.0263	0.0263	0.0263	0.0263	0.0263	0
2.0	0.0223	0.0223	0.0223	0.0223	0.0223	0
2.1	0.0185	0.0185	0.0185	0.0185	0.0185	0
2.2	0.0151	0.0151	0.0151	0.0151	0.0151	0
2.3	0.0121	0.0121	0.0121	0.0121	0.0121	0
2.4	0.0096	0.0096	0.0096	0.0096	0.0096	0
2.5	0.0075	0.0075	0.0075	0.0075	0.0075	0
2.6	0.0057	0.0057	0.0057	0.0057	0.0057	0
2.7	0.0042	0.0042	0.0042	0.0042	0.0042	0
2.8	0.0030	0.0030	0.0030	0.0030	0.0030	0
2.9	0.0021	0.0021	0.0021	0.0021	0.0021	0
3.0	0.0015	0.0015	0.0015	0.0015	0.0015	0
3.1	0.0011	0.0011	0.0011	0.0011	0.0011	0
3.2	0.0008	0.0008	0.0008	0.0008	0.0008	0
3.3	0.0006	0.0006	0.0006	0.0006	0.0006	0
3.4	0.0005	0.0005	0.0005	0.0005	0.0005	0
3.5	0.0003	0.0003	0.0003	0.0003	0.0003	0
3.6	0.0003	0.0003	0.0003	0.0003	0.0003	0
3.7	0.0002	0.0002	0.0002	0.0002	0.0002	0
3.8	0.0002	0.0002	0.0002	0.0002	0.0002	0
3.9	0.0001	0.0001	0.0001	0.0001	0.0001	0
4.0	0.0001	0.0001	0.0001	0.0001	0.0001	0

Now find the z -score corresponding to an area of 0.05 to the right, which means the area to the left is 0.95. From Table V, we find an area of 0.9495 and 0.9505, which correspond to 1.64 and 1.65. The approximate z -score, such that the area to the right is 0.05, is $z_2 = 1.645$.

Step 3 The SAT mathematics score that separates the bottom 5% from the top 95% of scores is

$$\begin{aligned}x_1 &= \mu + z_1\sigma \\&= 516 + (-1.645)(116) \\&= 325\end{aligned}$$

The SAT mathematics score that separates the bottom 95% from the top 5% of scores is

$$\begin{aligned}x_2 &= \mu + z_2\sigma \\&= 516 + (1.645)(116) \\&= 707\end{aligned}$$

Step 2 Figure 28 shows the results obtained from Excel with the values of x_1 and x_2 highlighted.

Figure 28

Function Arguments						
NORMINV						
Probability	0.05	= 0.05	<input checked="" type="text"/>	<input type="button" value="OK"/>	<input type="button" value="Cancel"/>	
Mean	516	= 516	<input checked="" type="text"/>	<input type="button" value="OK"/>	<input type="button" value="Cancel"/>	
Standard_dev	116	= 116	<input checked="" type="text"/>	<input type="button" value="OK"/>	<input type="button" value="Cancel"/>	
Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation.						
Standard_dev is the standard deviation of the distribution, a positive number.						
Formula result = 325.1969793						
Help on this function						

Function Arguments						
NORMINV						
Probability	0.95	= 0.95	<input checked="" type="text"/>	<input type="button" value="OK"/>	<input type="button" value="Cancel"/>	
Mean	516	= 516	<input checked="" type="text"/>	<input type="button" value="OK"/>	<input type="button" value="Cancel"/>	
Standard_dev	116	= 116	<input checked="" type="text"/>	<input type="button" value="OK"/>	<input type="button" value="Cancel"/>	
Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation.						
Probability is a probability corresponding to the normal distribution, a number between 0 and 1 inclusive.						
Formula result = 706.8030207						
Help on this function						

Interpretation SAT mathematics scores that separate the middle 90% of the scores from the bottom and top 5% are 325 and 707. Put another way, a student who scores 325 on the SAT math exam is at the 5th percentile. A student who scores 707 on the SAT math exam is at the 95th percentile. We might use these results to identify those scores that are unusual.

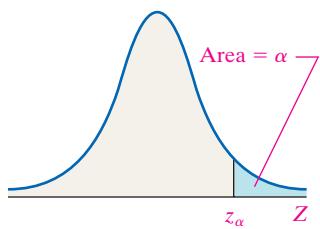
NW Now Work Problem 47(b)

We could also obtain the by-hand solution to Example 4 using symmetry. Because the normal curve is symmetric about its mean, the z -score that corresponds to an area of 0.05 to the left will be the additive inverse (i.e., the opposite) of the z -score that corresponds to an area of 0.05 to the right. Since the area to the left of $z = -1.645$ is 0.05, the area to the right of $z = 1.645$ is 0.05.

Important Notation for the Future

In upcoming chapters, we will need to find the z -score that has a specified area to the right. We have special notation to represent this situation.

Figure 29



The notation z_α (pronounced “ z sub alpha”) is the z -score such that the area under the standard normal curve to the right of z_α is α . Figure 29 illustrates the notation.

EXAMPLE 5**Finding the Value of z_α**

Problem Find the value of $z_{0.10}$.

Approach We wish to find the z -value such that the area under the standard normal curve to the right of the z -value is 0.10.

By-Hand Solution The area to the right of the unknown z -value is 0.10, so the area to the left of the z -value is $1 - 0.10 = 0.90$. We look in Table V for the area closest to 0.90. The closest area is 0.8997, which corresponds to a z -value of 1.28. Therefore, $z_{0.10} = 1.28$.

Technology Solution The area to the right of the unknown z -value is 0.10, so the area to the left is $1 - 0.10 = 0.90$. A TI-84 Plus CE is used to find that the z -value such that the area to the left of z is 0.90. See Figure 30. Therefore, $z_{0.10} = 1.28$.

Figure 30

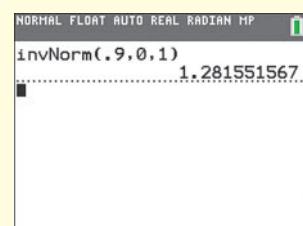
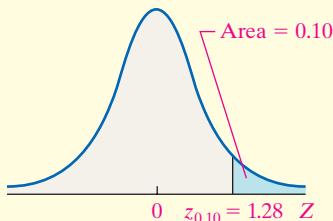


Figure 31 shows the z -value on the normal curve.

Figure 31

**NW Now Work Problem 19**

For any continuous random variable, the probability of observing a specific value of the random variable is 0. For example, for a normal random variable, $P(a) = 0$ for any value of a , because there is no area under the normal curve associated with a single value. Therefore, the following probabilities are equivalent:

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$$

Technology Step-by-Step**The Normal Distribution****TI-83/84 Plus****Finding Area under the Normal Curve**

- From the HOME screen, press 2^{nd} VARS to access the DISTRibution menu.
- Select 2:normalcdf(.
- Enter the lowerbound, upperbound, μ , and σ . Highlight Paste and hit ENTER. Hit ENTER again with the formula on the HOME screen.

Note: When there is no lowerbound, enter $-1E99$. When there is no upperbound, enter $1E99$. The E shown is scientific notation; it is selected by pressing 2^{nd} then ,.

Finding Normal Values Corresponding to an Area

- From the HOME screen, press 2^{nd} VARS to access the DISTRibution menu.
- Select 3:invNorm(.
- Enter the *area left*, μ , and σ . Highlight Paste and hit ENTER. Hit ENTER again with the formula on the HOME screen.

Minitab

Finding Area under the Normal Curve

1. Select the **Calc** menu, highlight **Probability Distributions**, and highlight **Normal . . .**
 2. Select **Cumulative Probability**. Enter the mean, μ , and the standard deviation, σ . Select **Input Constant**, and enter the observation. Click **OK**.

Finding Normal Values Corresponding to an Area

1. Select the **Calc** menu, highlight **Probability Distributions**, and highlight **Normal . . .**.
 2. Select **Inverse Cumulative Probability**. Enter the mean, μ , and the standard deviation, σ . Select **Input Constant**, and enter the area to the left of the unknown normal value. Click **OK**.

Excel

Finding Area under the Normal Curve

1. Select the Formulas tab. Click “More Functions”, then highlight “Statistical”. Select “NORM.DIST”.
 2. Enter the specified observation, μ , and σ , and set **Cumulative** to True. Click OK.

Finding Normal Values Corresponding to an Area

1. Select the Formulas tab. Click “More Functions”, then highlight “Statistical”. Select “NORM.INV”.
 2. Enter the area left of the unknown normal value, μ , and σ . Click OK.

StatCrunch

Finding Area under the Standard Normal Curve

1. Select **Stat**, highlight **Calculators**, select **Normal**.
 2. Enter the mean and the standard deviation. If you want to compute $P(X \leq x)$ or $P(X \geq x)$, select the Standard tab. In the pull-down menu, select $P(X \leq x)$ or $P(X \geq x)$. Finally, enter the value of x . Click Compute. If you want to compute $P(a \leq X \leq b)$, select the Between tab. Enter the values of a and b . Click Compute.

Finding Scores Corresponding to an Area

1. Select **Stat**, highlight **Calculators**, select **Normal**.
 2. Enter the mean and the standard deviation. Select the Standard tab. If you are given the area to the left of the unknown value, in the pull-down menu choose the \leq option; if given the area to the right, choose the \geq option. Finally, enter the area in the right-most cell. Click Compute.



7.2 Assess Your Understanding

Vocabulary and Skill Building

- 1.** A random variable Z that is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$ is said to have the _____.

2. The notation z_α is the z -score such that the area under the standard normal curve to the right of z_α is ____.

3. If X is a normal random variable with mean 40 and standard deviation 10 and $P(X > 45) = 0.3085$, then $P(X < 35) =$ _____.

4. If X is normal random variable with mean 40 and standard deviation 10 and $P(X < 38) = 0.4207$, then $P(X \leq 38) =$ _____.

In Problems 5–12, find the indicated areas. For each problem, be sure to draw a standard normal curve and shade the area that is to be found.

5. Determine the area under the standard normal curve that lies to the left of

(a) $z = -2.45$ (b) $z = -0.43$
(c) $z = 1.35$ (d) $z = 3.49$

6. Determine the area under the standard normal curve that lies to the left of

(a) $z = -3.49$ (b) $z = -1.99$
(c) $z = 0.92$ (d) $z = 2.90$

7. Determine the area under the standard normal curve that lies to the right of

(a) $z = -3.01$ (b) $z = -1.59$
(c) $z = 1.78$ (d) $z = 3.11$

In Problems 13–18, find the indicated z-score. Be sure to draw a standard normal curve that depicts the solution.

- 13.** Find the z -score such that the area under the standard normal curve to its left is 0.1.

14. Find the z -score such that the area under the standard normal curve to its left is 0.2.

- 15.** Find the z -score such that the area under the standard normal curve to its right is 0.25.
- 16.** Find the z -score such that the area under the standard normal curve to its right is 0.35.
- 17.** Find z -scores that separate the middle 99% of the distribution from the area in the tails of the standard normal distribution.
- 18.** Find the z -scores that separate the middle 94% of the distribution from the area in the tails of the standard normal distribution.

In Problems 19–22, find the value of z_α

- NW 19.** $z_{0.01}$ **20.** $z_{0.02}$
21. $z_{0.025}$ **22.** $z_{0.15}$

In Problems 23–32, assume that the random variable X is normally distributed, with mean $\mu = 50$ and standard deviation $\sigma = 7$. Compute the following probabilities. Be sure to draw a normal curve with the area corresponding to the probability shaded.

- 23.** $P(X > 35)$
24. $P(X > 65)$
25. $P(X \leq 45)$
26. $P(X \leq 58)$
27. $P(40 < X < 65)$
28. $P(56 < X < 68)$
29. $P(55 \leq X \leq 70)$
30. $P(40 \leq X \leq 49)$
31. $P(38 < X \leq 55)$
32. $P(56 \leq X < 66)$

In Problems 33–36, assume that the random variable X is normally distributed, with mean $\mu = 50$ and standard deviation $\sigma = 7$. Find each indicated percentile for X .

- 33.** The 9th percentile
34. The 90th percentile
35. The 81st percentile
36. The 38th percentile

Applying the Concepts

37. Egg Incubation Times The mean incubation time of fertilized chicken eggs kept at 100.5°F in a still-air incubator is 21 days. Suppose that the incubation times are approximately normally distributed with a standard deviation of 1 day.

Source: University of Illinois Extension

- (a) Draw a normal model that describes egg incubation times of fertilized chicken eggs.
(b) Find and interpret the probability that a randomly selected fertilized chicken egg hatches in less than 20 days.
(c) Find and interpret the probability that a randomly selected fertilized chicken egg takes over 22 days to hatch.
(d) Find and interpret the probability that a randomly selected fertilized chicken egg hatches between 19 and 21 days.
(e) Would it be unusual for an egg to hatch in less than 18 days? Why?

38. Reading Rates The reading speed of sixth-grade students is approximately normal, with a mean speed of 125 words per minute and a standard deviation of 24 words per minute.

- (a) Draw a normal model that describes the reading speed of sixth-grade students.
(b) Find and interpret the probability that a randomly selected sixth-grade student reads less than 100 words per minute.
(c) Find and interpret the probability that a randomly selected sixth-grade student reads more than 140 words per minute.
(d) Find and interpret the probability that a randomly selected sixth-grade student reads between 110 and 130 words per minute.
(e) Would it be unusual for a sixth grader to read more than 200 words per minute? Why?

NW 39. Chips Ahoy! Cookies The number of chocolate chips in an 18-ounce bag of Chips Ahoy! chocolate chip cookies is approximately normally distributed with a mean of 1262 chips and standard deviation 118 chips according to a study by cadets of the U.S. Air Force Academy.

Source: Brad Warner and Jim Rutledge, *Chance* 12(1): 10–14, 1999.

- (a) What is the probability that a randomly selected 18-ounce bag of Chips Ahoy! contains between 1000 and 1400 chocolate chips, inclusive?
(b) What is the probability that a randomly selected 18-ounce bag of Chips Ahoy! contains fewer than 1000 chocolate chips?
(c) What proportion of 18-ounce bags of Chips Ahoy! contains more than 1200 chocolate chips?
(d) What proportion of 18-ounce bags of Chips Ahoy! contains fewer than 1125 chocolate chips?
(e) What is the percentile rank of an 18-ounce bag of Chips Ahoy! that contains 1475 chocolate chips?
(f) What is the percentile rank of an 18-ounce bag of Chips Ahoy! that contains 1050 chocolate chips?

40. Growth Plates The ability to determine the age of some individuals can be difficult if there are not quality government records of birth. Bone growth takes place at the growth plates at the end of long bones. Once all growth plates fuse, growth stops, and an individual is considered a biological adult. The age at which growth plates fuse for males is approximately normally distributed with a mean of 19 years and a standard deviation of 15.4 months. *Source:* Tim Cole, “People Smugglers, Statistics, and Bone Age,” *Significance Magazine*, June 2012: 9(3).

- (a) What is the probability a male’s growth plates fuse after age 20?
(b) What is the probability a male’s growth plates fuse before age 17?
(c) What proportion of a male’s growth plates fuse between 15 and 17 years of age?
(d) Would it be unusual for a male’s growth plates to fuse when he is 23 years old? Explain.

41. Gestation Period The lengths of human pregnancies are approximately normally distributed, with mean $\mu = 266$ days and standard deviation $\sigma = 16$ days.

- (a) What proportion of pregnancies lasts more than 270 days?
(b) What proportion of pregnancies lasts less than 250 days?
(c) What proportion of pregnancies lasts between 240 and 280 days?

- (d) What is the probability that a randomly selected pregnancy lasts more than 280 days?
- (e) What is the probability that a randomly selected pregnancy lasts no more than 245 days?
- (f) A “very preterm” baby is one whose gestation period is less than 224 days. Are very preterm babies unusual?

42. Blackjack Using a technique referred to as the “Hi-Opt I System” in the casino game blackjack, a player can expect to earn \$1.65 per shoe of cards with a standard deviation of \$48.93 with a \$1 bet per hand. The earnings are approximately normally distributed. This assumes the shoe of cards contains 8 decks and the dealer reshuffles the shoe of cards when 80% of the cards have been dealt from the shoe.

Source: W. J. Hurley and Andrey Pavlov, “There Will Be Blood: On the Risk-Return Characteristics of a Blackjack Counting System,” *Chance*, 24:(2), Spring 2011.

- (a) What is the probability of earning at least \$10 from the shoe?
 - (b) What is the probability of losing more than \$10 from the shoe?
 - (c) What is the probability of being “up” after one shoe?
- 43. Manufacturing** Steel rods are manufactured with a mean length of 25 centimeters (cm). Because of variability in the manufacturing process, the lengths of the rods are approximately normally distributed, with a standard deviation of 0.07 cm.
- (a) What proportion of rods has a length less than 24.9 cm?
 - (b) Any rods that are shorter than 24.85 cm or longer than 25.15 cm are discarded. What proportion of rods will be discarded?
 - (c) Using the results of part (b), if 5000 rods are manufactured in a day, how many should the plant manager expect to discard?
 - (d) If an order comes in for 10,000 steel rods, how many rods should the plant manager manufacture if the order states that all rods must be between 24.9 cm and 25.1 cm?

- 44. Manufacturing** Ball bearings are manufactured with a mean diameter of 5 millimeters (mm). Because of variability in the manufacturing process, the diameters of the ball bearings are approximately normally distributed, with a standard deviation of 0.02 mm.
- (a) What proportion of ball bearings has a diameter more than 5.03 mm?
 - (b) Any ball bearings that have a diameter less than 4.95 mm or greater than 5.05 mm are discarded. What proportion of ball bearings will be discarded?
 - (c) Using the results of part (b), if 30,000 ball bearings are manufactured in a day, how many should the plant manager expect to discard?
 - (d) If an order comes in for 50,000 ball bearings, how many bearings should the plant manager manufacture if the order states that all ball bearings must be between 4.97 mm and 5.03 mm?

45. NCAA Basketball Point Spreads In sports betting, Las Vegas sports books establish winning margins for a team that is favored to win a game. An individual can place a wager on the game and will win if the team bet upon wins after accounting for the spread. For example, if Team A is favored by 5 points, and wins the game by 7 points, then a bet on Team A is a winning

bet. However, if Team A wins the game by only 3 points, then a bet on Team A is a losing bet. In games where a team is favored by 12 or fewer points, the margin of victory for the favored team relative to the spread is approximately normally distributed with a mean of 0 points and a standard deviation of 10.9 points.

Source: Justin Wolfers, “Point Shaving: Corruption in NCAA Basketball.”

- (a) Explain the meaning of “the margin of victory relative to the spread has a mean of 0 points.” Does this imply that the spreads are accurate for games in which a team is favored by 12 or fewer points?
- (b) In games where a team is favored by 12 or fewer points, what is the probability that the favored team wins by 5 or more points relative to the spread?
- (c) In games where a team is favored by 12 or fewer points, what is the probability that the favored team loses by 2 or more points relative to the spread?

46. NCAA Basketball Point Spreads Revisited See Problem 45. In games where a team is favored by more than 12 points, the margin of victory for the favored team relative to the spread is normally distributed with a mean of -1.0 point and a standard deviation of 10.9 points.

Source: Justin Wolfers, “Point Shaving: Corruption in NCAA Basketball”

- (a) In games where a team is favored by more than 12 points, what is the probability that the favored team wins by 5 or more points relative to the spread?
- (b) In games where a team is favored by more than 12 points, what is the probability that the favored team loses by 2 or more points relative to the spread?
- (c) In games where a team is favored by more than 12 points, what is the probability that the favored team “beats the spread”? Does this imply that the possible point shaving spreads are accurate for games in which a team is favored by more than 12 points?

NW 47. Egg Incubation Times The mean incubation time of fertilized chicken eggs kept at 100.5°F in a still-air incubator is 21 days. Suppose that the incubation times are approximately normally distributed with a standard deviation of 1 day.

Source: University of Illinois Extension.

- (a) Determine the 17th percentile for incubation times of fertilized chicken eggs.
- (b) Determine the incubation times that make up the middle 95% of fertilized chicken eggs.

48. Reading Rates The reading speed of sixth-grade students is approximately normal, with a mean speed of 125 words per minute and a standard deviation of 24 words per minute.

- (a) What is the reading speed of a sixth-grader whose reading speed is at the 90th percentile?
- (b) A school psychologist wants to determine reading rates for unusual students (both slow and fast). Determine the reading rates of the middle 95% of all sixth-grade students. What are the cutoff points for unusual readers?

49. Chips Ahoy! Cookies The number of chocolate chips in an 18-ounce bag of Chips Ahoy! chocolate chip cookies is approximately normally distributed, with a mean of 1262 chips and a standard deviation of 118 chips, according to a study by cadets of the U.S. Air Force Academy.

Source: Brad Warner and Jim Rutledge, *Chance* 12(1): 10–14, 1999.

- (a) Determine the 30th percentile for the number of chocolate chips in an 18-ounce bag of Chips Ahoy! cookies.
 (b) Determine the number of chocolate chips in a bag of Chips Ahoy! that make up the middle 99% of bags.
 (c) What is the interquartile range of the number of chips in Chips Ahoy! cookies?

50. Wendy's Drive-Thru Fast-food restaurants spend quite a bit of time studying the amount of time cars spend in their drive-thru. Certainly, the faster the cars get through the drive-thru, the more the opportunity for making money. *QSR Magazine* studied drive-thru times for fast-food restaurants, and found Wendy's had the best time, with a mean time a car spent in the drive-thru equal to 138.5 seconds. Assume that drive-thru times are normally distributed, with a standard deviation of 29 seconds. Suppose that Wendy's wants to institute a policy at its restaurants that it will not charge any patron that must wait more than a certain amount of time for an order. Management does not want to give away free meals to more than 1% of the patrons. What time would you recommend Wendy's advertise as the maximum wait time before a free meal is awarded?

51. Speedy Lube The time required for Speedy Lube to complete an oil change service on an automobile approximately follows a normal distribution, with a mean of 17 minutes and a standard deviation of 2.5 minutes.

- (a) Speedy Lube guarantees customers that the service will take no longer than 20 minutes. If it does take longer, the customer will receive the service for half-price. What percent of customers receive the service for half price?
 (b) If Speedy Lube does not want to give the discount to more than 3% of its customers, how long should it make the guaranteed time limit?

DATA 52. Putting It Together: Birth Weights The following data represent the distribution of birth weights (in grams) for babies in which the pregnancy went full term (37–41 weeks).

Birth Weight (g)	Number of Live Births
0–499	22
500–999	201
1000–1499	1,645
1500–1999	9,365
2000–2499	92,191
2500–2999	569,319
3000–3499	1,387,335
3500–3999	988,011
4000–4499	255,700
4500–4999	36,766
5000–5499	3,994

Source: National Vital Statistics Report.

- (a) Construct a relative frequency distribution for birth weight.
 (b) Draw a relative frequency histogram for birth weight. Describe the shape of the distribution.
 (c) Determine the mean and standard deviation birth weight.
 (d) Use the normal model to determine the proportion of babies in each class.
 (e) Compare the proportions predicted by the normal model to the relative frequencies found in part (a). Do you believe that the normal model is effective in describing the birth weights of babies?

DATA 53. Putting It Together: Home Run Distance The file 7_2_53 represents distance (in feet) of all home runs hit during the 2018 Major League baseball season. It is available at www.pearsonhighered.com/sullivanstats.

- (a) Draw a relative frequency histogram of the home run distances using a lower class limit of the first class of 310 and a class width of 20. Comment on the shape of the distribution.
 (b) Determine the population mean and standard deviation distance of a home run hit during the 2018 Major League baseball season.
 (c) Draw the normal curve with the mean and standard deviation found in part (b) on the relative frequency histogram from part (a). Do you think a normal model describes the variable "distance"? Explain.
 (d) Use the normal model to find the area under the normal curve between 350 and 370 feet. Compare this result to the relative frequency with which a home run distance between 350 and 369.9 feet is observed.
 (e) Use the normal model to find the area under the normal curve to the right of 410 feet. Compare this result to the relative frequency with which a home run distance exceeds 410 feet.

Explaining the Concepts

54. Give three interpretations for the area under a normal curve.

55. Explain why $P(X < 30)$ should be reported as <0.0001 if X is a normal random variable with mean 100 and standard deviation 15.

56. Explain why $P(X \leq 220)$ should be reported as >0.9999 if X is a normal random variable with mean 100 and standard deviation 15.

57. The ACT and SAT are two college entrance exams. The composite score on the ACT is approximately normally distributed with mean 21.1 and standard deviation 5.1. The composite score on the SAT is approximately normally distributed with mean 1026 and standard deviation 210. Suppose you scored 26 on the ACT and 1240 on the SAT. Which exam did you score better on? Justify your reasoning using the normal model.

7.3 Assessing Normality



Preparing for This Section Before getting started, review the following:

- Shape of a distribution (Section 2.2, pp. 82–83)
- Correlation coefficient (Section 4.1, pp. 173–177)

Objective ① Use normal probability plots to assess normality

Up to this point, we have said that a random variable X is normally distributed, or at least approximately normal, provided the histogram of the data is symmetric and bell shaped. This works well for large data sets, but the shape of a histogram drawn from a small sample of observations does not always accurately represent the shape of the population. For this reason, we need additional methods for assessing the normality of a random variable X when we are looking at a small set of sample data.

① Use Normal Probability Plots to Assess Normality

IN OTHER WORDS

Normal probability plots are used to assess the normality of a data set.

A **normal probability plot** is a graph that plots observed data versus *normal scores*. A **normal score** is the expected z -score of the data value, assuming that the distribution of the random variable is normal. The expected z -score of an observed value depends on the number of observations in the data set.

Drawing a normal probability plot requires the following steps:

Drawing a Normal Probability Plot

Step 1 Arrange the data in ascending order.

Step 2 Compute $f_i = \frac{i - 0.375}{n + 0.25}$,* where i is the index (the position of the data

value in the ordered list) and n is the number of observations. The expected proportion of observations less than or equal to the i th data value is f_i .

Step 3 Find the z -score corresponding to f_i from Table V.

Step 4 Plot the observed values on the horizontal axis and the corresponding z -scores on the vertical axis.

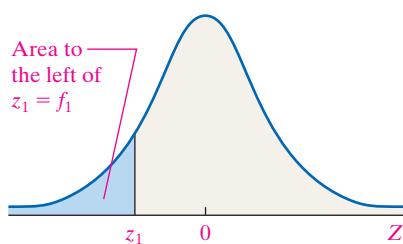
The idea behind finding the expected z -score is that, if the data come from a normally distributed population, we could predict the area to the left of each data value. The value of f_i represents the expected area to the left of the i th observation when the data come from a population that is normally distributed. For example, f_1 is the expected area to the left of the smallest data value, f_2 is the expected area to the left of the second-smallest data value, and so on. See Figure 32.

Once we determine each f_i , we find the z -scores corresponding to f_1, f_2 , and so on. The smallest observation in the data set will be the smallest expected z -score, and the largest observation in the data set will be the largest expected z -score. Also, because of the symmetry of the normal curve, the expected z -scores are always paired as positive and negative values.

Values of normal random variables and their z -scores are linearly related ($x = \mu + z\sigma$), so a plot of observations of normal variables against their expected z -scores will be linear. We conclude the following:

If sample data are taken from a population that is normally distributed, a normal probability plot of the observed values versus the expected z -scores will be approximately linear.

Figure 32



*The derivation of this formula is beyond the scope of this text.

It is difficult to determine whether a normal probability plot is “linear enough.” However, we can use a procedure based on the research of S. W. Looney and T. R. Gulledge in their paper “Use of the Correlation Coefficient with Normal Probability Plots,” published in the *American Statistician*. Basically, if the linear correlation coefficient between the observed values and expected z -scores is greater than the critical value found in Table VI in Appendix A, then it is reasonable to conclude that the data could come from a population that is normally distributed.

Normal probability plots are typically drawn using graphing calculators or statistical software. However, it is worthwhile to go through an example that constructs the plot by hand to better understand the results supplied by technology.

EXAMPLE 1

Constructing a Normal Probability Plot

Table 4

31.35	32.52
32.06	31.26
31.91	32.37

Source: Greyhound Park, Dubuque, IA.

Problem The data in Table 4 represent the finishing time (in seconds) for six randomly selected races of a greyhound named Barbies Bomber in the $\frac{5}{16}$ -mile race at Greyhound Park in Dubuque, Iowa. Is there evidence to support the belief that the variable “finishing time” is normally distributed?

Approach Follow Steps 1 through 4.

Solution

Step 1 Column 1 in Table 5 represents the index i . Column 2 represents the observed values in the data set, written in ascending order.

Table 5

Index, i	Observed Value	f_i	Expected z -score
1	31.26	$\frac{1 - 0.375}{6 + 0.25} = 0.10$	-1.28
2	31.35	$\frac{2 - 0.375}{6 + 0.25} = 0.26$	-0.64
3	31.91	0.42	-0.20
4	32.06	0.58	0.20
5	32.37	0.74	0.64
6	32.52	0.90	1.28

Step 2 Column 3 in Table 5 represents $f_i = \frac{i - 0.375}{n + 0.25}$ for each observation. This value is the expected area under the normal curve to the left of the i th observation, assuming the data come from a population that is normally distributed. For example, $i = 1$ corresponds to the finishing time of 31.26, and

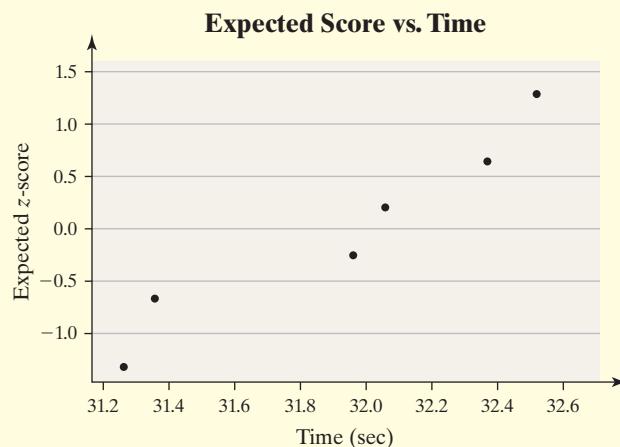
$$f_1 = \frac{1 - 0.375}{6 + 0.25} = 0.10$$

So the area under the normal curve to the left of 31.26 is 0.10 if the sample data come from a population that is normally distributed.

Step 3 Use Table V to find the z -scores that correspond to each f_i , then list them in Column 4 of Table 5. Look in Table V for the area closest to $f_1 = 0.1$. The expected z -score is -1.28. Notice that for each negative expected z -score there is a corresponding positive expected z -score, as a result of the symmetry of the normal curve.

Step 4 Plot the actual observations on the horizontal axis and the expected z -scores on the vertical axis. See Figure 33 on the next page.

Figure 33



Interpretation The linear correlation between the observed values and expected z -scores from the data in Table 5 is 0.970.

The critical value in Table VI for $n = 6$ observations is 0.888. Because the correlation coefficient is greater than the critical value ($0.970 > 0.888$), it is reasonable to conclude that the finishing times of Barbies Bomber in the 5/16-mile race are approximately normally distributed. 

Typically, normal probability plots are drawn using either a graphing calculator with advanced statistical features or statistical software.

EXAMPLE 2

Assessing Normality Using Technology

Problem Draw a normal probability plot of the data in Table 4 using technology. Is there evidence to support the belief that the variable “finishing time” is normally distributed?

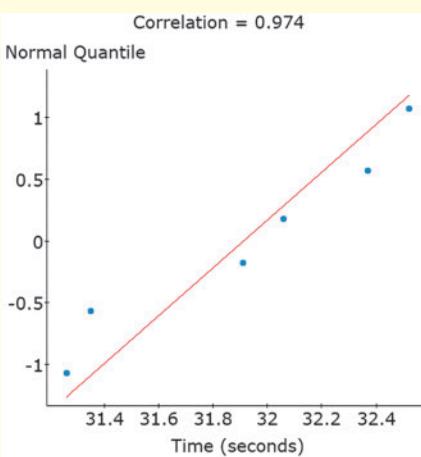
Approach We will use StatCrunch to draw the normal probability plot and find the correlation between the observed values and expected z -scores. If the correlation is greater than the critical value from Table VI, we conclude the data could come from a population that is normally distributed. The steps for constructing normal probability plots using TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch can be found on page 358.

Solution Figure 34 shows the normal probability plot. The correlation between the observed values and expected z -scores is 0.974.

NOTE

Minitab will draw “confidence bands” in its normal probability plots. If all the data lie within these bands, then it is reasonable to conclude the data come from a population that is normally distributed.

Figure 34



(continued)

NOTE

The correlations in Examples 1 and 2 differ due to rounding.

NW Now Work Problem 3

The critical value in Table VI for $n = 6$ observations is 0.888. Because the correlation coefficient is greater than the critical value ($0.974 > 0.888$), it is reasonable to conclude that the finishing times of Barbies Bomber in the 5/16-mile race are approximately normally distributed.

Technology Step-by-Step**Normal Probability Plots****TI-83/84 Plus**

- Enter the raw data into L1.
- Press 2^{nd} Y= to access STAT PLOTS.
- Select 1:Plot1.
- Turn Plot1 on by highlighting On and pressing ENTER. Press the down-arrow key. Highlight the *normal probability plot* icon. Press ENTER to select this plot type. The Data List should be set at L1. The Data Axis should be the x -axis.
- Press ZOOM, and select 9:ZoomStat. Once you have the graph, TRACE to find the values of the observations and the corresponding normal scores. Enter these observations into L1 and L2. Find the correlation coefficient for this data.

Minitab

- Enter the raw data into C1.
- Select the **Graph** menu. Choose **Probability Plot . . .**. Select “Single.” Click OK.

- In the Graph variables cell, enter the column that contains the raw data. Make sure Distribution is set to Normal. Click OK.

Excel

- Install XLSTAT.
- Enter the raw data into column A.
- Select the **XLSTAT** menu. Highlight **Visualizing Data** and select **Univariate Plots**.
- In the general tab, check Quantitative data. With the cursor in the Quantitative data box, highlight the raw data. Uncheck the box “Sample Labels.” Click the charts tab and check Normal P-P plots or Normal Q-Q plots. Click OK.

StatCrunch

- If necessary, enter the raw data into column var1. Name the column.
- Select **Graph** and highlight **QQ Plot**.
- Select the variable. Check the box to add the correlation statistic. Check the “Normal quantiles on y-axis” box. Click Compute!.

**7.3 Assess Your Understanding****Vocabulary and Skill Building**

- A _____ is a graph that plots observed data versus normal scores.
- True or False:** A normal score is the expected z -score of a data value, assuming the distribution of the random variable is normal.

In Problems 3–6, use the results in the table to (a) draw a normal probability plot, (b) determine the linear correlation between the observed values and expected z -scores, (c) determine the critical value in Table VI to assess the normality of the data.

NW 3.

Index, i	Observed Value	f_i	Expected z -score
1	39	0.08	-1.41
2	45	0.20	-0.84
3	48	0.32	-0.47
4	52	0.44	-0.15
5	54	0.56	0.15
6	56	0.68	0.47
7	60	0.80	0.84
8	62	0.92	1.41

4.

Index, i	Observed Value	f_i	Expected z -score
1	77	0.07	-1.48
2	80	0.18	-0.92
3	84	0.28	-0.58
4	91	0.39	-0.28
5	98	0.5	0
6	104	0.61	0.28
7	109	0.72	0.58
8	112	0.82	0.92
9	120	0.93	1.48

5.

Index, i	Observed Value	f_i	Expected z -score
1	1	0.09	-1.34
2	3	0.22	-0.77
3	6	0.36	-0.36
4	8	0.50	0
5	10	0.64	0.36
6	13	0.78	0.77
7	35	0.91	1.34

6.

Index, i	Observed Value	f_i	Expected z-score
1	1	0.10	-1.28
2	3	0.26	-0.64
3	19	0.42	-0.20
4	30	0.58	0.20
5	88	0.74	0.64
6	99	0.90	1.28

In Problems 7–10, use a normal probability plot to assess whether the sample data could have come from a population that is normally distributed.



- 7. O-Ring Thickness** A random sample of O-rings was obtained, and the wall thickness (in inches) of each was recorded.

0.276	0.274	0.275	0.274	0.277
0.273	0.276	0.276	0.279	0.274
0.273	0.277	0.275	0.277	0.277
0.276	0.277	0.278	0.275	0.276



- 8. Customer Service** A random sample of weekly work logs at an automobile repair station was obtained, and the average number of customers per day was recorded.

26	24	22	25	23
24	25	23	25	22
21	26	24	23	24
25	24	25	24	25
26	21	22	24	24



- 9. School Loans** A random sample of 20 undergraduate students receiving student loans was obtained, and the amount of their loans for the 2018–2019 school year was recorded.

2,500	1,000	2,000	14,000	1,800
3,800	10,100	2,200	29,000	16,000
5,000	2,200	6,200	9,100	2,800
2,500	1,400	13,200	750	12,000



- 10. Memphis Snowfall** A random sample of 25 years between 1890 and 2018 was obtained, and the amount of snowfall, in inches, for Memphis was recorded.

24.0	7.9	1.5	0.0	0.3
0.4	8.1	4.3	0.0	0.5
3.6	2.9	0.4	2.6	0.1
16.6	1.4	23.8	25.1	1.6
12.2	14.8	0.4	3.7	4.2

Source: National Oceanic and Atmospheric Administration.

Applying the Concepts



- 11. Chips per Bag** In a 1998 advertising campaign, Nabisco claimed that every 18-ounce bag of Chips Ahoy! cookies contained at least 1000 chocolate chips. Brad Warner and Jim Rutledge tried to verify the claim. The following data represent the number of chips in an 18-ounce bag of Chips Ahoy! based on their study.

1087	1098	1103	1121	1132
1185	1191	1199	1200	1213
1239	1244	1247	1258	1269
1307	1325	1345	1356	1363
1135	1137	1143	1154	1166
1214	1215	1219	1219	1228
1270	1279	1293	1294	1295
1377	1402	1419	1440	1514

Source: *Chance* 12(1): 10–14, 1999.

- Draw a normal probability plot to determine if the data could have come from a normal distribution.
- Determine the mean and standard deviation of the sample data.
- Using the sample mean and sample standard deviation obtained in part (b) as estimates for the population mean and population standard deviation, respectively, draw a graph of a normal model for the distribution of chips in a bag of Chips Ahoy!
- Using the normal model from part (c), find the probability that an 18-ounce bag of Chips Ahoy! selected at random contains at least 1000 chips.
- Using the normal model from part (c), determine the proportion of 18-ounce bags of Chips Ahoy! that contains between 1200 and 1400 chips, inclusive.



- 12. Hours of TV** A random sample of college students aged 18–24 years was obtained, and the number of hours of television watched in a typical week was recorded.

36.1	30.5	2.9	17.5	21.0
23.5	25.6	16.0	28.9	29.6
7.8	20.4	33.8	36.8	0.0
9.9	25.8	19.5	19.1	18.5
22.9	9.7	39.2	19.0	8.6

- Draw a normal probability plot to determine if the data could have come from a normal distribution.
- Determine the mean and standard deviation of the sample data.
- Using the sample mean and sample standard deviation obtained in part (b) as estimates for the population mean and population standard deviation, respectively, draw a graph of a normal model for the distribution of weekly hours of television watched.
- Using the normal model from part (c), find the probability that a college student aged 18–24 years, selected at random, watches between 20 and 35 hours of television each week.
- Using the normal model from part (c), determine the proportion of college students aged 18–24 years who watch more than 40 hours of television per week.



- 13. Putting It Together: Disney's Dinosaur Ride** Retrieve the data 7_3_13 at www.pearsonhighered.com/sullivanstats using the file format of your choice for the text you are using. The data represent the time spent waiting in line (in minutes) for the Dinosaur Ride at Walt Disney World for 100 randomly selected riders.

- Draw a relative frequency histogram with lower class limit of the first class equal to 0 and class width 20.
- Comment on the shape of the distribution.
- Draw a normal probability plot to assess the normality of the random variable “wait time.”

7.4 The Normal Approximation to the Binomial Probability Distribution



Preparing for This Section Before getting started, review the following:

- Binomial probability distribution (Section 6.2, pp. 312–323)

Objective ① Approximate binomial probabilities using the normal distribution

1 Approximate Binomial Probabilities Using the Normal Distribution

In Section 6.2, we discussed the binomial probability distribution. Now, we will review the criteria for a probability experiment to be a binomial experiment.

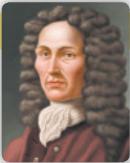
Criteria for a Binomial Probability Experiment

A probability experiment is a binomial experiment if all the following are true:

1. The experiment is performed n independent times. Each repetition of the experiment is called a **trial**. Independence means that the outcome of one trial will not affect the outcome of the other trials.
2. For each trial, there are two mutually exclusive outcomes—success or failure.
3. The probability of success, p , is the same for each trial of the experiment.

Historical Note

The normal approximation to the binomial was discovered by Abraham de Moivre in 1733. With the advance of computing technology, its importance has been diminished.



The binomial probability formula can be used to compute probabilities of events in a binomial experiment. A large number of trials of a binomial experiment, however, makes this formula difficult to use. For example, given 500 trials of a binomial experiment, to compute the probability of 400 or more successes requires that we compute the following probabilities:

$$P(X \geq 400) = P(400) + P(401) + \cdots + P(500)$$

This would be time consuming to compute by hand! Fortunately, we have an alternative means for approximating binomial probabilities, provided that certain conditions are met.

Recall the following fact from page 322:

For a fixed p , as the number of trials n in a binomial experiment increases, the probability distribution of the random variable X becomes more nearly symmetric and bell shaped. As a rule of thumb, if $np(1 - p) \geq 10$, the probability distribution will be approximately symmetric and bell shaped.

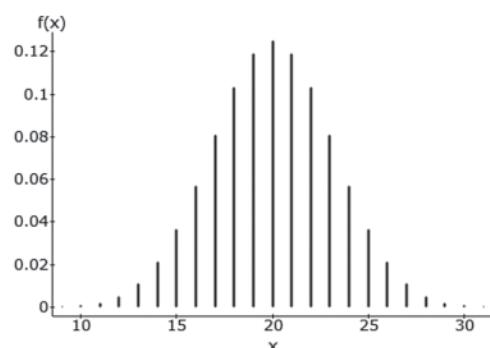
This result suggests that binomial probabilities can be approximated by the area under a normal curve, provided that $np(1 - p) \geq 10$.

The Normal Approximation to the Binomial Probability Distribution

If $np(1 - p) \geq 10$, the binomial random variable X is approximately normally distributed, with mean $\mu_X = np$ and standard deviation $\sigma_X = \sqrt{np(1 - p)}$.

Figure 35 shows a graph of the probability distribution for the binomial random variable X , with $n = 40$ and $p = 0.5$, drawn in StatCrunch. Because $np(1 - p) = 40(0.5)(1 - 0.5) = 10$, we can use a normal model with $\mu_X = np = 40(0.5) = 20$ and standard deviation $\sigma_X = \sqrt{np(1 - p)} = \sqrt{40(0.5)(0.5)} = \sqrt{10}$ to describe X .

Figure 35

**CAUTION!**

Don't forget about the correction for continuity. It is needed because we are using a continuous density function to approximate the probability of a discrete random variable.

To approximate the probability of a specific value of the binomial random variable, such as $P(18)$, we find the area under the normal curve from $x = 17.5$ to $x = 18.5$. We add and subtract 0.5 from $x = 18$ as a **correction for continuity** because we are using a continuous density function to approximate the probability of a discrete random variable.

To approximate $P(X \leq 18)$, compute the area under the normal curve for $X \leq 18.5$. Do you see why?

To approximate $P(X \geq 18)$, compute $P(X \geq 17.5)$. Do you see why? Table 6 summarizes how to use the correction for continuity.

Table 6

Exact Probability Using Binomial	Approximate Probability Using Normal	Graphical Depiction
$P(a)$	$P(a - 0.5 \leq X \leq a + 0.5)$	
$P(X \leq a)$	$P(X \leq a + 0.5)$	
$P(X \geq a)$	$P(X \geq a - 0.5)$	
$P(a \leq X \leq b)$	$P(a - 0.5 \leq X \leq b + 0.5)$	

A question remains, however. What do we do if the probability is of the form $P(X > a)$, $P(X < a)$, or $P(a < X < b)$? The solution is to rewrite the inequality in a form with \leq or \geq . For example, $P(X > 4) = P(X \geq 5)$ and $P(X < 4) = P(X \leq 3)$ for binomial random variables, because the values of the random variables must be whole numbers.

EXAMPLE 1 The Normal Approximation to a Binomial Random Variable

Problem According to the American Red Cross, 7% of people in the United States have blood type O-negative. What is the probability that, in a simple random sample of 500 people in the United States, fewer than 30 have blood type O-negative?

Approach

Step 1 Verify that this is a binomial experiment.

Step 2 Computing the probability by hand would be very tedious. Verify $np(1 - p) \geq 10$. Then we may use the normal distribution to approximate the binomial probability.

Step 3 Approximate $P(X < 30) = P(X \leq 29)$ by using the normal approximation to the binomial distribution.

Solution

Step 1 Each of the 500 independent trials has a probability of success equal to 0.07. This is a binomial experiment.

Step 2 Verify $np(1 - p) \geq 10$.

$$np(1 - p) = 500(0.07)(0.93) = 32.55 \geq 10$$

Use the normal distribution to approximate the binomial distribution.

Step 3 The probability that fewer than 30 people in the sample have blood type O-negative is $P(X < 30) = P(X \leq 29)$. This is approximately equal to the area under the normal curve to the left of $x = 29.5$, with $\mu_X = np = 500(0.07) = 35$ and $\sigma_X = \sqrt{np(1 - p)} = \sqrt{500(0.07)(1 - 0.07)} = \sqrt{32.55} \approx 5.71$. See Figure 36. Convert $x = 29.5$ to a z -score.

$$z = \frac{29.5 - 35}{\sqrt{32.55}} = -0.96$$

Figure 36

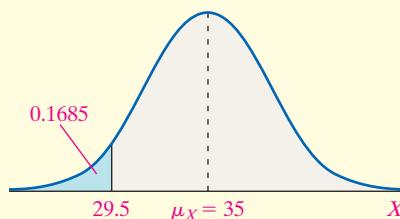
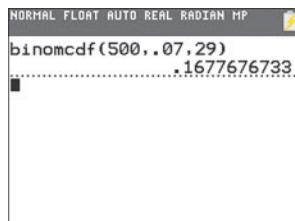


Figure 37



From Table V, we find that the area to the left of $z = -0.96$ is 0.1685. Therefore, the approximate probability that fewer than 30 people will have blood type O-negative is 0.1685.

NW Now Work Problem 21

Using the *binomcdf*(command on a TI-84 Plus CE graphing calculator, we find that the exact probability is 0.1678. See Figure 37. The approximate result is close indeed!

EXAMPLE 2 A Normal Approximation to the Binomial

Problem According to the Gallup Organization, 65% of adult Americans are in favor of the death penalty for individuals convicted of murder. Erica selects a random sample of 1000 adult Americans in Will County, Illinois, and finds that 630 of them are in favor of the death penalty for individuals convicted of murder.

- Assuming that 65% of adult Americans in Will County are in favor of the death penalty, what is the probability of obtaining a random sample of no more than 630 adult Americans in favor of the death penalty from a sample of size 1000?
- Does the result from part (a) contradict the Gallup Organization's findings? Explain.

Approach This is a binomial experiment with $n = 1000$ and $p = 0.65$. Erica needs to determine the probability of obtaining a random sample of no more than 630 adult Americans who favor the death penalty, assuming 65% of adult Americans favor the death penalty. Computing this probability using the binomial probability formula would be difficult, so Erica will use the normal approximation to the binomial, since $np(1 - p) = 1000(0.65)(1 - 0.65) = 227.5 \geq 10$. Approximate $P(X \leq 630)$ by computing the area under the normal curve to the left of $x = 630.5$ with $\mu_X = np = 650$ and $\sigma_X = \sqrt{np(1 - p)} = \sqrt{1000(0.65)(1 - 0.65)} \approx 15.083$.

Solution

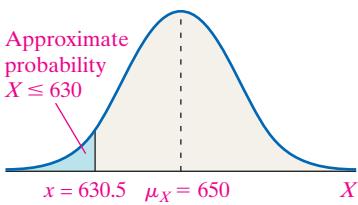
- Figure 38 shows the area we wish to compute. Convert $x = 630.5$ to a z -score.

$$z = \frac{630.5 - 650}{15.083} = -1.29$$

The area under the standard normal curve to the left of $z = -1.29$ is 0.0985. The probability of obtaining 630 or fewer adult Americans who favor the death penalty from a sample of 1000 adult Americans, assuming the proportion of adult Americans who favor the death penalty is 0.65 is 0.0985.

- The result of part (a) means that, if we had obtained 100 different simple random samples of size 1000, we would expect about 10 to result in 630 or fewer adult Americans favoring the death penalty if the true proportion is 0.65. Because the results obtained are not unusual under the assumption that $p = 0.65$, Erica finds that the results of her survey do not contradict those of Gallup.

NW Now Work Problem 27



7.4 Assess Your Understanding

Vocabulary and Skill Building

- In a binomial experiment with n trials and probability of success p , if _____, the binomial random variable X is approximately normal with $\mu_X =$ _____ and $\sigma_X =$ _____.
- When adding or subtracting 0.5 from x , we are making a correction for _____.
- Suppose X is a binomial random variable. To approximate $P(X < 5)$, compute _____.
- Suppose X is a binomial random variable. To approximate $P(3 \leq X \leq 10)$, compute _____.

In Problems 5–14, a discrete random variable is given. Assume the probability of the random variable will be approximated using the normal distribution. Describe the area under the normal curve

that will be computed. For example, if we wish to compute the probability of finding at least five defective items in a shipment, we would approximate the probability by computing the area under the normal curve to the right of $x = 4.5$.

- The probability that at least 40 households have a gas stove
- The probability no more than 20 people want to see *Roe v. Wade* overturned
- The probability that exactly eight defective parts are in the shipment
- The probability that exactly 12 students pass the course
- The probability that the number of people with blood type O-negative is between 18 and 24, inclusive
- The probability that the number of tornadoes that occur in the month of May is between 30 and 40, inclusive

11. The probability that more than 20 people want to see the marriage tax penalty abolished
12. The probability that fewer than 40 households have a pet
13. The probability that no more than 500 adult Americans support a bill proposing to extend daylight savings time
14. The probability that fewer than 35 people support the privatization of Social Security

In Problems 15–20, compute $P(x)$ using the binomial probability formula. Then determine whether the normal distribution can be used as an approximation for the binomial distribution. If so, approximate $P(x)$ and compare the result to the exact probability.

15. $n = 60, p = 0.4, x = 20$
16. $n = 80, p = 0.15, x = 18$
17. $n = 40, p = 0.25, x = 30$
18. $n = 100, p = 0.05, x = 50$
19. $n = 75, p = 0.75, x = 60$
20. $n = 85, p = 0.8, x = 70$

Applying the Concepts

- NW 21. On-Time Flights** According to American Airlines, Flight 215 from Orlando to Los Angeles is on time 90% of the time. Randomly select 150 flights and use the normal approximation to the binomial to

- (a) approximate the probability that exactly 130 flights are on time.
- (b) approximate the probability that at least 130 flights are on time.
- (c) approximate the probability that fewer than 125 flights are on time.
- (d) approximate the probability that between 125 and 135 flights, inclusive, are on time.

- 22. Morality** In a recent poll, the Gallup Organization found that 45% of adult Americans believe that the overall state of moral values in the United States is poor. Suppose a survey of a random sample of 500 adult Americans is conducted in which they are asked to disclose their feelings on the overall state of moral values in the United States. Use the normal approximation to the binomial to approximate the probability that

- (a) exactly 250 of those surveyed feel the state of morals is poor.
- (b) no more than 220 of those surveyed feel the state of morals is poor.
- (c) more than 250 of those surveyed feel the state of morals is poor.
- (d) between 220 and 250, inclusive, believe the state of morals is poor.
- (e) at least 260 adult Americans believe the overall state of moral values is poor. Would you find this result unusual? Why?

- 23. Toilet Flushing** In the Healthy Handwashing Survey conducted by Bradley Corporation, it was found that 64% of adult Americans operate the flusher of toilets in public restrooms with their foot. Suppose you survey a random sample of 740 adult American women aged 18–24 years. Use the normal approximation to the binomial to approximate the probability that

- (a) exactly 490 of those surveyed flush toilets in public restrooms with their foot.
- (b) no more than 490 of those surveyed flush toilets in public restrooms with their foot.
- (c) at least 503 of those surveyed flush toilets in public restrooms with their foot. What does this result suggest?

- 24. Sneeze** According to a study done by Nick Wilson of Otago University Wellington, the probability a randomly selected individual will not cover his or her mouth when sneezing is 0.267. Suppose you sit on a bench in a mall and observe 300 randomly selected individuals' habits as they sneeze. Use the normal approximation to the binomial to approximate the probability that of the 300 randomly observed individuals:

- (a) exactly 100 do not cover the mouth when sneezing.
- (b) fewer than 75 do not cover the mouth.
- (c) Would you be surprised if, after observing 300 individuals, more than 100 did not cover the mouth when sneezing? Why?

- 25. Males Living at Home** According to the *Current Population Survey* (Internet release date: September 15, 2004), 55% of males between the ages of 18 and 24 years lived at home in 2003. (Unmarried college students living in a dorm are counted as living at home.) Suppose a survey is administered today to 200 randomly selected males between the ages of 18 and 24 years, and 130 of them respond that they live at home.

- (a) Approximate the probability that such a survey will result in at least 130 of the respondents living at home under the assumption that the true percentage is 55%.
- (b) What does the result from part (a) suggest?

- 26. Females Living at Home** According to the *Current Population Survey* (Internet release date: September 15, 2004), 46% of females between the ages of 18 and 24 years lived at home in 2003. (Unmarried college students living in a dorm are counted as living at home.) Suppose a survey is administered today to 200 randomly selected females between the ages of 18 and 24 years, and 110 of them respond that they live at home.

- (a) Approximate the probability that such a survey will result in at least 110 of the respondents living at home under the assumption that the true percentage is 46%.
- (b) What does the result from part (a) suggest?

- NW 27. Views of Socialism** In a Pew Research poll, 42% of adult Americans had a positive view of socialism. You conduct a survey of 200 randomly selected students at your school and find that 103 have a positive view of socialism.

- (a) Approximate the probability that, in a random sample of 200 students, at least 103 would have a positive view of socialism, assuming the true percentage is 42%.

- (b) Explain what this result suggests.

- 28. Liars** According to a *USA Today* “Snapshot,” 3% of Americans surveyed lie frequently. You conduct a survey of 500 college students and find that 20 of them lie frequently.

- (a) Compute the probability that, in a random sample of 500 college students, at least 20 lie frequently, assuming the true percentage is 3%.
- (b) Does this result contradict the *USA Today* “Snapshot”? Explain.



Chapter 7 Review

Summary

In this chapter we introduced continuous random variables and the normal probability density function. A continuous random variable is said to be approximately normally distributed if a histogram of its values is symmetric and bell shaped. We use a normal curve to model normal random variables. The normal model is bell shaped with the mean representing the high point of the curve. The curve also has inflection points one standard deviation on either side of the mean.

The area under a normal curve for an interval of numbers may be interpreted as a proportion, probability, or percentile. We may also reverse the process, and find the value of a random variable that corresponds to a particular proportion, probability, or percentile.

One method for determining whether a random variable might come from a population that is normally distributed is to draw a normal probability plot. If a normal probability plot is approximately linear, we say the distribution of the random variable is approximately normal. The correlation coefficient between the raw data and normal scores may be used to assess the normality of the variable.

If X is a binomial random variable with $np(1 - p) \geq 10$, the area under the normal curve may be used to approximate the probability of a particular binomial random variable. The parameters of the normal curve are $\mu_X = np$ and $\sigma_X = \sqrt{np(1 - p)}$, where n is the number of trials of the binomial experiment and p is the probability of success on a single trial.

Vocabulary

Uniform probability distribution (p. 333)
Probability density function (p. 333)
Model (p. 335)
Normal curve (p. 335)
Normally distributed (p. 335)
Normal probability distribution (p. 335)
Density (p. 335)

Inflection points (p. 336)
Probability density function (p. 338)
Normal probability density function (p. 338)
Standard normal distribution (p. 343)
Standard normal random variable Z (p. 343)
Standard normal curve (p. 343)

Normal probability plot (p. 355)
Normal score (p. 355)
Trial (p. 360)
Normal approximation to the binomial distribution (p. 360)
Correction for continuity (p. 361)

Formulas

Standardizing a Normal Random Variable

$$z = \frac{x - \mu}{\sigma}$$

Finding the Score

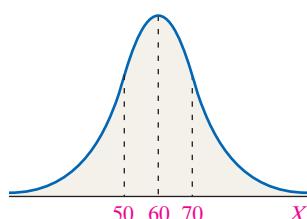
$$x = \mu + z\sigma$$

Objectives

Section	You should be able to . . .	Example(s)	Review Exercises
7.1	1 Use the uniform probability distribution (p. 333) 2 Graph a normal curve (p. 335) 3 State the properties of the normal curve (p. 336) 4 Explain the role of area in the normal density function (p. 337)	1 and 2 pages 335–336 pages 336–339 3 and 4	18 7–9 19 1
7.2	1 Find and interpret the area under a normal curve (p. 343) 2 Find the value of a normal random variable (p. 347)	1 and 2 3–5	7–9, 10(a)–(c), 11(a)–(d), 12 4–6, 10(d), 11(e)–(f)
7.3	1 Use normal probability plots to assess normality (p. 355)	1 and 2	14–16, 20
7.4	1 Approximate binomial probabilities using the normal distribution (p. 360)	1 and 2	13, 17

Review Exercises

1. Use the figure to answer the questions that follow.



- (a) What is μ ?
(b) What is σ ?
(c) Suppose that the area under the normal curve to the right of $x = 75$ is 0.0668. Provide two interpretations for this area.
(d) Suppose that the area under the normal curve between $x = 50$ and $x = 75$ is 0.7745. Provide two interpretations for this area.

In Problems 2 and 3, draw a standard normal curve and shade the area indicated. Then find the area of the shaded region.

2. The area to the left of $z = -1.04$.
3. The area between $z = -0.34$ and $z = 1.03$.
4. Find the z -score such that the area to the right of the z -score is 0.483.
5. Find the z -scores that separate the middle 92% of the data from the area in the tails of the standard normal distribution.
6. Find the value of $z_{0.20}$.

In Problems 7–9, draw the normal curve with the parameters indicated. Then find the probability of the random variable X . Shade the area that represents the probability.

7. $\mu = 50, \sigma = 6, P(X > 55)$
8. $\mu = 30, \sigma = 5, P(X \leq 23)$
9. $\mu = 70, \sigma = 10, P(65 < X < 85)$

10. Tire Wear Suppose that Dunlop Tire manufactures a tire with a lifetime that approximately follows a normal distribution with mean 70,000 miles and standard deviation 4400 miles.

- (a) What proportion of the tires will last at least 75,000 miles?
- (b) Suppose that Dunlop warrants the tires for 60,000 miles. What proportion of the tires will last 60,000 miles or less?
- (c) What is the probability that a randomly selected Dunlop tire lasts between 65,000 and 80,000 miles?
- (d) Suppose that Dunlop wants to warrant no more than 2% of its tires. What mileage should the company advertise as its warranty mileage?

11. Wechsler Intelligence Scale The Wechsler Intelligence Scale for Children is approximately normally distributed, with mean 100 and standard deviation 15.

- (a) What is the probability that a randomly selected test taker will score above 125?
- (b) What is the probability that a randomly selected test taker will score below 90?
- (c) What proportion of test takers will score between 110 and 140?
- (d) If a child is randomly selected, what is the probability that she scores above 150?
- (e) What intelligence score will place a child in the 98th percentile?
- (f) If normal intelligence is defined as scoring in the middle 95% of all test takers, figure out the scores that differentiate normal intelligence from abnormal intelligence.

12. Major League Baseballs According to Major League Baseball rules, the ball must weigh between 5 and 5.25 ounces. A factory produces baseballs whose weights are approximately normally distributed, with mean 5.11 ounces and standard deviation 0.062 ounce.

Source: www.baseball-almanac.com

- (a) What proportion of the baseballs produced by this factory are too heavy for use by Major League Baseball?
- (b) What proportion of the baseballs produced by this factory are too light for use by Major League Baseball?
- (c) What proportion of the baseballs produced by this factory can be used by Major League Baseball?
- (d) If 8000 baseballs are ordered, how many baseballs should be manufactured, knowing that some will need to be discarded?

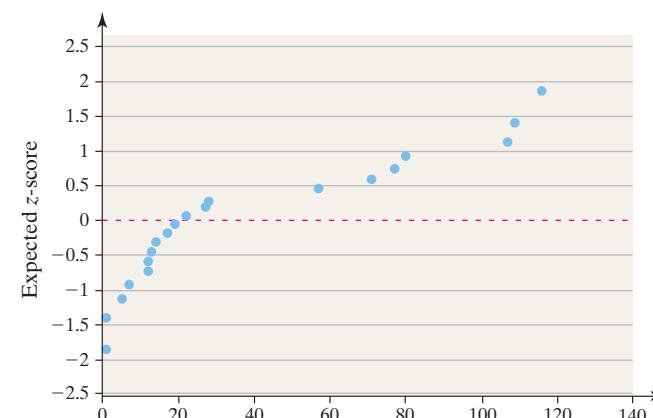
13. America Reads According to a Gallup poll, 46% of Americans 18 years old or older stated that they had read at least six books (fiction and nonfiction) within the past year. You conduct a random sample of 250 Americans 18 years old or older

and ask the individuals to disclose whether they read at least six books in the past year, or not.

- (a) Verify that the conditions for using the normal distribution to approximate the binomial distribution are met.
- (b) Approximate the probability that exactly 125 read at least six books within the past year. Interpret this result.
- (c) Approximate the probability that fewer than 120 read at least six books within the past year. Interpret this result.
- (d) Approximate the probability that at least 140 read at least six books within the past year. Interpret this result.
- (e) Approximate the probability that between 100 and 120, inclusive, read at least six books within the past year. Interpret this result.
14. Use the results in the table to (a) draw a normal plot, (b) determine the linear correlation between the observed values and expected z -scores, (c) determine the critical value in Table VI to assess the normality of the data.

Index, i	Observed Value	f_i	Expected z -score
1	48	0.09	-1.34
2	49	0.22	-0.77
3	51	0.36	-0.36
4	52	0.50	0
5	54	0.64	0.36
6	54	0.78	0.77
7	56	0.91	1.34

15. Hector obtained a random sample of twenty recent college graduates who own cars and asked each to disclose the age of their car (in months). Is it reasonable to conclude that age of car is normally distributed? The normal probability plot is shown below and the correlation between age of car and expected z -scores is 0.914.



- DATA 16. Density of Earth** In 1798, Henry Cavendish obtained 27 measurements of the density of Earth, using a torsion balance. The following data represent his estimates, given as a multiple of the density of water. Is it reasonable to conclude that the sample data come from a population that is normally distributed?

5.50	5.34	5.57	5.30	5.42	5.36	5.63	5.27	5.55
5.47	5.10	4.88	5.86	5.62	5.58	5.44	5.53	5.29
5.29	5.65	5.34	5.39	5.26	5.61	5.79	4.07	5.85

Source: S. M. Stigler. "Do Robust Estimators Work with Real Data?" *Annals of Statistics* 5(1977), 1055–1078.

17. Creative Thinking According to a *USA Today* “Snapshot,” 20% of adults surveyed do their most creative thinking while driving. You conduct a survey of 250 adults and find that 30 do their most creative thinking while driving.

- (a) Compute the probability that, in a random sample of 250 adults, 30 or fewer do their most creative thinking while driving.
- (b) Does this result contradict the *USA Today* “Snapshot”? Explain.

18. A continuous random variable X is uniformly distributed with $0 \leq X \leq 20$.

- (a) Draw a graph of the uniform density function.

(b) What is $P(0 \leq X \leq 5)$?

(c) What is $P(10 \leq X \leq 18)$?

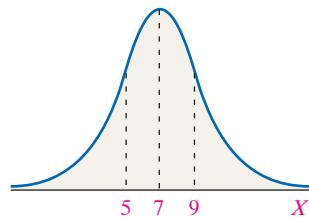
19. List the properties of the normal density curve.

20. Explain how to use a normal probability plot to assess normality.



Chapter Test

1. Use the figure to answer the questions that follow:



- (a) What is μ ?
- (b) What is σ ?
- (c) Suppose that the area under the normal curve to the left of $x = 10$ is 0.9332. Provide two interpretations for this area.
- (d) Suppose that the area under the normal curve between $x = 5$ and $x = 8$ is 0.5328. Provide two interpretations for this area.

2. Draw a standard normal curve and shade the area to the right of $z = 2.04$. Then find the area of the shaded region.

3. Find the z -scores that separate the middle 88% of the data from the area in the tails of the standard normal distribution.

4. Find the value of $z_{0.04}$.

5. (a) Draw a normal curve with $\mu = 20$ and $\sigma = 3$.
(b) Shade the region that represents $P(22 \leq X \leq 27)$ and find the probability.

6. Suppose that the talk time on the Apple iPhone is approximately normally distributed with mean 7 hours and standard deviation 0.8 hour.

- (a) What proportion of the time will a fully charged iPhone last at least 6 hours?
- (b) What is the probability a fully charged iPhone will last less than 5 hours?
- (c) What talk time would represent the cutoff for the top 5% of all talk times?
- (d) Would it be unusual for the phone to last more than 9 hours? Why?

7. The waist circumference of males 20–29 years old is approximately normally distributed, with mean 92.5 cm and standard deviation 13.7 cm.

Source: M. A. McDowell, C. D. Fryar, R. Hirsch, and C. L. Ogden. *Anthropometric Reference Data for Children and Adults: U.S. Population, 1999–2002*. Advance data from vital and health statistics: No. 361. Hyattsville, MD: National Center for Health Statistics, 2005.

- (a) Use the normal model to determine the proportion of 20- to 29-year-old males whose waist circumference is less than 100 cm.

- (b) What is the probability that a randomly selected 20- to 29-year-old male has a waist circumference between 80 and 100 cm?

- (c) Determine the waist circumferences that represent the middle 90% of all waist circumferences.

- (d) Determine the waist circumference that is at the 10th percentile.

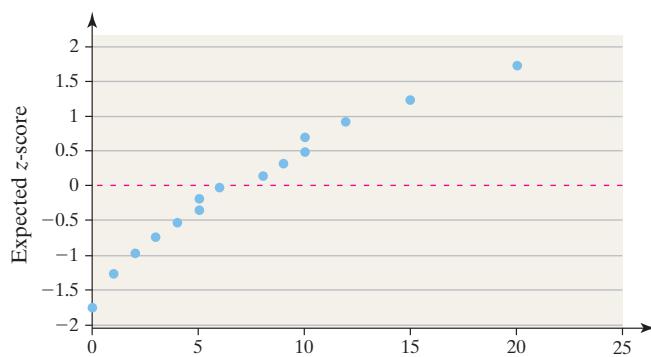
8. Suppose the scores earned on Professor McArthur’s third statistics exam are normally distributed with mean 64 and standard deviation 8. Professor McArthur wants to curve the exam scores as follows: The top 6% get an A, the next 14% get a B, the middle 60% get a C, the bottom 6% fail, and the rest earn a D. Any student who can determine these cut-offs earns five bonus points. Determine the cut-offs for Professor McArthur.

9. In a poll conducted by the Gallup organization, 16% of adult, employed Americans were dissatisfied with the amount of their vacation time. You conduct a survey of 500 adult, employed Americans.

- (a) Approximate the probability that exactly 100 are dissatisfied with their amount of vacation time.

- (b) Approximate the probability that less than 60 are dissatisfied with the amount of their vacation time.

10. Jane obtained a random sample of 15 college students and asked how many hours they studied last week. Is it reasonable to believe that hours studied is normally distributed? The normal probability plot is shown below and the correlation between hours studied and expected z -scores is 0.974.



11. A continuous random variable X is uniformly distributed with $10 \leq X \leq 50$.

- (a) Draw a graph of the uniform density function.

- (b) What is $P(20 \leq X \leq 30)$?

- (c) What is $P(X < 15)$?

Making an Informed Decision

Stock Picking

You are interested in modeling the behavior of stocks. In particular, you want to build a model that describes the rate of return on a basket of stocks, such as large capitalization companies.

(a) Go to a website that provides historical rates of return on a certain basket of stocks, such as www.morningstar.com. Decide on a certain sector of the economy you would like to model, such as consumer goods or energy. Choose the largest 100 companies in this sector and determine the time frame for which you want to build your model. For example, you might decide to build a model for the one-year rate of return on the stock.

(b) Enter your data into statistical software and construct a relative frequency histogram. Does the data appear bell-shaped? Do you think the normal model would be a good model to describe the rate of return for the sector you have chosen? Why or why not?

(c) Regardless of your answer to part (b), build a normal model by determining the mean and standard deviation for the rate of return. Draw the normal model on the relative frequency histogram from part (b).

(d) One purpose of financial models is to identify quality investments going forward. Without a crystal ball, all investment managers have is historical data. Use your historical data to determine a rate of return that falls into the top 20% of all companies within the sector. Use this rate of return as a criterion for choosing a company to invest in.



(e) Conduct further research on the stock you wish to invest in. For example, have the company's earnings been growing for the past five years? What is the company's market share? Does the company pay a dividend? If so, for how long has the company paid a dividend? Has the dividend been growing consistently?

(f) Write a report that lays out your recommendation regarding the particular stock you have researched. Perhaps include a few other models that might help to decide whether the company is a solid investment.



Inference: From Samples to Population

In Chapter 1, we presented the following process of statistics:

- Step 1:** Identify the research objective.
- Step 2:** Collect the data needed to answer the question(s) posed in Step 1.
- Step 3:** Describe the data.
- Step 4:** Perform inference.

The methods for conducting Steps 1 and 2 were discussed in Chapter 1. The methods for conducting Step 3 were discussed in Chapters 2 through 4. We took a break from the statistical process in Chapters 5 through 7 so that we could develop skills that allow us to tackle Step 4.

Since it is difficult to gain access to populations, the data found in Step 2 is often from a sample. Sample data are used to make inferences about the population. For example, we might compute a sample mean from the data collected in Step 2 and use this information to draw conclusions regarding the population mean. Part 4 focuses on how sample data are used to draw conclusions about populations.

- CHAPTER 8** Sampling Distributions
- CHAPTER 9** Estimating the Value of a Parameter
- CHAPTER 10** Hypothesis Tests Regarding a Parameter
- CHAPTER 11** Inference on Two Population Parameters
- CHAPTER 12** Additional Inferential Methods

8

Sampling Distributions

Outline

- 8.1** Distribution of the Sample Mean
- 8.2** Distribution of the Sample Proportion

Making an Informed Decision



The American Time Use Survey, conducted by the Bureau of Labor Statistics, investigates how adult Americans allocate their time during a day. As a reporter for the school newspaper, you wish to file a report that compares the typical student at your school to other Americans. See the Decisions project on page 394.

Putting It Together

In Chapters 6 and 7, we learned about random variables and their probability distributions. A random variable is a numerical measure of the outcome to a probability experiment. A probability distribution provides a way to assign probabilities to the possible values of the random variable. For discrete random variables, we discussed the binomial probability distribution where we assigned probabilities using a formula. For continuous random variables, we discussed the normal probability distribution. To compute probabilities for a normal random variable, we found the area under a normal density curve.

In this chapter, we continue our discussion of probability distributions where statistics, such as \bar{x} , will be the random variable. Statistics are random variables because the value of a statistic varies from sample to sample. For this reason, statistics have probability distributions associated with them. For example, there is a probability distribution for the sample mean, sample proportion, and so on. We use probability distributions to make probability statements regarding the statistic. So this chapter discusses the shape, center, and spread of statistics such as \bar{x} .

8.1 Distribution of the Sample Mean



Preparing for This Section Before getting started, review the following:

- Simple random sampling (Section 1.3, pp. 23–27)
- The mean (Section 3.1, pp. 108–110)
- The standard deviation (Section 3.2, pp. 123–128)
- Applications of the normal distribution (Section 7.2, pp. 343–350)

Objectives

- ① Describe the distribution of the sample mean: normal population
- ② Describe the distribution of the sample mean: nonnormal population

Suppose the government wanted to determine the mean income of all U.S. households. One approach the government could take is to survey every U.S. household to determine the population mean, μ . This would be a very expensive and time-consuming survey!

A second approach the government could (and does) take is to survey a random sample of U.S. households and use the results to estimate the mean household income. The Current Population Survey is administered to approximately 60,000 randomly selected households each month. Among the many questions on the survey, respondents are asked to report the income of each individual in the household. From this information, the federal government obtains a sample mean household income for U.S. households. For example, in 2017 the mean annual household income in the United States was estimated to be $\bar{x} = \$86,220$. The government might infer from this survey that the mean annual household income of *all* U.S. households in 2017 was $\mu = \$86,220$.

The households in the Current Population Survey were determined by chance (random sampling). A second random sample of households would likely lead to a different sample mean, such as $\bar{x} = \$85,839$, and a third random sample of households would likely lead to a third sample mean, such as $\bar{x} = \$86,941$. Because the households selected will vary from sample to sample, the sample mean of household income will also vary from sample to sample. For this reason, the sample mean \bar{x} is a random variable, so it has a probability distribution. Our goal in this section is to describe the distribution of the sample mean. Remember, when we describe a distribution, we do so in terms of its shape, center, and spread.

Definitions

The **sampling distribution** of a statistic is a probability distribution for all possible values of the statistic computed from a sample of size n .

The **sampling distribution of the sample mean** \bar{x} is the probability distribution of all possible values of the random variable \bar{x} computed from a sample of size n from a population with mean μ and standard deviation σ .

The idea behind obtaining the sampling distribution of the sample mean is as follows:

Step 1 Obtain a simple random sample of size n .

Step 2 Compute the sample mean.

Step 3 Assuming that we are sampling from a finite population, repeat Steps 1 and 2 until all distinct simple random samples of size n have been obtained.

Note: Once a particular sample is obtained, it cannot be obtained a second time.

① Describe the Distribution of the Sample Mean: Normal Population

The probability distribution of the sample mean is determined from statistical theory. We will use simulation to help justify the result that statistical theory provides. We

IN OTHER WORDS

If the number of individuals in a population is a positive integer, we say the population is finite. Otherwise, the population is infinite.

consider two possibilities. In the first case (Examples 1, 2, and 3), we sample from a population that is normally distributed. In the second case (Examples 4 and 5), we sample from a population that is not normally distributed.

EXAMPLE 1

Sampling Distribution of the Sample Mean: Normal Population

Problem An intelligence quotient, or IQ, is a measurement of intelligence derived from a standardized test, such as the Stanford Binet IQ test. Scores on this test are approximately normal with a mean score of 100 and a standard deviation of 15. What is the sampling distribution of the sample mean for a sample of size $n = 9$?

Approach The problem asks us to determine the shape, center, and spread of the distribution of the sample mean. Remember, the sampling distribution of the sample mean would be the distribution of *all* possible sample means of size $n = 9$. To get a sense of this distribution, use Minitab to simulate obtaining 1000 samples of size $n = 9$ by randomly generating 1000 rows of IQs over 9 columns. Each row represents a random sample of size 9. For each of the 1000 samples (the 1000 rows), we determine the mean IQ score. Draw a histogram to gauge the shape of the distribution of the sample mean, determine the mean of the 1000 sample means to approximate the mean of the sampling distribution, and determine the standard deviation of the 1000 sample means to approximate the standard deviation of the sampling distribution.

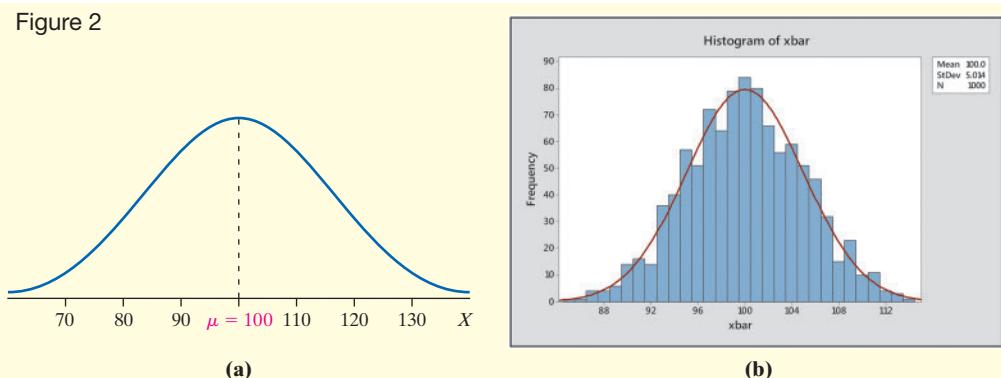
Solution Figure 1 shows random samples from Minitab. Row 1 contains the first sample, where the IQ scores of the nine individuals are 90, 74, 95, 88, 91, 91, 102, 91, and 96. The mean of these nine IQ scores is 90.9. Row 2 represents a second sample with nine different IQ scores; row 3 represents a third sample, and so on. Column C10 (xbar) lists the sample means for each of the different samples.

Figure 1

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	xbar
1	90	74	95	88	91	91	102	91	96	90.9	$\bar{x} = 90.9$
2	114	86	96	82	80	68	93	136	111	96.2	$\bar{x} = 96.2$
3	89	89	86	98	96	96	89	99	107	94.3	
4	87	94	89	116	92	124	115	83	111	101.3	
5	107	103	86	86	109	104	94	82	110	97.8	
6	113	84	101	89	92	71	89	86	106	92.7	
7	75	116	107	118	112	86	104	97	106	102.4	
8	118	117	81	96	86	94	109	96	104	100.1	
9	91	91	79	80	109	89	97	81	99	90.6	
10	112	98	115	73	95	104	76	95	87	95.0	
11	103	92	100	75	95	108	105	125	62	96.1	
12	75	124	101	113	91	85	121	115	85	101.2	
13	104	103	81	134	111	108	101	88	115	105.1	
14	121	85	118	88	96	84	103	77	102	97.1	
15	119	84	119	80	99	98	88	88	89	95.9	
16	71	86	94	101	95	84	124	105	83	93.7	

Figure 2(a) shows the distribution of the population, and Figure 2(b) shows the distribution of the sample means from column C10 (using Minitab). The shape of the distribution of the population is normal. The histogram in Figure 2(b) shows that the shape of the distribution of the sample means is approximately normal. In addition, we notice that the center of the distribution of the sample means is the same as the center of the distribution of the population, but the spread of the distribution of the sample means is smaller than the spread of the distribution of the population. In fact, the mean of the 1000 sample means is 100.02, which is close to the population mean, 100; the standard deviation of the sample means is 5.01, which is less than the population standard deviation, 15.

Figure 2



We draw the following conclusions:

- **Shape:** The shape of the distribution of the sample mean is approximately normal.
- **Center:** The mean of the distribution of the sample mean equals the mean of the population, 100.
- **Spread:** The standard deviation of the sample mean is less than the standard deviation of the population.

Why is the standard deviation of the sample mean less than the standard deviation of the population? Consider that, if we randomly select any one individual, according to the Empirical Rule, there is about a 68% chance that the individual's IQ score is between 85 and 115 (that is, within 1 standard deviation of the mean). If we had a sample of 9 individuals, we would not expect as much spread in the sample mean as there is for a single individual, since individuals with lower IQs will offset individuals in the sample with higher IQs, resulting in a sample mean closer to the expected value of 100. Look back at Figure 1. In the first sample (row 1), the low-IQ individual ($\text{IQ} = 74$) is offset by the higher-IQ individual ($\text{IQ} = 102$), which is why the sample mean is closer to 100. In the second sample (row 2), the low-IQ individual ($\text{IQ} = 68$) is offset by the higher-IQ individual ($\text{IQ} = 136$), so the sample mean of the second sample is closer to 100. Therefore, the spread in the distribution of sample means should be less than the spread in the population from which the sample is drawn.

Based on this, what role do you think n , the sample size, plays in the standard deviation of the distribution of the sample mean?

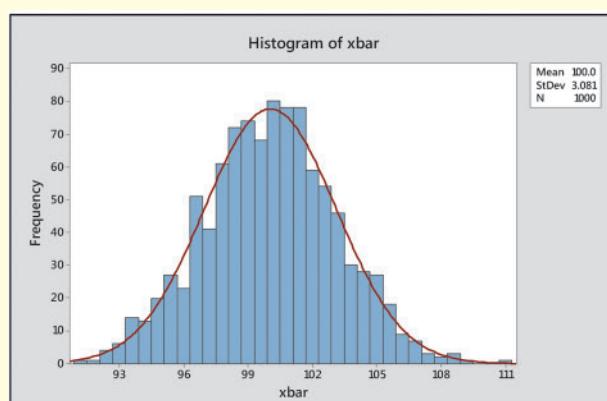
EXAMPLE 2 The Impact of Sample Size on Sampling Variability

Problem Repeat the problem in Example 1 with a sample of size $n = 25$.

Approach Use the approach presented in Example 1, but let $n = 25$ instead of $n = 9$.

Solution Figure 3 shows the histogram of the sample means. Notice that the sample means appear to be approximately normal with the center at 100. The histogram in Figure 3 shows less dispersion than the histogram in Figure 2(b). This implies that the distribution of \bar{x} with $n = 25$ has less variability than the distribution of \bar{x} with $n = 9$. In fact, the mean of the 1000 sample means is 100.05, and the standard deviation is 3.08.

Figure 3



From the results of Examples 1 and 2, we conclude that, as the sample size n increases, the standard deviation of the distribution of \bar{x} decreases. Although the proof is beyond the scope of this text, we should be convinced that the following result is reasonable.

IN OTHER WORDS

Regardless of the distribution of the population, the sampling distribution of \bar{x} will have a mean equal to the mean of the population and a standard deviation equal to the standard deviation of the population divided by the square root of the sample size!

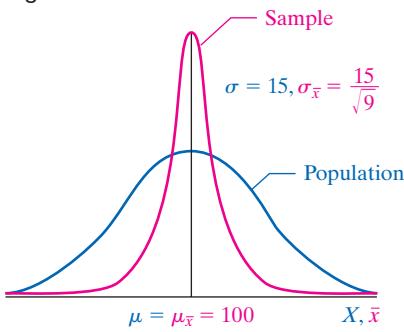
CAUTION!

It is important that two assumptions are satisfied with regard to sampling from a population.

1. The sample must be a random sample.
2. The sampled values must be independent. When sampling without replacement (which is the case when obtaining simple random samples), we verify this assumption by checking that the size is less than 5% of the population size ($n < 0.05N$).

NW Now Work Problem 9

Figure 4



The Mean and Standard Deviation of the Sampling Distribution of \bar{x}

Suppose that a simple random sample of size n is drawn from a population* with mean μ and standard deviation σ . The sampling distribution of \bar{x} has mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. The standard deviation of the sampling distribution of \bar{x} , $\sigma_{\bar{x}}$, is called the **standard error of the mean**.

For the population presented in Example 1, if we draw a simple random sample of size $n = 9$, the sampling distribution \bar{x} will have mean $\mu_{\bar{x}} = 100$ and standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{9}} = 5$$

This standard error of the mean is close to the approximate standard error of 5.01 found in our simulation in Example 1.

In Example 2, where the simple random sample was of size $n = 25$, the sampling distribution of \bar{x} will have mean $\mu_{\bar{x}} = 100$ and standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3$$

This standard error of the mean is close to the approximate standard error of 3.08 found in our simulation in Example 2.

Now that we can find the mean and standard deviation for any sampling distribution of \bar{x} , we can concentrate on the shape of the distribution. Refer back to Figures 2(b) and 3 from Examples 1 and 2. Both histograms appear to be approximately normal. Recall that the population from which the sample was drawn was normal. This leads us to believe that, if the population is normal, then the distribution of the sample mean is also approximately normal.

The Shape of the Sampling Distribution of \bar{x} If X Is Normal

If a random variable X is approximately normally distributed, the sampling distribution of the sample mean, \bar{x} , is approximately normally distributed.

For example, the IQ scores of individuals are modeled by a normal random variable with mean $\mu = 100$ and standard deviation $\sigma = 15$. The sampling distribution of the sample mean, \bar{x} , the mean IQ of a simple random sample of $n = 9$ individuals, is approximately normal, with mean $\mu_{\bar{x}} = 100$ and standard deviation $\sigma_{\bar{x}} = \frac{15}{\sqrt{9}}$. See Figure 4.

EXAMPLE 3

Describing the Distribution of the Sample Mean

Problem The IQ, X , of humans is approximately normal with mean $\mu = 100$ and standard deviation $\sigma = 15$. Compute the probability that a simple random sample of size $n = 10$ results in a sample mean greater than 110. That is, compute $P(\bar{x} > 110)$.

*Technically, we assume that we are drawing a simple random sample from an infinite population. For populations of finite size N , $\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \cdot \frac{\sigma}{\sqrt{n}}$. However, if the sample size is less than 5% of the population size ($n < 0.05N$), the effect of $\sqrt{\frac{N-n}{N-1}}$ (the finite population correction factor) can be ignored without significantly affecting the results.

Approach The random variable X is approximately normally distributed, so the sampling distribution of \bar{x} will be approximately normally distributed. Verify the randomness and independence requirements. The mean of the sampling distribution is $\mu_{\bar{x}} = \mu$, and its standard deviation is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Convert the sample mean $\bar{x} = 110$ to a z -score and then find the area under the standard normal curve to the right of this z -score.

Solution The sample mean is normally distributed, with mean $\mu_{\bar{x}} = 100$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{10}} = 4.743$. The sample is a random sample and the sample size is definitely less than 5% of the population size.

Figure 5 displays the normal curve with the area we want to compute shaded. To find the area by hand, convert $\bar{x} = 110$ to a z -score and obtain

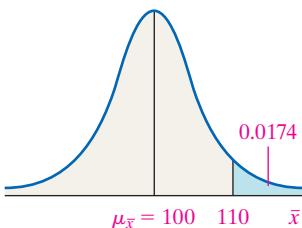
$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} = \frac{110 - 100}{\frac{15}{\sqrt{10}}} = 2.11$$

The area to the right of $z = 2.11$ is $1 - 0.9826 = 0.0174$.

Using technology, the area to the right of $\bar{x} = 110$ is 0.0175.

Interpretation The probability of obtaining a sample mean IQ greater than 110 from a population whose mean is 100 is approximately 0.02. That is, $P(\bar{x} > 110) = 0.0174$ (or 0.0175 using technology). If we take 100 simple random samples of $n = 10$ individuals from this population and if the population mean is 100, about 2 of the samples will result in a mean IQ that is greater than 110.

Figure 5


NW Now Work Problem 19

② Describe the Distribution of the Sample Mean: Nonnormal Population

Now we explore the distribution of the sample mean assuming the population from which the sample is drawn is not normal. Again we use simulation.

EXAMPLE 4

Sampling from a Population That Is Not Normal

Problem The data in Table 1 represent the probability distribution of the number of people living in households in the United States. Figure 6 shows a graph of the probability distribution. From the data in Table 1, we determine the mean and standard deviation number of people living in households in the United States to be $\mu = 2.9$ and $\sigma = 1.48$.

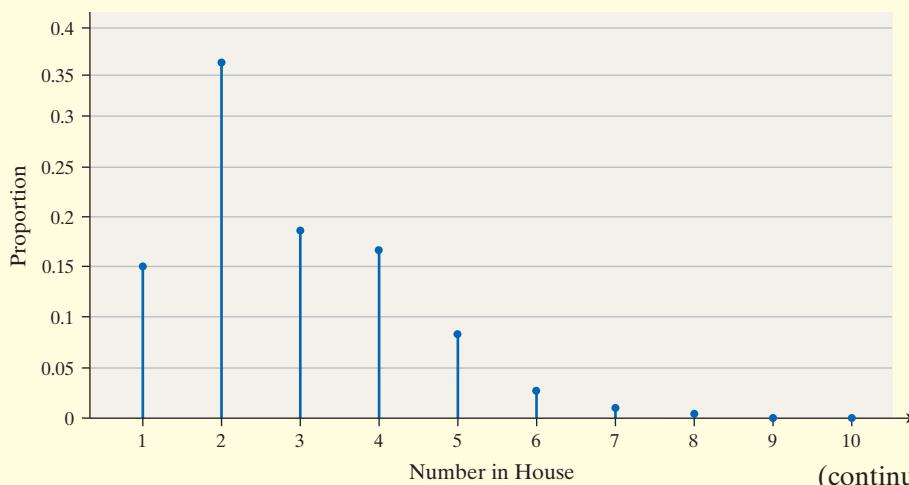
Clearly, the distribution is not normal. In fact, the random variable is discrete! Approximate the sampling distribution of the sample mean \bar{x} by obtaining, through simulation, 1000 samples of size (a) $n = 4$, (b) $n = 10$, and (c) $n = 30$ from the population.

Table 1

Number in Household	Proportion
1	0.147
2	0.361
3	0.187
4	0.168
5	0.083
6	0.034
7	0.012
8	0.004
9	0.002
10	0.002

Source: General Social Survey.

Figure 6

Number of People in Households

(continued)

Approach Use Minitab to obtain 1000 random samples of size $n = 4$ from the population. This simulates going to 4 households 1000 times and determining the number of people living in the household. Next, compute the mean of each of the 1000 random samples. Finally, draw a histogram, determine the mean and standard deviation of the 1000 sample means. Repeat this for samples of size $n = 10$ and $n = 30$.

Solution Figure 7 shows partial output from Minitab for random samples of size $n = 4$. Columns 1 and 2 represent the probability distribution. Each row in Columns 3 through 6 lists the number of individuals in the household for each sample. Column 7 (xbar) lists the sample mean for each sample (each row). For example, in the first sample (row 1), there are 4 individuals in the first house surveyed, and 2 individuals in the second, third, and fourth houses surveyed. The mean number of individuals in the household for the first sample is 2.5.

Figure 7

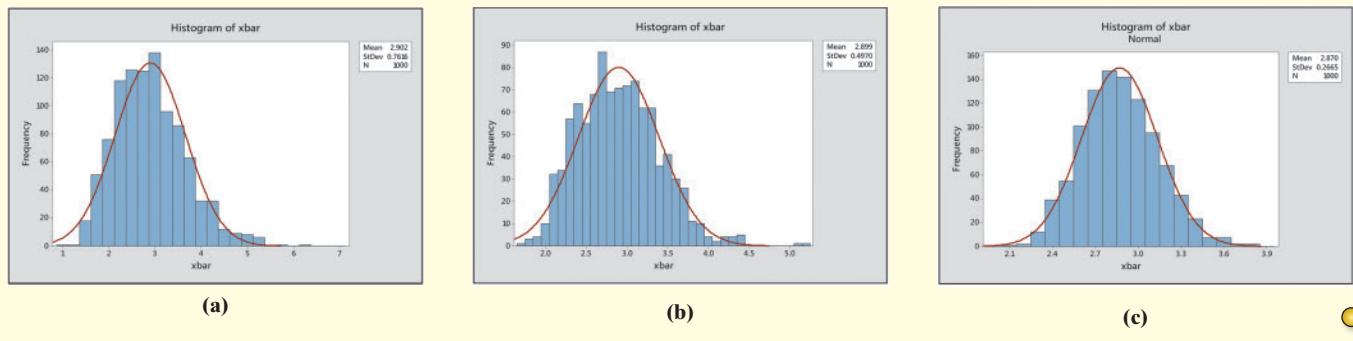
	C1	C2	C3	C4	C5	C6	C7
	Number	Proportion					xbar
1	1	0.147	4	2	2	2	2.50
2	2	0.361	1	3	4	2	2.50
3	3	0.187	4	4	2	4	3.50
4	4	0.168	5	4	2	6	4.25
5	5	0.083	7	2	4	4	4.25
6	6	0.034	2	2	6	2	3.00
7	7	0.012	4	8	3	1	4.00
8	8	0.004	3	2	2	2	2.25
9	9	0.002	2	2	4	2	2.50
10	10	0.002	2	2	4	1	2.25
11			2	1	7	2	3.00
12			3	7	2	1	3.25
13			5	3	2	3	3.25
14			1	5	2	2	2.50
15			1	1	4	7	3.25

Figure 8(a) shows the histogram of the 1000 sample means for a sample of size $n = 4$. The distribution of sample means is skewed right (just like the parent population, but not as strongly). The mean of the 1000 samples is 2.9, and the standard deviation is 0.76. So, the mean of the 1000 samples, $\mu_{\bar{x}}$, equals the population mean μ , and the standard deviation of the 1000 samples, $\sigma_{\bar{x}}$, is close to $\frac{\sigma}{\sqrt{n}} = \frac{1.48}{\sqrt{4}} = 0.74$.

Figure 8(b) shows the histogram of the 1000 sample means for a sample of size $n = 10$. The distribution of these sample means is also skewed right, but not as skewed as the distribution in Figure 8(a). The mean of the 1000 samples is 2.9, and the standard deviation is 0.50. So, the mean of the 1000 samples, $\mu_{\bar{x}}$, equals the population mean, μ , and the standard deviation of the 1000 samples, $\sigma_{\bar{x}}$, is close to $\frac{\sigma}{\sqrt{n}} = \frac{1.48}{\sqrt{10}} = 0.47$.

Figure 8(c) shows the histogram of the 1000 sample means for a sample of size $n = 30$. The distribution of sample means is approximately normal! The mean of the 1000 samples is 2.9, and the standard deviation is 0.27. So, the mean of the 1000 samples, $\mu_{\bar{x}}$, equals the population mean, μ , and the standard deviation of the 1000 samples, $\sigma_{\bar{x}}$, equals $\frac{\sigma}{\sqrt{n}} = \frac{1.48}{\sqrt{30}} = 0.27$.

Figure 8



There are two key concepts to understand in Example 4.

1. The mean of the sampling distribution of the sample mean is equal to the mean of the underlying population, and the standard deviation of the sampling distribution of the sample mean is $\frac{\sigma}{\sqrt{n}}$, regardless of the size of the sample.
2. The shape of the distribution of the sample mean becomes approximately normal as the sample size n increases, regardless of the shape of the underlying population.

We formally state point 2 as the *Central Limit Theorem*.

IN OTHER WORDS

For any population, regardless of its shape, as the sample size increases, the shape of the distribution of the sample mean becomes more “normal.”

CAUTION!

The Central Limit Theorem only has to do with the shape of the distribution of \bar{x} , not the center or spread. Regardless of the size of the sample, $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Historical Note



Pierre-Simon Laplace was born on March 23, 1749, in Normandy, France. At age 16, Laplace attended Caen University, where he studied theology. While there, his mathematical talents were discovered, which led him to Paris, where he obtained a job as professor of mathematics at the École Militaire. In 1773, Laplace was elected to the Académie des Sciences. Laplace was not humble. It is reported that, in 1780, he stated that he was the best mathematician in Paris. In 1799, Laplace published the first two volumes of *Mécanique céleste*, in which he discussed methods for calculating the motion of the planets. On April 9, 1810, Laplace presented the Central Limit Theorem to the Academy.

The Central Limit Theorem

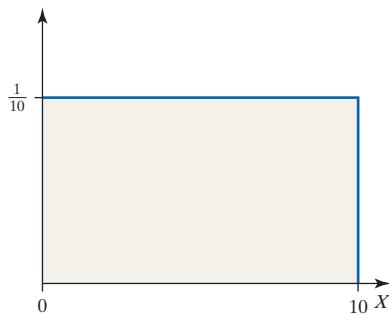
Regardless of the shape of the underlying population, the sampling distribution of \bar{x} becomes approximately normal as the sample size, n , increases.

How large does the sample size need to be before we can say that the sampling distribution of \bar{x} is approximately normal? The answer depends on the shape of the distribution of the underlying population. Distributions that are highly skewed will require a larger sample size for the distribution of \bar{x} to become approximately normal.

For example, the right-skewed distribution in Example 4 required a sample size of about 30 before the distribution of the sample mean became approximately normal. However, Figure 9(a) shows a uniform distribution for $0 \leq X \leq 10$. Figure 9(b) shows the distribution of the sample mean obtained via simulation using StatCrunch for $n = 4$. Even for samples as small as $n = 4$, the distribution of the sample mean is approximately normal.

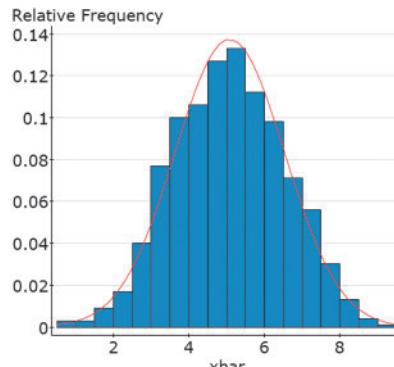
Figure 9

Uniform Distribution



(a)

Distribution of Sample Mean With $n = 4$

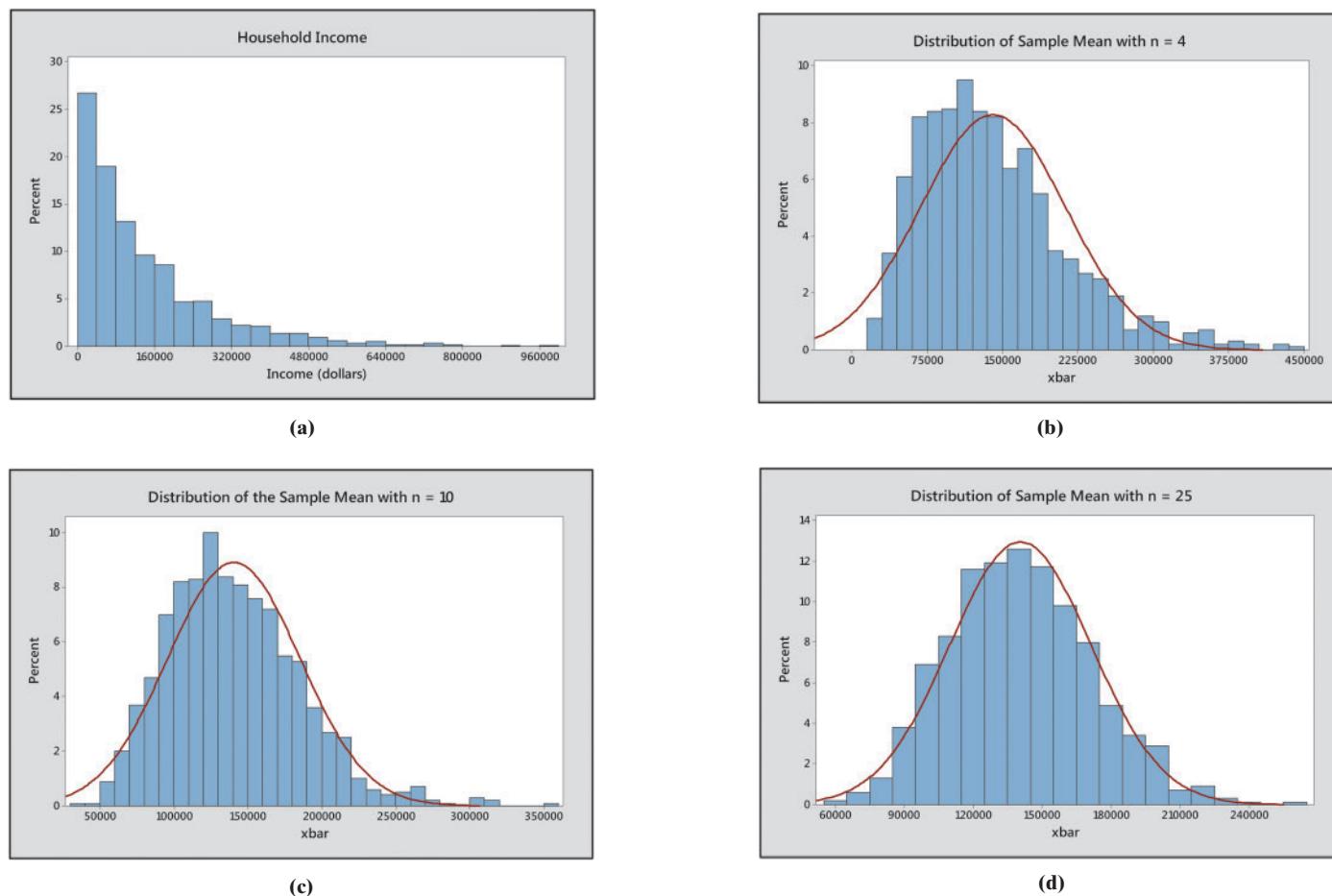


(b)

Figure 10(a) on the next page shows a distribution of household incomes for a town. Figure 10(b) shows the distribution of the sample mean for a random sample of $n = 4$ households from Minitab. Figure 10(c) shows the distribution of the sample mean for a random sample of $n = 10$ households, and Figure 10(d) shows the distribution of the

sample mean for a random sample of $n = 25$ households, also from Minitab. Notice the distribution of the sample mean is approximately normal for $n = 25$.

Figure 10



The results of Example 4, Figure 9, and Figure 10 confirm that the shape of the distribution of the population dictates the size of the sample required for the distribution of the sample mean to be approximately normal. The more skewed the distribution of the population is, the larger the sample size needed to invoke the Central Limit Theorem. We will err on the side of caution and use the following rule of thumb in deciding whether the distribution of the sample mean is approximately normal.

If the distribution of the population is unknown or not normal, then the distribution of the sample mean is approximately normal provided that the sample size is greater than or equal to 30.

EXAMPLE 5

Weight Gain During Pregnancy

Problem The mean weight gain during pregnancy is 30 pounds, with a standard deviation of 12.9 pounds. Weight gain during pregnancy is skewed right. An obstetrician obtains a random sample of 35 low-income patients and determines their mean weight gain during pregnancy was 36.2 pounds. Does this result suggest anything unusual?

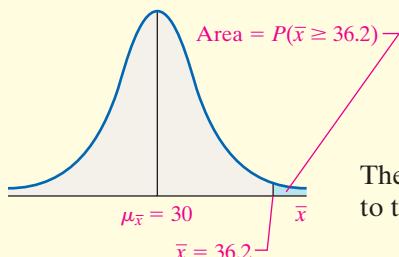
Approach We want to know whether the sample mean obtained is unusual. Therefore, determine the likelihood of obtaining a sample mean of 36.2 pounds or higher (if a 36.2-pound weight gain is unusual, certainly any weight gain above 36.2 pounds is also

unusual). Assume that the patients come from the population whose mean weight gain is 30 pounds. Verify the randomness and independence requirements. Use the normal model to obtain the probability since the sample size is large enough to use the Central Limit Theorem. Determine the area under the normal curve to the right of 36.2 pounds

$$\text{with } \mu_{\bar{x}} = \mu = 30 \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12.9}{\sqrt{35}}.$$

Solution The sample is a random sample. It seems reasonable there are at least 700 low-income pregnant women in the population. So the sample size is less than 5% of the population size. The probability is represented by the area under the normal curve to the right of 36.2. See Figure 11.

Figure 11



The area under the standard normal curve to the left of $z = 2.84$ is 0.9977. So the area to the right is 0.0023. Therefore, $P(\bar{x} \geq 36.2) = 0.0023$.

If we use technology to find the area to the right of $\bar{x} = 36.2$, we obtain 0.0022.

Interpretation If the population from which this sample is drawn has a mean weight gain of 30 pounds, the probability that a random sample of 35 women has a sample mean weight gain of 36.2 pounds (or more) is approximately 0.002. This means that about 2 samples in 1000 will result in a sample mean of 36.2 pounds or higher if the population mean is 30 pounds. We can conclude one of two things based on this result:

1. The mean weight gain for low-income patients is 30 pounds, and we happened to select women who, on average, gained more weight.
2. The mean weight gain for low-income patients is more than 30 pounds.

We are inclined to accept the second explanation over the first since our sample was obtained randomly. Therefore, the obstetrician should be concerned. Perhaps she should look at the diets and/or lifestyles of low-income patients while they are pregnant.

NW Now Work Problem 25

Summary: Shape, Center, and Spread of the Sampling Distribution of \bar{x}

Shape, Center, and Spread of the Population	Distribution of the Sample Mean		
	Shape	Center	Spread
Population is normal with mean μ and standard deviation σ	Regardless of the sample size n , the shape of the distribution of the sample mean is approximately normal	$\mu_{\bar{x}} = \mu$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Population is not normal with mean μ and standard deviation σ	As the sample size n increases, the distribution of the sample mean becomes approximately normal	$\mu_{\bar{x}} = \mu$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$



8.1 Assess Your Understanding

Vocabulary and Skill Building

1. The _____ of the sample mean, \bar{x} , is the probability distribution of all possible values of the random variable \bar{x} computed from a sample of size n from a population with mean μ and standard deviation σ .

2. Suppose a simple random sample of size n is drawn from a large population with mean μ and standard deviation σ . The sampling distribution of \bar{x} has mean $\mu_{\bar{x}} =$ _____ and standard deviation $\sigma_{\bar{x}} =$ _____.
3. The standard deviation of the sampling distribution of \bar{x} , $\sigma_{\bar{x}}$, is called the _____ of the _____.

4. True or False: The distribution of the sample mean, \bar{x} , will be approximately normally distributed if the sample is obtained from a population that is normally distributed, regardless of the sample size.

5. True or False: The distribution of the sample mean, \bar{x} , will be approximately normally distributed if the sample is obtained from a population that is not normally distributed, regardless of the sample size.

6. True or False: To cut the standard error of the mean in half, the sample size must be doubled.

7. A simple random sample of size $n = 10$ is obtained from a population that is normally distributed with $\mu = 30$ and $\sigma = 8$. What is the sampling distribution of \bar{x} ?

8. A simple random sample of size $n = 40$ is obtained from a population with $\mu = 50$ and $\sigma = 4$. Does the population need to be normally distributed for the sampling distribution of \bar{x} to be approximately normally distributed? Why? What is the sampling distribution of \bar{x} ?

In Problems 9–12, determine $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ from the given parameters of the population and the sample size.

NW 9. $\mu = 80$, $\sigma = 14$, $n = 49$

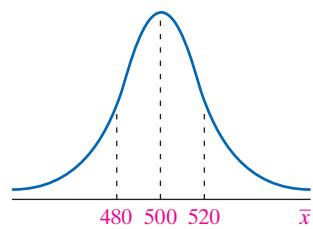
10. $\mu = 64$, $\sigma = 18$, $n = 36$

11. $\mu = 52$, $\sigma = 10$, $n = 21$

12. $\mu = 27$, $\sigma = 6$, $n = 15$

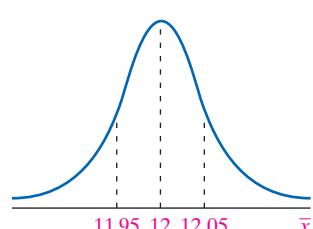
13. Answer the following questions for the sampling distribution of the sample mean shown to the right.

- (a) What is the value of $\mu_{\bar{x}}$?
- (b) What is the value of $\sigma_{\bar{x}}$?
- (c) If the sample size is $n = 16$, what is likely true about the shape of the population?
- (d) If the sample size is $n = 16$, what is the standard deviation of the population from which the sample was drawn?



14. Answer the following questions for the sampling distribution of the sample mean shown to the right.

- (a) What is the value of $\mu_{\bar{x}}$?
- (b) What is the value of $\sigma_{\bar{x}}$?
- (c) If the sample size is $n = 9$, what is likely true about the shape of the population?
- (d) If the sample size is $n = 9$, what is the standard deviation of the population from which the sample was drawn?



15. A simple random sample of size $n = 49$ is obtained from a population with $\mu = 80$ and $\sigma = 14$.

- (a) Describe the sampling distribution of \bar{x} .
- (b) What is $P(\bar{x} > 83)$?
- (c) What is $P(\bar{x} \leq 75.8)$?
- (d) What is $P(78.3 < \bar{x} < 85.1)$?

16. A simple random sample of size $n = 36$ is obtained from a population with $\mu = 64$ and $\sigma = 18$.

- (a) Describe the sampling distribution of \bar{x} .
- (b) What is $P(\bar{x} < 62.6)$?
- (c) What is $P(\bar{x} \geq 68.7)$?
- (d) What is $P(59.8 < \bar{x} < 65.9)$?

17. A simple random sample of size $n = 12$ is obtained from a population with $\mu = 64$ and $\sigma = 17$.

- (a) What must be true regarding the distribution of the population in order to use the normal model to compute

probabilities involving the sample mean? Assuming that this condition is true, describe the sampling distribution of \bar{x} .

- (b) Assuming that the requirements described in part (a) are satisfied, determine $P(\bar{x} < 67.3)$.
- (c) Assuming that the requirements described in part (a) are satisfied, determine $P(\bar{x} \geq 65.2)$.

18. A simple random sample of size $n = 20$ is obtained from a population with $\mu = 64$ and $\sigma = 17$.

- (a) What must be true regarding the distribution of the population in order to use the normal model to compute probabilities involving the sample mean? Assuming that this condition is true, describe the sampling distribution of \bar{x} .
- (b) Assuming that the requirements described in part (a) are satisfied, determine $P(\bar{x} < 67.3)$.
- (c) Assuming that the requirements described in part (a) are satisfied, determine $P(\bar{x} \geq 65.2)$.
- (d) Compare the results obtained in parts (b) and (c) with the results obtained in parts (b) and (c) in Problem 17. What effect does increasing the sample size have on the probabilities? Why do you think this is the case?

Applying the Concepts

NW 19. **Gestation Period** The length of human pregnancies is approximately normal with mean $\mu = 266$ days and standard deviation $\sigma = 16$ days.

- (a) What is the probability a randomly selected pregnancy lasts less than 260 days?
- (b) Suppose a random sample of 20 pregnancies is obtained. Describe the sampling distribution of the sample mean length of human pregnancies.
- (c) What is the probability that a random sample of 20 pregnancies has a mean gestation period of 260 days or less?
- (d) What is the probability that a random sample of 50 pregnancies has a mean gestation period of 260 days or less?
- (e) What might you conclude if a random sample of 50 pregnancies resulted in a mean gestation period of 260 days or less?
- (f) What is the probability a random sample of size 15 will have a mean gestation period within 10 days of the mean?

20. **Upper Leg Length** The upper leg length of 20- to 29-year-old males is approximately normal with a mean length of 43.7 cm and a standard deviation of 4.2 cm.

Source: "Anthropometric Reference Data for Children and Adults: U.S. Population, 1999–2002"; Volume 361, July 7, 2005.

- (a) What is the probability that a randomly selected 20- to 29-year-old male has an upper leg length that is less than 40 cm?
- (b) A random sample of 9 males who are 20 to 29 years old is obtained. What is the probability that the mean upper leg length is less than 40 cm?
- (c) What is the probability that a random sample of 12 males who are 20–29 years old results in a mean upper leg length that is less than 40 cm?
- (d) What effect does increasing the sample size have on the probability? Provide an explanation for this result.
- (e) A random sample of 15 males who are 20–29 years old results in a mean upper leg length of 46 cm. Do you find this result unusual? Why?

21. **Reading Rates** The reading speed of second-grade students is approximately normal, with a mean of 90 words per minute (wpm) and a standard deviation of 10 wpm.

- (a) What is the probability a randomly selected student will read more than 95 words per minute?

- (b) What is the probability that a random sample of 12 second-grade students results in a mean reading rate of more than 95 words per minute?
- (c) What is the probability that a random sample of 24 second-grade students results in a mean reading rate of more than 95 words per minute?
- (d) What effect does increasing the sample size have on the probability? Provide an explanation for this result.
- (e) A teacher instituted a new reading program at school. After 10 weeks in the program, it was found that the mean reading speed of a random sample of 20 second-grade students was 92.8 wpm. What might you conclude based on this result?
- (f) There is a 5% chance that the mean reading speed of a random sample of 20 second-grade students will exceed what value?

22. Old Faithful The most famous geyser in the world, Old Faithful in Yellowstone National Park, has a mean time between eruptions of 85 minutes. If the interval of time between eruptions is approximately normal with standard deviation 21.25 minutes, answer the following questions: *Source: www.unmuseum.org*

- (a) What is the probability that a randomly selected time interval between eruptions is longer than 95 minutes?
- (b) What is the probability that a random sample of 20 time intervals between eruptions has a mean longer than 95 minutes?
- (c) What is the probability that a random sample of 30 time intervals between eruptions has a mean longer than 95 minutes?
- (d) What effect does increasing the sample size have on the probability? Provide an explanation for this result.
- (e) What might you conclude if a random sample of 30 time intervals between eruptions has a mean longer than 95 minutes?
- (f) On a certain day, suppose there are 22 time intervals for Old Faithful. Treating these 22 eruptions as a random sample, the likelihood the mean length of time between eruptions exceeds _____ minutes is 0.20.

23. Rates of Return in Stocks The S&P 500 is a collection of 500 stocks of publicly traded companies. Using data obtained from Yahoo! Finance, the monthly rates of return of the S&P 500 since 1950 are approximately normal. The mean rate of return is 0.007233 (0.7233%), and the standard deviation for rate of return is 0.04135 (4.135%).

- (a) What is the probability that a randomly selected month has a positive rate of return? That is, what is $P(x > 0)$?
- (b) Treating the next 12 months as a simple random sample, what is the probability that the mean monthly rate of return will be positive? That is, with $n = 12$, what is $P(\bar{x} > 0)$?
- (c) Treating the next 24 months as a simple random sample, what is the probability that the mean monthly rate of return will be positive?
- (d) Treating the next 36 months as a simple random sample, what is the probability that the mean monthly rate of return will be positive?
- (e) Use the results of parts (b)–(d) to describe the likelihood of earning a positive rate of return on stocks as the investment time horizon increases.

24. Winning Poker A very good poker player is expected to earn \$1 per hand in \$100/\$200 Texas Hold'em. The standard deviation is approximately \$32.

- (a) What is the probability a very good poker player earns a profit (more than \$0) after playing 50 hands in \$100/\$200 Texas Hold'em?

- (b) What is the probability a very good poker player loses (earns less than \$0) after playing 100 hands in \$100/\$200 Texas Hold'em?
- (c) What proportion of the time can a very good poker player expect to earn at least \$500 after playing 100 hands in \$100/\$200 Texas Hold'em? (**Hint:** \$500 after 100 hands is a mean of \$5 per hand.)
- (d) Would it be unusual for a very good poker player to lose at least \$1000 after playing 100 hands in \$100/\$200 Texas Hold'em?
- (e) Suppose twenty hands are played per hour. What is the probability that a very good poker player earns a profit during a twenty-four hour marathon session?

NW 25. Oil Change The shape of the distribution of the time required to get an oil change at a 10-minute oil-change facility is skewed right. However, records indicate that the mean time for an oil change is 11.4 minutes, and the standard deviation for oil-change time is 3.2 minutes.

- (a) To compute probabilities regarding the sample mean using the normal model, what size sample would be required?
- (b) What is the probability that a random sample of $n = 40$ oil changes results in a sample mean time of less than 10 minutes?
- (c) Suppose the manager agrees to pay each employee a \$50 bonus if they meet a certain goal. On a typical Saturday, the oil-change facility will perform 40 oil changes between 10 A.M. and 12 P.M. Treating this as a random sample, what mean oil-change time would there be a 10% chance of being at or below? This will be the goal established by the manager.

26. Time Spent in the Drive-Thru The quality-control manager of a Long John Silver's restaurant wants to analyze the length of time that a car spends at the drive-thru window waiting for an order. It is determined that the mean time spent at the window is 59.3 seconds with a standard deviation of 13.1 seconds. The distribution of time at the window is skewed right (data based on information provided by Danica Williams, student at Joliet Junior College).

- (a) To obtain probabilities regarding a sample mean using the normal model, what size sample is required?
- (b) The quality-control manager wishes to use a new delivery system designed to get cars through the drive-thru system faster. A random sample of 40 cars results in a sample mean time spent at the window of 56.8 seconds. What is the probability of obtaining a sample mean of 56.8 seconds or less, assuming that the population mean is 59.3 seconds? Do you think that the new system is effective?
- (c) Treat the next 50 cars that arrive as a simple random sample. There is a 15% chance the mean time will be at or below _____ seconds.

27. Insect Fragments The Food and Drug Administration sets Food Defect Action Levels (FDALs) for some of the various foreign substances that inevitably end up in the food we eat and liquids we drink. For example, the FDAL for insect filth in peanut butter is 3 insect fragments (larvae, eggs, body parts, and so on) per 10 grams. A random sample of 50 ten-gram portions of peanut butter is obtained and results in a sample mean of $\bar{x} = 3.6$ insect fragments per ten-gram portion.

- (a) Why is the sampling distribution of \bar{x} approximately normal?
- (b) What is the mean and standard deviation of the sampling distribution of \bar{x} assuming that $\mu = 3$ and $\sigma = \sqrt{3}$?
- (c) What is the probability that a simple random sample of 50 ten-gram portions results in a mean of at least 3.6 insect fragments? Is this result unusual? What might we conclude?

28. Burger King's Drive-Thru Suppose that cars arrive at Burger King's drive-thru at the rate of 20 cars every hour between 12:00 noon and 1:00 P.M. A random sample of 40 one-hour time periods between 12:00 noon and 1:00 P.M. is selected and has 22.1 as the mean number of cars arriving.

- (a) Why is the sampling distribution of \bar{x} approximately normal?
- (b) What is the mean and standard deviation of the sampling distribution of \bar{x} assuming that $\mu = 20$ and $\sigma = \sqrt{20}$?
- (c) What is the probability that a simple random sample of 40 one-hour time periods results in a mean of at least 22.1 cars? Is this result unusual? What might we conclude?

29. Watching Television The amount of time Americans spend watching television is closely monitored by firms such as AC Nielsen because this helps determine advertising pricing for commercials.

- (a) Do you think the variable "weekly time spent watching television" would be normally distributed? If not, what shape would you expect the variable to have?
- (b) According to the American Time Use Survey, adult Americans spend 2.35 hours per day watching television on a weekday. Assume that the standard deviation for "time spent watching television on a weekday" is 1.93 hours. If a random sample of 40 adult Americans is obtained, describe the sampling distribution of \bar{x} , the mean amount of time spent watching television on a weekday.
- (c) Determine the probability that a random sample of 40 adult Americans results in a mean time watching television on a weekday of between 2 and 3 hours.
- (d) One consequence of the popularity of the Internet is that it is thought to reduce television watching. Suppose that a random sample of 35 individuals who consider themselves to be avid Internet users results in a mean time of 1.89 hours watching television on a weekday. Determine the likelihood of obtaining a sample mean of 1.89 hours or less from a population whose mean is presumed to be 2.35 hours. Based on the result obtained, do you think avid Internet users watch less television?

30. ATM Withdrawals According to Crown ATM Network, the mean ATM withdrawal is \$67. Assume that the standard deviation for withdrawals is \$35.

- (a) Do you think the variable "ATM withdrawal" is normally distributed? If not, what shape would you expect the variable to have?
- (b) If a random sample of 50 ATM withdrawals is obtained, describe the sampling distribution of \bar{x} , the mean withdrawal amount.
- (c) Determine the probability of obtaining a sample mean withdrawal amount between \$70 and \$75.

31. Sampling Distributions The following data represent the ages of the winners of the Academy Award for Best Actor for the years 2012–2017.

2012: Daniel Day-Lewis	55
2013: Matthew McConaughey	44
2014: Eddie Redmayne	33
2015: Leonardo DiCaprio	41
2016: Casey Affleck	41
2017: Gary Oldman	59

Source: awardsdatabase.oscars.org

- (a) Compute the population mean, μ .

- (b) List all possible samples with size $n = 2$. There should be ${}_6C_2 = 15$ samples.
- (c) Construct a sampling distribution for the mean by listing the sample means and their corresponding probabilities.
- (d) Compute the mean of the sampling distribution.
- (e) Compute the probability that the sample mean is within 3 years of the population mean age.
- (f) Repeat parts (b)–(e) using samples of size $n = 3$. Comment on the effect of increasing the sample size.

32. Sampling Distributions The following data represent the running lengths (in minutes) of the winners of the Academy Award for Best Picture for the years 2012–2017.

2012: Argo	120
2013: 12 Years a Slave	134
2014: Birdman	119
2015: Spotlight	129
2016: Moonlight	111
2017: The Shape of Water	123

Source: The Internet Movie Database.

- (a) Compute the population mean, μ .
- (b) List all possible samples with size $n = 2$. There should be ${}_6C_2 = 15$ samples.
- (c) Construct a sampling distribution for the mean by listing the sample means and their corresponding probabilities.
- (d) Compute the mean of the sampling distribution.
- (e) Compute the probability that the sample mean is within 5 minutes of the population mean running time.
- (f) Repeat parts (b)–(e) using samples of size $n = 3$. Comment on the effect of increasing the sample size.

33. Threaded Problem: Tornado The data set "Tornadoes_2017" located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a) Draw a relative histogram of the variable "Length." Describe the shape of the distribution.
- (b) Determine the population mean and standard deviation length of all tornadoes in 2017.
- (c) Based on the shape of the distribution of the variable "Length" from the histogram in part (a), what must be true about the sample size in order for the distribution of the sample mean to be approximately normal?
- (d) Suppose a random sample of $n = 35$ tornadoes is obtained from the population of all tornadoes in 2017. Describe the sampling distribution of the sample mean of the length of the tornadoes.
- (e) The following data represent the length (in miles) of a simple random sample of 35 tornadoes whose F scale is 0. What is the sample mean? **Note:** The data file 8_1_33e is available at www.pearsonhighered.com/sullivanstats.



2.26	0.30	5.30	1.02	2.99	0.88	3.00
1.81	6.50	0.20	0.61	6.06	0.81	0.91
1.91	1.20	2.69	4.00	1.25	0.07	0.16
0.90	0.70	6.48	3.00	13.36	4.76	1.40
0.09	4.53	0.23	0.60	0.01	0.10	0.03

- (f) Determine the probability of obtaining a sample mean less than that obtained in part (e) from the population of all tornadoes in 2017 using the model described in part (d).
 (g) Interpret the result from part (f).

DATA **34. Putting It Together: Bike Sharing** Bicycle sharing exists in a variety of cities around the country. Los Angeles has the Metro Bike Share system. Users pick up a bike from one station, go for a ride, and return the bike to any station. Go to www.pearsonhighered.com/sullivanstats and download the file 8_1_34. The data represent the duration (in minutes) of all rides in the fourth quarter (October through December) of 2018.

- (a) Draw a relative frequency histogram of the data with a first class having a lower class limit of 0 and a class width of 50. Describe the shape of the distribution.
 (b) Determine the population mean and standard deviation duration.
 (c) Based on the shape of the distribution of the variable “Duration” from the histogram in part (a), what must be true about the sample size in order for the distribution of the sample mean to be approximately normal?
 (d) Suppose a random sample of $n = 40$ bike rides is obtained from the population of all bike rides in the fourth quarter of 2018. Describe the sampling distribution of the sample mean of the duration of all bike rides.
 (e) The column “Sample1” represents the duration (in minutes) of a simple random sample of 40 bike rides where the ride was a “Round Trip” (that is, return the bike to the same location it was rented from). What is the sample mean?
 (f) Determine the probability of obtaining a sample mean greater than that obtained in part (e) from the population of all bike rides in the fourth quarter of 2018 using the model described in part (d).
 (g) The column “Sample2” represents the duration (in minutes) of a second (independent) simple random sample of 40 bike rides where the ride was a “Round Trip” (that is, return the bike to the same location it was rented from). What is the sample mean?
 (h) Determine the probability of obtaining a sample mean greater than that obtained in part (g) from the population of all bike rides in the fourth quarter of 2018 using the model described in part (d).
 (i) Explain why the probabilities in parts (f) and (h) differ. What does this suggest?

35. Putting It Together: Playing Roulette In the game of roulette, a wheel consists of 38 slots numbered 0, 00, 1, 2, ..., 36. (See the photo.) To play the game, a metal ball is spun around the wheel and is allowed to fall into one of the numbered slots. If the number of the slot the ball falls into matches the number you selected, you win \$35; otherwise you lose \$1.

- (a) Construct a probability distribution for the random variable X , the winnings of each spin.
 (b) Determine the mean and standard deviation of the random variable X . Round your results to the nearest penny.
 (c) Suppose that you play the game 100 times so that $n = 100$. Describe the sampling distribution of \bar{x} , the mean amount won per game.
 (d) What is the probability of being ahead after playing the game 100 times? That is,

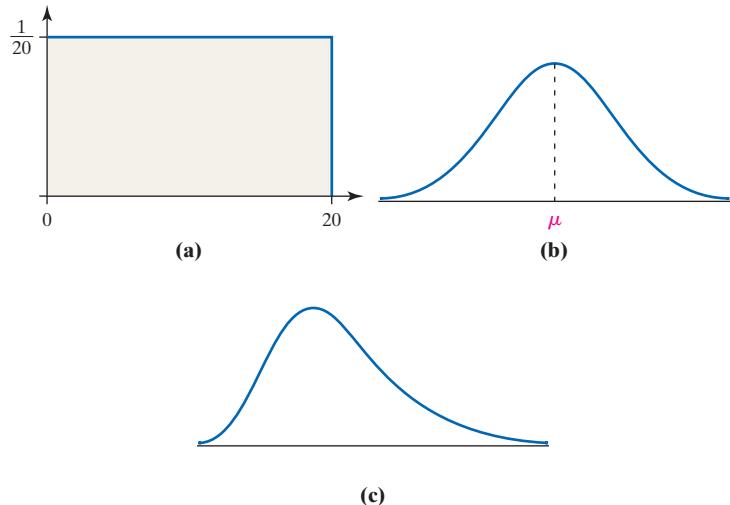


what is the probability that the sample mean is greater than 0 for $n = 100$?

- (e) What is the probability of being ahead after playing the game 200 times?
 (f) What is the probability of being ahead after playing the game 1000 times?
 (g) Compare the results of parts (d) and (e). What lesson does this teach you?

Explaining the Concepts

36. Explain what a sampling distribution is.
 37. State the Central Limit Theorem.
 38. We assume that we are obtaining simple random samples from infinite populations when obtaining sampling distributions. If the size of the population is finite, we technically need a finite population correction factor. However, if the sample size is small relative to the size of the population, this factor can be ignored. Explain what an “infinite population” is. What is the finite population correction factor? How small must the sample size be relative to the size of the population so that we can ignore the factor? Finally, explain why the factor can be ignored for such samples.
 39. Without doing any computation, decide which has a higher probability, assuming each sample is from a population that is normally distributed with $\mu = 100$ and $\sigma = 15$. Explain your reasoning.
 (a) $P(90 \leq \bar{x} \leq 110)$ for a random sample of size $n = 10$.
 (b) $P(90 \leq \bar{x} \leq 110)$ for a random sample of size $n = 20$.
 40. For the three probability distributions shown, rank each distribution from lowest to highest in terms of the sample size required for the distribution of the sample mean to be approximately normally distributed. Justify your choice.



41. Suppose Jack and Diane are each using simulation to describe the sampling distribution from a population that is skewed left with mean 50 and standard deviation 10. Jack obtains 1000 random samples of size $n = 3$ from the population, finds the mean of the 1000 samples, draws a histogram of the means, finds the mean of the means, and determines the standard deviation of the means. Diane does the same simulation, but obtains 1000 random samples of size $n = 30$ from the population.
 (a) Describe the shape you expect for Jack’s distribution of sample means. Describe the shape you expect for Diane’s distribution of sample means.

- (b) What do you expect the mean of Jack's distribution to be?
What do you expect the mean of Diane's distribution to be?
- (c) What do you expect the standard deviation of Jack's distribution to be? What do you expect the standard deviation of Diane's distribution to be?
- 42. Sleepy** Suppose you want to study the number of hours of sleep you get each evening. To do so, you look at the calendar and randomly select 10 days out of the next 300 days and record the number of hours you sleep.
- (a) Explain why number of hours of sleep in a night by you is a random variable.
- (b) Is the random variable "number of hours of sleep in a night" quantitative or qualitative?
- (c) After you obtain your ten nights of data, you compute the mean number of hours of sleep. Is this a statistic or a parameter? Why?
- (d) Is the mean number of hours computed in part (c) a random variable? Why? If it is a random variable, what is the source of variation?

Retain Your Knowledge

DATA 44. Bull Markets A bull market is defined as a market condition in which the price of a security rises for an extended period of time. A bull market in the stock market is often defined as a condition in which a market rises by 20% or more without a 20% decline. The data to the right represent the number of months and percentage change in the S&P 500 (a group of 500 stocks) during the 25 bull markets dating back to 1929 (the year of the famous market crash).

- (a) Treating the length of the bull market as the explanatory variable, draw a scatter diagram of the data.
- (b) Determine the linear correlation coefficient between months and percent change.
- (c) Does a linear relation exist between duration of the bull market and market performance?
- (d) Find the least-squares regression line treating length of the bull market as the explanatory variable.
- (e) Interpret the slope.
- (f) Did the bull market that lasted 50.4 months have a percent change above or below what would be expected? Explain.
- (g) Draw a residual plot. Any outliers?
- (h) Would you consider the bull market from December 4, 1987 through March 24, 2000, which lasted 149.8 months and saw

43. Sleepy Again Suppose you want to study the number of hours of sleep full-time college students at your college get each evening. To do so, you obtain a list of full-time students at your college, obtain a simple random sample of ten students, and ask each of them to disclose how many hours of sleep they obtained the most recent Monday.

- (a) What is the population of interest in this study? What is the sample?
- (b) Explain why number of hours of sleep in this study is a random variable.
- (c) After you obtain your ten observations, you compute the mean number of hours of sleep. Is this a statistic or a parameter? Why?
- (d) Is the mean number of hours computed in part (c) a random variable? Why? If it is a random variable, what is the source of variation? How does the source of variation in this study differ from that of Problem 42?

a 582.15% rise in stock prices, influential? Explain. Note: After this bull market, the market entered a bear market that lasted 18.2 months and saw the stock market decline 37%. This era is often referred to as the "Tech Bubble."

Bull Months	Percent Change	Bull Months	Percent Change
4.9	46.77	86.9	267.08
2.3	25.83	50.4	86.35
0.8	25.82	44.1	79.78
1.2	30.61	26.1	48.05
2.3	111.59	32.0	73.53
4.7	120.61	74.9	125.63
3.7	37.28	61.3	228.81
24.2	131.64	149.8	582.15
7.4	62.24	3.5	21.40
6.6	26.78	60.9	101.50
5.0	20.91	1.6	24.22
49.7	157.70	48.6	127.85
13.1	23.89		

8.2 Distribution of the Sample Proportion



Preparing for This Section Before getting started, review the following:

- Applications of the normal distribution (Section 7.2, pp. 343–350)

Objectives

- ① Describe the sampling distribution of a sample proportion
- ② Compute probabilities of a sample proportion

In Section 8.1, we described the sampling distribution of the sample mean. That is, we determined the shape, center, and spread of the sampling distribution of the sample mean, \bar{x} . In this section, we describe the sampling distribution of the *sample proportion*—the proportion of individuals in a sample who have a specified characteristic.

① Describe the Sampling Distribution of a Sample Proportion

Suppose we want to determine the proportion of households in a 100-house homeowners association that favor an increase in the annual assessments to pay for neighborhood improvements. We could survey all households to learn which are in favor of higher assessments. If 65 of the 100 households favor the higher assessment, the population proportion, p , of households in favor of a higher assessment is

$$p = \frac{65}{100} = 0.65$$

Of course, gaining access to all the individuals in a population is rare, so we usually obtain estimates of population parameters such as p .

Definition

Suppose that a random sample of size n is obtained from a population in which each individual either does or does not have a certain characteristic. The **sample proportion**, denoted \hat{p} (read “ p -hat”), is given by

$$\hat{p} = \frac{x}{n}$$

where x is the number of individuals in the sample with the specified characteristic.* The sample proportion, \hat{p} , is a statistic that estimates the population proportion, p .

EXAMPLE 1 Computing a Sample Proportion

Problem The Harris Poll conducted a survey of 1200 adult Americans who vacation during the summer and asked whether the individuals plan to work while on summer vacation. Of those surveyed, 552 indicated that they plan to work while on vacation. Find the sample proportion of individuals surveyed who plan to work while on summer vacation.

Approach Use the formula $\hat{p} = \frac{x}{n}$, where x is the number of individuals who plan to work on summer vacation and n is the sample size.

Solution Substituting $x = 552$ and $n = 1200$ into $\hat{p} = \frac{x}{n}$, we find that $\hat{p} = \frac{552}{1200} = 0.46$, so Harris estimates that 0.46 or 46% of adult Americans who vacation during the summer plan to work while on vacation. 

A second survey of 1200 American adults would likely have a different estimate of the proportion of Americans who plan to work while on summer vacation because different individuals would be in the sample. Because the value of \hat{p} varies from sample to sample, it is a random variable and has a probability distribution.

To get a sense of the shape, center, and spread of the sampling distribution of \hat{p} , we could repeat the exercise of obtaining simple random samples of 1200 adult Americans over and over. This would lead to a list of sample proportions. A histogram of the sample proportions will give us a feel for the shape of the distribution of the sample proportion. The mean of the sample proportions will give us an idea of the center of the distribution, and the standard deviation of the sample proportions will give us an idea of the spread of the distribution.

Rather than literally surveying 1200 adult Americans over and over again, we will use simulation to get an idea of the shape, center, and spread of the sampling distribution of the proportion.

*For those who studied Section 6.2 on binomial probabilities, x can be thought of as the number of successes in n trials of a binomial experiment.

EXAMPLE 2 Using Simulation to Describe the Distribution of the Sample Proportion

Problem Based on a study conducted by the Gallup organization, 76% of Americans believe that the state of moral values in the United States is getting worse. Describe the sampling distribution of the sample proportion for samples of size (a) $n = 10$, (b) $n = 25$, (c) $n = 60$.

Approach Describing a distribution means finding its shape, center, and spread. The actual sampling distribution of the sample proportion would be the distribution of *all* possible sample proportions of size $n = 10$. It is virtually impossible to find all possible samples of size $n = 10$ from the population of Americans. To get a sense as to the shape, center, and spread of the sampling distribution of the sample proportion, we will use StatCrunch's Bernoulli Random Data command with event probability 0.76 to simulate obtaining 2000 samples of size $n = 10$ by randomly generating 2000 rows of responses over 10 columns. Each row consists of a 0 or 1, with 0 representing a failure (individual does not believe the state of moral values is getting worse) and 1 representing a success (individual believes the state of moral values is getting worse). For each of the 2000 samples (the 2000 rows), determine the mean number of successes (the sample proportion). Draw a histogram of the 2000 sample proportions to gauge the shape of the distribution of the sample proportions. Determine the mean of the 2000 sample proportions to approximate the mean of the sampling distribution. Determine the standard deviation of the 2000 sample proportions to approximate the standard deviation of the sampling distribution. Repeat this process for samples of size $n = 25$ and $n = 60$.

Solution Figure 12 shows partial output from StatCrunch. Row 1 contains the first sample, where the results of the survey are 1 (success), 1 (success), 1 (success), . . . , 0 (failure), 1 (success). The mean number of successes, that is, the sample proportion, from the first sample of $n = 10$ adult Americans is 0.7.

Figure 12

Row	Bernoulli1	Bernoulli2	Bernoulli3	Bernoulli4	Bernoulli5	Bernoulli6	Bernoulli7	Bernoulli8	Bernoulli9	Bernoulli10	p-hat
1	1	1	1	1	0	1	0	1	0	1	0.7
2	1	1	1	1	0	1	0	1	0	1	0.7
3	1	0	1	1	0	1	1	1	1	1	0.8
4	1	1	0	1	1	1	1	1	0	1	0.8
5	1	1	1	1	1	1	1	1	1	1	1
6	0	1	1	0	0	0	1	1	1	1	0.6
7	1	0	1	1	1	1	1	0	1	1	0.8
8	1	1	1	1	1	1	0	0	1	1	0.8
9	1	1	1	0	1	1	1	0	0	0	0.6

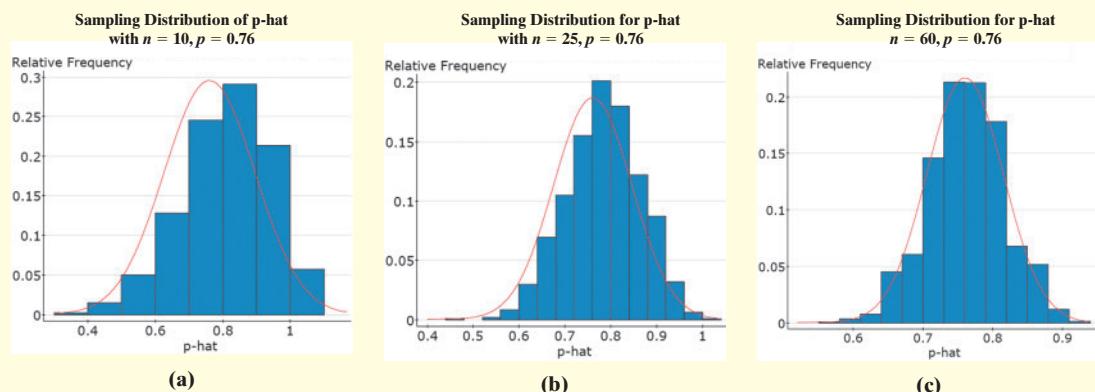
Sample 1 → $\hat{p} = 0.7$
 Sample 2 → $\hat{p} = 0.7$

Figure 13(a) shows the histogram of the 2000 sample proportions from column p-hat. Notice that the shape of the distribution is skewed left. The mean of the 2000 sample proportions is 0.76 and the standard deviation is 0.136. Notice that the mean of the sample proportions equals the population proportion.

Figure 13(b) shows the histogram for 2000 sample proportions from samples of size $n = 25$. Notice that the histogram is slightly skewed left (although not as skewed as the histogram with $n = 10$). The mean of the 2000 sample proportions for a sample of size $n = 25$ is 0.76 and the standard deviation is 0.086.

Figure 13(c) shows the histogram for 2000 sample proportions from samples of size $n = 60$. Notice that the histogram is approximately normal. The mean of the 2000 sample proportions is 0.76 and the standard deviation is 0.054.

Figure 13



Notice the following points regarding the sampling distribution of the sample proportion:

- **Shape:** As the size of the sample increases, the shape of the distribution of the sample proportion becomes approximately normal.
- **Center:** The mean of the distribution of the sample proportion equals the population proportion, p .
- **Spread:** The standard deviation of the distribution of the sample proportion decreases as the sample size increases.



Although the proof is beyond the scope of this text, we should be convinced that the following results are reasonable.

Sampling Distribution of \hat{p}

For a simple random sample of size n with a population proportion p ,

- The shape of the sampling distribution of \hat{p} is approximately normal provided $np(1 - p) \geq 10$.
- The mean of the sampling distribution of \hat{p} is $\mu_{\hat{p}} = p$.
- The standard deviation of the sampling distribution of \hat{p} is $\sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}}$.

IN OTHER WORDS

The sample size cannot be more than 5% of the population size because the success or failure of identifying an individual in the population that has the specified characteristic should not be affected by earlier observations. For example, in a population of size 100 where 14 of the individuals have brown hair, the probability that a randomly selected individual has brown hair is $14/100 = 0.14$. The probability that a second randomly selected individual has brown hair is $13/99 = 0.13$. The probability changes because the sampling is done without replacement.

As with the sampling distribution of the sample mean, the data must be obtained randomly. A second requirement of this model is that the sampled values must be independent of each other—that is, one outcome cannot affect the success or failure of any other outcome. When sampling from finite populations, we verify the independence requirement by checking that the sample size n is no more than 5% of the population size, N ($n \leq 0.05N$).

Also, regardless of whether $np(1 - p) \geq 10$, the mean of the sampling distribution of \hat{p} is p , and the standard deviation of the sampling distribution of \hat{p} is $\sqrt{\frac{p(1 - p)}{n}}$.

EXAMPLE 3

Describing the Sampling Distribution of the Sample Proportion

Problem Suppose the proportion of Americans who believe that the state of moral values in the United States is getting worse is 0.76 (Based on a study conducted by the Gallup organization). Suppose we obtain a simple random sample of $n = 60$ Americans and determine which believe that the state of the moral values in the United States is getting worse. Describe the sampling distribution of the sample proportion for Americans with this belief.

Approach If the sample size is less than 5% of the population size and $np(1 - p)$ is at least 10, the sampling distribution of \hat{p} is approximately normal with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}}$.

Solution The United States has over 300 million people, so the sample of $n = 60$ is less than 5% of the population size. Also, $np(1 - p) = 60(0.76)(1 - 0.76) = 10.944 \geq 10$. The distribution of \hat{p} is approximately normal with mean $\mu_{\hat{p}} = 0.76$ and standard deviation

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}} = \sqrt{\frac{0.76(1 - 0.76)}{60}} = 0.055$$



Notice that the standard deviation found in Example 3 (0.055) is very close to the standard deviation of the sample proportion found using simulation with $n = 60$ in Example 2 (0.054).

② Compute Probabilities of a Sample Proportion

The sampling distribution of the sample proportion may be used to compute probabilities involving sample proportions.

EXAMPLE 4

Compute Probabilities of a Sample Proportion

Problem According to the National Center for Health Statistics, 15% of all Americans have hearing trouble.

- In a random sample of 120 Americans, what is the probability at most 12% have hearing trouble?
- Suppose that a random sample of 120 Americans who regularly listen to music using headphones results in 26 having hearing trouble. What might you conclude?

Approach First, determine whether the data was obtained randomly. Then, determine whether the sampling distribution is approximately normal by verifying that the sample size is less than 5% of the population size and that $np(1 - p) \geq 10$. Then use the normal distribution to determine the probabilities.

Solution The data was obtained by a random sample. There are over 300 million people in the United States, so the sample size of $n = 120$ is definitely less than 5% of the population size. We are told that $p = 0.15$. Because $np(1 - p) = 120(0.15)(1 - 0.15) = 15.3 \geq 10$, the shape of the distribution of the sample proportion is approximately normal. The mean of the distribution of the sample proportion \hat{p} is $\mu_{\hat{p}} = 0.15$, and the standard deviation is $\sigma_{\hat{p}} = \sqrt{\frac{0.15(1 - 0.15)}{120}}$.

- We want to know the probability that a random sample of 120 Americans will result in a sample proportion of at most 0.12 (or 12%). That is, we want to know $P(\hat{p} \leq 0.12)$. Figure 14 shows the normal curve with the area to the left of 0.12 shaded.

To find this area by hand, convert $\hat{p} = 0.12$ to a z -score by subtracting the mean and dividing by the standard deviation. Don't forget that we round z to two decimal places.

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.12 - 0.15}{\sqrt{\frac{0.15(1 - 0.15)}{120}}} = -0.92$$

The area under the standard normal curve left of $z = -0.92$ is 0.1788. Remember, the area to the left of $z = -0.92$ is the same as the area to the left of $\hat{p} = 0.12$, so $P(\hat{p} \leq 0.12) = 0.1788$.

If we use technology to find the area to the left of $\hat{p} = 0.12$, we obtain 0.1787, so $P(\hat{p} \leq 0.12) = 0.1787$.

Interpretation The probability that a random sample of $n = 120$ Americans results in at most 12% having hearing trouble is approximately 0.18. This means that about 18 out of 100 random samples of size 120 will result in at most 12% having hearing trouble if the population proportion of Americans with hearing trouble is 0.15.

- A random sample of 120 Americans who regularly listen to music using headphones results in 26 having hearing trouble. The sample proportion is $\hat{p} = \frac{26}{120} = 0.217$.

We want to know if obtaining a sample proportion of 0.217 from a population whose proportion is assumed to be 0.15 is unusual. We compute $P(\hat{p} \geq 0.217)$, because if a sample proportion of 0.217 is unusual, then any sample proportion more than 0.217 is also unusual. Figure 15 shows the normal curve with the area to the right of 0.217 shaded.

Figure 14

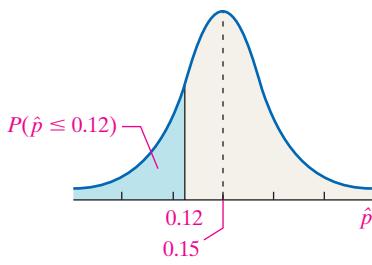
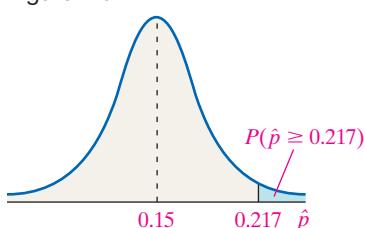


Figure 15



To find this area by hand, convert $\hat{p} = 0.217$ to a z -score.

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.217 - 0.15}{\sqrt{\frac{0.15(1 - 0.15)}{120}}} = 2.06$$

The area under the standard normal curve to the right of $z = 2.06$ is 0.0197. The area to the right of $z = 2.06$ is the same as the area to the right of $\hat{p} = 0.217$, so $P(\hat{p} \geq 0.217) = 0.0197$.

If we use technology to find the area to the right of $\hat{p} = 0.217$, we obtain 0.0199, so $P(\hat{p} \geq 0.217) = 0.0199$.

Interpretation About 2 samples in 100 will result in a sample proportion of 0.217 or more from a population whose proportion is 0.15. We obtained a result that should only happen about 2 times in 100, so the results obtained are unusual. We could make one of two conclusions:

- The population proportion of Americans with hearing trouble who regularly listen to music using headphones is 0.15, and we just happen to randomly select a higher proportion that have hearing trouble.
- The population proportion of Americans with hearing trouble who regularly listen to music using headphones is more than 0.15.

The second conclusion is more reasonable. We conclude that the proportion of Americans who regularly listen to music using headphones who have hearing trouble is higher than the general population.

NW Now Work Problem 17



8.2 Assess Your Understanding

Vocabulary and Skill Building

1. In a town of 500 households, 220 have a dog. The population proportion of dog owners in this town (expressed as a decimal) is $p = \underline{\hspace{2cm}}$.
2. The $\underline{\hspace{2cm}}$, denoted \hat{p} , is given by the formula $\hat{p} = \underline{\hspace{2cm}}$, where x is the number of individuals with a specified characteristic in a sample of n individuals.
3. *True or False:* The population proportion and sample proportion always have the same value.
4. *True or False:* The mean of the sampling distribution of \hat{p} is p .
5. Describe the circumstances under which the shape of the sampling distribution of \hat{p} is approximately normal.
6. What happens to the standard deviation of \hat{p} as the sample size increases? If the sample size is increased by a factor of 4, what happens to the standard deviation of \hat{p} ?

In Problems 7–10, describe the sampling distribution of \hat{p} .

Assume that the size of the population is 25,000 for each problem.

- NW** 7. $n = 500, p = 0.4$ 8. $n = 300, p = 0.7$
 9. $n = 1000, p = 0.103$ 10. $n = 1010, p = 0.84$
 11. A simple random sample of size $n = 75$ is obtained from a population whose size is $N = 10,000$ and whose population proportion with a specified characteristic is $p = 0.8$.

- (a) Describe the sampling distribution of \hat{p} .
- (b) What is the probability of obtaining $x = 63$ or more individuals with the characteristic? That is, what is $P(\hat{p} \geq 0.84)$?
- (c) What is the probability of obtaining $x = 51$ or fewer individuals with the characteristic? That is, what is $P(\hat{p} \leq 0.68)$?
12. A simple random sample of size $n = 200$ is obtained from a population whose size is $N = 25,000$ and whose population proportion with a specified characteristic is $p = 0.65$.
 - (a) Describe the sampling distribution of \hat{p} .
 - (b) What is the probability of obtaining $x = 136$ or more individuals with the characteristic? That is, what is $P(\hat{p} \geq 0.68)$?
 - (c) What is the probability of obtaining $x = 118$ or fewer individuals with the characteristic? That is, what is $P(\hat{p} \leq 0.59)$?
13. A simple random sample of size $n = 1000$ is obtained from a population whose size is $N = 1,000,000$ and whose population proportion with a specified characteristic is $p = 0.35$.
 - (a) Describe the sampling distribution of \hat{p} .
 - (b) What is the probability of obtaining $x = 390$ or more individuals with the characteristic?
 - (c) What is the probability of obtaining $x = 320$ or fewer individuals with the characteristic?

14. A simple random sample of size $n = 1460$ is obtained from a population whose size is $N = 1,500,000$ and whose population proportion with a specified characteristic is $p = 0.42$.

- (a) Describe the sampling distribution of \hat{p} .
- (b) What is the probability of obtaining $x = 657$ or more individuals with the characteristic?
- (c) What is the probability of obtaining $x = 584$ or fewer individuals with the characteristic?

Applying the Concepts

15. Foreign Language According to a study done by Wakefield Research, the proportion of Americans who can order a meal in a foreign language is 0.47.

- (a) Suppose a random sample of 200 Americans is asked to disclose whether they can order a meal in a foreign language. Is the response to this question qualitative or quantitative? Explain.
- (b) Explain why the sample proportion, \hat{p} , is a random variable. What is the source of the variability?
- (c) Describe the sampling distribution of \hat{p} , the proportion of Americans who can order a meal in a foreign language. Be sure to verify the model requirements.
- (d) In the sample obtained in part (a), what is the probability the proportion of Americans who can order a meal in a foreign language is greater than 0.5?
- (e) Would it be unusual that, in a survey of 200 Americans, 80 or fewer Americans can order a meal in a foreign language? Why?

16. Are You Satisfied? According to a study done by the Gallup organization, the proportion of Americans who are satisfied with the way things are going in their lives is 0.82.

- (a) Suppose a random sample of 100 Americans is asked, "Are you satisfied with the way things are going in your life?" Is the response to this question qualitative or quantitative? Explain.
- (b) Explain why the sample proportion, \hat{p} , is a random variable. What is the source of the variability?
- (c) Describe the sampling distribution of \hat{p} , the proportion of Americans who are satisfied with the way things are going in their life. Be sure to verify the model requirements.
- (d) In the sample obtained in part (a), what is the probability the proportion who are satisfied with the way things are going in their life is at least 0.85?
- (e) Would it be unusual for a survey of 100 Americans to reveal that 75 or fewer are satisfied with the way things are going in their life? Why?

NW 17. Marriage Obsolete? According to a study done by the Pew Research Center, 39% of adult Americans believe that marriage is now obsolete.

- (a) Suppose a random sample of 500 adult Americans is asked whether marriage is obsolete. Describe the sampling distribution of \hat{p} , the proportion of adult Americans who believe marriage is obsolete.
- (b) What is the probability that in a random sample of 500 adult Americans less than 38% believe that marriage is obsolete?
- (c) What is the probability that in a random sample of 500 adult Americans between 40% and 45% believe that marriage is obsolete?
- (d) Would it be unusual for a random sample of 500 adult Americans to result in 210 or more who believe marriage is obsolete?

18. Credit Cards According to creditcard.com, 29% of adults do not own a credit card.

(a) Suppose a random sample of 500 adults is asked, "Do you own a credit card?" Describe the sampling distribution of \hat{p} , the proportion of adults who do not own a credit card.

- (b) What is the probability that in a random sample of 500 adults more than 30% do not own a credit card?
- (c) What is the probability that in a random sample of 500 adults between 25% and 30% do not own a credit card?
- (d) Would it be unusual for a random sample of 500 adults to result in 125 or fewer who do not own a credit card? Why?

19. Afraid to Fly According to a study conducted by the Gallup organization, the proportion of Americans who are afraid to fly is 0.10. A random sample of 1100 Americans results in 121 indicating that they are afraid to fly. Explain why this is not necessarily evidence that the proportion of Americans who are afraid to fly has increased since the time of the Gallup study.

20. Having Children? The Pew Research Center recently reported that 18% of women 40–44 years of age have never given birth. Suppose a random sample of 250 adult women 40–44 years of age results in 52 indicating they have never given birth. Explain why this is not necessarily evidence that the proportion of women 40–44 years of age who have not given birth has increased since the time of the Pew study.

21. Election Prediction Exit polling is a popular technique used to determine the outcome of an election prior to results being tallied. Suppose a referendum to increase funding for education is on the ballot in a large town (voting population over 100,000). An exit poll of 310 voters finds that 164 voted for the referendum. How likely are the results of your sample if the population proportion of voters in the town in favor of the referendum is 0.49? Based on your result, comment on the dangers of using exit polling to call elections. Include a discussion of the potential nonsampling error that could disrupt your findings.

22. Acceptance Sampling A shipment of 50,000 transistors arrives at a manufacturing plant. The quality control engineer at the plant obtains a random sample of 500 resistors and will reject the entire shipment if 10 or more of the resistors are defective. Suppose that 4% of the resistors in the whole shipment are defective. What is the probability the engineer accepts the shipment? Do you believe the acceptance policy of the engineer is sound?

23. Social Security Reform A researcher studying public opinion of proposed Social Security changes obtains a simple random sample of 50 adult Americans and asks them whether or not they support the proposed changes. To say that the distribution of \hat{p} , the sample proportion of adults who respond yes, is approximately normal, how many more adult Americans does the researcher need to sample if

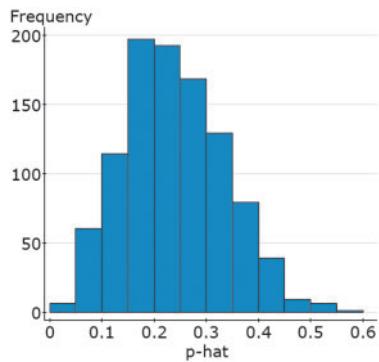
- (a) 10% of all adult Americans support the changes?
- (b) 20% of all adult Americans support the changes?

24. ADHD A researcher studying ADHD among teenagers obtains a simple random sample of 100 teenagers aged 13–17 and asks them whether or not they have ever been prescribed medication for ADHD. To say that the distribution of \hat{p} , the sample proportion of teenagers who respond no, is approximately normal, how many more teenagers aged 13–17 does the researcher need to sample if

- (a) 90% of all teenagers aged 13–17 have never been prescribed medication for ADHD?
- (b) 95% of all teenagers aged 13–17 have never been prescribed medication for ADHD?

25. Reincarnation Suppose 21% of all American teens (age 13–17 years) believe in reincarnation.

- (a) Bob and Alicia both obtain a random sample of 100 American teens and ask each participant to disclose whether they believe in reincarnation or not. Is “belief in reincarnation” qualitative or quantitative? Explain.
- (b) Explain why Bob’s sample of 100 randomly selected American teens might result in 18 who believe in reincarnation, while Alicia’s independent sample of 100 randomly selected American teens might result in 22 who believe in reincarnation.
- (c) Why is it important to randomly select American teens to estimate the population proportion who believe in reincarnation?
- (d) In a survey of 100 American teens, how many would you expect to believe in reincarnation?
- (e) Below is the histogram of the sample proportion of 1000 different surveys in which $n = 20$ American teens were asked to disclose whether they believed in reincarnation. Explain why the normal model should not be used to describe the distribution of the sample proportion.



- (f) What minimum sample size would you require in order for the distribution of the sample proportion to be modeled by the normal distribution?

26. Assessments Consider the homeowners association presented at the beginning of this section. A random sample of 20 households resulted in 15 indicating that they would favor an increase in assessments. Explain why the normal model could not be used to determine if a sample proportion of $\frac{15}{20} = 0.75$ or higher from a population whose proportion is 0.65 is unusual.

27. Airline Reservations In Chapter 6, we learned that the proportion of passengers who miss a flight for which they have a reservation is 0.0995.

- (a) Suppose a flight has 290 reservations. What is the probability that 25 or more passengers will miss the flight?
- (b) Suppose a flight has 320 reservations, but only 300 seats on the plane. What is the probability that 300 or fewer passengers show up for the flight?
- (c) Suppose a popular flight has 300 seats and 300 reservations. Because this flight is popular, the proportion of passengers with a reservation who miss the flight is only 0.04. You are booked on a later flight and put yourself on the stand-by list. There are 14 passenger names ahead of you. What is the probability you get on this flight?

28. Finite Population Correction Factor In this section, we assumed that the sample size was less than 5% of the size of the population. When sampling without replacement from a finite population in which $n > 0.05N$, the standard deviation of the distribution of \hat{p} is given by

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n-1} \cdot \left(\frac{N-n}{N}\right)}$$

where N is the size of the population. A survey is conducted at a college having an enrollment of 6502 students. The student council wants to estimate the percentage of students in favor of establishing a student union. In a random sample of 500 students, it was determined that 410 were in favor of establishing a student union.

- (a) Obtain the sample proportion, \hat{p} , of students surveyed who favor establishing a student union.
- (b) Calculate the standard deviation of the sampling distribution of \hat{p} using \hat{p} as an estimate of p .

Retain Your Knowledge

DATA **29. Fumbles** The New England Patriots made headlines prior to the 2015 Super Bowl for allegedly playing with underinflated footballs. An underinflated ball is easier to grip, and therefore, less likely to be fumbled. What do the data say? The following data represent the number of plays a team has per fumble. For example, the Chicago Bears run 48 offensive plays for every fumble.

- (a) Draw a boxplot of plays per fumble for all teams in the National Football League. Describe the shape of the distribution. Are there any outliers? If so, which team(s)?
- (b) Playing in a dome (inside) removes the effect of weather (such as rain) on the game. Draw a boxplot of plays per fumble for teams who do not play in a dome. Are there any outliers? If so, which team(s)?

Team	Dome	Plays
Atlanta Falcons	Yes	83
New Orleans Saints	Yes	80
New England Patriots	No	78
Houston Texans	Yes	61
Minnesota Vikings	Yes	59
Baltimore Ravens	No	58
Carolina Panthers	No	57
San Diego Chargers	No	57
Indianapolis Colts	Yes	56
Green Bay Packers	No	55
New York Giants	No	55
Cincinnati Bengals	No	53
Pittsburgh Steelers	No	52
Kansas City Chiefs	No	52
Jacksonville Jaguars	No	51
Cleveland Browns	No	50
Saint Louis Rams	Yes	50
Seattle Seahawks	No	50
Detroit Lions	Yes	49
Chicago Bears	No	48
San Francisco 49ers	No	47
New York Jets	No	46
Denver Broncos	No	46
Tampa Bay Buccaneers	No	45
Dallas Cowboys	Yes	45
Tennessee Titans	No	45
Oakland Raiders	No	44
Miami Dolphins	No	44
Buffalo Bills	No	43
Arizona Cardinals	Yes	43
Philadelphia Eagles	No	42
Washington Redskins	No	37

Source: Advanced Football Analytics.



Chapter 8 Review

Summary

Now that we have completed our study of probability distributions, we can begin to transition to inferential statistics. Remember, inferential statistics represents taking information learned from a sample and generalizing it to a population. Often, we accompany the generalization with a measure of reliability in the results. This chapter introduces models that are used to make the generalizations. With these models in hand, we spend the rest of the course performing inference.

In Section 8.1, we discussed the sampling distribution of the sample mean. The mean of the distribution of the sample mean equals the mean of the population ($\mu_{\bar{x}} = \mu$) and the standard deviation of the distribution of the sample mean is the standard deviation of the population divided by the square root of the sample size ($\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$) assuming the sample size is no more than 5% of the population size.

If a sample is obtained from a population that is known to be normally distributed, the shape of the distribution of

the sample mean is approximately normal. If the sample is obtained from a population that is not normal, the shape of the distribution of the sample mean becomes approximately normal as the sample size increases. This result is known as the Central Limit Theorem.

In Section 8.2, we discussed the sampling distribution of the sample proportion. If a sample of size n is obtained from a population and x successes are obtained, then $\hat{p} = \frac{x}{n}$. If $np(1 - p) \geq 10$, then the shape of the distribution of \hat{p} is approximately normal. The mean of the distribution of the sample proportion is the population proportion ($\mu_{\hat{p}} = p$) and the standard deviation of the distribution of the sample proportion is $\sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}}$. Because the sampled values must be independent, it must be the case that the sample size is no more than 5% of the population size.

Vocabulary

Sampling distribution (p. 371)

Sampling distribution of the sample mean (p. 371)

Standard error of the mean (p. 374)

Central Limit Theorem (p. 377)

Sample proportion (p. 385)

Sampling distribution of \hat{p} (p. 387)

Formulas

Mean and Standard Deviation of the Sampling Distribution of \bar{x}

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Sample Proportion

$$\hat{p} = \frac{x}{n}$$

Mean and Standard Deviation of the Sampling Distribution of \hat{p}

$$\mu_{\hat{p}} = p \quad \text{and} \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}}$$

Standardizing a Normal Random Variable

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \text{or} \quad z = \frac{\hat{p} - p}{\sqrt{\frac{p(1 - p)}{n}}}$$

Objectives

Section	You should be able to . . .	Example(s)	Review Exercises
8.1	1 Describe the distribution of the sample mean: normal population (p. 371) 2 Describe the distribution of the sample mean: nonnormal population (p. 375)	1–3 4 and 5	2, 4, 5 2, 6, 7
8.2	1 Describe the sampling distribution of a sample proportion (p. 385) 2 Compute probabilities of a sample proportion (p. 388)	2 and 3 4	3, 4, 8(b), 10(a) 8(c), (d), 9, 10(b), (c)

Review Exercises

1. In your own words, explain what a sampling distribution is.
2. Under what conditions is the sampling distribution of \bar{x} normal?
3. Under what conditions is the sampling distribution of \hat{p} approximately normal?

4. What are the mean and standard deviation of the sampling distribution of \bar{x} ? What are the mean and standard deviation of the sampling distribution of \hat{p} ?

5. **Energy Need during Pregnancy** The total energy need during pregnancy is normally distributed, with mean $\mu = 2600$ kcal/day and standard deviation $\sigma = 50$ kcal/day.

Source: American Dietetic Association.

- (a) Is total energy need during pregnancy qualitative or quantitative?
- (b) What is the probability that a randomly selected pregnant woman has an energy need of more than 2625 kcal/day? Is this result unusual?
- (c) Describe the sampling distribution of \bar{x} , the sample mean daily energy requirement for a random sample of 20 pregnant women. What might be the source of variability in the sample mean?
- (d) What is the probability that a random sample of 20 pregnant women has a mean energy need of more than 2625 kcal/day? Is this result unusual?

6. Copper Tubing A machine at K&A Tube & Manufacturing Company produces a certain copper tubing component in a refrigeration unit. The tubing components produced by the manufacturer have a mean diameter of 0.75 inch with a standard deviation of 0.004 inch. The quality-control inspector takes a random sample of 30 components once each week and calculates the mean diameter of these components. If the mean is either less than 0.748 inch or greater than 0.752 inch, the inspector concludes that the machine needs an adjustment.

- (a) Describe the sampling distribution of \bar{x} , the sample mean diameter, for a random sample of 30 such components.
- (b) What is the probability that, based on a random sample of 30 such components, the inspector will conclude that the machine needs an adjustment when, in fact, the machine is correctly calibrated?

7. Number of Televisions Based on data obtained from AC Nielsen, the mean number of televisions in a household in the United States is 2.24. Assume that the population standard deviation number of television sets in the United States is 1.38.

- (a) Do you believe the shape of the distribution of number of television sets follows a normal distribution? Why or why not?
- (b) A random sample of 40 households results in a total of 102 television sets. What is the mean number of televisions in these 40 households?
- (c) What is the probability of obtaining the sample mean obtained in part (b) if the population mean is 2.24? Does the statistic from part (b) contradict the results reported by AC Nielsen?

8. Entrepreneurship A Gallup survey indicated that 72% of 18- to 29-year-olds, if given a choice, would prefer to start their own business rather than work for someone else. A random sample of 600 18- to 29-year-olds is obtained today.

- (a) Is the variable start own business versus work for someone else qualitative or quantitative?
- (b) Describe the sampling distribution of \hat{p} , the sample proportion of 18- to 29-year-olds who would prefer to start their own business. Explain the source of variability in the sample proportion.
- (c) In a random sample of 600 18- to 29-year-olds, what is the probability that no more than 70% would prefer to start their own business?
- (d) Would it be unusual if a random sample of 600 18- to 29-year-olds resulted in 450 or more who would prefer to start their own business?

9. Advanced Degrees According to the U.S. Census Bureau, in 2009, 10% of adults 25 years and older in the United States had advanced degrees. A researcher with the U.S. Department of Education surveys 500 randomly selected adults 25 years of age or older and finds that 60 of them have an advanced degree. Explain why this is not necessarily evidence that the proportion of adults 25 years of age or older with advanced degrees has increased.

10. Variability in Baseball Suppose, during the course of a typical season, a batter has 500 at-bats. This means the player has the opportunity to get a hit 500 times during the course of a season. Further, suppose a batter is a career 0.280 hitter (he averages 280 hits every 1000 at-bats or he has 280 successes in 1000 trials of the experiment), so the population proportion of hits is 0.280.

- (a) Assuming each at-bat is an independent event, describe the sampling distribution of \hat{p} , the proportion of hits in 500 at-bats over the course of a season.
- (b) Would it be unusual for a player who is a career 0.280 hitter to have a season in which he hits at least 0.310?
- (c) Would it be unusual for the player who hit 0.310 one season to hit below 0.255 the following season?
- (d) Explain why a career 0.280 hitter could easily have a batting average between 0.260 and 0.300.
- (e) Use the result of part (d) to explain that a player who hit 0.260 in a season may not be a worse player than one who hit 0.300.



Chapter Test

- State the Central Limit Theorem.
- If a random sample of size 36 is obtained from a population with mean 50 and standard deviation 24, what is the mean and standard deviation of the sampling distribution of the sample mean?
- The charge life of a certain lithium ion battery for camcorders is normally distributed, with mean 90 minutes and standard deviation 35 minutes.
 - What is the probability that a randomly selected battery of this type lasts more than 100 minutes on a single charge? Is this result unusual?
 - Describe the sampling distribution of \bar{x} , the sample mean charge life for a random sample of 10 such batteries.

- What is the probability that a random sample of 10 such batteries has a mean charge life of more than 100 minutes? Is this result unusual?
- What is the probability that a random sample of 25 such batteries has a mean charge life of more than 100 minutes?
- Explain what causes the probabilities in parts (c) and (d) to be different.
- A machine used for filling plastic bottles with a soft drink has a known standard deviation of $\sigma = 0.05$ liter. The target mean fill volume is $\mu = 2.0$ liters.
 - Describe the sampling distribution of \bar{x} , the sample mean fill volume, for a random sample of 45 such bottles.

- (b)** A quality-control manager obtains a random sample of 45 bottles. He will shut down the machine if the sample mean of these 45 bottles is less than 1.98 liters or greater than 2.02 liters. What is the probability that the quality-control manager will shut down the machine even though the machine is correctly calibrated?
5. According to the National Center for Health Statistics, 22.4% of adults are smokers. A random sample of 300 adults is obtained.
- Describe the sampling distribution of \hat{p} , the sample proportion of adults who smoke.
 - In a random sample of 300 adults, what is the probability that at least 50 are smokers?
 - Would it be unusual if a random sample of 300 adults results in 54 or fewer being smokers?
6. Peanut and tree nut allergies are considered to be the most serious food allergies. According to the National Institute of Allergy and Infectious Diseases, roughly 1% of

Americans are allergic to peanuts or tree nuts. A random sample of 1500 Americans is obtained.

- Explain why a large sample is needed for the distribution of the sample proportion to be approximately normal.
 - Would it be unusual if a random sample of 1500 Americans results in fewer than 10 with peanut or tree nut allergies?
7. Net worth is defined as total assets (value of house, cars, money, etc.) minus total liabilities (mortgage balance, credit card debt, etc.). According to a recent study by TNS Financial Services, 7% of American households had a net worth in excess of \$1 million (excluding their primary residence). A random sample of 1000 American households results in 82 having a net worth in excess of \$1 million. Explain why the results of this survey do not necessarily imply that the proportion of households with a net worth in excess of \$1 million has increased.

Making an Informed Decision

How Much Time Do You Spend in a Day . . . ?

The American Time Use Survey is a survey of adult Americans conducted by the Bureau of Labor Statistics. The purpose of the survey is to learn how Americans allocate their time in a day. As a reporter for the school newspaper, you wish to file a report that compares the typical student at your school to other Americans.

- Go to the American Time Use Survey web page. Research variables of interest to you (or that you believe would be of interest to your readers). Determine the mean amount of time Americans who are enrolled in school spend doing the activities you find interesting.
- Use StatCrunch or an online polling site (such as surveymonkey.com) to conduct a survey of a random sample of students at your school. Write questions to learn how students at your school use their time. For example, how much time does a student at your school spend attending class

each day? Be very careful about the wording in your survey questions to avoid confusion about what is being asked.

- For each question, describe the sampling distribution of the sample mean. Use the national norms as the population mean for each variable. Use the sample standard deviation from your survey sample as the population standard deviation.
- Compute probabilities regarding values of the statistics obtained from the study. Are any of the results unusual?
- Write an article for your newspaper reporting your findings. Be sure any interpretations are written so they are statistically accurate, but understandable for the statistically untrained reader.





Estimating the Value of a Parameter

Outline

- 9.1** Estimating a Population Proportion
- 9.2** Estimating a Population Mean
- 9.3** Putting It Together: Which Method Do I Use?
- 9.4** Estimating a Population Standard Deviation (eText only)
- 9.5** Estimating with Bootstrapping (eText only)

Making an Informed Decision



A home purchase is one of the biggest investment decisions we make in our lifetime. Buying a home involves consideration of location, price, features, and neighborhoods. This decision could also affect your family's social life. See the Decision Project on page 433.

Putting It Together

Chapters 1 through 7 laid the groundwork for the remainder of the course. These chapters dealt with data collection (Chapter 1), descriptive statistics (Chapters 2 through 4), and probability (Chapters 5 through 7). Chapter 8 formed a bridge between probability and statistical inference by giving us models we can use to make probability statements about the sample mean and sample proportion.

We know from Section 8.1 that the sample mean is a random variable and has a distribution associated with it. This distribution is called the sampling distribution of the sample mean. The mean of this distribution is equal to the mean of the population, μ , and the standard deviation is $\frac{\sigma}{\sqrt{n}}$. The shape of the distribution of the sample mean is approximately normal if the population from which the sample is drawn is approximately normal or if the sample size is large. We learned in Section 8.2 that \hat{p} is also a random variable whose mean is p and whose standard deviation is $\sqrt{\frac{p(1-p)}{n}}$. If $np(1-p) \geq 10$, the distribution of the random variable \hat{p} is approximately normal. For both the distribution of the sample mean and sample proportion, the sample size must be small relative to the size of the population ($n \leq 0.05N$) and the sampled values must be obtained randomly. This requirement is necessary so that the sampled values are independent of each other.

We now discuss inferential statistics—the process of generalizing information obtained from a sample to a population. We will study two areas of inferential statistics: (1) estimation—sample data are used to estimate the value of unknown parameters such as μ or p , and (2) hypothesis testing—statements regarding a characteristic of one or more populations are tested using sample data. In this chapter, we discuss estimation of an unknown parameter, and in the next chapter we discuss hypothesis testing.

9.1 Estimating a Population Proportion



Preparing for This Section Before getting started, review the following:

- Parameter versus statistic (Section 1.1, p. 5)
- Simple random sampling (Section 1.3, pp. 23–27)
- Sampling error (Section 1.5, p. 41)
- z_α notation (Section 7.2, pp. 349–350)
- Finding the value of a normal random variable (Section 7.2, pp. 347–349)
- Distribution of the sample proportion (Section 8.2, pp. 384–387)

Objectives

- ① Obtain a point estimate for the population proportion
- ② Construct and interpret a confidence interval for the population proportion
- ③ Determine the sample size necessary for estimating a population proportion within a specified margin of error

1 Obtain a Point Estimate for the Population Proportion

Suppose we want to estimate the proportion of adult Americans who believe that the amount they pay in federal income taxes is too high. It is unreasonable to expect that we could survey every adult American. Instead, we use a sample of adult Americans to arrive at an estimate of the proportion. We call this estimate a *point estimate*.

Definition

A **point estimate** is the value of a statistic that estimates the value of a parameter.

For example, the point estimate for the population proportion is $\hat{p} = \frac{x}{n}$, where x is the number of individuals in the sample with a specified characteristic and n is the sample size.

EXAMPLE 1

Obtaining a Point Estimate of a Population Proportion

Problem The Gallup Organization conducted a poll in which a simple random sample of 1015 Americans 18 and older were asked, “Do you consider the amount of federal income tax you have to pay is too high?” Of the 1015 adult Americans surveyed, 458 said yes. Obtain a point estimate for the proportion of Americans 18 and older who believe the amount of federal income tax they pay is too high.

Approach The point estimate of the population proportion is $\hat{p} = \frac{x}{n}$, where $x = 458$ and $n = 1015$.

Solution Substituting into the formula, we get $\hat{p} = \frac{x}{n} = \frac{458}{1015} = 0.451$. We estimate the proportion of Americans 18 and older believe that the amount of federal income tax they have to pay is too high is 0.451.

NW Now Work Problem 25(a)

Note: We agree to round proportions to three decimal places.

2 Construct and Interpret a Confidence Interval for the Population Proportion

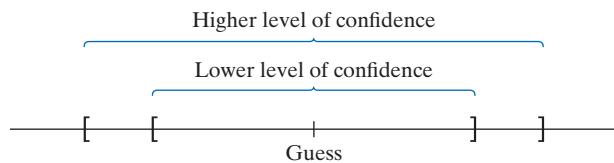
Based on the point estimate of Example 1, can we conclude that less than a majority (less than 50%) of the United States adult population believes that the amount of federal income tax they have to pay is too high? Or is it possible that more than a majority of adult Americans believe that their federal income tax is too high, and we just

happened to sample more folks who do not believe their taxes are too high? After all, statistics such as \hat{p} vary from sample to sample. So a different random sample of adult Americans might result in a different point estimate of the population proportion, such as $\hat{p} = 0.514$. If the method used to select the adult Americans was done appropriately, both point estimates would be good guesses of the population proportion. Due to variability in the sample proportion, we need to report a range (or *interval*) of values, including a measure of the likelihood that the interval includes the unknown population proportion.

To understand the idea of this interval, consider the following situation. Suppose you were asked to guess the proportion of students on your campus who use Instagram. If a survey of 80 students results in 60 who use Instagram, then $\hat{p} = 0.75$. From this, you might guess that the proportion of *all* students on your campus who use Instagram is 0.75 but since you did not survey every student on campus, your estimate may be incorrect. To account for this error, you might adjust your guess by stating that the proportion of students on your campus who use Instagram is 0.75, give or take 0.05 (the *margin of error*). Mathematically, we write this as 0.75 ± 0.05 . If asked how confident you are that the proportion is between 0.70 and 0.80, you might respond, “I am 90% confident that the proportion of students on my campus who use Instagram is between 0.70 and 0.80.” If you want an interval for which your confidence increases to, say, 95%, what do you think will happen to the interval? Having more confidence that the interval will capture the unknown population proportion requires that the width of the interval will need to increase to, say, 0.65 to 0.85.

In statistics, we construct an interval for a population parameter based on a guess along with a level of confidence. The guess is the point estimate of the population parameter, and the level of confidence plays a role in the width of the interval. See Figure 1.

Figure 1



Definitions

IN OTHER WORDS

A confidence interval is a range of numbers, such as 22–30. The level of confidence is the proportion of intervals that will contain the unknown parameter if repeated samples are obtained.

- A **confidence interval** for an unknown parameter consists of an interval of numbers based on a point estimate.
- The **level of confidence** represents the expected proportion of intervals that will contain the parameter if a large number of different samples is obtained. The level of confidence is denoted $(1 - \alpha) \cdot 100\%$.

For example, a 95% level of confidence ($\alpha = 0.05$) implies that if 100 different confidence intervals are constructed, each based on a different sample from the same population, then we expect 95 of the intervals to include the parameter and 5 not to include the parameter.

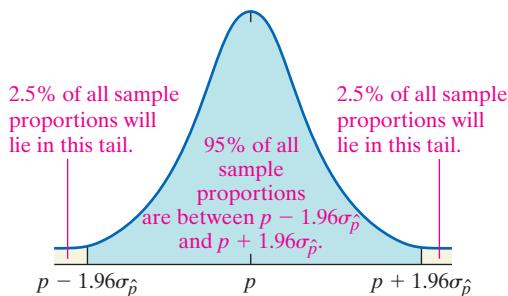
To understand how to construct a confidence interval, we need to review what we know about the model that describes the sampling distribution of \hat{p} , the sample proportion.

- The shape of the distribution of all possible sample proportions is approximately normal provided $np(1 - p) \geq 10$, the sample size is no more than 5% of the population size, and the data are obtained randomly.
- The mean of the distribution of the sample proportion equals the population proportion. That is, $\mu_{\hat{p}} = p$.
- The standard deviation of the distribution of the sample proportion (the standard error) is $\sigma_{\hat{p}} = \sqrt{\frac{p(1 - p)}{n}}$.

Because the distribution of the sample proportion is approximately normal, we know that 95% of all sample proportions will lie within 1.96 standard deviations of the

population proportion, p , and 2.5% of the sample proportions will lie in each tail. See Figure 2. The 1.96 comes from the fact that $z_{0.025}$ is the z -value such that 2.5% of the area under the standard normal curve is to its right. Recall that $z_{0.025} = 1.96$ and $-z_{0.025} = -1.96$.

Figure 2



From Figure 2, we see that 95% of all sample proportions are in the following inequality:

$$p - 1.96\sigma_{\hat{p}} < \hat{p} < p + 1.96\sigma_{\hat{p}}$$

parameter - 1.96 standard error < point estimate < parameter + 1.96 standard error

With a little algebraic manipulation, we can rewrite this inequality with p in the middle and obtain the following:

$$\hat{p} - 1.96\sigma_{\hat{p}} < p < \hat{p} + 1.96\sigma_{\hat{p}} \quad (1)$$

point estimate - 1.96 standard error < parameter < point estimate + 1.96 standard error

This inequality states that 95% of *all* sample proportions will result in confidence interval estimates that contain the population proportion, whereas 5% of *all* sample proportions will result in confidence interval estimates that do not contain the population proportion.

It is common to write confidence interval estimates for the population proportion as

$$\text{point estimate} \pm \text{margin of error}$$

So, write a 95% confidence interval for the population proportion as

$$\hat{p} \pm 1.96\sigma_{\hat{p}}$$

point estimate \pm 1.96 standard error

point estimate \pm margin of error

The **margin of error** for a 95% confidence interval for the population proportion is $1.96\sigma_{\hat{p}}$. This determines the width of the interval.

To visually illustrate the idea of a confidence interval, draw the sampling distribution of \hat{p} . Now create a “slider” that shows $\hat{p} - 1.96\sigma_{\hat{p}}$ and $\hat{p} + 1.96\sigma_{\hat{p}}$. As \hat{p} leaves the blue-shaded region, the corresponding confidence interval does not capture the population proportion, p . See Figure 3.

Figure 3

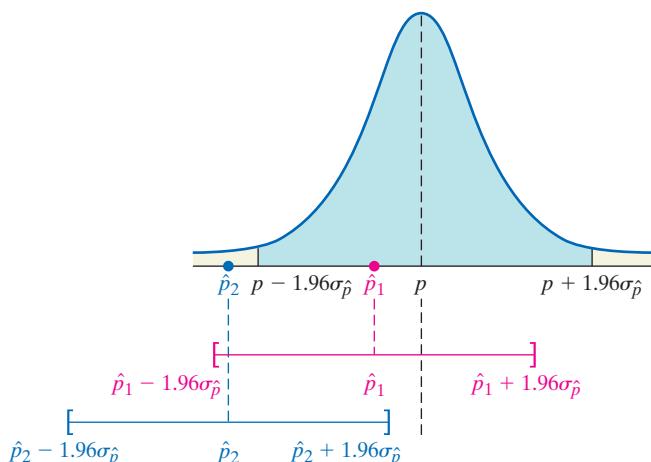


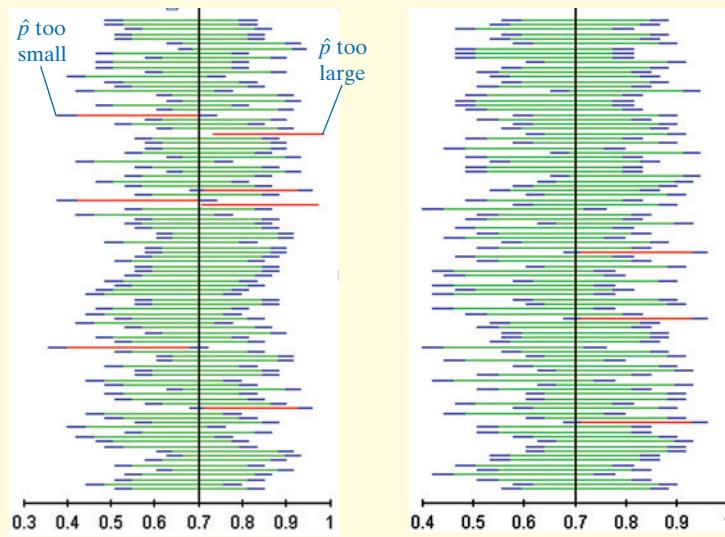
Figure 3 tells us that for a 95% confidence interval, 95% of all sample proportions will result in a confidence interval that includes the population proportion, while 5% of all sample proportions (those in the tails) will result in a confidence interval that does not include the population proportion.

EXAMPLE 2**Illustrating the Meaning of Level of Confidence Using Simulation**

Let's illustrate what "95% confidence" means in a 95% confidence interval in another way. Simulate obtaining 200 different random samples of size $n = 50$ from a population with $p = 0.7$. Figure 4 shows the confidence intervals in groups of 100. A green interval is a 95% confidence interval that includes the population proportion, 0.7. A red interval is a confidence interval that does not include the population proportion. (For now, ignore the blue intervals.) Notice that the red intervals that do not capture the population proportion 0.7 have centers that are "far away" (more than 1.96 standard errors) from 0.7. Of the 200 confidence intervals obtained, 10 (the red intervals) do not include the population proportion. For example, the first interval to miss has a sample proportion that is too small to result in an interval that captures 0.7. The second interval to miss has a sample proportion that is too large to result in an interval that captures 0.7. So $190/200 = 0.95$ (or 95%) of the samples have intervals that capture the population proportion.

A 95% level of confidence means that 95% of all possible samples result in confidence intervals that include the parameter (and 5% of all possible samples result in confidence intervals that do not include the parameter).

Figure 4

**CAUTION!**

A 95% confidence interval does *not* mean that there is a 95% probability that the interval contains the parameter (such as p or μ). Remember, probability describes the likelihood of undetermined events. Therefore, it does not make sense to talk about the probability that the interval contains the parameter since the parameter is a fixed value. Think of it this way: I flip a coin and obtain a head. If I ask you to determine the probability that the flip resulted in a head, it would not be 0.5, because the outcome has already been determined. Instead, the probability is 0 or 1. Confidence intervals work the same way. Because p or μ are fixed quantities whose value is unknown, we do not say that there is a 95% probability that the interval contains p or μ . Instead, the probability that the interval includes the parameter is either 0 (interval does not include parameter) or 1 (interval includes parameter).

We can express the results of Example 2 in a different way. For a 95% confidence interval, any sample proportion that lies within 1.96 standard errors of the population proportion will result in a confidence interval that includes p , and any sample proportion that is more than 1.96 standard errors from the population proportion will result in a confidence interval that does not contain p .

Whether a confidence interval contains the population parameter depends solely on the value of the sample statistic. Any sample statistic that is in the tails of the sampling distribution will result in a confidence interval that does not include the population parameter.

In practice, we construct only one confidence interval. We do not know whether the sample results in a confidence interval that includes the parameter, but we do know that

if we construct a 95% confidence interval, it will include the unknown parameter in 95% of all samples. This is why we say that we are 95% confident the interval includes the unknown parameter—it is because the method “works” in 95% of all samples.

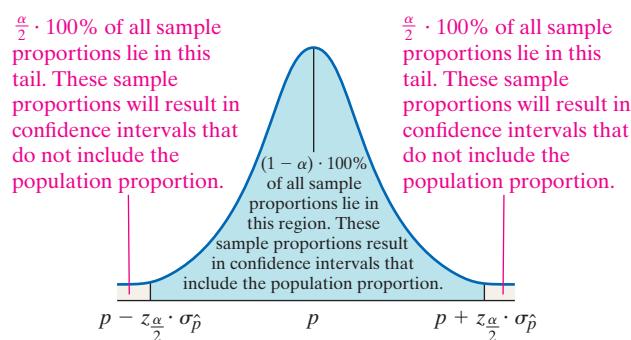
We need a method for constructing any $(1 - \alpha) \cdot 100\%$ confidence interval. [When $\alpha = 0.05$, we are constructing a $(1 - 0.05) \cdot 100\% = 95\%$ confidence interval.]

We generalize Formula (1) on page 398 by first noting that $(1 - \alpha) \cdot 100\%$ of all sample proportions are in the interval

$$p - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} < \hat{p} < p + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}$$

as shown in Figure 5.

Figure 5



Rewrite this inequality with p in the middle and obtain

$$\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}$$

So $(1 - \alpha) \cdot 100\%$ of all sample proportions will result in confidence intervals that contain the population proportion. The sample proportions that are in the tails of the distribution in Figure 5 will not result in confidence intervals that contain the population proportion.

The value $z_{\frac{\alpha}{2}}$ is called the **critical value** of the distribution. It represents the number of standard deviations the sample statistic can be from the parameter and still result in an interval that includes the parameter. Table 1 shows some of the common critical values used in the construction of confidence intervals. Notice that higher levels of confidence correspond to higher critical values. After all, if your level of confidence that the interval includes the unknown parameter increases, the width of your interval (using the margin of error) should increase.

Table 1

Level of Confidence, $(1 - \alpha) \cdot 100\%$	Area in Each Tail, $\frac{\alpha}{2}$	Critical Value, $z_{\frac{\alpha}{2}}$
90%	0.05	1.645
95%	0.025	1.96
99%	0.005	2.576

NOTE

If you use technology to find the critical value for a 99% confidence interval, it will be 2.576.

Interpretation of a Confidence Interval

A $(1 - \alpha) \cdot 100\%$ confidence interval indicates that $(1 - \alpha) \cdot 100\%$ of all simple random samples of size n from the population whose parameter is unknown will result in an interval that contains the parameter.

For example, a 90% confidence interval for a parameter suggests that 90% of all possible samples will result in an interval that includes the unknown parameter and 10% of the samples will result in an interval that does not capture the parameter.

IN OTHER WORDS

The interpretation of a confidence interval is this: We are (*insert level of confidence*) confident that the population proportion is between (*lower bound*) and (*upper bound*). This is an abbreviated way of saying that the method results in an interval that includes the population proportion in $(1 - \alpha) \cdot 100\%$ of all samples.

Look back at Figure 4 on page 399 from Example 2. The intervals, *including* the blue parts, represent 99% confidence intervals. Any interval entirely in red is an interval whose 99% confidence interval does not include the population proportion 0.7. We can see that only 2 of the 200 intervals constructed do not include 0.7. [They both are in the figure on the left with sample proportions that are too large to result in an interval that includes 0.7] So 198/200 or 99% of the 200 intervals do include the population proportion. Therefore, a 99% confidence interval for a parameter means that 99% of all possible samples result in an interval that includes the parameter and 1% of the samples result in an interval that does not capture the parameter.

An extremely important point is that the level of confidence refers to the confidence in the *method*, not in the specific interval. A 90% confidence interval means the method “works” (that is, the interval includes the unknown parameter) for 90% of all samples. So, we do not know whether a particular sample statistic obtained is one of the 90% with an interval that includes the parameter, or one of the 10% whose interval does not include the parameter. **A 90% level of confidence does not tell us there is a 90% probability the parameter lies between the lower and upper bound.**

EXAMPLE 3**Interpreting a Confidence Interval**

Problem When the Gallup Organization conducted the poll introduced in Example 1, 45.1% of those surveyed considered the amount of federal income tax they have to pay as too high. Gallup reported its “survey methodology” as follows:

Results are based on telephone interviews with a random sample of 1015 national adults, aged 18 and older. For results based on the total sample of national adults, one can say with 95% confidence that the maximum margin of sampling error is 4 percentage points.

Determine and interpret the confidence interval for the proportion of Americans aged 18 and older who believe the amount of federal income tax they have to pay is too high.

Approach Confidence intervals for a proportion are of the form point estimate \pm margin of error. So add and subtract the margin of error from the point estimate to obtain the confidence interval. Interpret the confidence interval, “We are 95% confident that the proportion of Americans aged 18 and older who believe that the amount of federal income tax they have to pay is too high is between *lower bound* and *upper bound*.”

Solution The point estimate is 0.451, and the margin of error is 0.04. The confidence interval is 0.451 ± 0.04 . Therefore, the lower bound of the confidence interval is $0.451 - 0.04 = 0.411$ and the upper bound of the confidence interval is $0.451 + 0.04 = 0.491$. We are 95% confident that the proportion of Americans aged 18 and older who believe that the amount of federal income tax they have to pay is too high is between 0.411 and 0.491.

NW Now Work Problem 21

We are now prepared to present a method for constructing a $(1 - \alpha) \cdot 100\%$ confidence interval about the population proportion, p .

Constructing a $(1 - \alpha) \cdot 100\%$ Confidence Interval for a Population Proportion

Suppose that a simple random sample of size n is taken from a population or the data are the result of a randomized experiment. A $(1 - \alpha) \cdot 100\%$ confidence interval for p is given by the following quantities:

$$\text{Lower bound: } \hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad \text{Upper bound: } \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (2)$$

Note: It must be the case that $n\hat{p}(1 - \hat{p}) \geq 10$ and $n \leq 0.05N$ to construct this interval.

Notice that we use \hat{p} in place of p in the standard deviation. This is because p is unknown, and \hat{p} is the best point estimate of p .

The width of the interval is determined by the margin of error.

Definition

The **margin of error**, E , in a $(1 - \alpha) \cdot 100\%$ confidence interval for a population proportion is given by

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (3)$$

EXAMPLE 4 Constructing a Confidence Interval for a Population Proportion

Problem In the Parent–Teen Cell Phone Survey conducted by Princeton Survey Research Associates International, 800 randomly sampled 16- to 17-year-olds living in the United States were asked whether they have ever used their cell phone to text while driving. Of the 800 teenagers surveyed, 272 indicated that they text while driving. Obtain a 95% confidence interval for the proportion of 16- to 17-year-olds who text while driving.

By Hand Approach

Step 1 Compute the value of \hat{p} .

Step 2 Verify the data was obtained through a random sample. Verify that $n\hat{p}(1 - \hat{p}) \geq 10$ (the normality condition) and $n \leq 0.05N$ (the sample size is no more than 5% of the population size – the independence condition).

Step 3 Determine the critical value $z_{\frac{\alpha}{2}}$.

Step 4 Use Formula (2) to determine the lower and upper bounds of the confidence interval.

Step 5 Interpret the result.

By-Hand Solution

Step 1 There are $x = 272$ successes (teens who text while driving) out of $n = 800$ individuals in the survey, so

$$\hat{p} = \frac{x}{n} = \frac{272}{800} = 0.34$$

Step 2 The individuals were obtained through a random sample. $n\hat{p}(1 - \hat{p}) = 800(0.34)(1 - 0.34) = 179.52 \geq 10$. There are more than 1,000,000 teenagers 16 to 17 years of age in the United States, so our sample size is definitely less than 5% of the population size. The independence requirement is satisfied.

Step 3 Because we want a 95% confidence interval, we have $\alpha = 1 - 0.95 = 0.05$, so $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$.

Step 4 Substituting into Formula (2) with $n = 800$, we obtain the lower and upper bounds of the confidence interval:

Lower bound:

$$\begin{aligned} \hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} &= 0.34 - 1.96 \cdot \sqrt{\frac{0.34(1 - 0.34)}{800}} \\ &= 0.34 - 0.033 \\ &= 0.307 \end{aligned}$$

Technology Approach

Step 1 By hand, compute the value of \hat{p} .

Step 2 Verify the data was obtained through a random sample. Verify that $n\hat{p}(1 - \hat{p}) \geq 10$ (the normality condition) and $n \leq 0.05N$ (the sample size is no more than 5% of the population size – the independence condition).

Step 3 Use a statistical spreadsheet or graphing calculator with advanced statistical features to obtain the confidence interval. We will use StatCrunch. The steps for constructing confidence intervals using StatCrunch, Minitab, Excel, and the TI-83/84 Plus graphing calculators are given in the Technology Step-by-Step on pages 405–406.

Step 4 Interpret the result.

Technology Solution

Step 1 There are $x = 272$ successes (teens who text while driving) out of $n = 800$ individuals in the survey, so

$$\hat{p} = \frac{x}{n} = \frac{272}{800} = 0.34$$

Step 2 The individuals were obtained through a random sample. $n\hat{p}(1 - \hat{p}) = 800(0.34)(1 - 0.34) = 179.52 \geq 10$. There are more than 1,000,000 teenagers 16 to 17 years of age in the United States, so our sample size is definitely less than 5% of the population size. The independence requirement is satisfied.

Step 3 Figure 6 shows the results obtained from StatCrunch.

Figure 6

95% Confidence Interval Results

p : proportion of successes for population

Method: Standard-Wald

Proportion	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
p	272	800	0.34	0.016748134	0.30717427	0.37282574

Upper bound:

$$\begin{aligned}\hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} &= 0.34 + 1.96 \cdot \sqrt{\frac{0.34(1 - 0.34)}{800}} \\ &= 0.34 + 0.033 \\ &= 0.373\end{aligned}$$

The margin of error is 0.033.

Step 5 We are 95% confident that the proportion of 16- to 17-year-olds who text while driving is between 0.307 and 0.373.

The lower bound (L. Limit) is 0.307 and the upper bound (U. Limit) is 0.373. The margin of error is found by determining the difference between the point estimate and upper bound. Here, $E = 0.373 - 0.34 = 0.033$.

Step 4 We are 95% confident that the proportion of 16- to 17-year-olds who text while driving is between 0.307 and 0.373.

Using Technology

Confidence intervals constructed by hand may differ from those using technology because of rounding.

NW Now Work Problem 25 (b), (c), and (d)

It is important to remember the correct interpretation of a confidence interval. The statement “95% confident” means that if 1000 samples of size 800 were taken, we would expect 950 of the intervals to contain the parameter p , while 50 will not. Unfortunately, we cannot know whether the interval we constructed in Example 4 is one of the 950 intervals that contains p or one of the 50 that does not contain p .

Note: In this text we report the interval as *Lower bound: 0.307; Upper bound: 0.373*. Some texts will use interval notation as in $(0.307, 0.373)$.

The Effect of Level of Confidence on the Margin of Error

We stated on page 397 that logic suggests that a higher level of confidence leads to a wider interval.

EXAMPLE 5

Role of the Level of Confidence in the Margin of Error

Problem For the problem of estimating the proportion of 16- to 17-year-old teenagers who text while driving, determine the effect on the margin of error by increasing the level of confidence from 95% to 99%.

By-Hand Approach

With a 99% level of confidence, $\alpha = 1 - 0.99 = 0.01$. So, to compute the margin of error, E , determine the value of $z_{\frac{\alpha}{2}} = z_{\frac{0.01}{2}} = z_{0.005}$. Then substitute this value into Formula (3) with $\hat{p} = 0.34$ and $n = 800$.

By-Hand Solution

After consulting Table V (Appendix A), we determine that $z_{0.005} = 2.575$. Substituting into Formula (3),

$$\begin{aligned}E &= z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 2.575 \cdot \sqrt{\frac{0.34(1 - 0.34)}{800}} \\ &= 0.043\end{aligned}$$

Technology Approach

Construct a 99% confidence interval using a statistical spreadsheet or graphing calculator with advanced statistical features. Then use the fact that a confidence interval about the population proportion is of the form $\hat{p} \pm E$. The midpoint of the upper and lower bound gives the point estimate. The difference between the point estimate and the lower bound is the margin of error.

Technology Solution

Figure 7 shows the 99% confidence interval using a TI-84 Plus CE graphing calculator.

Figure 7



The lower bound is 0.297 and the upper bound is 0.383.

The midpoint is $\hat{p} = \frac{0.297 + 0.383}{2} = 0.34$. The margin of error is $E = 0.34 - 0.297 = 0.043$.

The margin of error for the 95% confidence interval found in Example 4 is 0.033, so increasing the level of confidence increases the margin of error, resulting in a wider confidence interval.

IN OTHER WORDS

As the level of confidence increases, the margin of error also increases.

The Effect of Sample Size on the Margin of Error

We know that larger sample sizes produce more precise estimates (the Law of Large Numbers). Given that the margin of error is $z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$, we can see that increasing the sample size n decreases the standard error; so the margin of error decreases. This means that larger sample sizes will result in narrower confidence intervals.

To illustrate this idea, suppose the survey conducted in Example 4 resulted in $\hat{p} = 0.34$ for the proportion of 16- to 17-year-old teenagers who text while driving, but the sample size is only 200. The margin of error would be

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 1.96 \cdot \sqrt{\frac{0.34(1 - 0.34)}{200}} = 0.066$$

IN OTHER WORDS

As the sample size increases, the margin of error decreases.

So a sample size that is one-fourth the original size causes the margin of error to double. Put another way, if the sample size is quadrupled, the margin of error will be cut in half.

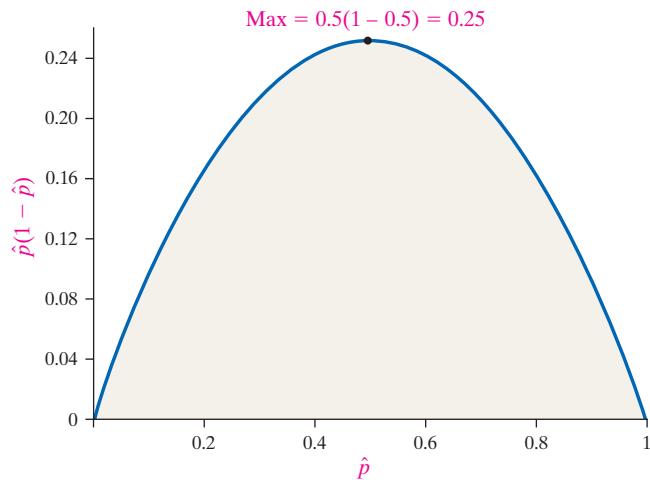
③ Determine the Sample Size Necessary for Estimating a Population Proportion within a Specified Margin of Error

Suppose you want to estimate a proportion with a 3% (0.03) margin of error and 95% confidence. How many individuals should be in your sample? From Formula (3), the margin of error is $E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$.

$$n = \hat{p}(1 - \hat{p}) \cdot \left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2.$$

If we solve this formula for n , we obtain The problem with this formula is that it depends on \hat{p} , and $\hat{p} = \frac{x}{n}$ depends on the sample size, n , which is what we are trying to determine in the first place! How do we resolve this issue? There are two possibilities: (1) We could determine a preliminary value for \hat{p} based on a pilot study or an earlier study, or (2) we could let $\hat{p} = 0.5$. When $\hat{p} = 0.5$, the maximum value of $\hat{p}(1 - \hat{p}) = 0.25$ is obtained, as illustrated in Figure 8. Using the maximum value gives the largest possible value of n for a given level of confidence and a given margin of error.

Figure 8



The disadvantage of the second option is that it could lead to a larger sample size than is necessary. Because of the time and expense of sampling, it is desirable to avoid too large a sample.

Sample Size Needed for Estimating the Population Proportion p

The sample size required to obtain a $(1 - \alpha) \cdot 100\%$ confidence interval for p with a margin of error E is given by

$$n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \quad (4)$$

CAUTION!

Rounding up is different from rounding off. We round 12.3 up to 13; we round 12.3 off to 12.

rounded up to the next integer, where \hat{p} is a prior estimate of p .

If a prior estimate of p is unavailable, the sample size required is

$$n = 0.25 \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \quad (5)$$

rounded up to the next integer.

The margin of error should always be expressed as a decimal when using Formulas (4) and (5).

EXAMPLE 6 Determining Sample Size

Problem An economist wants to know if the proportion of the U.S. population who commutes to work via carpooling is on the rise. What size sample should be obtained if the economist wants an estimate within 2 percentage points of the true proportion with 90% confidence if

- (a) the economist uses the 2009 estimate of 10% obtained from the American Community Survey?
- (b) the economist does not use any prior estimates?

Approach In both cases, $E = 0.02$ ($2\% = 0.02$) and $z_{\frac{\alpha}{2}} = z_{\frac{0.1}{2}} = z_{0.05} = 1.645$. To answer part (a), let $\hat{p} = 0.10$ in Formula (4). To answer part (b), use Formula (5).

Solution

- (a) Substituting $E = 0.02$, $z_{0.05} = 1.645$, and $\hat{p} = 0.10$ into Formula (4), we obtain

$$n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 = 0.10(1 - 0.10) \left(\frac{1.645}{0.02} \right)^2 = 608.9$$

Round this value up to 609, so the economist must survey 609 randomly selected residents of the United States.

- (b) Substituting $E = 0.02$ and $z_{0.05} = 1.645$ into Formula (5), we obtain

$$n = 0.25 \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 = 0.25 \left(\frac{1.645}{0.02} \right)^2 = 1691.3$$

Round this value up to 1692, so the economist must survey 1692 randomly selected residents of the United States.

The effect of not having a prior estimate of p is that the sample size more than doubled!

CAUTION!

We always round up when determining sample size.

 **Using Technology**

StatCrunch has the ability to determine sample size. See the Technology Step-by-Step.

NW Now Work Problem 35

Technology Step-by-Step

Confidence Intervals about p

TI-83/84 Plus

1. Press STAT, highlight TESTS, and select A: 1-PropZInt
2. Enter the values of x and n .
3. Enter the confidence level following C-Level.
4. Highlight Calculate: press ENTER.

Minitab

1. If you have raw data, enter the data in column C1.
2. Select the Stat menu, then Basic Statistics, then highlight 1 Proportion
3. If you have raw data, select “One or more samples, each in a column” from the pull-down menu. Place the cursor

(continued)

- in the box, highlight the column of the data, and click Select. If you have summary statistics, select “Summarized data” from the pull-down menu. Enter the number of events (successes), x , and the number of trials, n .
- Click the **Options . . .** button. Enter a confidence level. Select “Normal approximation” from the drop-down menu (provided that the assumptions stated are satisfied). Click OK twice.

Excel

- Load the XLSTAT Add-in.
- Select the XLSTAT menu. Highlight **Parametric tests**, then select **Tests for one proportion**.
- In the Frequency cell, enter the number of successes. In the Sample size cell, enter the sample size. Be sure the Frequency radio button is checked for Data format and the z test box is checked. Click the Options tab.
- Be sure the Sample radio button is checked under Variance and the Wald radio button is selected under Confidence Interval. For a 90% confidence interval, enter 10 for Significance Level; for a 95% confidence interval, enter 5 for Significance Level, and so on. Click OK.

StatCrunch**Confidence Intervals**

- If you have raw data, enter the data into the spreadsheet. Name the column variable.
- Select **Stat**, then **Proportion Stats**, then **One Sample**, and then choose either **With Data** or **With Summary**.
- If you chose **With Data**, select the column that has the observations, choose which outcome represents a success. If you chose **With Summary**, enter the number of successes and the number of trials. Choose the confidence interval radio button. Enter the level of confidence. Leave the Method as the Standard-Wald. Click Compute!.

Determining Sample Size

- Select **Stat**, then **Proportion Stats**, then **One Sample**, and then **Width/Sample Size**.
- Enter the Confidence level. For the target proportion, enter \hat{p} or enter 0.5 if there is no prior estimate of p . The width is the difference between the lower bound and upper bound in the confidence interval. Therefore, the width is two times the margin of error. Clear any entry in the sample size cell. Click Compute.



9.1 Assess Your Understanding

Vocabulary and Skill Building

- Define each of the following.
 - Point estimate
 - Confidence interval
 - Level of confidence
 - Margin of error
- If you constructed one hundred 95% confidence intervals based on one hundred different simple random samples of size n , how many of the intervals would you expect to include the unknown parameter? Assume all model requirements are satisfied.
- Put the following in order from narrowest to widest interval. Assume the sample size and sample proportion is the same for all four confidence intervals.
 - 95% confidence interval
 - 80% confidence interval
 - 99% confidence interval
 - 90% confidence interval
- If a sample proportion is 0.55, which of the following could be a 90% confidence interval for the population proportion? Select all that apply.
 - Lower bound: 0.50; Upper bound: 0.60
 - Lower bound: 0.53; Upper bound: 0.59
 - Lower bound: 0.52; Upper bound: 0.58
 - Lower bound: 0.60; Upper bound: 0.70
 - Lower bound: 0.45; Upper bound: 0.60
- True or False:** A 95% confidence interval for a population proportion with lower bound 0.45 and upper bound 0.51 means there is a 95% probability the population proportion is between 0.45 and 0.51.
- As the level of confidence of a confidence interval increases, the margin of error _____ (increases/decreases). As the

sample size used to obtain a confidence interval increases, the margin of error _____ (increases/decreases).

In Problems 7–10, determine the critical value $z_{\alpha/2}$ that corresponds to the given level of confidence.

- 90%
- 99%
- 98%
- 92%

In Problems 11–14, determine the point estimate of the population proportion, the margin of error for each confidence interval, and the number of individuals in the sample with the specified characteristic, x , for the sample size provided.

- Lower bound: 0.201, upper bound: 0.249, $n = 1200$
- Lower bound: 0.051, upper bound: 0.074, $n = 1120$
- Lower bound: 0.462, upper bound: 0.509, $n = 1680$
- Lower bound: 0.853, upper bound: 0.871, $n = 10,732$

In Problems 15–20, construct a confidence interval of the population proportion at the given level of confidence.

- $x = 30, n = 150$, 90% confidence
- $x = 80, n = 200$, 98% confidence
- $x = 120, n = 500$, 99% confidence
- $x = 400, n = 1200$, 95% confidence
- $x = 860, n = 1100$, 94% confidence
- $x = 540, n = 900$, 96% confidence

Applying the Concepts

NW 21. You Explain It! New Deal Policies In response to the Great Depression, Franklin D. Roosevelt enacted many New Deal policies. One such policy was the enactment of the National Recovery Administration (NRA), which required businesses to agree to wages and prices within their particular industry. The thought was that this would encourage higher wages among the working class, thereby spurring consumption. In a Gallup survey conducted in 1933 of 2025 adult Americans, 55% thought that wages paid to workers in industry were too low. The margin of error was 3 percentage points with 95% confidence. Which of the following represents a reasonable interpretation of the survey results? For those that are not reasonable, explain the flaw.

- (a) We are 95% confident 55% of adult Americans during the Great Depression felt wages paid to workers in industry were too low.
- (b) We are 92% to 98% confident 55% of adult Americans during the Great Depression felt wages paid to workers in industry were too low.
- (c) We are 95% confident the proportion of adult Americans during the Great Depression who believed wages paid to workers in industry were too low was between 0.52 and 0.58.
- (d) In 95% of samples of adult Americans during the Great Depression, the proportion who believed wages paid to workers in industry were too low is between 0.52 and 0.58.

22. You Explain It! Superstition A *USA Today/Gallup* poll asked 1006 adult Americans how much it would bother them to stay in a room on the 13th floor of a hotel. Interestingly, 13% said it would bother them. The margin of error was 3 percentage points with 95% confidence. Which of the following represents a reasonable interpretation of the survey results? For those not reasonable, explain the flaw.

- (a) We are 95% confident that the proportion of adult Americans who would be bothered to stay in a room on the 13th floor is between 0.10 and 0.16.
- (b) We are between 92% and 98% confident that 13% of adult Americans would be bothered to stay in a room on the 13th floor.
- (c) In 95% of samples of adult Americans, the proportion who would be bothered to stay in a room on the 13th floor is between 0.10 and 0.16.
- (d) We are 95% confident that 13% of adult Americans would be bothered to stay in a room on the 13th floor.

23. You Explain It! Valentine's Day A Rasmussen Reports national survey of 1000 adult Americans found that 18% dreaded Valentine's Day. The margin of error for the survey was 4.5 percentage points with 95% confidence. Explain what this means.

24. You Explain It! A Stressful Commute? A Gallup poll of 547 adult Americans employed full or part time asked, “Generally speaking, would you say your commute to work is—very stressful, somewhat stressful, not that stressful, or not stressful at all?” Gallup reported that 24% of American workers said that their commute was “very” or “somewhat” stressful. The margin of error was 4 percentage points with 95% confidence. Explain what this means.

NW 25. Giving Blood A survey of 2306 adult Americans aged 18 and older conducted by Harris Interactive found that 417 have donated blood in the past two years.

- (a) Obtain a point estimate for the population proportion of adult Americans aged 18 and older who have donated blood in the past two years.
- (b) Verify that the requirements for constructing a confidence interval about p are satisfied.

(c) Construct a 90% confidence interval for the population proportion of adult Americans who have donated blood in the past two years.

- (d) Interpret the interval.

26. Education Relative to other nations, how do fourth-graders in the United States rank in terms of reading and math ability? Are they in the bottom 50% or in the top 50%? In a survey of 700 randomly sampled registered voters in the United States conducted by Conquest Communications Group, 258 correctly answered that fourth-graders are in the top 50%. **Note:** By the age of 15, U.S. students drop to the 50th percentile in reading and below the 25th percentile in mathematics.

- (a) Obtain a point estimate for the population proportion of U.S. registered voters who believe fourth-graders rank in the top 50% in reading and math ability.
- (b) Verify that the requirements for constructing a confidence interval about p are satisfied.
- (c) Construct a 95% confidence interval for the population proportion of U.S. registered voters who answered correctly.
- (d) Interpret the interval.

27. Luxury or Necessity? A random sample of 1003 adult Americans was asked, “Do you pretty much think televisions are a necessity or a luxury you could do without?” Of the 1003 adults surveyed, 521 indicated that televisions are a luxury they could do without.

- (a) Obtain a point estimate for the population proportion of adult Americans who believe that televisions are a luxury they could do without.
- (b) Verify that the requirements for constructing a confidence interval about p are satisfied.
- (c) Construct and interpret a 95% confidence interval for the population proportion of adult Americans who believe that televisions are a luxury they could do without.
- (d) Is it possible that a supermajority (more than 60%) of adult Americans believe that television is a luxury they could do without? Is it likely?
- (e) Use the results of part (c) to construct a 95% confidence interval for the population proportion of adult Americans who believe that televisions are a necessity.

28. Family Values In a *USA Today/Gallup* poll, 768 of 1024 randomly selected adult Americans aged 18 or older stated that a candidate’s positions on the issue of family values are extremely or very important in determining their vote for president.

- (a) Obtain a point estimate for the proportion of adult Americans aged 18 or older for which the issue of family values is extremely or very important in determining their vote for president.
- (b) Verify that the requirements for constructing a confidence interval for p are satisfied.
- (c) Construct a 99% confidence interval for the proportion of adult Americans aged 18 or older for which the issue of family values is extremely or very important in determining their vote for president.
- (d) Is it possible that the proportion of adult Americans aged 18 or older for which the issue of family values is extremely or very important in determining their vote for president is below 70%? Is this likely?
- (e) Use the results of part (c) to construct a 99% confidence interval for the proportion of adult Americans aged 18 or older for which the issue of family values is not extremely or very important in determining their vote for president.

29. AndroGel Low levels of testosterone in adult males may be treated using AndroGel 1.62%. In clinical studies of 234 adult males who were being treated with AndroGel 1.62%, it was found that 26 saw their prostate-specific antigen (PSA) elevated.

The PSA is a protein produced by cells of the prostate gland.

Source: <http://www.androgelpro.com/clinical-studies/default.aspx>

- (a) Determine a 95% confidence interval for the proportion of adult males treated with AndroGel 1.62% who will experience elevated levels of PSA.
- (b) Determine a 99% confidence interval for the proportion of adult males treated with AndroGel 1.62% who will experience elevated levels of PSA.
- (c) What is the impact of increasing the level of confidence on the margin of error?

30. Reading In a survey of 700 community college students, 481 indicated that they have read a book for personal enjoyment during the school year (based on data from the Community College Survey of Student Engagement).

- (a) Determine a 90% confidence interval for the proportion of community college students who have read a book for personal enjoyment during the school year.
- (b) Determine a 95% confidence interval for the proportion of community college students who have read a book for personal enjoyment during the school year.
- (c) What is the impact of increasing the level of confidence on the margin of error?

31. Phone in the John In a survey conducted by the marketing agency 11mark, 241 of 1000 adults 19 years of age or older confessed to bringing and using their cell phone every trip to the bathroom (confessions included texting and answering phone calls).

- (a) What is the sample in this study? What is the population of interest?
- (b) What is the variable of interest in this study? Is it qualitative or quantitative?
- (c) Based on the results of this survey, obtain a point estimate for the proportion of adults 19 years of age or older who bring their cell phone every trip to the bathroom.
- (d) Explain why the point estimate found in part (c) is a statistic. Explain why it is a random variable. What is the source of variability in the random variable?
- (e) Construct and interpret a 95% confidence interval for the population proportion of adults 19 years of age or older who bring their cell phone every trip to the bathroom.
- (f) What ensures that the results of this study are representative of all adults 19 years of age or older?

DATA 32. Deficit Reduction The Sullivan Statistics Survey I asks, “Would you be willing to pay higher taxes if the tax revenue went directly toward deficit reduction?” Treat the survey respondents as a random sample of adult Americans. Go to www.pearsonhighered.com/sullivanstats to obtain the data file SullivanStatsSurveyI using the file format of your choice for the version of the text you are using. The column “Deficit” has survey responses. Construct and interpret a 90% confidence interval for the proportion of adult Americans who would be willing to pay higher taxes if the revenue went directly toward deficit reduction.

33. Threaded Problem: Tornado The data set “Tornadoes_2017” located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a) The F scale is used to describe the range of wind speeds of a tornado. A tornado whose F scale is 0 has wind speeds less

than 73 miles per hour. The column F0 indicates whether a tornado had an F scale of 0 (Yes) or not (No). Determine the population proportion of tornadoes in 2017 in which the F scale was 0.

- DATA**
- (b) Open the data file 9_1_33b, which represents a random sample of 50 tornadoes from 2017. What is the sample proportion of tornadoes in which the F scale was 0?
 - (c) Verify the requirements for constructing a confidence interval about the population proportion of F0 tornadoes in 2017 are satisfied.
 - (d) Construct a 90% confidence interval for the population proportion of F0 tornadoes in 2017.
 - (e) Does the confidence interval include the population proportion found in part (a)?
 - (f) What proportion of 90% confidence intervals would you expect to include the population proportion found in part (a)?
 - (g) What would cause a confidence interval to result in an interval that does not include the population proportion?
 - DATA** (h) Open the data file 9_1_33h, which represents a random sample of 50 tornadoes from 2017 (obtained independently from the sample in part (b)). Construct a 90% confidence interval for the population proportion of F0 tornadoes in 2017 based on this sample. Are the lower and upper bounds the same as the interval from part (d)? Why or why not? Does the interval include the population proportion?

34. Random Walk Go to www.pearsonhighered.com/sullivanstats to obtain the data file 9_1_34 using the file format of your choice for the version of the text you are using. The data represent the daily (for example, Monday to Tuesday) movement of Johnson & Johnson (JNJ) stock for 200 randomly selected trading days. “Up” means the stock price increased for the time period. “Down” means the stock price decreased (or was unchanged) for the time period. Construct and interpret a 95% confidence interval for the proportion of days JNJ stock increases.

NW 35. High-Speed Internet Access A researcher wishes to estimate the proportion of households that have broadband Internet access. What size sample should be obtained if she wishes the estimate to be within 0.03 with 99% confidence if

- (a) she uses a 2009 estimate of 0.635 obtained from the National Telecommunications and Information Administration?
- (b) she does not use any prior estimates?

36. Home Ownership An urban economist wishes to estimate the proportion of Americans who own their homes. What size sample should be obtained if he wishes the estimate to be within 0.02 with 90% confidence if

- (a) he uses a 2010 estimate of 0.669 obtained from the U.S. Census Bureau?
- (b) he does not use any prior estimates?

37. A Penny for Your Thoughts A researcher for the U.S. Department of the Treasury wishes to estimate the percentage of Americans who support abolishing the penny. What size sample should be obtained if he wishes the estimate to be within 2 percentage points with 98% confidence if

- (a) he uses a 2006 estimate of 15% obtained from a Coinstar National Currency Poll?
- (b) he does not use any prior estimate?

38. Credit-Card Debt A school administrator is concerned about the amount of credit-card debt that college students have. She wishes to conduct a poll to estimate the percentage of full-time college students who have credit-card debt of \$2000 or more. What size sample should be obtained if she wishes the estimate to be within 2.5 percentage points with 94% confidence if

- (a) a pilot study indicates that the percentage is 34%?
- (b) no prior estimates are used?

39. Football Fans A television sports commentator wants to estimate the proportion of Americans who follow professional football. What sample size should be obtained if he wants to be within 3 percentage points with 95% confidence if

- (a) he uses a 2010 estimate of 53% obtained from a Harris poll?
- (b) he does not use any prior estimates?
- (c) Why are the results from parts (a) and (b) so close?

40. Affirmative Action A sociologist wishes to conduct a poll to estimate the percentage of Americans who favor affirmative action programs for women and minorities for admission to colleges and universities. What sample size should be obtained if she wishes the estimate to be within 4 percentage points with 90% confidence if

- (a) she uses a 2003 estimate of 55% obtained from a Gallup Youth Survey?
- (b) she does not use any prior estimates?
- (c) Why are the results from parts (a) and (b) so close?

41. Death Penalty In a Gallup poll, 64% of the people polled answered yes to the following question: “Are you in favor of the death penalty for a person convicted of murder?” The margin of error in the poll was 3%, and the estimate was made with 95% confidence. At least how many people were surveyed?

42. Gun Control In a Gallup Poll, 44% of the people polled answered “more strict” to the following question: “Do you feel that the laws covering the sale of firearms should be made more strict, less strict, or kept as they are now?” Suppose the margin of error in the poll was 3.5% and the estimate was made with 95% confidence. At least how many people were surveyed?

43. Senate Race CNN polled 702 likely voters immediately preceding the 2018 Arizona senate race. The results of the survey indicated that Kyrsten Sinema had the support of 51% of respondents, while Martha McSally had support of 47%. The poll’s margin of error was 4.4%. CNN suggested the race was too close to call. Use the concept of a confidence interval to explain what this means.

44. Simulation—When Model Requirements Fail A Bernoulli random variable is a variable that is either 0 (a failure) or 1 (a success). The probability of success is denoted p .

- (a) Use a statistical spreadsheet to generate 1000 Bernoulli samples of size $n = 20$ with $p = 0.15$.
- (b) Determine the sample proportion for each of the 1000 Bernoulli samples.
- (c) Draw a histogram of the 1000 proportions from part (b). What is the shape of the histogram?
- (d) Construct a 95% confidence interval for each of the 1000 Bernoulli samples using the normal model.
- (e) What proportion of the intervals do you expect to include the population proportion, p ? What proportion of the intervals actually captures the population proportion? Explain any differences.

To deal with issues such as the distribution of \hat{p} not following a normal distribution (Problem 44), A. Agresti and B. Coull proposed a modified approach to constructing confidence intervals for a proportion. If x is the number of successes in n trials and $n = n + z_{\alpha/2}^2$ and $\tilde{p} = \frac{1}{n}\left(x + \frac{1}{2}z_{\alpha/2}^2\right)$, then a $(1 - \alpha) \cdot 100\%$ confidence interval for p is given by

$$\text{Lower bound: } \tilde{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{n} \cdot \tilde{p}(1 - \tilde{p})}$$

$$\text{Upper bound: } \tilde{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{n} \cdot \tilde{p}(1 - \tilde{p})}$$

45. Cauliflower? Jane wants to estimate the proportion of students on her campus who eat cauliflower. After surveying 20 students, she finds 2 who eat cauliflower. Obtain and interpret a 95% confidence interval for the proportion of students who eat cauliflower on Jane’s campus using Agresti and Coull’s method.

46. Walk to Work Alan wants to estimate the proportion of adults who walk to work. In a survey of 10 adults, he finds 1 who walks to work. Explain why a 95% confidence interval using the normal model yields silly results. Then compute and interpret a 95% confidence interval for the proportion of adults who walk to work using Agresti and Coull’s method.

47. Putting It Together: Handwashing The American Society for Microbiology (ASM) and the Soap and Detergent Association (SDA) jointly commissioned two separate studies, both of which were conducted by Harris Interactive. In one of the studies, 1001 adults were interviewed by telephone and asked about their handwashing habits. In the telephone interviews, 921 of the adults said they always wash their hands in public restrooms. In the other study, the handwashing behavior of 6076 adults was inconspicuously observed within public restrooms in four U.S. cities and 4679 of the 6076 adults were observed washing their hands.

- (a) In the telephone survey, what is the variable of interest? Is it qualitative or quantitative?
- (b) What is the sample in the telephone survey? What is the population to which this study applies?
- (c) Verify that the requirements for constructing a confidence interval for the population proportion of adults who say they always wash their hands in public restrooms are satisfied.
- (d) Using the results from the telephone interviews, construct a 95% confidence interval for the proportion of adults who say they always wash their hands in public restrooms.
- (e) In the study where handwashing behavior was observed, what is the variable of interest? Is it qualitative or quantitative?
- (f) We are told that 6076 adults were inconspicuously observed, but were not told how these adults were selected. We know randomness is a key ingredient in statistical studies that allows us to generalize results from a sample to a population. Suggest some ways randomness might have been used to select the individuals in this study.
- (g) Verify the requirements for constructing a confidence interval for the population proportion of adults who actually washed their hands while in a public restroom.
- (h) Using the results from the observational study, construct a 95% confidence interval for the proportion of adults who wash their hands in public restrooms.

- (i) Based on your findings in parts (a) through (h), what might you conclude about the proportion of adults who say they always wash their hands versus the proportion of adults who actually wash their hands in public restrooms?
- (j) Cite some sources of variability in both studies.

Explaining the Concepts

- 48.** Explain what “95% confidence” means in a 95% confidence interval.
- 49.** What type of variable is required to construct a confidence interval for a population proportion?
- 50.** Explain why quadrupling the sample size causes the margin of error to be cut in half.
- 51.** Why do polling companies often survey 1060 individuals when they wish to estimate a population proportion with a margin of error of 3% with 95% confidence?
- 52.** Katrina wants to estimate the proportion of adult Americans who read at least 10 books last year. To do so, she obtains a simple random sample of 100 adult Americans and constructs a 95% confidence interval. Matthew also wants to estimate the proportion of adult Americans who read at least 10 books last

year. He obtains a simple random sample of 400 adult Americans and constructs a 99% confidence interval. Assuming both Katrina and Matthew obtained the same point estimate, whose estimate will have the smaller margin of error? Justify your answer.

- 53.** Two researchers, Jaime and Mariya, are each constructing confidence intervals for the proportion of a population who is left-handed. They find the point estimate is 0.13. Each independently constructed a confidence interval based on the point estimate, but Jaime’s interval has a lower bound of 0.097 and an upper bound of 0.163, while Mariya’s interval has a lower bound of 0.117 and an upper bound of 0.173. Which interval is wrong? Why?
- 54.** The 116th House of Representatives of the United States of America has 435 members, of which 106 are women. An alien lands near the U.S. Capitol and treats members of the House as a random sample of the human race. He reports to his superiors that a 95% confidence interval for the proportion of the human race that is female has a lower bound of 0.203 and an upper bound of 0.284. What is wrong with the alien’s approach to estimating the proportion of the human race that is female?

9.2 Estimating a Population Mean



Preparing for This Section Before getting started, review the following:

- Parameter versus statistic (Section 1.1, p. 5)
- z -scores (Section 3.4, pp. 146–147)
- Simple random sampling (Section 1.3, pp. 23–27)
- Degrees of freedom (Section 3.2, p. 125)

- Normal probability plots (Section 7.3, pp. 355–358)
- Distribution of the sample mean (Section 8.1, pp. 371–379)

Objectives ① Obtain a point estimate for the population mean

- ② State properties of Student’s t -distribution
- ③ Determine t -values
- ④ Construct and interpret a confidence interval for a population mean
- ⑤ Determine the sample size needed to estimate a population mean within a specified margin of error

1 Obtain a Point Estimate for the Population Mean

Remember, the goal of statistical inference is to use information obtained from a sample and generalize the results to the population being studied. As with estimating the population proportion, the first step is to obtain a point estimate of the parameter. The point estimate of the population mean, μ , is the sample mean, \bar{x} .

EXAMPLE 1

Computing a Point Estimate of the Population Mean

Problem The website fueleconomy.gov allows drivers to report the miles per gallon of their vehicle. The data in Table 2 show the reported miles per gallon of 2014 Toyota Camry automobiles for 16 different owners. Obtain a point estimate of the population mean miles per gallon of a 2014 Toyota Camry.

Table 2

25.9	27.2	27.4	31.2
26.2	30.0	27.4	25.1
32.5	25.5	27.0	30.6
32.0	26.4	28.0	27.2

Source: www.fueleconomy.gov

Approach Treat the 16 entries as a simple random sample of all 2014 Toyota Camry automobiles. To find a point estimate of the population mean, compute the sample mean miles per gallon of the 16 cars.

Solution The sample mean is

$$\bar{x} = \frac{25.9 + 26.2 + \dots + 27.2}{16} = \frac{449.6}{16} = 28.1 \text{ miles per gallon}$$

The point estimate of μ is 28.1 miles per gallon. Remember, round statistics to one more decimal point than the raw data if necessary.

② State Properties of Student's *t*-distribution

In Example 1, a different random sample of 16 cars would likely result in a different point estimate of μ . For this reason, we want to construct a confidence interval for the population mean, just as we did for the population proportion.

A confidence interval for the population mean is of the form point estimate \pm margin of error (just like the confidence interval for a population proportion). To determine the margin of error, we need to know the sampling distribution of the sample mean. Recall that the distribution of \bar{x} is approximately normal if the population from which the sample is drawn is normal or the sample size is sufficiently large. In addition, the distribution of \bar{x} has the same mean as the parent population, $\mu_{\bar{x}} = \mu$, and a standard deviation equal to the parent population's standard deviation divided by the square root of the sample size, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Following the same logic used in constructing a confidence interval about a population proportion, our confidence interval would be

point estimate \pm margin of error

$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

This presents a problem because we need to know the population standard deviation to construct this interval. It does not seem likely that we would know the population standard deviation but not know the population mean. So what can we do? A logical option is to use the sample standard deviation, s , as an estimate of σ . Then the standard deviation of the distribution of \bar{x} would be estimated by $\frac{s}{\sqrt{n}}$ and our confidence interval would be

$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \quad (1)$$

Unfortunately, there is a problem with this approach. The sample standard deviation, s , is a statistic and therefore will vary from sample to sample. Using the normal model to determine the critical value, $z_{\frac{\alpha}{2}}$, in the margin of error does not take into account the additional variability introduced by using s in place of σ . This is not much of a problem for large samples because the variability in the sample standard deviation decreases as the sample size increases (Law of Large Numbers), but for small samples, we have a real

problem. Put another way, the z -score of \bar{x} , $\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$, is normally distributed with mean 0

and standard deviation 1 (provided \bar{x} is normally distributed). However, $\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ is not

normally distributed with mean 0 and standard deviation 1. So a new model must be used to determine the margin of error in a confidence interval that accounts for this additional variability. This leads to the story of William Gosset.

NOTE

Recall, the z -score is computed as follows:

$$z = \frac{\text{Statistic} - \text{Parameter}}{\text{Standard Deviation}}$$

In the early 1900s, William Gosset worked for the Guinness brewery. Gosset was in charge of conducting experiments at the brewery to identify the best barley variety. When working with beer, Gosset was limited to small data sets. At the time, the model used for constructing confidence intervals about a mean was the normal model, regardless of whether the population standard deviation was known. Gosset did not know the population standard deviation, so he simply substituted the sample standard deviation for the population standard deviation as suggested by Formula (1). While doing this, he was finding that his confidence intervals did not include the population mean at the rate expected. This led Gosset to develop a model that accounts for the additional variability introduced by using s in place of σ when determining the margin of error. Guinness would not allow Gosset to publish his results under his real name (Guinness was very secretive about its brewing practices), but did allow the results to be published under a pseudonym. Gosset chose Student. So we have Student's t -distribution.

Student's t -Distribution

Suppose a simple random sample of size n is taken from a population that follows a normal distribution. The distribution of

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

follows Student's t -distribution with $n - 1$ degrees of freedom,* where \bar{x} is the sample mean and s is the sample standard deviation.

The interpretation of t is the same as that of the z -score. The t -statistic represents the number of *sample* standard errors \bar{x} is from the population mean, μ . It turns out that the shape of the t -distribution depends on the sample size, n .

To help see how the t -distribution differs from the standard normal (or z -) distribution and the role that the sample size n plays, we will go through the following simulation.

EXAMPLE 2 Comparing the Standard Normal Distribution to the t -Distribution Using Simulation

- (a) Use statistical software such as Minitab or StatCrunch to obtain 2000 simple random samples of size $n = 5$ from a normal population with $\mu = 50$ and $\sigma = 10$. Calculate the sample mean and sample standard deviation for each sample.

Compute $z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}}$ and $t = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}}$ for each sample. Draw histograms for both z and t .

- (b) Repeat part (a) for 2000 simple random samples of size $n = 10$.

Solution

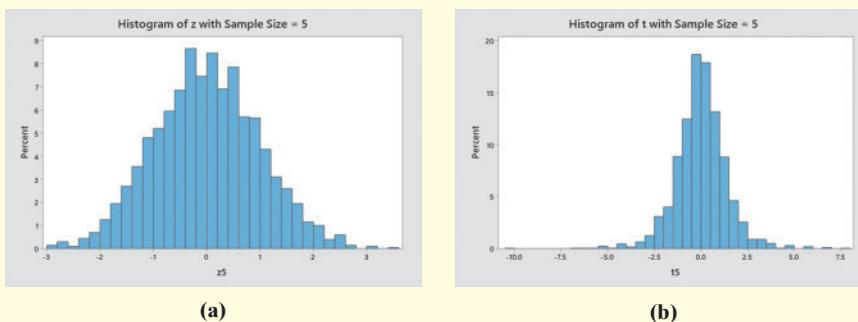
- (a) We use Minitab to obtain the 2000 simple random samples and compute the 2000 sample means and sample standard deviations. We then compute

$z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{x} - 50}{\frac{10}{\sqrt{5}}}$ and $t = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}} = \frac{\bar{x} - 50}{\frac{s}{\sqrt{5}}}$ for each of the 2000 samples.

Figure 9(a) shows the histogram for z , and Figure 9(b) shows the histogram for t .

*The reader may wish to review the discussion of degrees of freedom in Section 3.2 on p. 125.

Figure 9



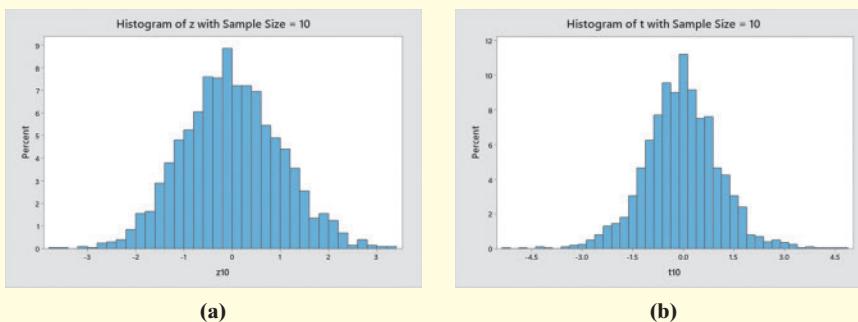
(a)

(b)

Notice that the histogram in Figure 9(a) is symmetric and bell shaped, with the histogram centered at 0, and virtually all the rectangles lying between -3 and 3 . In other words, z with $n = 5$ follows a standard normal distribution. The histogram of t with $n = 5$ is also symmetric, bell shaped, and centered at 0, but the histogram of t has longer tails (that is, t is more dispersed), so it is unlikely that t follows a standard normal distribution. This additional spread is due to the fact that we divided by $\frac{s}{\sqrt{n}}$ to find t instead of by $\frac{\sigma}{\sqrt{n}}$.

- (b) Repeat part (a) for samples of size $n = 10$. Figure 10(a) shows the histogram for z , and Figure 10(b) shows the histogram for t . What do you notice?

Figure 10



(a)

(b)

The histogram in Figure 10(a) is symmetric, bell shaped, centered at 0, and virtually all the rectangles lie between -3 and 3 . In other words, z with $n = 10$ follows a standard normal distribution. The histogram of t with $n = 10$ is also symmetric, bell shaped, and centered at 0, but the histogram has longer tails (that is, t is more dispersed) than the histogram of z with $n = 10$. So t with $n = 10$ also does not appear to follow the standard normal distribution.

One very important distinction must be made. The distribution of t with $n = 10$ (Figure 10(b)) is less dispersed than the distribution of t with $n = 5$ (Figure 9(b)).

We conclude that there are different t distributions for different sample sizes. In addition, the spread in the distribution of t decreases as the sample size n increases. In fact, it can be shown that as the sample size n increases, the distribution of t behaves more and more like the standard normal distribution. ●

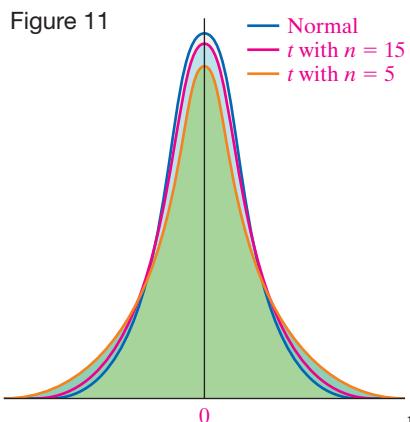
The results of the simulation in Example 2 lead to the following properties of Student's t -distribution.

Properties of the t -Distribution

1. The t -distribution is different for different degrees of freedom.
2. The t -distribution is centered at 0 and is symmetric about 0.
3. The area under the curve is 1. The area under the curve to the right of 0 equals the area under the curve to the left of 0, which equals $\frac{1}{2}$.

(continued)

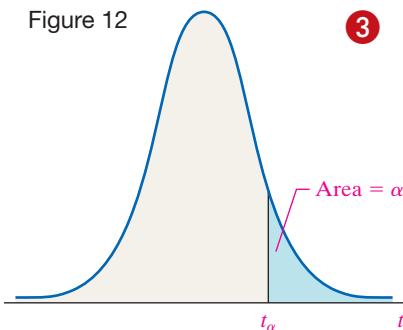
Figure 11



4. As t increases or decreases without bound, the graph approaches, but never equals, zero.
5. The area in the tails of the t -distribution is a little greater than the area in the tails of the standard normal distribution, because we are using s as an estimate of σ , thereby introducing further variability into the t -statistic.
6. As the sample size n increases, the density curve of t gets closer to the standard normal density curve. This result occurs because, as the sample size increases, the values of s get closer to the value of σ , by the Law of Large Numbers.

In Figure 11, we show the t -distribution for the sample sizes $n = 5$ and $n = 15$, along with the standard normal density curve.

Figure 12



③ Determine t -Values

Recall that the notation z_α is used to represent the z -score whose area under the normal curve to the right of z_α is α . Similarly let t_α represent the t -value whose area under the t -distribution to the right of t_α is α . See Figure 12.

The shape of the t -distribution depends on the sample size, n . Therefore, the value of t_α depends not only on α , but also on the degrees of freedom, $n - 1$. In Table VII in Appendix A, the far left column gives the degrees of freedom. The top row represents the area under the t -distribution to the right of some t -value.

EXAMPLE 3 Finding t -Values

Problem Find the t -value such that the area under the t -distribution to the right of the t -value is 0.10, assuming 15 degrees of freedom (df). That is, find $t_{0.10}$ with 15 degrees of freedom.

Approach

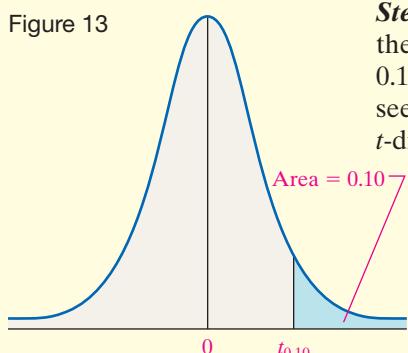
Step 1 Draw a t -distribution with the unknown t -value labeled. Shade the area under the curve to the right of the t -value, as in Figure 12.

Step 2 Find the row in Table VII corresponding to 15 degrees of freedom and the column corresponding to an area in the right tail of 0.10. Identify where the row and column intersect. This is the unknown t -value.

Solution

Step 1 Figure 13 shows the graph of the t -distribution with 15 degrees of freedom. The unknown value of t is labeled, and the area under the curve to the right of t is shaded.

Figure 13



Step 2 A portion of Table VII is shown in Figure 14 on the next page. We have enclosed the row that represents 15 degrees of freedom and the column that represents the area 0.10 in the right tail. The value where the row and column intersect is the t -value we are seeking. The value of $t_{0.10}$ with 15 degrees of freedom is 1.341; that is, the area under the t -distribution to the right of $t = 1.341$ with 15 degrees of freedom is 0.10.

Figure 14

df	Area in Right Tail											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.894	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015

The critical value of z with an area to the right of 0.10 is approximately 1.28, which is smaller than the critical value of t with 15 degrees of freedom. This is because the t -distribution has more spread than the z -distribution.

NW Now Work Problem 7

Using Technology

The TI-84 Plus graphing calculator has an invT feature, which finds the value of t given an area to the left of the unknown t -value and the degrees of freedom. Minitab, Excel, and StatCrunch all have the ability to find the value of t given an area to the left as well.

If the degrees of freedom we desire are not listed in Table VII, choose the closest number in the “df” column. For example, if we have 43 degrees of freedom, use 40 degrees of freedom from Table VII. In addition, the last row of Table VII lists the z -values from the standard normal distribution. Use these values when the degrees of freedom are more than 1000 because the t -distribution starts to behave like the standard normal distribution as n increases.

4 Construct and Interpret a Confidence Interval for a Population Mean

We are now ready to construct a confidence interval for a population mean.

Constructing a $(1 - \alpha) \cdot 100\%$ Confidence Interval for μ

Provided

- sample data come from a simple random sample or randomized experiment,
- sample size is small relative to the population size ($n \leq 0.05N$), and
- the data come from a population that is normally distributed, or the sample size is large.

A $(1 - \alpha) \cdot 100\%$ confidence interval for μ is given by

$$\text{Lower bound: } \bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \quad \text{Upper bound: } \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \quad (2)$$

where $t_{\frac{\alpha}{2}}$ is the critical value with $n - 1$ degrees of freedom.

Because this confidence interval uses the t -distribution, it is often referred to as a **t -interval**.

The **margin of error** for constructing confidence intervals about a population mean is

$$E = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

The Normality Condition

When Gosset developed the t -distribution, he assumed the sample data came from a population that is normally distributed. Most distributions are not *exactly* normal, so we need to verify that the sample data appear “normal enough” before using the t -distribution to construct confidence intervals for the population mean. How can we do this? We have two options.

Option 1 (The Better Option)

- $n < 30$: Draw a normal probability plot (Section 7.3) to check the normality condition and a boxplot to check for outliers. If the data appear to come from a population that is approximately normal with no outliers, then it is reasonable to use Student’s t -distribution to construct a confidence interval for a population mean.

Option 2 This option* relies on the *robustness* of constructing confidence intervals using Student’s t -distribution. An inferential method is **robust** if it is accurate despite minor departures from its underlying assumptions (such as the assumption of normality in Student’s t -distribution).

- $n < 15$: Use Student’s t -distribution to construct confidence intervals about a population mean if the sample data are symmetric with no outliers. The data should not be skewed left or right. This condition may be verified with a boxplot. The median should be in the middle of the box and the whiskers should be of equal length.
- $15 \leq n < 30$: Use Student’s t -distribution to construct confidence intervals about a population mean provided the sample data do not have “extreme” skewness and no outliers. For example, look back at Figure 10 in Section 8.1, which shows population data that is continuous and skewed right. A sample size of 25 was needed for the distribution of the sample mean to be approximately normal.

If $n \geq 30$, use Student’s t -distribution to construct confidence intervals about a population mean even for skewed distributions. This relies on the Central Limit Theorem (for the distribution of the sample mean to be approximately normal) and the Law of Large Numbers (for the sample standard deviation to be close to the population standard deviation).

The advantage of Option 1 is that it removes any subjectivity in determining whether a particular boxplot is showing too much skewness. This is why Option 1 is the preferred method for verifying the model requirements in constructing a confidence interval using Student’s t -distribution.

EXAMPLE 4

Constructing a Confidence Interval about a Population Mean

Table 3

25.9	27.2	27.4	31.2
26.2	30.0	27.4	25.1
32.5	25.5	27.0	30.6
32.0	26.4	28.0	27.2

Source: www.fueleconomy.gov

Problem The website fueleconomy.gov allows drivers to report the miles per gallon of their vehicle. The data in Table 3 show the reported miles per gallon of 2014 Toyota Camry automobiles for 16 different owners. Treat the sample as a simple random sample of all 2014 Toyota Camry automobiles. Construct a 95% confidence interval for the mean miles per gallon of a 2014 Toyota Camry. Interpret the interval.

Approach

Step 1 Verify the data are obtained randomly and the sample size is small relative to the population size. Because the sample size is small, draw a normal probability plot to verify the data come from a population that is normally distributed and a boxplot to verify that there are no outliers.

*E.S. Pearson, N.W. Please, “Relation between the Shape of the Population Distribution and the Robustness of Four Simple Test Statistics,” *Biometrika*, 62(2): 223–241, 1 August 1975, <https://doi.org/10.1093/biomet/62.2.223>

By-Hand Approach

Step 2 Compute the value of \bar{x} and s .

Step 3 Determine the critical value $t_{\frac{\alpha}{2}}$ with $n - 1$ degrees of freedom.

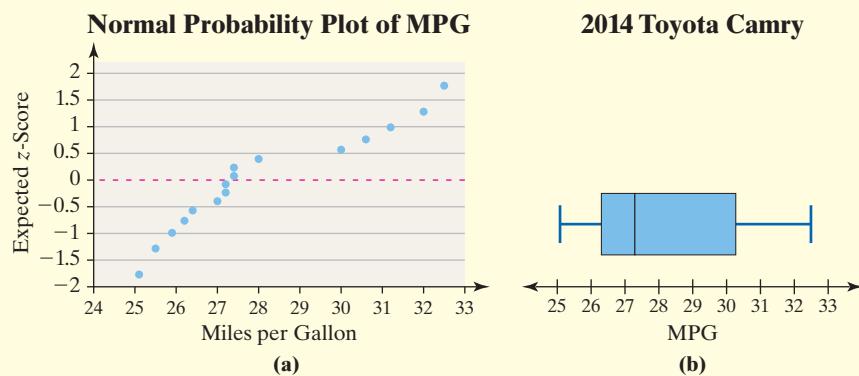
Step 4 Use Formula (2) to determine the lower and upper bounds of the confidence interval.

Step 5 Interpret the result.

Solution

Step 1 The data are obtained from a simple random sample. In addition, there are likely thousands of 2014 Toyota Camry vehicles on the road, so the sample size is small relative to the population size. Figure 15 shows a normal probability plot and boxplot for the data in Table 3. The correlation between MPG and the expected z -scores is 0.954. Because $0.954 > 0.941$ (Table VI), it is reasonable to conclude the sample data come from a population that is normally distributed. The boxplot does not reveal any outliers. The requirements for constructing the confidence interval are satisfied using Option 1.

Figure 15



If you are using Option 2 to assess the model requirements, the boxplot suggests the distribution of sample data is skewed right (because the median is left of center in the box) with no outliers. Because the sample size is 16, and the skewness is not extreme, the requirements for constructing the confidence interval are satisfied using Option 2.

By-Hand Solution

Step 2 We determined the sample mean in Example 1 to be $\bar{x} = 28.1$ mpg. Using a calculator, the sample standard deviation is $s = 2.38$ mpg.

Step 3 For a 95% confidence interval, $\alpha = 1 - 0.95 = 0.05$. The sample size is $n = 16$. So we find $t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = t_{0.025}$ with $16 - 1 = 15$ degrees of freedom. Table VII shows that $t_{0.025} = 2.131$.

Step 4 Substituting into Formula (2), we obtain:

Lower bound:

$$\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 28.1 - 2.131 \cdot \frac{2.38}{\sqrt{16}} = 26.83$$

Upper bound:

$$\bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 28.1 + 2.131 \cdot \frac{2.38}{\sqrt{16}} = 29.37$$

Step 5 We are 95% confident that the mean miles per gallon of all 2014 Toyota Camry cars is between 26.83 and 29.37 mpg.

Technology Approach

Step 2 Use a statistical spreadsheet or graphing calculator with advanced statistical features to obtain the confidence interval. We will use Minitab to construct the confidence interval. The steps for constructing confidence intervals using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on page 419.

Step 3 Interpret the result.

Technology Solution

Step 2 Figure 16 shows the results from Minitab.

Figure 16

One-Sample T: MPG

Descriptive Statistics

N	Mean	StDev	SE Mean	95% CI for μ
16	28.100	2.380	0.595	(26.832, 29.368)

μ : mean of MPG

Minitab presents confidence intervals in the form *(lower bound, upper bound)*. The lower bound is 26.83 and the upper bound is 29.37.

Step 3 We are 95% confident that the mean miles per gallon of all 2014 Toyota Camry cars is between 26.83 and 29.37 mpg.



Notice that $t_{0.025} = 2.131$ for 15 degrees of freedom, while $z_{0.025} = 1.96$. The t -distribution gives a larger critical value, so the width of the interval is wider. Remember, this larger critical value is necessary to account for the increased variability due to using s as an estimate of σ .

Remember, 95% confidence refers to our confidence in the method. If we obtained 100 samples of size $n = 16$ from the population of 2014 Toyota Camry cars, we would expect about 95 of the samples to result in confidence intervals that include μ . We do not know whether the interval in Example 4 includes μ or does not include μ .

What should we do if the requirements to compute a t -interval are not met? We could increase the sample size beyond 30 observations, or we could try to use *nonparametric procedures*. **Nonparametric procedures** typically do not require normality, and the methods are resistant to outliers. A third option is to use resampling methods, such as bootstrapping.

5 Determine the Sample Size Needed to Estimate a Population Mean within a Specified Margin of Error

The margin of error in constructing a confidence interval about the population mean is $E = t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$. Solving this for n , we obtain $n = \left(\frac{t_{\frac{\alpha}{2}} \cdot s}{E} \right)^2$. The problem with this formula is that the critical value $t_{\frac{\alpha}{2}}$ requires that we know the sample size to determine the degrees of freedom, $n - 1$. Obviously, if we do not know n we cannot know the degrees of freedom. The solution to this problem lies in the fact that the t -distribution approaches the standard normal z -distribution as the sample size increases. To convince yourself of this, look at the last few rows of Table VII and compare them with the corresponding z -scores for 95% or 99% confidence. Now if we use z in place of t and a sample standard deviation, s , from previous or pilot studies, we can write the margin of error formula as $E = z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$ and solve it for n to obtain a formula for determining sample size.

Determining the Sample Size n

The sample size required to estimate the population mean, μ , with a level of confidence $(1 - \alpha) \cdot 100\%$ within a specified margin of error, E , is given by

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot s}{E} \right)^2 \quad (3)$$

where n is *rounded up* to the nearest whole number.

CAUTION!

Rounding *up* is different from rounding *off*. We round 5.32 *up* to 6 and *off* to 5.

EXAMPLE 5

Determining Sample Size

Problem We again consider the problem of estimating the miles per gallon of a 2014 Toyota Camry. How large a sample is required to estimate the mean miles per gallon within 0.5 mile per gallon with 95% confidence?

Approach Use Formula (3) with $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$, $s = 2.38$, and $E = 0.5$ to find the required sample size.

Solution Substitute the values of z , s , and E into Formula (3) and obtain

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot s}{E} \right)^2 = \left(\frac{1.96 \cdot 2.38}{0.5} \right)^2 = 87.04$$

Round 87.04 up to 88. A sample size of $n = 88$ results in an interval estimate of the population mean miles per gallon of a 2014 Toyota Camry with a margin of error of 0.5 mile per gallon with 95% confidence.

Using Technology

StatCrunch has the ability to determine sample size. See the Technology Step-by-Step on the next page.

CAUTION!

Don't forget to round up when determining sample size.



Technology Step-by-Step

Confidence Intervals for μ

TI-83/84 Plus

1. If necessary, enter raw data in L1.
2. Press STAT, highlight TESTS, and select 8:TInterval.
3. If the data are raw, highlight Data. Make sure List is set to L1 and Freq to 1. If summary statistics are known, highlight Stats and enter the summary statistics.
4. Enter the confidence level following C-Level:.
5. Highlight Calculate; press ENTER.

Minitab

1. If you have raw data, enter them in column C1.
2. Select the Stat menu, then Basic Statistics, then highlight 1-Sample t....
3. If you have raw data, select “One or more samples, each in a column” from the pull-down menu. Place the cursor in the box, highlight the column containing the raw data, and click “Select”. If you have summarized data, select “Summarized data” from the pull-down menu and enter the summarized data. Select Options ... and enter a confidence level. Click OK twice.

Excel

1. Load the XLSTAT Add-in.
2. Enter the raw data in Column A.
3. Select the XLSTAT menu, highlight Parametric tests. Select One-sample t-test and z-test.
4. Place the cursor in the Data cell. Highlight the raw data in the spreadsheet. Be sure the box for Student's t test is checked and the radio button for One sample is selected.

Click the Options tab. For a 90% confidence interval, let the Significance level (%) equal 10; for a 95% confidence interval, let the Significance level (%) equal 5, and so on. Click OK.

StatCrunch

Constructing a Confidence Interval for the Population Mean

1. If necessary, enter the raw data into column var1. Name the column.
2. Select Stat, highlight T Stats, highlight One Sample. Choose With Data if you have raw data, choose With Summary if you have summarized data.
3. If you chose With Data, highlight the column that contains the data in “Select column(s)”. If you chose With Summary, enter the sample mean, sample standard deviation, and sample size. Choose the confidence interval radio button. Enter the level of confidence. Click Compute!.

Determining Sample Size When Estimating a Population Mean

1. Select Stat, highlight Z Stats, highlight One Sample, and highlight Width/Sample Size. Note: You may also highlight T Stats and follow the same steps.
2. Enter the Confidence level and standard deviation. The width is the difference between the lower bound and the upper bound in the confidence interval. Therefore, the width is two times the margin of error. Clear any entry in the sample size cell. Click Compute.



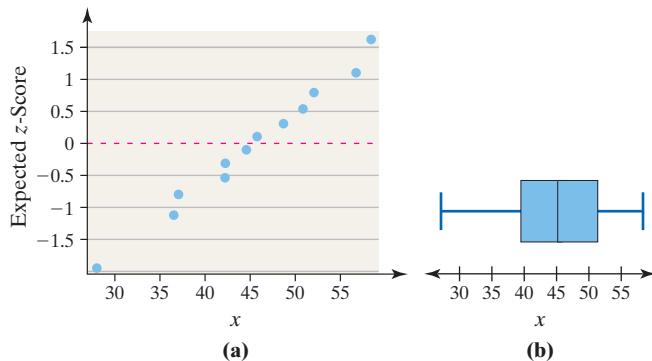
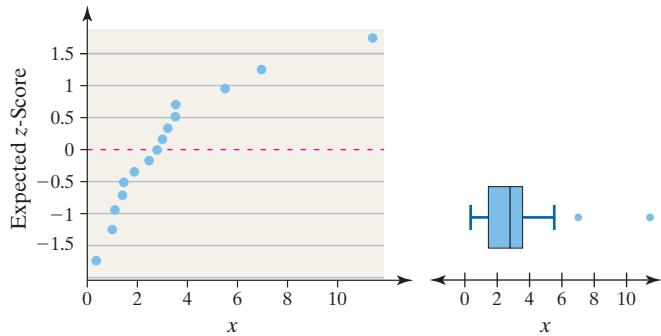
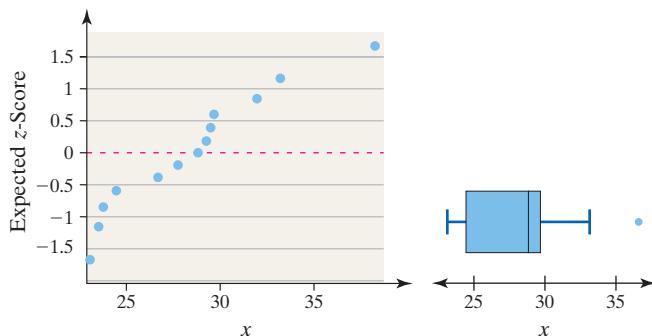
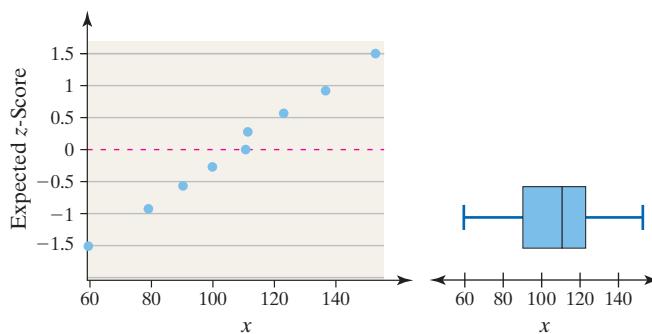
9.2 Assess Your Understanding

Vocabulary and Skill Building

- As the number of degrees of freedom in the t -distribution increases, the spread of the distribution _____ (increases/decreases).
- The notation t_α is the t -value such that the area under the t -distribution to the right of t_α is _____.
- State the properties of Student's t -distribution.
- Put the following in order from least to greatest.
 - $t_{0.10}$ with 5 degrees of freedom
 - $t_{0.10}$ with 15 degrees of freedom
 - $z_{0.10}$
- True or False:** To construct a confidence interval about the mean, the population from which the sample is drawn must be approximately normal.
- The procedure for constructing a confidence interval about a mean is _____, which means minor departures from normality do not affect the accuracy of the interval.
- NW** 7. (a) Find the t -value such that the area in the right tail is 0.10 with 25 degrees of freedom.

- (b) Find the t -value such that the area in the right tail is 0.05 with 30 degrees of freedom.
- (c) Find the t -value such that the area left of the t -value is 0.01 with 18 degrees of freedom. (**Hint:** Use symmetry.)
- (d) Find the critical t -value that corresponds to 90% confidence. Assume 20 degrees of freedom.
8. (a) Find the t -value such that the area in the right tail is 0.02 with 19 degrees of freedom.
- (b) Find the t -value such that the area in the right tail is 0.10 with 32 degrees of freedom.
- (c) Find the t -value such that the area left of the t -value is 0.05 with 6 degrees of freedom. (**Hint:** Use symmetry.)
- (d) Find the critical t -value that corresponds to 95% confidence. Assume 16 degrees of freedom.

In Problems 9–12, a simple random sample of size $n < 30$ for a quantitative variable has been obtained. Using the normal probability plot, the correlation between the variable and expected z -score, and the boxplot, judge whether a t -interval should be constructed.

9. $n = 12$; Correlation = 0.98710. $n = 15$; Correlation = 0.89311. $n = 13$; Correlation = 0.96612. $n = 9$; Correlation = 0.997

In Problems 13–16, determine the point estimate of the population mean and margin of error for each confidence interval.

13. Lower bound: 18, upper bound: 24

14. Lower bound: 20, upper bound: 30

15. Lower bound: 5, upper bound: 23

16. Lower bound: 15, upper bound: 35

17. A simple random sample of size n is drawn from a population that is normally distributed. The sample mean, \bar{x} ,

is found to be 108, and the sample standard deviation, s , is found to be 10.

(a) Construct a 96% confidence interval for μ if the sample size, n , is 25.

(b) Construct a 96% confidence interval for μ if the sample size, n , is 10. How does decreasing the sample size affect the margin of error, E ?

(c) Construct a 90% confidence interval for μ if the sample size, n , is 25. Compare the results to those obtained in part (a). How does decreasing the level of confidence affect the size of the margin of error, E ?

(d) Could we have computed the confidence intervals in parts (a)–(c) if the population had not been normally distributed? Why?

18. A simple random sample of size n is drawn from a population that is normally distributed. The sample mean, \bar{x} , is found to be 50, and the sample standard deviation, s , is found to be 8.

(a) Construct a 98% confidence interval for μ if the sample size, n , is 20.

(b) Construct a 98% confidence interval for μ if the sample size, n , is 15. How does decreasing the sample size affect the margin of error, E ?

(c) Construct a 95% confidence interval for μ if the sample size, n , is 20. Compare the results to those obtained in part (a). How does decreasing the level of confidence affect the margin of error, E ?

(d) Could we have computed the confidence intervals in parts (a)–(c) if the population had not been normally distributed? Why?

19. A simple random sample of size n is drawn. The sample mean, \bar{x} , is found to be 18.4, and the sample standard deviation, s , is found to be 4.5.

(a) Construct a 95% confidence interval for μ if the sample size, n , is 35.

(b) Construct a 95% confidence interval for μ if the sample size, n , is 50. How does increasing the sample size affect the margin of error, E ?

(c) Construct a 99% confidence interval for μ if the sample size, n , is 35. Compare the results to those obtained in part (a). How does increasing the level of confidence affect the margin of error, E ?

(d) If the sample size is $n = 15$, what conditions must be satisfied to compute the confidence interval?

20. A simple random sample of size n is drawn. The sample mean, \bar{x} , is found to be 35.1, and the sample standard deviation, s , is found to be 8.7.

(a) Construct a 90% confidence interval for μ if the sample size, n , is 40.

(b) Construct a 90% confidence interval for μ if the sample size, n , is 100. How does increasing the sample size affect the margin of error, E ?

(c) Construct a 98% confidence interval for μ if the sample size, n , is 40. Compare the results to those obtained in part (a). How does increasing the level of confidence affect the margin of error, E ?

(d) If the sample size is $n = 18$, what conditions must be satisfied to compute the confidence interval?

Applying the Concepts

21. You Explain It! Hours Worked In a survey conducted by the Gallup Organization, 1100 adult Americans were asked how

many hours they worked in the previous week. Based on the results, a 95% confidence interval for mean number of hours worked was lower bound: 42.7 hours and upper bound: 44.5 hours. Which of the following represents a reasonable interpretation of the result? For those that are not reasonable, explain the flaw.

- (a) There is a 95% probability the mean number of hours worked by adult Americans in the previous week was between 42.7 hours and 44.5 hours.
- (b) We are 95% confident that the mean number of hours worked by adult Americans in the previous week was between 42.7 hours and 44.5 hours.
- (c) 95% of adult Americans worked between 42.7 hours and 44.5 hours last week.
- (d) We are 95% confident that the mean number of hours worked by adults in Idaho in the previous week was between 42.7 hours and 44.5 hours.

22. You Explain It! Sleeping A 90% confidence interval for the number of hours that full-time college students sleep during a weekday is lower bound: 7.8 hours and upper bound: 8.8 hours. Which of the following represents a reasonable interpretation of the result? For those that are not reasonable, explain the flaw.

- (a) 90% of full-time college students sleep between 7.8 hours and 8.8 hours.
- (b) We are 90% confident that the mean number of hours of sleep that full-time college students get any day of the week is between 7.8 hours and 8.8 hours.
- (c) There is a 90% probability that the mean hours of sleep that full-time college students get during a weekday is between 7.8 hours and 8.8 hours.
- (d) We are 90% confident that the mean hours of sleep that full-time college students get during a weekday is between 7.8 hours and 8.8 hours.

23. You Explain It! Drive-Through Service Time The trade magazine QSR routinely checks the drive-through service times of fast-food restaurants. A 90% confidence interval that results from examining 607 customers in Taco Bell's drive-through has a lower bound of 161.5 seconds and an upper bound of 164.7 seconds.

- (a) What is the mean service time from the 607 customers?
- (b) What is the margin of error for the confidence interval?
- (c) Interpret the confidence interval.

24. You Explain It! Screen Time Michael Sullivan, the author of this text, determined the mean amount of weekly screen time spent on his phone was 81.3 minutes. A 95% confidence interval for the mean amount of time spent on his phone weekly has a lower bound of 72.9 minutes.

- (a) What is the margin of error for the confidence interval?
- (b) What is the upper bound for the confidence interval?
- (c) Interpret the confidence interval.

25. Hours Worked Revisited For the "Hours Worked" survey conducted by Gallup in Problem 21, provide two recommendations for decreasing the width of the interval.

26. Sleeping Revisited Refer to the "Sleeping" results from Problem 22. What could be done to decrease the width of the confidence interval?

27. Blood Alcohol Concentration A random sample of 51 fatal crashes in 2017 in which the driver had a positive blood alcohol

concentration (BAC) from the National Highway Traffic Safety Administration results in a mean BAC of 0.167 gram per deciliter (g/dL) with a standard deviation of 0.010 g/dL.

- (a) A histogram of blood alcohol concentrations in fatal accidents shows that BACs are highly skewed right. Explain why a large sample size is needed to construct a confidence interval for the mean BAC of fatal crashes with a positive BAC.
- (b) In 2017, there were approximately 25,000 fatal crashes in which the driver had a positive BAC. Explain why this, along with the fact that the data were obtained using a simple random sample, satisfies the requirements for constructing a confidence interval.
- (c) Determine and interpret a 90% confidence interval for the mean BAC in fatal crashes in which the driver had a positive BAC.
- (d) All 50 states and the District of Columbia use a BAC of 0.08 g/dL as the legal intoxication level. Is it possible that the mean BAC of all drivers involved in fatal accidents who are found to have positive BAC values is less than the legal intoxication level? Explain.

28. Hungry or Thirsty? How much time do Americans spend eating or drinking? Suppose for a random sample of 1001 Americans age 15 or older, the mean amount of time spent eating or drinking per day is 1.22 hours with a standard deviation of 0.65 hour.

Source: American Time Use Survey conducted by the Bureau of Labor Statistics.

- (a) A histogram of time spent eating and drinking each day is highly skewed right. Use this result to explain why a large sample size is needed to construct a confidence interval for the mean time spent eating and drinking each day.
- (b) There are over 200 million Americans age 15 or older. Explain why this, along with the fact that the data were obtained using a random sample, satisfies the requirements for constructing a confidence interval.
- (c) Determine and interpret a 95% confidence interval for the mean amount of time Americans age 15 or older spend eating and drinking each day.
- (d) Could the interval be used to estimate the mean amount of time a 9-year-old American spends eating and drinking each day? Explain.

29. Tootsie Pops A Tootsie Pop is a sucker with a candy center. A famous commercial for Tootsie Pops once asked, "How many licks to the center of a Tootsie Pop?" In an attempt to answer this question, Cory Heid of Siena Heights University asked 92 volunteers to count the number of licks required before reaching the chocolate center. The mean number of licks required was 356.1 with a standard deviation of 185.7. Find and interpret a 95% confidence interval for the number of licks required to reach the candy center of a Tootsie Pop.

Source: Heid, Cory. "Tootsie Pops: How Many Licks to the Chocolate?" *Significance*, October, 2013 Volume 10 Issue 5.

30. How Much Do You Read? A recent Gallup poll asked 1006 Americans, "During the past year, about how many books, either hardcover or paperback, did you read either all or part of the way through?" Results of the survey indicated that $\bar{x} = 13.4$ books and $s = 16.6$ books. Construct a 99% confidence interval for the mean number of books that Americans read either all or part of during the preceding year. Interpret the interval.

NW 31. pH of Rain The following data represent the pH of rain for a random sample of 12 rain dates in Tucker County, West

Virginia. A normal probability plot suggests the data could come from a population that is normally distributed. A boxplot indicates there are no outliers.

4.58	5.19	5.05	4.80	4.77	4.77
5.72	4.75	5.02	4.74	4.76	4.56

Source: National Atmospheric Deposition Program.

- (a) Determine a point estimate for the population mean pH of rainwater in Tucker County.
- (b) Construct and interpret a 95% confidence interval for the mean pH of rainwater in Tucker County, West Virginia.
- (c) Construct and interpret a 99% confidence interval for the mean pH of rainwater in Tucker County, West Virginia.
- (d) What happens to the interval as the level of confidence is increased? Explain why this is a logical result.

32. Travel Taxes Travelers pay taxes for flying, car rentals, and hotels. The following data represent the total travel tax for a 3-day business trip in 8 randomly selected cities. **Note:** Chicago travel taxes are the highest in the country at \$101.27. A normal probability plot suggests the data could come from a population that is normally distributed. A boxplot indicates there are no outliers.

67.81	78.69	68.99	84.36
80.24	86.14	101.27	99.29

- (a) Determine a point estimate for the population mean travel tax.
- (b) Construct and interpret a 95% confidence interval for the mean tax paid for a three-day business trip.
- (c) What would you recommend to a researcher who wants to increase the precision of the interval, but does not have access to additional data?

33. Crash Test Results The following data represent the repair cost for a low-impact collision in a simple random sample of mini- and micro-vehicles (such as the Chevrolet Aveo or Mini Cooper).

\$3148	\$1758	\$1071	\$3345	\$743
\$2057	\$663	\$2637	\$773	\$1370

Source: Insurance Institute for Highway Safety.

- (a) Use either Option 1 or Option 2 to verify the requirements for constructing a confidence interval for the population mean are satisfied.
- (b) Construct and interpret a 95% confidence interval for the population mean cost of repair.
- (c) Suppose you obtain a simple random sample of size $n = 10$ of a Mini Cooper that was in a low-impact collision and determine the cost of repair. Do you think a 95% confidence interval would be wider or narrower? Explain.

34. Crawling Babies The following data represent the age (in weeks) at which babies first crawl based on a survey of 12 mothers conducted by Essential Baby.

52	30	44	35	39	26
47	37	56	26	39	28

Source: www.essentialbaby.com

- (a) Use either Option 1 or Option 2 to verify the requirements for constructing a confidence interval for the population mean are satisfied.
- (b) Construct and interpret a 95% confidence interval for the mean age at which a baby first crawls.

DATA 35. Wait Time The following data represent the wait time (in minutes) for a random sample of 40 visitors to Disney's Dinosaur Ride in Animal Kingdom.

6	31	8	0	21	16	0	7
15	6	44	27	7	52	3	7
4	5	10	5	21	3	6	14
5	24	10	9	9	10	12	8
4	8	39	5	28	30	4	15

Source: touringplans.com

- (a) Draw a relative frequency histogram of the data. Comment on the shape of the distribution.
- (b) Draw a boxplot of the data. Are there any outliers?
- (c) Discuss the need for a large sample size in order to use Student's t -distribution to obtain a confidence interval for the population mean wait time at Disney's Dinosaur Ride.
- (d) Construct and interpret a 95% confidence interval for the population mean wait time at Disney's Dinosaur Ride.

DATA 36. PepsiCo Stock Volume The trade volume of a stock is the number of shares traded on a given day. The following data, in millions (so that 6.16 represents 6,160,000 shares traded), represent the volume of PepsiCo stock traded for a random sample of 40 trading days in 2018.

6.16	6.39	5.05	4.41	4.16	4.00	2.37	7.71
4.98	4.02	4.95	4.97	7.54	6.22	4.84	7.29
5.55	4.35	4.42	5.07	8.88	4.64	4.13	3.94
4.28	6.69	3.25	4.80	7.56	6.96	6.67	5.04
7.28	5.32	4.92	6.92	6.10	6.71	6.23	2.42

Source: TD Ameritrade.

- (a) Use the data to compute a point estimate for the population mean number of shares traded per day in 2018.
- (b) Construct a 95% confidence interval for the population mean number of shares traded per day in 2018. Interpret the confidence interval.
- (c) A second random sample of 40 days in 2018 resulted in the data shown next. Construct another 95% confidence interval for the population mean number of shares traded per day in 2018. Interpret the confidence interval.

6.12	5.73	6.85	5.00	4.89	3.79	5.75	6.04
4.49	6.34	5.90	5.44	10.96	4.54	5.46	6.58
8.57	3.65	4.52	7.76	5.27	4.85	4.81	6.74
3.65	4.80	3.39	5.99	7.65	8.13	6.69	4.37
6.89	5.08	8.37	5.68	4.96	5.14	7.84	3.71

- (d) Explain why the confidence intervals obtained in parts (b) and (c) are different.

DATA 37. Threaded Problem: Tornado The data set

"Tornadoes_2017" located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a) Compute the population mean length of a tornado in the United States in 2017.
- (b) Open the data file 9_2_37b. The data represent the length of a random sample of 50 tornadoes in the United States in 2017. Draw a relative frequency histogram of the variable length. Comment on the shape of the distribution.

- (c) Draw a boxplot of the length data from part (b). Are there any outliers?
- (d) Explain why a large sample size is necessary to construct a confidence interval for the mean length of tornadoes in the United States in 2017.
- (e) Construct and interpret a 95% confidence interval for the mean length of a tornado in the United States in 2017. Does your interval include the population mean? What proportion of intervals would you expect to include the population mean?

DATA **38. Tax Rate** The Sullivan Statistics Survey II asks, “What percent of one’s income should an individual pay in federal income taxes?” Go to www.pearsonhighered.com/sullivanstats to obtain the data file SullivanStatsSurveyII using the file format of your choice for the version of the text you are using. The data is in the column “Tax Rate.”

- (a) Draw a relative frequency histogram of the variable “Tax Rate” using a lower class limit of the first class of 0 and a class width of 5. Comment on the shape of the distribution.
- (b) Draw a boxplot of the variable “Tax Rate.” Are there any outliers?
- (c) Explain why a large sample is necessary to construct a confidence interval for the mean tax rate.
- (d) Treat the respondents of this survey as a simple random sample of U.S. residents. Use statistical software to construct and interpret a 90% confidence interval for the mean tax rate U.S. residents feel an individual should pay in federal income taxes.

NW 39. Sample Size Dr. Paul Owsiecmiski wants to estimate the mean serum HDL cholesterol of all 20- to 29-year-old females. How many subjects are needed to estimate the mean serum HDL cholesterol of all 20- to 29-year-old females within 2 points with 99% confidence assuming that $s = 13.4$ based on earlier studies? Suppose that Dr. Owsiecmiski would be content with 95% confidence. How does the decrease in confidence affect the sample size required?

40. Sample Size Dr. Paul Owsiecmiski wants to estimate the mean serum HDL cholesterol of all 20- to 29-year-old males. How many subjects are needed to estimate the mean serum HDL cholesterol of all 20- to 29-year-old males within 1.5 points with 90% confidence, assuming that $s = 12.5$ based on earlier studies? Suppose that Dr. Owsiecmiski would prefer 95% confidence. How does the increase in confidence affect the sample size required?

41. Reading A recent Gallup poll asked Americans to disclose the number of books they read during the previous year. Initial survey results indicate that $s = 16.6$ books.

- (a) How many subjects are needed to estimate the number of books Americans read the previous year within four books with 95% confidence?
- (b) How many subjects are needed to estimate the number of books Americans read the previous year within two books with 95% confidence?
- (c) What effect does doubling the required accuracy have on the sample size?
- (d) How many subjects are needed to estimate the number of books Americans read the previous year within four books with 99% confidence? Compare this result to part (a). How does increasing the level of confidence in the estimate affect sample size? Why is this reasonable?

42. Television A researcher wanted to determine the mean number of hours per week (Sunday through Saturday) the typical person watches television. Results from the Sullivan Statistics Survey I indicate that $s = 7.5$ hours.

- (a) How many people are needed to estimate the number of hours people watch television per week within 2 hours with 95% confidence?
- (b) How many people are needed to estimate the number of hours people watch television per week within 1 hour with 95% confidence?
- (c) What effect does doubling the required accuracy have on the sample size?
- (d) How many people are needed to estimate the number of hours people watch television per week within 2 hours with 90% confidence? Compare this result to part (a). How does decreasing the level of confidence in the estimate affect sample size? Why is this reasonable?

DATA **43. Resistance and Robustness** The data sets represent simple random samples from a population whose mean is 100.

Data Set I

106	122	91	127	88
74	77	108		

Data Set II

106	122	91	127	88
74	77	108	87	88
111	86	113	115	97
122	99	86	83	102

Data Set III

106	122	91	127	88
74	77	108	87	88
111	86	113	115	97
122	99	86	83	102
88	111	118	91	102
80	86	106	91	116

- (a) Compute the sample mean of each data set.
- (b) For each data set, construct a 95% confidence interval about the population mean.
- (c) What effect does the sample size n have on the width of the interval?

For parts (d)–(e), suppose that the data value 106 was accidentally recorded as 016.

- (d) For each data set, construct a 95% confidence interval about the population mean using the incorrectly entered data.
- (e) Which intervals, if any, still capture the population mean, 100? What concept does this illustrate?

DATA **44. Effect of Outliers** The following small data set represents a simple random sample from a population whose mean is 50.

43	63	53	50	58	44
53	53	52	41	50	43

- (a) A normal probability plot indicates that the data could come from a population that is normally distributed with no outliers. Compute a 95% confidence interval for this data set.

- (b) Suppose that the observation, 41, is inadvertently entered into the computer as 14. Verify that this observation is an outlier.
- (c) Construct a 95% confidence interval on the data set with the outlier. What effect does the outlier have on the confidence interval?
- DATA** (d) Consider the following data set, which represents a simple random sample of size 36 from a population whose mean is 50. Verify that the sample mean for the large data set is the same as the sample mean for the small data set from part (a).

43	63	53	50	58	44
53	53	52	41	50	43
47	65	56	58	41	52
49	56	57	50	38	42
59	54	57	41	63	37
46	54	42	48	53	41

- (e) Compute a 95% confidence interval for the large data set. Compare the results to part (a). What effect does increasing the sample size have on the confidence interval?
- (f) Suppose that the last observation, 41, is inadvertently entered as 14. Verify that this observation is an outlier.
- (g) Compute a 95% confidence interval for the large data set with the outlier. Compare the results to part (e). What effect does an outlier have on a confidence interval when the data set is large?

45. Simulation: Normal Distribution IQ scores based on the Wechsler Intelligence Scale for Children (WISC) are known to be approximately normally distributed with $\mu = 100$ and $\sigma = 15$.

- (a) Use StatCrunch, Minitab, or some other statistical software to simulate obtaining 100 simple random samples of size $n = 5$ from this population.
- (b) Obtain the sample mean and sample standard deviation for each of the 100 samples.
- (c) Construct 95% t -intervals for each of the 100 samples.
- (d) How many of the intervals do you expect to include the population mean? How many of the intervals actually include the population mean?

46. Simulation: Exponential Distribution The *exponential probability distribution* can be used to model waiting time in line or the lifetime of electronic components. Its density function is skewed right. Suppose the wait-time in a line can be modeled by the exponential distribution with $\mu = \sigma = 5$ minutes.

- (a) Use StatCrunch, Minitab, or some other statistical software to generate 100 random samples of size $n = 4$ from this population.
- (b) Construct 95% t -intervals for each of the 100 samples found in part (a).
- (c) How many of the intervals do you expect to include the population mean? How many of the intervals actually contain the population mean? Explain what your results mean.
- (d) Repeat parts (a)–(c) for samples of size $n = 15$ and $n = 25$. Explain what your results mean.

47. Putting It Together: Smoking Cessation Study Researchers Havar Brendryen and Pal Kraft conducted a study in which 396 subjects were randomly assigned to either an experimental smoking cessation program or control group. The experimental program consisted of the Internet and phone-based Happy

Ending Intervention, which lasted 54 weeks and consisted of more than 400 contacts by e-mail, web pages, interactive voice response, and short message service (SMS) technology. The control group received a self-help booklet. Both groups were offered free nicotine replacement therapy. Abstinence was defined as “not even a puff of smoke, for the last 7 days,” and assessed by means of Internet surveys or telephone interviews. The response variable was abstinence after 12 months. Of the participants in the experimental program, 22.3% reported abstinence; of the participants in the control group, 13.1% reported abstinence.

Source: “Happy Ending: A Randomized Controlled Trial of a Digital MultiMedia Smoking Cessation Intervention.” Havar Brendryen and Pal Kraft. *Addiction* 103(3):478–484, 2008.

- (a) What type of experimental design is this?
- (b) What is the treatment? How many levels does it have?
- (c) What is the response variable?
- (d) What are the statistics reported by the authors?
- (e) An **odds ratio** is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. These groups might be men and women, an experimental group and a control group, or any other dichotomous classification. If the probabilities of the event in each of the groups are p (first group) and q (second group), then the odds ratio is

$$\frac{\frac{p}{(1-p)}}{\frac{q}{(1-q)}} = \frac{p(1-q)}{q(1-p)}$$

An odds ratio of 1 indicates that the condition or event under study is equally likely in both groups. An odds ratio greater than 1 indicates that the condition or event is more likely in the first group. And an odds ratio less than 1 indicates that the condition or event is less likely in the first group. The odds ratio must be greater than or equal to zero. As the odds of the first group approach zero, the odds ratio approaches zero. As the odds of the second group approach zero, the odds ratio approaches positive infinity. Verify that the odds ratio for this study is 1.90. What does this mean?

- (f) The authors of the study reported a 95% confidence interval for the odds ratio to be *lower bound*: 1.12 and *upper bound* 3.26. Interpret this result.
- (g) Write a conclusion that generalizes the results of this study to the population of all smokers.

Retain Your Knowledge

48. How Many Drinks? A question on the General Social Survey was, “When you drink, how many drinks do you have?” The survey was administered to a random sample of 243 adult Americans aged 21 or older. Go to www.pearsonhighered.com/sullivanstats to obtain the data file 9_2_48 using the file format of your choice for the version of the text you are using.

- (a) What type of variable is “number of drinks”?
- (b) Draw a histogram of the data and comment on the shape of the distribution.
- (c) Determine the mean and standard deviation for number of drinks.
- (d) What is the mode number of drinks?

- (e) What is the probability a randomly selected individual consumes two drinks?
- (f) Would it be unusual to observe an individual who consumes at least eight drinks? Why?
- (g) Describe the shape of the distribution of the sample mean number of drinks. What is the source of variability in the sampling distribution?
- (h) Construct a 95% confidence interval for the mean number of drinks. Interpret this interval.

Explaining the Concepts

49. Explain why the t -distribution has less spread as the number of degrees of freedom increases.
50. The procedure for constructing a t -interval is *robust*. Explain what this means.
51. Explain what is meant by *degrees of freedom*.
52. The mean age of the 45 presidents of the United States (as of 2018) on the day of inauguration is 55.0 years, with a standard deviation of 6.6 years. A researcher constructed a 95%

confidence interval for the mean age of presidents on inauguration day. He wrote that he was 95% confident that the mean age of the president on inauguration day is between 53.0 and 57.0 years of age. Is there anything wrong with the researcher's analysis? Explain.

53. Suppose you have two populations: Population A—All students at Illinois State University ($N = 21,000$) and Population B—All residents of the city of Homer Glen, IL ($N = 21,000$). You want to estimate the mean age of each population using two separate samples each of size $n = 75$. If you construct a 95% confidence interval for each population mean, will the margin of error for population A be larger, the same, or smaller than the margin of error for population B? Justify your reasoning.

54. Population A has standard deviation $\sigma_A = 5$, and population B has standard deviation $\sigma_B = 10$. How many times larger than Population A's sample size does Population B's need to be to estimate μ with the same margin of error? (Hint: Compute n_B/n_A .)

9.3 Putting It Together: Which Method Do I Use?

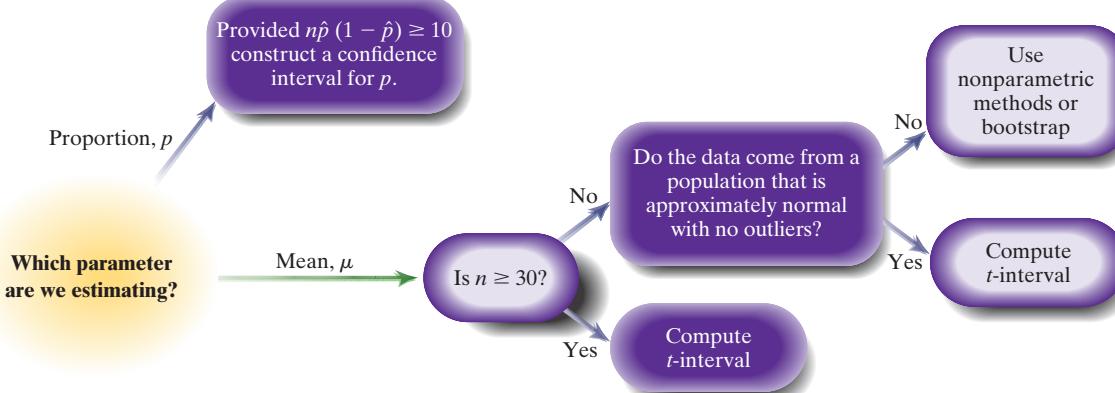


Objective ① Determine the appropriate confidence interval to construct

① Determine the Appropriate Confidence Interval to Construct

Perhaps the most difficult aspect of constructing a confidence interval is determining which type to construct. To assist in your decision making, we present Figure 17.

Figure 17



EXAMPLE 1

Constructing a Confidence Interval: Which Method Do I Use?

Table 4

323.9	326.8	370.6
450.7	368.8	423.8
398.8	417.5	
382.7	343.1	

Problem Robert wishes to estimate the mean number of miles that his Buick LaCrosse can be driven on a full tank of gas. He fills up his car with regular unleaded gasoline from the same gas station 10 times and records the number of miles that he drives until his low-tank indicator light comes on. He obtains the data shown in Table 4. Construct a 95% confidence interval for the mean number of miles driven on a full tank of gas.

Approach Follow the flow chart given in Figure 17.

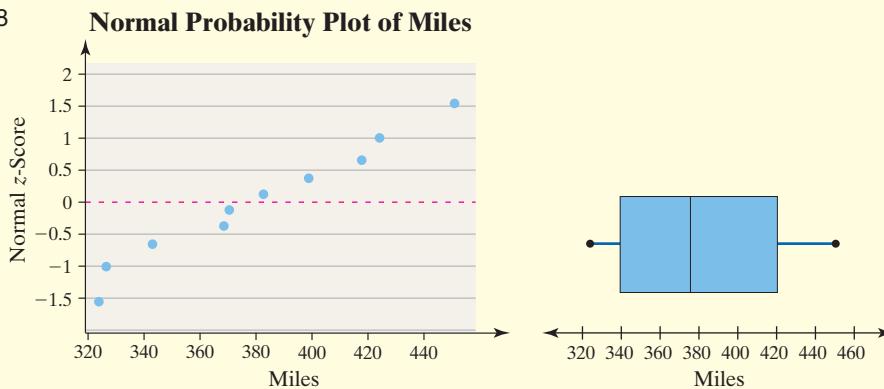
(continued)

NOTE

The boxplot in Figure 18 suggests the sample data is roughly symmetric with no outliers. Due to the robustness in constructing confidence intervals for a population mean, we may proceed with construction of the confidence interval.

Solution We are asked to construct a 95% confidence interval for the *mean* number of miles driven. Treat the data as a simple random sample from a large population. Because the sample size is small, verify that the data come from a population that is normally distributed with no outliers by drawing a normal probability plot and boxplot. See Figure 18. The correlation between miles and expected *z*-scores in the normal probability plot is 0.987. Because $0.987 > 0.918$ (Table VI), conclude the data come from a population that is normally distributed. The boxplot shows there are no outliers. We may proceed with constructing the confidence interval for the population mean.

Figure 18

**By-Hand Solution**

From the sample data in Table 4, we have $n = 10$, $\bar{x} = 380.67$, and $s = 42.47$. For a 95% confidence interval with $n - 1 = 10 - 1 = 9$ degrees of freedom, we have

$$t_{\frac{\alpha}{2}} = t_{\frac{0.05}{2}} = t_{0.025} = 2.262$$

Lower bound:

$$\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 380.67 - 2.262 \cdot \frac{42.47}{\sqrt{10}} = 350.29$$

Upper bound:

$$\bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 380.67 + 2.262 \cdot \frac{42.47}{\sqrt{10}} = 411.05$$

Technology Solution

Figure 19 shows the results from StatCrunch.

Figure 19

95% Confidence Interval Results

μ : mean of Variable

Variable	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
Miles	380.67	13.429769	9	350.28976	411.05023

So

Lower bound: 350.29 Upper bound: 411.05

Interpretation We are 95% confident that the population mean miles driven on a full tank of gas is between 350.29 and 411.05.



9.3 Assess Your Understanding

Skill Building

- For what type of variable does it make sense to construct a confidence interval about a population proportion?
- For what type of variable does it make sense to construct a confidence interval about a population mean?
- What requirements must be satisfied in order to construct a confidence interval about a population proportion?
- What requirements must be satisfied in order to construct a confidence interval about a population mean?

In Problems 5–12, construct the appropriate confidence interval.

- A simple random sample of size $n = 300$ individuals who are currently employed is asked if they work at home at least once per week. Of the 300 employed individuals surveyed, 35 responded that they did work at home at least once per week. Construct a 99% confidence interval for the population proportion of employed individuals who work at home at least once per week.
- A simple random sample of size $n = 785$ adults was asked if they follow college football. Of the 785 surveyed, 275 responded

that they did follow college football. Construct a 95% confidence interval for the population proportion of adults who follow college football.

7. A simple random sample of size $n = 12$ is drawn from a population that is normally distributed. The sample mean is found to be $\bar{x} = 45$, and the sample standard deviation is found to be $s = 14$. Construct a 90% confidence interval for the population mean.

8. A simple random sample of size $n = 17$ is drawn from a population that is normally distributed. The sample mean is found to be $\bar{x} = 3.25$, and the sample standard deviation is found to be $s = 1.17$. Construct a 95% confidence interval for the population mean.

9. A simple random sample of size $n = 40$ is drawn from a population. The sample mean is found to be $\bar{x} = 120.5$, and the sample standard deviation is found to be $s = 12.9$. Construct a 99% confidence interval for the population mean.

10. A simple random sample of size $n = 210$ is drawn from a population. The sample mean is found to be $\bar{x} = 20.1$, and the sample standard deviation is found to be $s = 3.2$. Construct a 90% confidence interval for the population mean.

Applying the Concepts

11. Autonomous Vehicles Self-driving vehicles periodically suffer from a disengagement of the self-driving feature. In these cases, it is important to know how long it takes for the driver to manually take control of the vehicle. In a study of 487 instances where a self-driving vehicle became disengaged from the self-driving feature, the mean time for the driver to take control was 0.84 second with a standard deviation of 0.90 second.

Source: Dixit VV, Chand S, Nair DJ (2016) "Autonomous Vehicles: Disengagements, Accidents, and Reaction Times," *PLoS ONE*, 11(12):e0168054. <https://doi.org/10.1371/journal.pone.0168054>

- (a) An analysis of the reaction times indicates that reaction time to disengagement is highly skewed right. Explain why a large sample size would be necessary to construct a confidence interval for the mean reaction time.
- (b) Construct and interpret a 90% confidence interval for the mean reaction time to disengagement.

12. Click It Based on a poll conducted by the Centers for Disease Control, 862 of 1013 randomly selected adults said that they always wear seat belts. Construct and interpret a 95% confidence interval for the proportion of adults who always wear seat belts.

13. Estate Tax Returns In a random sample of 100 estate tax returns that was audited by the Internal Revenue Service, it was determined that the mean amount of additional tax owed was \$3421 with a standard deviation of \$2583. Construct and interpret a 90% confidence interval for the mean additional amount of tax owed for estate tax returns.

14. Muzzle Velocity Fifty rounds of a new type of ammunition were fired from a test weapon, and the muzzle velocity of the projectile was measured. The sample had a mean muzzle velocity of 863 meters per second and a standard deviation of 2.7 meters per second. Construct and interpret a 99% confidence interval for the mean muzzle velocity.

15. Worried about Retirement? In a survey of 1008 adult Americans, the Gallup organization asked, "When you retire, do you think you will have enough money to live comfortably or not?" Of the 1008 surveyed, 526 stated that they were worried about having enough money to live comfortably in retirement.

Construct a 90% confidence interval for the proportion of adult Americans who are worried about having enough money to live comfortably in retirement.

16. Theme Park Spending In a random sample of 40 visitors to a certain theme park, it was determined that the mean amount of money spent per person at the park (including ticket price) was \$93.43 per day with a standard deviation of \$15. Construct and interpret a 99% confidence interval for the mean amount spent daily per person at the theme park.

17. Fastball Clayton Kershaw of the Los Angeles Dodgers is one of the premier pitchers in baseball. His most popular pitch is a four-seam fastball. The data below represent the pitch speed (in miles per hour) for a random sample of 18 of his four-seam fastball pitches.

93.63	93.83	94.18	94.71	95.52	95.07
95.12	95.35	94.15	94.62	96.08	93.86
94.75	94.70	95.28	95.49	95.77	93.34

Source: Brooksbaseball.net

- (a) Is "pitch speed" a quantitative or qualitative variable? Why is it important to know this when determining the type of confidence interval you may construct?
- (b) Use either Option 1 or Option 2 to verify the requirements for constructing a confidence interval for the population mean are satisfied.
- (c) Construct and interpret a 95% confidence interval for the mean pitch speed of a Clayton Kershaw four-seam fastball.
- (d) Do you believe that a 95% confidence interval for the mean pitch speed of all major league pitchers' four-seam fastball would be narrower or wider? Why?

18. Annoying Behavior Harris Interactive conducted a poll of a random sample of 2234 adult Americans 18 years of age or older and asked, "Which is more annoying to you, tailgaters or slow drivers who stay in the passing lane?" Among those surveyed, 1184 were more annoyed by tailgaters.

- (a) Explain why the variable of interest is qualitative with two possible outcomes. What are the two outcomes?
- (b) Verify the requirements for constructing a 90% confidence interval for the population proportion of all adult Americans who are more annoyed by tailgaters than slow drivers in the passing lane.
- (c) Construct a 90% confidence interval for the population proportion of all adult Americans who are more annoyed by tailgaters than slow drivers in the passing lane.

19. Sleep Apnea and Gum Disease Sleep apnea is a disorder in which you have one or more pauses in breathing or shallow breaths while you sleep. In a cross-sectional study of 320 individuals who suffer from sleep apnea, it was found that 192 had gum disease. *Note:* In the general population, about 17.5% of individuals have gum disease.

Source: Seo, WH et al. "The Association between Periodontitis and Obstructive Sleep Apnea: A Preliminary Study," *J Periodontal Res* 2013 Aug; 48 (4).

- (a) What does it mean for this study to be cross-sectional?
- (b) What is the variable of interest in this study? Is it qualitative or quantitative? Explain.
- (c) Estimate the proportion of individuals who suffer from sleep apnea who have gum disease with 95% confidence. Interpret your result.

- DATA** **20.** The following data represent the property tax for a random sample of 50 single-family homes in the city of Houston.

Suppose you want to estimate the typical property real estate tax for a single-family home in the city of Houston.

5171.13	631.14	1418.64	456.69	1512.15
481.26	299.82	168.31	0	4570.1
1299.73	741.78	1153.19	613.64	0
2254.20	614.02	2188.74	2183.94	172.40
12718	363.97	677.40	1056.59	834.17
1942.26	582.71	1191.99	3481.21	0
1681.39	0	2419.75	495.72	505.34
0	485.85	206.48	2112.28	2244.02
135.19	549.83	503.01	298.71	521.50
399.17	160.40	1425.93	9183.38	796.47

Source: Open Data Houston.

- (a) What type of variable is “Tax” in the data set?
- (b) What type of confidence interval should be constructed for this variable?
- (c) Draw a boxplot of the sample data.
- (d) Construct a 90% confidence interval for the typical property tax on a single-family home in Houston. Be sure the model requirements have been satisfied.

- DATA** **21. Short-Term Rentals** Go to www.pearsonhighered.com/sullivanstats to obtain the data file 9_4_21. The data represent the status of licenses for short-term rentals in New Orleans, where “Yes” indicates the license is expired and “No” indicates the license is not expired. Source: City of New Orleans Open Data.

- (a) What type of variable is the variable “Expired” in the data set?
- (b) Estimate with 95% confidence the likelihood that you will select a short-term rental in New Orleans where the license is expired. Be sure all model requirements for obtaining the estimate are satisfied. **Hint:** There are 13,422 short-term rentals in the city of New Orleans.

Retain Your Knowledge

- 22. Weight Gain** Researchers conducted a study to see the effect of specific lifestyle and dietary changes for preventing long-term weight gain. The study involved the consolidation of three cohorts from the Nurses’ Health Study (NHS): (1) cohort of 121,701 female registered nurses, who enrolled in 1976; (2) the Nurses’ Health Study II (NSH II), a cohort of 116,686 younger female nurses, who enrolled in 1989; and (3) the Health Professionals Follow-up Study (HPFS), a cohort of 51,529 male health professionals, who enrolled in 1986. Participants were followed with biennial questionnaires concerning medical history, lifestyle, and health practices. The baseline year of the study was used to determine diet, physical activity, and smoking habits of the participants. The final analysis included 50,422 from the NHS, 49,898 from the NHS II, and 22,557 from the HPFS (those with health issues or obesity

were excluded from the study). Individuals were categorized based on various lifestyle choices. For example, those who disclosed an increase in consumption of french fries had a mean weight gain of 3.35 pounds over a 4-year period with a 95% confidence interval of 2.28 pounds to 4.42 pounds. Among those who disclosed a change in smoking status from former smoker to current smoker the mean weight gain was −2.47 pounds over a 4-year period with a 95% confidence interval of −3.82 pounds to −1.12 pounds. Source: Dariush Mozaffarian, M.D. et al. “Changes in Diet and Lifestyle and Long-Term Weight Gain in Women and Men,” *The New England Journal of Medicine* 364;25.

- (a) Explain what it means for this study to be a cohort study.
- (b) Is the variable weight gain quantitative or qualitative? How was the variable measured?
- (c) Within the cohort “individuals who increased their consumption of french fries during the observation period,” there was a mean weight gain of 3.35 pounds over the 4-year period. Is the variable “increase in consumption of french fries” quantitative or qualitative? Explain.
- (d) What is the margin of error for the 95% confidence interval for increase in weight gain among those who increased their consumption of french fries?
- (e) What does a mean weight gain of −2.47 pounds among those whose smoking status changed from former to current smoker suggest?
- (f) Interpret the 95% confidence interval for the mean weight gain among those whose smoking status changed from former to current smoker.
- (g) Describe the population to which the results of this study apply.

In Problems 23 through 26, indicate whether a confidence interval for a proportion or mean should be constructed to estimate the value of the variable of interest. Justify your response.

- 23.** A developmental mathematics instructor wishes to estimate the typical amount of time students dedicate to studying mathematics in a week. She asks a random sample of 50 students enrolled in developmental mathematics at her school to report the amount of time spent studying mathematics in the past week.
- 24.** Researchers at the Gallup Organization asked a random sample of 1016 adult Americans aged 21 years or older, “Right now, do you think the state of moral values in the country as a whole is getting better, or getting worse?”
- 25.** A researcher wanted to know whether consumption of green tea on a daily basis reduces LDL (bad) cholesterol. She obtains a random sample of 500 subjects. Each subject consumes at least 1 cup of green tea daily for 1 year. After 1 year, the researcher determines whether the subjects LDL cholesterol decreased, or not.
- 26.** Does chewing your food for a longer period of time reduce one’s caloric intake of food at dinner? A researcher requires a sample of 75 healthy males to chew their food twice as long as they normally do. The researcher then records the calorie consumption at dinner.



Chapter 9 Review

Summary

In this chapter, we discussed estimation methods. We estimated the values of the parameters p and μ . We started by estimating the value of a population proportion, p . A confidence interval is an interval of numbers for an unknown parameter that is reported with a level of confidence. The level of confidence represents the proportion of intervals that will contain the parameter if a large number of different samples is obtained and is denoted $(1 - \alpha) \cdot 100\%$. For example, when constructing a 95% confidence interval, we would expect 95 of 100 different random samples from the same population to include the unknown parameter.

In Section 9.1, we learned how to estimate a population proportion. A $(1 - \alpha) \cdot 100\%$ confidence interval for the population proportion p is given by

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}},$$

provided the data are obtained using simple random sampling or through a randomized

experiment, $n\hat{p}(1 - \hat{p}) \geq 10$, and the sample size is no more than 5% of the population size (independence requirement).

In Section 9.2 we learned how to estimate the population mean, μ . To construct this interval, the sample must come from a population that is normally distributed or the sample size is large (so that the distribution of \bar{x} is approximately normal). We also required that the sample size be no more than 5% of the population size (independence requirement) and that the data be obtained using simple random sampling or through a randomized experiment. Because the population standard deviation is likely unknown (since we do not know μ , how could we expect to know σ ?), we use Student's t -distribution with $n - 1$ degrees of freedom to construct a confidence interval for the population mean. A $(1 - \alpha) \cdot 100\%$ confidence interval for the population mean, μ , is given by

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ has $n - 1$ degrees of freedom.

Vocabulary

Point estimate (p. 396)

Margin of error (pp. 398, 402, 415)

t -interval (p. 415)

Confidence interval (p. 397)

Critical value (p. 400)

Robust (p. 416)

Level of confidence (p. 397)

Student's t -distribution (p. 412)

Nonparametric procedures (p. 418)

Formulas

Confidence Intervals

- A $(1 - \alpha) \cdot 100\%$ confidence interval for p is

$$\hat{p} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}},$$

provided that $n\hat{p}(1 - \hat{p}) \geq 10$ and $n \leq 0.05N$.

- A $(1 - \alpha) \cdot 100\%$ confidence interval for μ is

$$\bar{x} \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}},$$

where $t_{\alpha/2}$ has $n - 1$ degrees of freedom,

provided that the population from which the sample was drawn is normal or that the sample size is large ($n \geq 30$) and $n \leq 0.05N$.

Sample Size

- To estimate the population proportion within a margin of error E at a $(1 - \alpha) \cdot 100\%$ level of confidence requires a sample of size

$$n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2$$

(rounded up to the next integer), where \hat{p} is a prior estimate of the population proportion.

- To estimate the population proportion within a margin of error E at a $(1 - \alpha) \cdot 100\%$ level of confidence requires a sample of size $n = 0.25 \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2$ (rounded up to the next integer) when no prior estimate is available.

- To estimate the population mean within a margin of error E at a $(1 - \alpha) \cdot 100\%$ level of confidence

$$\text{requires a sample of size } n = \left(\frac{z_{\frac{\alpha}{2}} \cdot s}{E} \right)^2$$

(rounded up to the next integer).

Objectives

Section	You should be able to ...	Example(s)	Review Exercises
9.1	1 Obtain a point estimate for the population proportion (p. 396) 2 Construct and interpret a confidence interval for the population proportion (p. 396) 3 Determine the sample size necessary for estimating a population proportion within a specified margin of error (p. 404)	1 2–5 6	15(a) 15(b) 15(c), 15(d)
9.2	1 Obtain a point estimate for the population mean (p. 410) 2 State properties of Student's <i>t</i> -distribution (p. 411) 3 Determine <i>t</i> -values (p. 414) 4 Construct and interpret a confidence interval for a population mean (p. 415) 5 Determine the sample size needed to estimate a population mean within a specified margin of error (p. 418)	1 2 3 4 5	14(a) 5, 6, 7 1 8, 9(a)–(c), 10(b), 11(b), 12(b), 12(d), 13(b), 13(c), 14(c) 10(c)
9.3	1 Determine the appropriate confidence interval to construct (p. 425)	1	8–15

Review Exercises

1. Find the critical *t*-value for constructing a confidence interval for a population mean at the given level of confidence for the given sample size, n .

- (a) 99% confidence; $n = 18$
(b) 90% confidence; $n = 27$

2. **IQ Scores** Many of the examples and exercises in the text have dealt with IQ scores. We now know that IQ scores based on the Stanford–Binet IQ test are approximately normally distributed with a mean of 100 and standard deviation 15. If you were to obtain 100 different simple random samples of size 20 from the population of all adult humans and determine 95% confidence intervals for each of them, how many of the intervals would you expect to include 100? What causes a particular interval to not include 100?

3. What does the 95% represent in a 95% confidence interval?

4. For what proportion of samples will a 90% confidence interval for a population mean not capture the true population mean?

5. The area under the *t*-distribution with 18 degrees of freedom to the right of $t = 1.56$ is 0.0681. What is the area under the *t*-distribution with 18 degrees of freedom to the left of $t = -1.56$? Why?

6. Which is larger, the area under the *t*-distribution with 10 degrees of freedom to the right of $t = 2.32$ or the area under the standard normal distribution to the right of $z = 2.32$? Why?

7. State the properties of Student's *t*-distribution.

8. A simple random sample of size n is drawn from a population. The sample mean, \bar{x} , is 54.8 and the sample standard deviation is 10.5.

- (a) Construct the 90% confidence interval for the population mean if the sample size, n , is 30.
(b) Construct the 90% confidence interval for the population mean if the sample size, n , is 51. How does increasing the sample size affect the width of the interval?
(c) Construct the 99% confidence interval for the population mean if the sample size, n , is 30. Compare the results to

those obtained in part (a). How does increasing the level of confidence affect the confidence interval?

9. A simple random sample of size n is drawn from a population that is known to be normally distributed. The sample mean, \bar{x} , is determined to be 104.3 and the sample standard deviation, s , is determined to be 15.9.

- (a) Construct the 90% confidence interval for the population mean if the sample size, n , is 15.
(b) Construct the 90% confidence interval for the population mean if the sample size, n , is 25. How does increasing the sample size affect the width of the interval?
(c) Construct the 95% confidence interval for the population mean if the sample size, n , is 15. Compare the results to those obtained in part (a). How does increasing the level of confidence affect the confidence interval?

10. **Long Life?** In a survey of 35 adult Americans, it was found that the mean age (in years) that people would like to live to is 87.9 with a standard deviation of 15.5. An analysis of the raw data indicates the distribution is skewed left.

- (a) Explain why a large sample size is necessary to construct a confidence interval for the mean age that people would like to live.
(b) Construct and interpret a 95% confidence interval for the mean.
(c) How many adult Americans would need to be surveyed to estimate the mean age that people would like to live to within 2 years with 95% confidence?

11. **E-mail** The General Social Survey asked: “How many e-mails do you send in a day?” The results of 928 respondents indicate that the mean number of e-mails sent in a day is 10.4, with a standard deviation of 28.5.

- (a) Given the fact that 1 standard deviation to the left of the mean results in a negative number of e-mails being sent, what shape would you expect the distribution of e-mails sent to have?
(b) Construct and interpret a 90% confidence interval for the mean number of e-mails sent per day.

12. Caffeinated Sports Drinks Researchers conducted an experiment to determine the effectiveness of a commercial caffeinated carbohydrate-electrolyte sports drink compared with a placebo. Sixteen highly trained cyclists each completed two trials of prolonged cycling in a warm environment, one while receiving the sports drink and another while receiving a placebo. For a given trial, one beverage treatment was administered throughout a 2-hour variable-intensity cycling bout followed by a 15-minute performance ride. Total work (in kilojoules) performed during the final 15 minutes was used to measure performance. The beverage order for individual subjects was randomly assigned with a period of at least five days separating the trials. Assume that the researchers verified the normality of the population of total work performed for each treatment. *Source: Kirk J. Cureton, Gordon L. Warren, et al., "Caffeinated Sports Drink: Ergogenic Effects and Possible Mechanisms." International Journal of Sport Nutrition and Exercise Metabolism 17(1):35–55, 2007.*

- (a) Why do you think the sample size was small ($n = 16$) for this experiment?
- (b) For the sports-drink treatment, the mean total work performed during the performance ride for the $n = 16$ riders was 218 kilojoules, with standard deviation 31 kilojoules. Construct and interpret a 95% confidence interval for the population mean total work performed.
- (c) Is it possible for the population mean total work performed for the sports-drink treatment to be less than 198 kilojoules? Do you think this is likely?
- (d) For the placebo treatment, the mean total work performed during the performance ride for the $n = 16$ riders was 178 kilojoules, with standard deviation 31 kilojoules. Construct and interpret a 95% confidence interval for the population mean total work performed.
- (e) Is it possible for the population mean total work performed for the placebo treatment to be more than 198 kilojoules? Do you think this is likely?
- (f) The researchers concluded that the caffeinated carbohydrate-electrolyte sports drink substantially enhanced physical performance during prolonged exercise compared with the placebo. Do your findings in parts (b) and (d) support the researchers' conclusion? Explain.

DATA 13. Family Size A random sample of 60 married couples who have been married 7 years was asked the number of children they have. The results of the survey are as follows:

0	0	0	3	3	3	1	3	2	2	3	1
3	2	4	0	3	3	3	1	0	2	3	3
1	4	2	3	1	3	3	5	0	2	3	0
4	4	2	2	3	2	2	2	2	3	4	3
2	2	1	4	3	2	4	2	1	2	3	2

Note: $\bar{x} = 2.27$, $s = 1.22$.

- (a) What is the shape of the distribution of the sample mean? Why?
- (b) Compute a 95% confidence interval for the mean number of children of all couples who have been married 7 years. Interpret this interval.
- (c) Compute a 99% confidence interval for the mean number of children of all couples who have been married 7 years. Interpret this interval.

DATA 14. Diameter of Douglas Fir Trees The diameter of the Douglas fir tree is measured at a height of 1.37 meters. The following data represent the diameter in centimeters of a

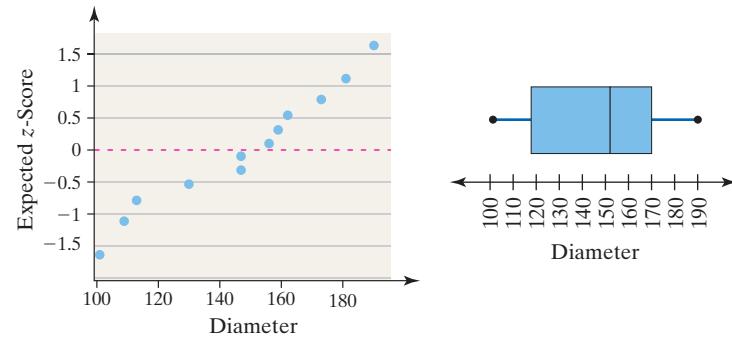
random sample of 12 Douglas firs in the western Washington Cascades.

156	190	147	173	159	181
162	130	101	147	113	109

Source: L. Winter. "Live Tree and Tree-Ring Records to Reconstruct the Structural Development of an Old-Growth Douglas Fir/Western Hemlock Stand in the Western Washington Cascades." Corvallis, OR: Forest Science Data Bank, 2005.

- (a) Obtain a point estimate for the mean and standard deviation diameter of a Douglas fir tree in the western Washington Cascades.
- (b) Because the sample size is small, we must verify that the data come from a population that is normally distributed and that the data do not contain any outliers. The figures show the normal probability plot and boxplot. The correlation between the tree diameters and expected z -scores is 0.982. Use Option 1 or Option 2 to determine if the conditions for constructing a confidence interval for the population mean diameter are satisfied.

Normal Probability Plot of Diameter of Douglas Fir Trees



- (c) Construct a 95% confidence interval for the mean diameter of a Douglas fir tree in the western Washington Cascades.

15. Hypertension In a random sample of 678 adult males 20 to 34 years of age, it was determined that 58 of them have hypertension (high blood pressure).

Source: The Centers for Disease Control.

- (a) Obtain a point estimate for the proportion of adult males 20 to 34 years of age who have hypertension.
- (b) Construct a 95% confidence interval for the proportion of adult males 20 to 34 years of age who have hypertension. Interpret the confidence interval.
- (c) You wish to conduct your own study to determine the proportion of adult males 20 to 34 years old who have hypertension. What sample size would be needed for the estimate to be within 3 percentage points with 95% confidence if you use the point estimate obtained in part (a)?
- (d) You wish to conduct your own study to determine the proportion of adult males 20 to 34 years old who have hypertension. What sample size would be needed for the estimate to be within 3 percentage points with 95% confidence if you don't have a prior estimate?



Chapter Test

- Find the critical t -value for constructing a confidence interval about a population mean at the given level of confidence for the given sample size, n .
 - 96% confidence; $n = 26$
 - 98% confidence; $n = 18$
 - Determine the point estimate of the population mean and margin of error if the confidence interval has lower bound: 125.8 and upper bound: 152.6.
 - A question on the General Social Survey was this: "How many family members do you know that are in prison?" The results of 499 respondents indicate that the mean number of family members in jail is 1.22, with a standard deviation of 0.59.
 - What shape would you expect the distribution of this variable to have? Why?
 - Construct and interpret a 99% confidence interval for the mean number of family members in jail.
 - A random sample of 50 recent college graduates results in a mean time to graduate of 4.58 years, with a standard deviation of 1.10 years.
- Source:* Based on data from *The Toolbox Revisited* by Clifford Adelman, U.S. Department of Education.

- Compute and interpret a 90% confidence interval for time to graduate with a bachelor's degree.
- Does this evidence contradict the widely held belief that it takes 4 years to complete a bachelor's degree? Why?

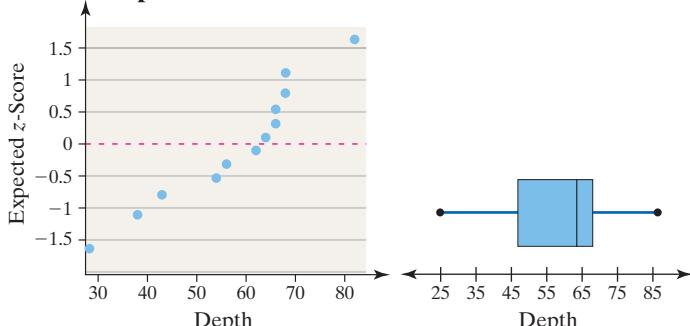
- DATA** 5. The campus at Joliet Junior College has a lake. A student used a Secchi disk to measure the clarity of the lake's water by lowering the disk into the water and measuring the distance below the water surface at which the disk is no longer visible. The following measurements (in inches) were taken on the lake at various points in time over the course of a year.

82	64	62	66	68	43
38	26	68	56	54	66

Source: Virginia Piekarski, Joliet Junior College.

- Use the data to compute a point estimate for the population mean and population standard deviation.
- Because the sample size is small, we must verify that the data are normally distributed and do not contain any outliers. The figures below show the normal probability plot and boxplot. The correlation between depth and expected z -scores is 0.960. Are the conditions for constructing a confidence interval about μ satisfied?

Normal Probability Plot of Depth of Secchi Disk



- Construct a 95% confidence interval for the mean Secchi disk measurement. Interpret this interval.

- Construct a 99% confidence interval for the mean Secchi disk measurement. Interpret this interval.
- From a random sample of 1201 Americans, it was discovered that 1139 of them lived in neighborhoods with acceptable levels of carbon monoxide.
Source: Environmental Protection Agency.
 - Obtain a point estimate for the proportion of Americans who live in neighborhoods with acceptable levels of carbon monoxide.
 - Construct a 99% confidence interval for the proportion of Americans who live in neighborhoods with acceptable levels of carbon monoxide.
 - You wish to conduct your own study to determine the proportion of Americans who live in neighborhoods with acceptable levels of carbon monoxide. What sample size would be needed for the estimate to be within 1.5 percentage points with 90% confidence if you use the estimate obtained in part (a)?
 - You wish to conduct your own study to determine the proportion of Americans who live in neighborhoods with acceptable levels of carbon monoxide. What sample size would be needed for the estimate to be within 1.5 percentage points with 90% confidence if you do not have a prior estimate?

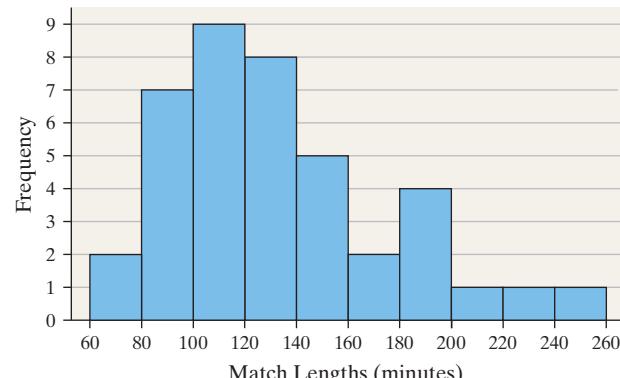
- DATA** 7. **Wimbledon Match Lengths** A tennis enthusiast wants to estimate the mean length of men's singles matches held during the Wimbledon tennis tournament. From the Wimbledon history archives, he randomly selects 40 matches played during the tournament since the year 1968 (when professional players were first allowed to participate). The lengths of the 40 selected matches, in minutes, follow:

110	76	84	231	122	115	87	137
101	119	138	132	136	111	194	92
198	153	256	146	149	103	116	163
182	132	123	178	140	151	115	107
202	128	60	89	94	95	89	182

Source: www.wimbledon.org

- Obtain a point estimate of the population mean length of men's singles matches during Wimbledon.
- A frequency histogram of the data is shown below. Explain why a large sample is necessary to construct a confidence interval for the population mean length of men's singles matches during Wimbledon.

Wimbledon Match Lengths



- (c) Construct and interpret a 99% confidence interval for the population mean length of men's singles matches during Wimbledon.
- (d) Construct and interpret a 95% confidence interval for the population mean length of men's singles matches during Wimbledon.

- (e) What effect does increasing the level of confidence have on the interval?
- (f) Do the confidence intervals computed in parts (c) and (d) represent an estimate for the population mean length of men's singles matches during all professional tennis tournaments? Why?

Making an Informed Decision

How Much Should I Spend for this House?

One of the biggest purchases we make in our lifetimes is for a home. Questions that we all ask are these:

- How much should I spend for a particular home?
- How many bathrooms are there?
- How long should I expect a home to be on the market?
- What is the cost per square foot?

The purpose of this project is to help you make an informed decision about housing values. This will help to ensure you receive a good deal when purchasing a home.

(a) Go to a real estate website such as www.realtor.com or www.zillow.com and enter the particular zip code you are interested in moving to. Randomly select at least 30 homes for sale and record the following information:

- Asking price
- Square footage
- Number of days on the market
- Cost per square foot (asking price divided by square footage)

(b) For each of the variables identified, determine a 95% confidence interval. Interpret the interval.

(c) Now randomly select 30 recently sold homes and determine the percentage discount from the asking price. This is, determine discount by computing

$$\frac{\text{asking price} - \text{closing price}}{\text{asking price}}$$



(d) Determine a 95% confidence interval for percentage discount. Interpret the interval.

(e) For the type of house you are considering (such as a 2400 square foot 3-bedroom/2-bath home), identify at least 20 homes that are for sale in the neighborhood you are considering. Compute a 95% confidence interval for the asking price of this type of home.

(f) Write a report that details how much you should expect to pay for the type of house you are considering.



Hypothesis Tests Regarding a Parameter

Outline

- 10.1** The Language of Hypothesis Testing
- 10.2** Hypothesis Tests for a Population Proportion
- 10.3** Hypothesis Tests for a Population Mean
- 10.4** Putting It Together: Which Method Do I Use?
- 10.5** Hypothesis Tests for a Population Standard Deviation (eText only)

Making an Informed Decision



Suppose you have just received a \$1000 bonus at your job. Rather than waste the money on frivolous items, you decide to invest the money so you can use it later to buy a home. You have many investment options. Your family and friends who have some experience investing recommend mutual funds. Which fund should you choose? See the Decision project on page 476.

Putting It Together

In Chapter 9, we mentioned there are two types of inferential statistics: (1) estimation and (2) hypothesis testing. We have already discussed the procedures for estimating the population proportion and the population mean.

We now focus our attention on hypothesis testing. Hypothesis testing is used to test statements regarding a characteristic of one or more populations. In this chapter, we will test various hypotheses regarding a single population parameter. The hypotheses that we test concern the population proportion and the population mean.

10.1 The Language of Hypothesis Testing



Preparing for This Section Before getting started, review the following:

- Parameter versus statistic (Section 1.1, p. 5)
- Simple random sampling (Section 1.3, pp. 23–27)
- Table 9 (Section 6.2, p. 316)
- Sampling distribution of \bar{x} (Section 8.1, pp. 371–379)

- Objectives**
- ① Determine the null and alternative hypotheses
 - ② Explain Type I and Type II errors
 - ③ State conclusions to hypothesis tests

We begin with an example.

EXAMPLE 1 Is Your Friend Cheating?

Problem A friend of yours wants to play a simple coin-flipping game. If the coin comes up heads, you win; if it comes up tails, your friend wins. Suppose the outcome of five plays of the game is T, T, T, T, T. Is your friend cheating?

Approach To decide whether your friend is cheating, determine the likelihood of obtaining five tails in a row. Assume that the coin is fair so $P(\text{tail}) = P(\text{head}) = \frac{1}{2}$ and the flips of the coin are independent. Next ask, “Is it unusual to obtain five tails in a row with a fair coin?”

Solution We will determine the probability of getting five tails in a row, assuming that the coin is fair. The flips of the coin are independent, so

$$\begin{aligned} P(\text{five tails in a row}) &= P(\text{T and T and T and T and T}) \\ &= P(\text{T}) \cdot P(\text{T}) \cdot P(\text{T}) \cdot P(\text{T}) \cdot P(\text{T}) \\ &= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \\ &= \left(\frac{1}{2}\right)^5 \\ &= 0.03125 \end{aligned}$$

If we flipped a *fair* coin 5 times, 100 different times, we would expect about 3 of the 100 experiments to result in all tails. So what we observed is possible but not likely. You can make one of two conclusions:

1. Your friend is not cheating and happens to be lucky.
2. Your friend is not using a fair coin (that is, the probability of obtaining a tail on one flip is greater than $\frac{1}{2}$) and is cheating.

Is your friend cheating, or did you just happen to experience an unusual result from a fair coin?

This is at the heart of *hypothesis testing*. We make an assumption about reality (in this case, the probability of obtaining a tail is $\frac{1}{2}$). We then look at (or gather) sample evidence to determine whether it contradicts our assumption.

① Determine the Null and Alternative Hypotheses

According to dictionary.com, a **hypothesis** is a proposition assumed as a premise in an argument. The word *hypothesis* comes from the Greek word *hypothēnai*, which means “to suppose.” The definition of *hypothesis* in statistics is given next.

Definition

A hypothesis is a statement regarding a characteristic of one or more populations.

In this chapter, we look at hypotheses regarding a single population parameter. Consider the following:

(A) According to a Gallup poll conducted in 2016, 85% of Americans felt satisfied with the way things were going in their personal lives. A researcher wonders if the percentage of satisfied Americans is different today (a statement regarding a population proportion).

(B) The packaging on a light bulb states that the bulb will last 500 hours under normal use. A consumer advocate would like to know if the mean lifetime of a bulb is less than 500 hours (a statement regarding the population mean).

(C) The standard deviation of the rate of return for a certain class of mutual funds is 0.08 percent. A mutual fund manager believes the standard deviation of the rate of return for his fund is less than 0.08 percent (a statement regarding the population standard deviation).

CAUTION!

If population data are available, there is no need for inferential statistics.

We test these types of statements using sample data because it is usually impossible or impractical to gain access to the entire population. The procedure (or process) we use to test such statements is called *hypothesis testing*.

Definition

Hypothesis testing is a procedure, based on sample evidence and probability, used to test statements regarding a characteristic of one or more populations.

The basic steps in conducting a hypothesis test are these:

Steps in Hypothesis Testing

1. Make a statement regarding the nature of the population.
2. Collect evidence (sample data) to test the statement.
3. Analyze the data to assess the plausibility of the statement.

Because we use sample data to test hypotheses, we cannot state with 100% certainty that the statement is true. Instead, we can only determine whether the sample data support the statement, or not. Because the statement can be either true or false, hypothesis testing is based on two types of hypotheses. In this chapter, the hypotheses will be statements regarding the value of a population parameter.

Definitions

The **null hypothesis**, denoted H_0 (read “H-naught”), is a statement to be tested. The null hypothesis is a statement of no change, no effect, or no difference and is assumed true until evidence indicates otherwise.

The **alternative hypothesis**, denoted H_1 (read “H-one”), is a statement that we are trying to find evidence to support.

In this chapter, there are three ways to set up the null and alternative hypotheses.

1. Equal hypothesis versus not equal hypothesis (**two-tailed test**)

$$H_0: \text{parameter} = \text{some value}$$

$$H_1: \text{parameter} \neq \text{some value}$$

2. Equal versus less than (**left-tailed test**)

$$H_0: \text{parameter} = \text{some value}$$

$$H_1: \text{parameter} < \text{some value}$$

3. Equal versus greater than (**right-tailed test**)

$$H_0: \text{parameter} = \text{some value}$$

$$H_1: \text{parameter} > \text{some value}$$

IN OTHER WORDS

The null hypothesis is a statement of **status quo** or **no difference** and always contains a statement of equality. The null hypothesis is assumed to be true until we have evidence to the contrary. We seek evidence that supports the statement in the alternative hypothesis.

Left- and right-tailed tests are referred to as **one-tailed tests**. Notice that in the left-tailed test the direction of the inequality sign in the alternative hypothesis points to the left ($<$), while in the right-tailed test the direction of the inequality sign in the alternative hypothesis points to the right ($>$). In all three tests, the null hypothesis contains a statement of equality.

Refer to the three hypotheses made on the previous page. In Situation A, the null hypothesis is expressed using the notation $H_0: p = 0.85$. This is a statement of *status quo* or no difference. The Latin phrase *status quo* means “the existing state or condition.” So, the statement in the null hypothesis means that American opinions have not changed from 2016. We are trying to show that the proportion is different today, so the alternative hypothesis is $H_1: p \neq 0.85$. In Situation B, the null hypothesis is $H_0: \mu = 500$ hours. This is a statement of no difference between the population mean and the lifetime stated on the label. We are trying to show that the mean lifetime is less than 500 hours, so the alternative hypothesis is $H_1: \mu < 500$ hours. In Situation C, the null hypothesis is $H_0: \sigma = 0.08$ percent. This is a statement of no difference between the population standard deviation rate of return of the manager’s mutual fund and all mutual funds. The alternative hypothesis is $H_1: \sigma < 0.08$ percent. Do you see why?

The statement we are trying to gather evidence for, which is dictated by the researcher before any data are collected, determines the structure of the alternative hypothesis (two-tailed, left-tailed, or right-tailed). For example, the label on a can of soda states that the can contains 12 ounces of liquid. A consumer advocate would be concerned only if the mean contents are less than 12 ounces, so the alternative hypothesis is $H_1: \mu < 12$ ounces. However, a quality-control engineer for the soda manufacturer would be concerned if there is too little or too much soda in the can, so the alternative hypothesis would be $H_1: \mu \neq 12$ ounces. In both cases, however, the null hypothesis is a statement of no difference between the manufacturer’s assertion on the label and the actual mean contents of the can, so the null hypothesis is $H_0: \mu = 12$ ounces.

EXAMPLE 2 Forming Hypotheses

Problem Determine the null and alternative hypotheses. State whether the test is two-tailed, left-tailed, or right-tailed.

- The Medco pharmaceutical company has just developed a new antibiotic for children. Two percent of children taking competing antibiotics experience headaches as a side effect. A researcher for the Food and Drug Administration wishes to know if the percentage of children taking the new antibiotic who experience headaches as a side effect is more than 2%.
- The *Blue Book* price of a used three-year-old Chevy Corvette Z06 is \$59,083. Grant wonders if the mean price of a used three-year-old Chevy Corvette Z06 in the Miami metropolitan area is different from \$59,083.
- The standard deviation of the contents in a 64-ounce bottle of detergent using an old filling machine is 0.23 ounce. The manufacturer wants to know if a new filling machine has less variability.

Approach In each case, we must determine the parameter to be tested, the statement of no change or no difference (*status quo*), and the statement we are attempting to gather evidence for.

Solution

IN OTHER WORDS

Structuring the null and alternative hypotheses:

- Identify the parameter to be tested.
- Determine the status quo value of the parameter. This gives the null hypothesis.
- Determine the statement that reflects what we are trying to gather evidence for. This gives the alternative hypothesis.

- The hypothesis deals with a population proportion, p . If the new drug is no different from competing drugs, the proportion of individuals taking it who experience a headache will be 0.02 (because 2% = 0.02). This is the statement of no effect. We want to determine if the proportion of individuals who experience a headache is more than 0.02. This is the statement we are attempting to find evidence to support.

(continued)

	In Words	Symbolically
Null Hypothesis	H_0 : The proportion of individuals taking the new drugs who experience a headache as a side effect is 0.02	$H_0: p = 0.02$
Alternative Hypothesis	H_1 : The proportion of individuals taking the new drugs who experience a headache as a side effect is more than 0.02	$H_1: p > 0.02$

This is a right-tailed test because the alternative hypothesis contains a $>$ symbol.

- (b) The hypothesis deals with a population mean, μ . If the mean price of a three-year-old Corvette Z06 in Miami is no different from the *Blue Book* price, then the population mean in Miami will be \$59,083. This is the statement of no difference. Grant wants to know if the mean price is different from \$59,083. This is the statement we are attempting to find evidence to support.

	In Words	Symbolically
Null Hypothesis	H_0 : The mean price of a three-year-old Corvette Z06 in Miami is no different from the <i>Blue Book</i> price, \$59,083	$H_0: \mu = \$59,083$
Alternative Hypothesis	H_1 : The mean price of a three-year-old Corvette Z06 in Miami is different from the <i>Blue Book</i> price, \$59,083	$H_1: \mu \neq \$59,083$

This is a two-tailed test because the alternative hypothesis contains a \neq symbol.

- (c) The hypothesis deals with a population standard deviation, σ . If the new machine is no different from the old machine, the standard deviation of the amount in the bottles filled by the new machine will be 0.23 ounce. This is the statement of no change. The company wants to know if the new machine has *less* variability than the old machine. This is the statement we are attempting to find evidence to support.

	In Words	Symbolically
Null Hypothesis	H_0 : The standard deviation amount in the bottles is no different from the old machine, 0.23 ounce	$H_0: \sigma = 0.23 \text{ ounce}$
Alternative Hypothesis	H_1 : The standard deviation amount in the bottles is less than the old machine, 0.23 ounce	$H_1: \sigma < 0.23 \text{ ounce}$

This is a left-tailed test because the alternative hypothesis contains a $<$ symbol. 

② Explain Type I and Type II Errors

IN OTHER WORDS

When you are testing a hypothesis, there is always the possibility that your conclusion will be wrong. To make matters worse, you won't know whether you are wrong or not! Don't fret, however; we have tools to help manage these incorrect conclusions.

Four Outcomes from Hypothesis Testing

1. Reject the null hypothesis when the alternative hypothesis is true. This decision would be correct.
2. Do not reject the null hypothesis when the null hypothesis is true. This decision would be correct.
3. Reject the null hypothesis when the null hypothesis is true. This decision would be incorrect. This type of error is called a **Type I error**.
4. Do not reject the null hypothesis when the alternative hypothesis is true. This decision would be incorrect. This type of error is called a **Type II error**.

Figure 1 illustrates the two types of errors that can be made in hypothesis testing.

Figure 1

		Reality	
		H_0 Is True	H_1 Is True
Conclusion	Do Not Reject H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion



IN OTHER WORDS

A Type I error is like putting an innocent person in jail. A Type II error is like letting a guilty person go free.

We illustrate the idea of Type I and Type II errors by looking at hypothesis testing from the point of view of a criminal trial. In any trial, the defendant is assumed to be innocent. (We give the defendant the benefit of the doubt.) The district attorney must collect and present evidence proving that the defendant is guilty beyond all reasonable doubt.

Because we are seeking evidence for guilt, it becomes the alternative hypothesis. Innocence is assumed, so it is the null hypothesis.

$$\begin{aligned} H_0: & \text{the defendant is innocent} \\ H_1: & \text{the defendant is guilty} \end{aligned}$$

In a trial, the jury obtains information (sample data). It then deliberates about the evidence (the data analysis). Finally, it either convicts the defendant (rejects the null hypothesis) or declares the defendant not guilty (fails to reject the null hypothesis).

Note that the defendant is never declared innocent. That is, the null hypothesis is never declared true. The two correct decisions are to declare an innocent person not guilty or declare a guilty person to be guilty. The two incorrect decisions are to convict an innocent person (a Type I error) or to let a guilty person go free (a Type II error). It is helpful to think in this way when trying to remember the difference between a Type I and a Type II error.

EXAMPLE 3

Type I and Type II Errors

Problem The Medco pharmaceutical company has just developed a new antibiotic. Two percent of children taking competing antibiotics experience headaches as a side effect. A researcher for the Food and Drug Administration wishes to know if the percentage of children taking the new antibiotic who experience a headache as a side effect is more than 2%. The researcher conducts a hypothesis test with $H_0: p = 0.02$ and $H_1: p > 0.02$.

- (a) Explain what it would mean to make a Type I Error.
- (b) Explain what it would mean to make a Type II Error.

Approach A Type I error occurs if we reject the null hypothesis when the null hypothesis is true. A Type II error occurs if we do not reject the null hypothesis when the alternative hypothesis is true.

Solution

- (a) A Type I error is made if the sample evidence leads the researcher to believe that $p > 0.02$ (that is, we reject the null hypothesis) when, in fact, the proportion of children who experience a headache is not greater than 0.02.
- (b) A Type II error is made if the researcher does not reject the null hypothesis that the proportion of children experiencing a headache is equal to 0.02 when, in fact, the proportion of children who experience a headache is more than 0.02. In other words, the sample evidence led the researcher to believe $p = 0.02$ when, in fact, the true proportion is some value larger than 0.02.



The Probability of Making a Type I or Type II Error

When we studied confidence intervals, we learned that we never know whether a confidence interval contains the unknown parameter. We only know the likelihood that a confidence interval captures the parameter. Similarly, we never know whether the conclusion of a hypothesis test is correct. However, just as we place a level of confidence in the construction of a confidence interval, we can assign probabilities to making Type I or Type II errors when testing hypotheses. The following notation is commonplace:

$$\alpha = P(\text{Type I error}) = P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true})$$

$$\beta = P(\text{Type II error}) = P(\text{not rejecting } H_0 \text{ when } H_1 \text{ is true})$$

The symbol β is the Greek letter beta (pronounced “BAY tah”). The probability of making a Type I error, α , is chosen by the researcher *before* the sample data are collected. This probability is referred to as the *level of significance*.

Definition

The **level of significance**, α , is the probability of making a Type I error.

The choice of the level of significance depends on the consequences of making a Type I error. If the consequences are severe, the level of significance should be small (say, $\alpha = 0.01$). However, if the consequences are not severe, a higher level of significance can be chosen (say $\alpha = 0.05$ or $\alpha = 0.10$).

Why is the level of significance not always set at $\alpha = 0.01$? Reducing the probability of making a Type I error increases the probability of making a Type II error, β . Using our court analogy, a jury is instructed that the prosecution must provide proof of guilt “beyond all reasonable doubt.” This implies that we are choosing to make α small so that the probability of convicting an innocent person is very small. The consequence of the small α , however, is a large β , which means many guilty defendants will go free. For now, we are content to recognize the inverse relation between α and β (as one goes up the other goes down).

IN OTHER WORDS

As the probability of a Type I error increases, the probability of a Type II error decreases, and vice versa.

③ State Conclusions to Hypothesis Tests

CAUTION!

We never accept the null hypothesis, because, without having access to the entire population, we don't know the exact value of the parameter stated in the null hypothesis. Rather, we say that we do not reject the null hypothesis. This is just like the court system. We never declare a defendant innocent, but rather say the defendant is not guilty.

Once the decision whether or not to reject the null hypothesis is made, the researcher must state his or her conclusion. It is important to recognize that we never *accept* the null hypothesis. Again, the court system analogy helps to illustrate the idea. The null hypothesis is H_0 : innocent. When the evidence presented to the jury is not enough to convict beyond all reasonable doubt, the jury's verdict is “not guilty.”

Notice that the verdict does not state that the null hypothesis of innocence is true; it states that there is not enough evidence to conclude guilt. This is a huge difference. Being told that you are not guilty is very different from being told that you are innocent!

So, sample evidence can never prove the null hypothesis to be true. By not rejecting the null hypothesis, we are saying that the evidence indicates the null hypothesis *could* be true. That is, there is not enough evidence to reject our assumption that the null hypothesis is true.

EXAMPLE 4

Stating the Conclusion

Problem The Medco pharmaceutical company has just developed a new antibiotic. Two percent of children taking competing antibiotics experience a headache as a side effect. A researcher for the Food and Drug Administration believes that the proportion of children taking the new antibiotic who experience a headache as a side effect is more than 0.02. From Example 2(a), we know the null hypothesis is $H_0: p = 0.02$ and the alternative hypothesis is $H_1: p > 0.02$.

IN OTHER WORDS

The conclusion to a hypothesis test is always as follows: There (*is/is not*) sufficient evidence to conclude that ***insert statement in alternative hypothesis.***

NW Now Work Problem 23

Suppose that the sample evidence indicates that

- (a) the null hypothesis is rejected. State the conclusion.
- (b) the null hypothesis is not rejected. State the conclusion.

Approach When the null hypothesis is rejected, we say that there is sufficient evidence to support the statement in the alternative hypothesis. When the null hypothesis is not rejected, we say that there is not sufficient evidence to support the statement in the alternative hypothesis. We never say that the null hypothesis is true!

Solution

- (a) The statement in the alternative hypothesis is that the proportion of children taking the new antibiotic who experience a headache as a side effect is more than 0.02. Because the null hypothesis ($p = 0.02$) is rejected, there is sufficient evidence to conclude that the proportion of children who experience a headache as a side effect is more than 0.02.
- (b) Because the null hypothesis is not rejected, there is not sufficient evidence to say that the proportion of children who experience a headache as a side effect is more than 0.02.



10.1 Assess Your Understanding

Vocabulary and Skill Building

1. (a) A _____ is a statement regarding a characteristic of one or more populations.
- (b) The _____ is a statement of no change, no effect, or no difference.
- (c) The _____ is a statement we are trying to find evidence to support.
2. _____ is a procedure, based on sample evidence and probability, used to test statements regarding a characteristic of one or more populations.
3. If we reject the null hypothesis when the statement in the null hypothesis is true, we have made a Type _____ error.
4. If we do not reject the null hypothesis when the statement in the alternative hypothesis is true, we have made a Type _____ error.
5. The _____ is the probability of making a Type I error.
6. *True or False:* Sample evidence can prove a null hypothesis is true.

In Problems 7–12, the null and alternative hypotheses are given. Determine whether the hypothesis test is left-tailed, right-tailed, or two-tailed. What parameter is being tested?

- | | |
|------------------------|-------------------------|
| 7. $H_0: \mu = 5$ | 8. $H_0: p = 0.2$ |
| $H_1: \mu > 5$ | $H_1: p < 0.2$ |
| 9. $H_0: \sigma = 4.2$ | 10. $H_0: p = 0.76$ |
| $H_1: \sigma \neq 4.2$ | $H_1: p > 0.76$ |
| 11. $H_0: \mu = 120$ | 12. $H_0: \sigma = 7.8$ |
| $H_1: \mu < 120$ | $H_1: \sigma \neq 7.8$ |

In Problems 13–20, (a) state the null and alternative hypotheses in words, (b) state the null and alternative hypotheses symbolically,

- (c) explain what it would mean to make a Type I error, and
- (d) explain what it would mean to make a Type II error.

13. **Complete College** For students who first enrolled in two-year public institutions in fall 2013, the proportion who earned a bachelor's degree within six years was 0.236. The president of Joliet Junior College believes that the proportion of students who enroll in her institution have a higher completion rate.
14. **Pizza** Historically, the time to order and deliver a pizza at Jimbo's pizza was 48 minutes. Jim, the owner, implements a new system for ordering and delivering pizzas that he believes will reduce the time required to get a pizza to his customers.

- NW 15. Single-Family Home Price** According to the National Association of Home Builders, the mean price of an existing single-family home in 2018 was \$395,000. A real estate broker believes that existing home prices in her neighborhood are lower.

16. **Fair Packaging and Labeling** Federal law requires that a jar of peanut butter that is labeled as containing 32 ounces must contain at least 32 ounces. A consumer advocate feels that a certain peanut butter manufacturer is shorting customers by underfilling the jars.

17. **Valve Pressure** The standard deviation in the pressure required to open a certain valve is known to be $\sigma = 0.7$ psi. Due to changes in the manufacturing process, the quality-control manager feels that the pressure variability has been reduced.

18. **Overweight** According to the Centers for Disease Control and Prevention, 19.6% of children aged 6 to 11 years are overweight. A school nurse thinks that the percentage of 6- to 11-year-olds who are overweight is different in her school district.

19. **Cell Phone Service** According to the *CTIA-The Wireless Association*, the mean monthly revenue per cell phone was \$38.66 in 2017. A researcher suspects the mean monthly revenue per cell phone is different today.

20. SAT Reading Scores In 2017, the standard deviation of SAT score on the Evidence-based Reading and Writing Test for all students taking the exam was 100. A teacher believes that, due to changes to high school curricula, the standard deviation of SAT score has decreased.

In Problems 21–32, state the conclusion based on the results of the test.

- 21.** For the hypotheses in Problem 13, the null hypothesis is rejected.
22. For the hypotheses in Problem 14, the null hypothesis is not rejected.

NW 23. For the hypotheses in Problem 15, the null hypothesis is not rejected.

24. For the hypotheses in Problem 16, the null hypothesis is rejected.

25. For the hypotheses in Problem 17, the null hypothesis is not rejected.

26. For the hypotheses in Problem 18, the null hypothesis is not rejected.

27. For the hypotheses in Problem 19, the null hypothesis is rejected.

28. For the hypotheses in Problem 20, the null hypothesis is not rejected.

29. For the hypotheses in Problem 13, the null hypothesis is not rejected.

30. For the hypotheses in Problem 14, the null hypothesis is rejected.

31. For the hypotheses in Problem 15, the null hypothesis is rejected.

32. For the hypotheses in Problem 16, the null hypothesis is not rejected.

Applying the Concepts

33. Quality Control A can of soda is labeled as containing 12 fluid ounces. The quality control manager wants to verify that the filling machine is neither over-filling nor under-filling the cans.

- (a) Determine the null and alternative hypotheses that would be used to determine if the filling machine is calibrated correctly.
(b) The quality control manager obtains a sample of 75 cans and measures the contents. The sample evidence leads the manager to reject the null hypothesis. Write a conclusion for this hypothesis test.
(c) Suppose, in fact, the machine is not out of calibration. Has a Type I or Type II error been made?
(d) Management has informed the quality control department that it does not want to shut down the filling machine unless the evidence is overwhelming that the machine is out of calibration. What level of significance would you recommend the quality control manager use? Explain.

34. Popcorn Consumption According to popcorn.org, the mean consumption of popcorn annually by Americans is 54 quarts. The marketing division of popcorn.org unleashes an aggressive campaign designed to get Americans to consume even more popcorn.

- (a) Determine the null and alternative hypotheses that would be used to test the effectiveness of the marketing campaign.
(b) A sample of 800 Americans provides enough evidence to conclude that the marketing campaign was effective. Provide a statement that should be put out by the marketing department.

(c) Suppose, in fact, that the mean annual consumption of popcorn after the marketing campaign is 53.4 quarts. Has a Type I or Type II error been made by the marketing department? If they tested the hypothesis at the $\alpha = 0.05$ level of significance, what is the probability of making a Type I error?

35. E-Cigs According to the Centers for Disease Control and Prevention, 20.8% of high school students currently use electronic cigarettes. A high school counselor is concerned the use of e-cigs at her school is higher.

- (a) Determine the null and alternative hypotheses.
(b) If the sample data indicate that the null hypothesis should not be rejected, state the conclusion of the school counselor.
(c) Suppose, in fact, that the proportion of students at the counselor's high school who use electronic cigarettes is 0.217. Was a Type I or Type II error committed?

36. Migraines According to the Centers for Disease Control, 15.2% of American adults experience migraine headaches. Stress is a major contributor to the frequency and intensity of headaches. A massage therapist feels that she has a technique that can reduce the frequency and intensity of migraine headaches.

- (a) Determine the null and alternative hypotheses that would be used to test the effectiveness of the massage therapist's techniques.
(b) A sample of 500 American adults who participated in the massage therapist's program results in data that indicate that the null hypothesis should be rejected. Provide a statement that supports the massage therapist's program.
(c) Suppose, in fact, that the percentage of patients in the program who experience migraine headaches is 15.3%. Was a Type I or Type II error committed?

37. Prolong Engine Treatment The manufacturer of Prolong Engine Treatment claims that if you add one 12-ounce bottle of its \$20 product, your engine will be protected from excessive wear. An infomercial claims that a woman drove 4 hours without oil, thanks to Prolong. *Consumer Reports* magazine tested engines in which they added Prolong to the motor oil, ran the engines, drained the oil, and then determined the time until the engines seized.

- (a) Determine the null and alternative hypotheses that *Consumer Reports* will test.
(b) Both engines took exactly 13 minutes to seize. What conclusion might *Consumer Reports* draw based on this evidence?
38. Refer to Problem 16. Researchers must choose the level of significance based on the consequences of making a Type I error. In your opinion, is a Type I error or Type II error more serious? Why? On the basis of your answer, decide on a level of significance, α . Be sure to support your opinion.

Retain Your Knowledge

39. Retirement Savings Designed by Bill Bengen, the 4 percent rule says that a retiree may withdraw 4% of savings during the first year of retirement, and then each year after that withdraw the same amount plus an adjustment for inflation. Under this rule, your retirement savings should be expected to last 30 years, which is longer than most retirements.

- (a) If your retirement savings is \$750,000, how much may you withdraw in your first year of retirement if you want the retirement savings to last 30 years?

- (b) According to the American College of Financial Services, the proportion of people 60 to 75 years of age who believe it would be safe to withdraw 6 to 8 percent of their retirement savings annually is 0.16. Suppose you conduct a survey of twenty 60 to 75 year olds and ask them if it is safe to withdraw 6 to 8 percent of retirement savings annually if they wish their retirement savings to last 30 years. Explain why this is a binomial experiment. What are the values of n and p ?
- (c) In a random sample of twenty 60 to 75 year olds, what is the probability exactly 8 individuals will believe it is safe to withdraw 6 to 8 percent of retirement savings annually if they wish their retirement savings to last 30 years.
- (d) In a random sample of twenty 60 to 75 year olds, what is the probability fewer than 8 individuals will believe it is safe to withdraw 6 to 8 percent of retirement savings annually if they wish their retirement savings to last 30 years.
- (e) Suppose you obtain a random sample of five hundred 60 to 75 year olds. Explain why the normal model may be used to describe the sampling distribution of \hat{p} the sample proportion of 60 to 75 year olds who believe it is safe to withdraw 6 to 8 percent of their retirement savings annually. Describe this sampling distribution. That is, find the shape, center, and spread of the sampling distribution of the sample proportion.
- (f) Use the normal model from part (e) to approximate the probability of obtaining a random sample of at least one hundred (out of five hundred) 60 to 75 years olds who believe it would be safe to withdraw 6 to 8 percent of their retirement savings annually assuming the true proportion is 0.16. Is this result unusual? Explain.

Explaining the Concepts

40. If the consequences of making a Type I error are severe, would you choose the level of significance, α , to equal 0.01, 0.05, or 0.10? Why?

41. What happens to the probability of making a Type II error, β , as the level of significance, α , decreases? Why?

42. The following is a quotation from Sir Ronald A. Fisher, a famous statistician.

For the logical fallacy of believing that a hypothesis has been proved true, merely because it is not contradicted by the available facts, has no more right to insinuate itself in statistics than in other kinds of scientific reasoning. . . . It would, therefore, add greatly to the clarity with which the tests of significance are regarded if it were generally understood that tests of significance, when used accurately, are capable of rejecting or invalidating hypotheses, in so far as they are contradicted by the data; but that they are never capable of establishing them as certainly true. . . .

Source: Letter by Ronald A Fisher in Nature. Copyright © by Nature Publishing Group.

In your own words, explain what this quotation means.

43. In your own words, explain the difference between “beyond all reasonable doubt” and “beyond all doubt.” Use these phrases to explain why we never “accept” the statement in the null hypothesis.

10.2 Hypothesis Tests for a Population Proportion



Preparing for This Section Before getting started, review the following:

- Using probabilities to identify unusual events (Section 5.1, p. 230)
- z_α notation (Section 7.2, pp. 349–350)
- Sampling distribution of the sample proportion (Section 8.2, pp. 384–389)
- Computing normal probabilities (Section 7.2, pp. 343–347)
- Binomial probability distribution (Section 6.2, pp. 314–320)

Objectives

- ① Explain the logic of hypothesis testing
- ② Test hypotheses about a population proportion
- ③ Test hypotheses about a population proportion using the binomial probability distribution

1 Explain the Logic of Hypothesis Testing

Recall that the best point estimate of p , the proportion of the population with a certain characteristic, is given by

$$\hat{p} = \frac{x}{n}$$

where x is the number of individuals in the sample with the specified characteristic and n is the sample size. We learned in Section 8.2 that the sampling distribution of \hat{p} is

approximately normal, with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, provided that the following requirements are satisfied.

1. The sample is a simple random sample.
2. $np(1-p) \geq 10$.
3. The sampled values are independent of each other ($n \leq 0.05N$).

We will present three methods for testing hypotheses. The first method is called the classical (traditional) approach, the second method is the P -value approach, and the third method uses confidence intervals. Your instructor may choose to cover one, two, or all three approaches to hypothesis testing.

First, we lay out a scenario that will be used to understand both the classical and P -value approaches. Suppose a politician wants to know if a majority (more than 50%) of her constituents are in favor of a certain policy. We are therefore testing the following hypotheses:

$$H_0: p = 0.5 \text{ versus } H_1: p > 0.5$$

The politician hires a polling firm to obtain a random sample of 1000 registered voters in her district and finds that 534 are in favor of the policy, so $\hat{p} = \frac{534}{1000} = 0.534$. Do

these results suggest that among *all* registered voters more than 50% favor the policy? Or is it possible that the true proportion of registered voters who favor the policy is 0.5 and we just happened to survey a majority in favor of the policy? In other words, would it be unusual to obtain a sample proportion of 0.534 or higher from a population whose proportion is 0.5? What is convincing, or *statistically significant*, evidence?

Definition

When observed results are unlikely under the assumption that the null hypothesis is true, we say the result is **statistically significant** and we reject the null hypothesis.

To determine if a sample proportion of 0.534 is statistically significant, we build a probability model. After all, a second random sample of 1000 registered voters will likely result in a different sample proportion, and we want to describe this variability so we can determine if the results we obtained are unusual assuming the population proportion of voters in favor of the policy is $p = 0.5$. Since the sample data are from a simple random sample, $np(1-p) = 1000(0.5)(1-0.5) = 250 \geq 10$, and the sample size ($n = 1000$) is sufficiently smaller than the population size (provided there are at least $N = 20,000$ registered voters in the politician's district), we can use a normal model to describe the variability in \hat{p} . The mean of the distribution of \hat{p} is $\mu_{\hat{p}} = 0.5$ (since we assume the statement in the null hypothesis is true) and the standard deviation of the distribution of \hat{p} is $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5(1-0.5)}{1000}} \approx 0.016$. Figure 2 shows the sampling distribution of the sample proportion for the "politician" example.

Now that we have a model that describes the distribution of the sample proportion, we can use it to look at the logic of the classical and P -value approaches to test if a majority of the politician's constituents are in favor of the policy.

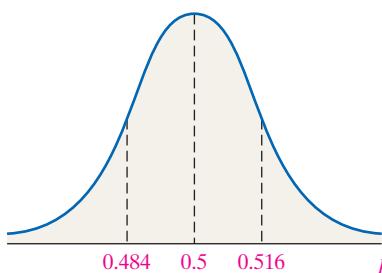
The Logic of the Classical Approach

We may consider the sample evidence to be statistically significant (or sufficient) if the sample proportion is too many standard deviations, say 2 standard deviations, above the assumed population proportion of 0.5.

Recall that $z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}}$ represents the number of standard deviations that \hat{p} is from the population proportion, p . Our simple random sample of 1000 registered voters results in a sample proportion of 0.534, so under the assumption that the null hypothesis is true we have

$$z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{0.534 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{1000}}} = 2.15$$

Figure 2



The sample proportion is 2.15 standard deviations above the hypothesized population proportion of 0.5, which is more than 2 standard deviations (that is, “too far”) above the hypothesized population proportion. So we will reject the null hypothesis. There is statistically significant (sufficient) evidence to conclude that a majority of registered voters are in favor of the policy.

Why does it make sense to reject the null hypothesis if the sample proportion is more than 2 standard deviations away from the hypothesized proportion? The area under the standard normal curve to the right of $z = 2$ is 0.0228, as shown in Figure 3.

Figure 3

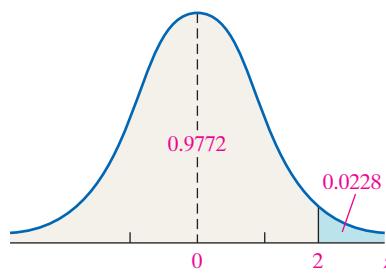


Figure 4

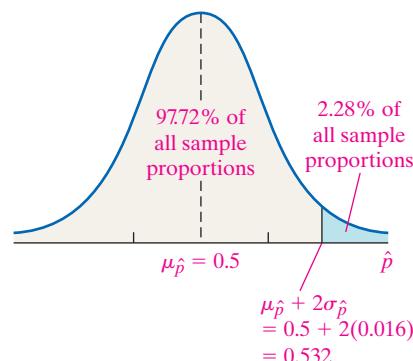


Figure 4 shows that if the null hypothesis is true (that is, if the population proportion is 0.5), then 97.72% of all sample proportions will be 0.532 or less, and only 2.28% of the sample proportions will be more than 0.532 (0.532 is 2 standard deviations above the hypothesized proportion of 0.5). If a sample proportion lies in the blue region, we are inclined to believe it came from a population whose proportion is greater than 0.5, rather than believe that the population proportion equals 0.5 and our sample just happened to result in a proportion of registered voters much higher than 0.5.

Notice that our criterion for rejecting the null hypothesis will lead to making a Type I error (rejecting a true null hypothesis) 2.28% of the time. This is because 2.28% of all sample proportions are more than 0.532, even though the population proportion is 0.5.

This discussion illustrates the following point.

Hypothesis Testing Using the Classical Approach

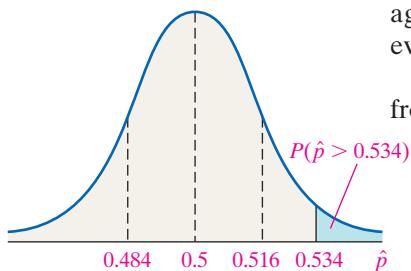
If the sample statistic is too many standard deviations from the population parameter stated in the null hypothesis, we reject the statement in the null hypothesis.

The Logic of the P-Value Approach

A second criterion we may use for testing hypotheses is to determine how likely it is to obtain a sample proportion of 0.534 or higher from a population whose proportion is 0.5. If a sample proportion of 0.534 or higher is unlikely (or unusual), we have evidence against the statement in the null hypothesis. Otherwise, we do not have sufficient evidence against the statement in the null hypothesis.

We can compute the probability of obtaining a sample proportion of 0.534 or higher from a population whose proportion is 0.5 using the normal model. See Figure 5.

Figure 5



$$P(\hat{p} > 0.534) = P\left(z > \frac{0.534 - 0.5}{\sqrt{\frac{0.5(1 - 0.5)}{100}}}\right) = P(z > 2.15) = 0.0158$$

NOTE

$$0.0158 \approx 0.02 = \frac{2}{100}$$

The value 0.0158 is called the *P-value*, which means about 2 samples in 100 will give a sample proportion as high or higher than the one we obtained if the population proportion really is 0.5. Because the observed results are unusual, we take this as evidence against the statement in the null hypothesis.

Definition

A **P-value** is the probability of observing a sample statistic as extreme or more extreme than one observed under the assumption that the statement in the null hypothesis is true. Put another way, the *P*-value is the likelihood or probability that a sample will result in a statistic such as the one obtained if the statement in the null hypothesis is true.

This discussion illustrates the idea behind hypothesis testing using the *P*-value approach.

Hypothesis Testing Using the P-Value Approach

If the probability of getting a sample statistic as extreme or more extreme than the one obtained is small under the assumption the statement in the null hypothesis is true, reject the null hypothesis.

How small should the *P*-value be to reject the statement in the null hypothesis? It depends on the situation and the consequences of making a Type I error. For example, in an educational study to determine if the proportion of students passing a class using a new teaching method increased, we might be willing to reject the null hypothesis that the method is the same as old teaching methods for any *P*-value less than 0.05. However, suppose a certain drug resulted in a high proportion of individuals contracting cancer, so the company reformulates the drug. A very small *P*-value showing a significantly lower proportion of individuals contracting cancer would likely be required to convince regulatory agencies the new formulation is safe. After all, making a Type I error would mean rejecting the null that the proportion of individuals contracting cancer is the same as it was before reformulation, when, in fact, the proportion of individuals contracting cancer *is* the same as it was before reformulation. A very dangerous error.

The bottom line is this—**the lower the *P*-value, the stronger the evidence against the statement in the null hypothesis.** The following table provides a methodology for describing the strength of evidence against the statement in the null hypothesis using the *P*-value.

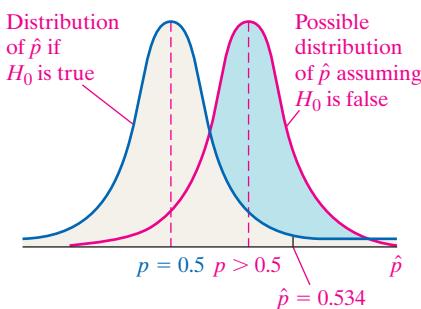
P-Value	Conclusion
$P\text{-value} \geq 0.10$	Do not reject the statement in the null hypothesis. The sample evidence is consistent with the statement in the null hypothesis.
$0.05 \leq P\text{-value} < 0.10$	There is some evidence against the statement in the null hypothesis.
$0.01 \leq P\text{-value} < 0.05$	There is moderate evidence against the statement in the null hypothesis.
$0.001 \leq P\text{-value} < 0.01$	There is strong evidence against the statement in the null hypothesis.
$P\text{-value} < 0.001$	There is very strong evidence against the statement in the null hypothesis.

Most professional journals use a method similar to the one above when reporting *P*-values. It is common for journals to use * to represent statistical significance at the 0.05 level; ** to represent statistical significance at the 0.01 level; and *** to represent statistical significance at the 0.001 level. For example, a study in which the *P*-value was 0.032 would be reported 0.032* to indicate significance at the 0.05 level (suggesting moderate evidence against the statement in the null hypothesis).

For the remainder of the text, however, we will always provide a level of significance, α , for all hypothesis tests. The decision rule will be to reject the statement in the null hypothesis if the *P*-value is less than the level of significance.

Figure 6 illustrates the “politician” situation for both the classical and *P*-value approaches to hypothesis testing. The distribution in blue shows the distribution of the sample proportion assuming the statement in the null hypothesis is true. The sample proportion of 0.534 is “too far” from the assumed population proportion of 0.5. Therefore, we reject the null hypothesis that $p = 0.5$ and conclude that $p > 0.5$, as indicated by the distribution in red. We do not know what the population proportion of registered voters who are in favor of the policy is, but we have evidence to say it is greater than 0.5 (a majority).

Figure 6



2 Test Hypotheses about a Population Proportion

We now formalize the procedure for testing hypotheses regarding a population proportion.

Testing Hypotheses Regarding a Population Proportion, p

Step 1 Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways:

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: p = p_0$	$H_0: p = p_0$	$H_0: p = p_0$
$H_1: p \neq p_0$	$H_1: p < p_0$	$H_1: p > p_0$

Note: p_0 is the assumed value of the population proportion.

Step 2 Select a level of significance α , depending on the seriousness of making a Type I error.

Continue with Steps 3 through 5, provided that

- the sample is obtained by simple random sampling or the data result from a randomized experiment.
- $np_0(1 - p_0) \geq 10$.
- the sampled values are independent of each other. That is, the sample size is less than 5% of the population size.

Classical Approach

Step 3 Compute the **test statistic**

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

Use Table V to determine the critical value.

	Two-Tailed	Left-Tailed	Right-Tailed
Critical value	$-z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$	$-z_\alpha$	z_α

Step 4 Compare the critical value with the test statistic.

Two-Tailed	Left-Tailed	Right-Tailed
If $z_0 < -z_{\frac{\alpha}{2}}$ or $z_0 > z_{\frac{\alpha}{2}}$, reject the null hypothesis.	If $z_0 < -z_\alpha$, reject the null hypothesis.	If $z_0 > z_\alpha$, reject the null hypothesis.

P-Value Approach

By Hand Step 3 Compute the **test statistic**

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

Use Table V to determine the *P*-value.

Two-Tailed	Left-Tailed	Right-Tailed
The sum of the area in the tails is the <i>P</i> -value	The area left of z_0 is the <i>P</i> -value	The area right of z_0 is the <i>P</i> -value

Technology Step 3 Use a statistical spreadsheet or calculator with statistical capabilities to obtain the *P*-value. The directions for obtaining the *P*-value using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch are in the Technology Step-by-Step on page 453.

Step 4 If P -value $< \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

Notice in Step 3 that we are using p_0 (the proportion stated in the null hypothesis) in computing the standard error rather than \hat{p} , as we did in constructing confidence intervals about p . This is because H_0 is assumed to be true when performing a hypothesis test, so the assumed mean of the distribution of \hat{p} is $\mu_{\hat{p}} = p_0$ and the assumed standard

$$\text{error is } \sigma_{\hat{p}} = \sqrt{\frac{p_0(1 - p_0)}{n}}.$$

EXAMPLE 1 Testing a Hypothesis about a Population Proportion: Right-Tailed Test

Problem Humira is a medication used to treat rheumatoid arthritis (RA). In clinical trials of Humira, 705 subjects diagnosed with RA were administered 40 mg of Humira every other week. Of the 705 subjects, 66 reported nausea as a side effect. It is known that the proportion of RA subjects in similar studies receiving a placebo who report nausea as a side effect is 0.08. Does the sample evidence represent significant evidence that a higher proportion of subjects receiving Humira experience nausea as a side effect than those taking a placebo? Use the $\alpha = 0.05$ level of significance. *Source: rxabbvie.com*

Approach The variable of interest is whether the subject taking Humira experiences nausea as a side effect, or not. Therefore, use a hypothesis test of a proportion to analyze the problem. Follow Steps 1 through 5 provided the model conditions are satisfied.

Solution

Step 1 The null hypothesis would be that the proportion of subjects taking Humira who experience nausea is 0.08 because this is the proportion of subjects taking a placebo who experience nausea. Remember, the null hypothesis is always a statement of “no difference.” We want to determine if the sample evidence suggests that a higher proportion of subjects taking Humira experience nausea as a side effect. Symbolically, the null and alternative hypotheses are

$$H_0: p = 0.08 \quad \text{versus} \quad H_1: p > 0.08$$

Step 2 The level of significance is $\alpha = 0.05$.

- The data result from a randomized experiment.
- $np_0(1 - p_0) = 705(0.08)(1 - 0.08) = 51.888 \geq 10$
- About 1,500,000 people suffer from RA in the United States alone, so the sample size is less than 5% of the population size.

Classical Approach

Step 3 Assume the statement in the null hypothesis is true so that $p_0 = 0.08$. The sample proportion experiencing nausea is $\hat{p} = \frac{x}{n} = \frac{66}{705} = 0.094$. We want to know if it is unusual to obtain a sample proportion of 0.094 or higher from a population whose proportion is 0.08.

The test statistic is

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.094 - 0.08}{\sqrt{\frac{0.08(1 - 0.08)}{705}}} = 1.37$$

P-Value Approach

By Hand **Step 3** Assume the statement in the null hypothesis is true so that $p_0 = 0.08$. The sample proportion experiencing nausea is $\hat{p} = \frac{x}{n} = \frac{66}{705} = 0.094$. We want to know how likely it is to obtain a sample proportion of 0.094 or higher from a population whose proportion is 0.08.

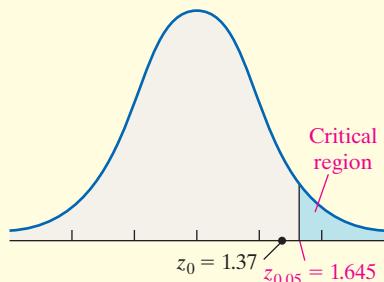
The test statistic is

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.094 - 0.08}{\sqrt{\frac{0.08(1 - 0.08)}{705}}} = 1.37$$

Because this is a right-tailed test, the *P*-value is the area under the standard normal distribution to the right of the test statistic, $z_0 = 1.37$, as shown in Figure 8.

Because this is a right-tailed test, determine the critical value at the $\alpha = 0.05$ level of significance to be $z_{0.05} = 1.645$. The critical region is shown in Figure 7.

Figure 7



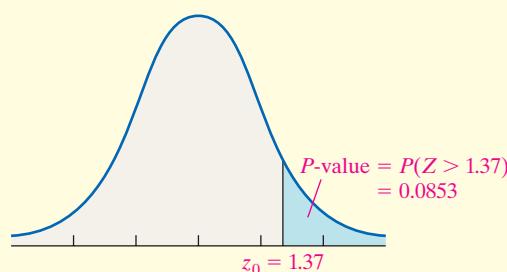
Step 4 The test statistic, $z_0 = 1.37$, is labeled in Figure 7. Because the test statistic is less than the critical value ($1.37 < 1.645$), do not reject the statement in the null hypothesis. Notice the test statistic does not fall into the critical region.

NOTE

While we did not reject the statement in the null hypothesis that the proportion of Humira subjects experiencing nausea is 0.08, the P -value suggests that there is moderate evidence to suggest that Humira users are experiencing an adverse reaction.

So, $P\text{-value} = P(Z > z_0) = P(Z > 1.37) = 0.0853$.

Figure 8



Technology Step 3 Using Minitab, we find the P -value is 0.091. See Figure 9.

Figure 9

Test

Null hypothesis	$H_0: p = 0.08$
Alternative hypothesis	$H_1: p > 0.08$
Z-Value	1.33
P-Value	0.091

Step 4 The P -value of 0.085 (by hand) means that if the statement in the null hypothesis that $p = 0.08$ is true, we would expect to observe 66 or more subjects who experience nausea in about 85 out of 1000 repetitions of this study. The observed results are moderately unusual. However, because the P -value is greater than the level of significance $\alpha = 0.05$ ($0.085 > 0.05$), do not reject the statement in the null hypothesis.

Step 5 There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude the proportion of subjects taking Humira who experience nausea is greater than 0.08. ●

NW Now Work Problem 17**CAUTION!**

In Example 1, we do not have enough evidence to reject the statement in the null hypothesis. In other words, it is not unusual to obtain a sample proportion of 0.094 from a population whose proportion is 0.08. However, this does not imply that we are accepting the statement in the null hypothesis (that is, we are not saying that the proportion equals 0.08). We are only saying we do not have enough evidence to conclude that the proportion is greater than 0.08. Be sure that you understand the difference between “accepting” and “not rejecting.” It is similar to the difference between being declared “innocent” versus “not guilty.”

Also, be sure you understand that the P -value is the probability of obtaining a sample statistic as extreme or more extreme than the one observed if the statement in the null hypothesis is true. The P -value does not represent the probability that the null hypothesis is true. The statement in the null hypothesis is either true or false, we just don’t know which.

EXAMPLE 2 Testing Hypotheses about a Population Proportion: Two-Tailed Test

Problem Which do you think is easier to raise—a boy or a girl? When asked this question in 1947, 24% of all Americans said raising a girl was easier. In June 2018, the Gallup Organization surveyed 1500 adult Americans, of which 408 felt it was easier to raise a girl. Does this result suggest the proportion of adult Americans who believe it is easier to raise a girl has changed since 1947?

Approach The variable of interest is whether an adult American thinks it is easier to raise a girl, or not. Therefore, use a hypothesis test of a proportion to analyze the problem. Follow Steps 1 through 5 provided the model conditions are satisfied.

(continued)

Solution

Step 1 The null hypothesis would be that the proportion of adult Americans who believe it is easier to raise a girl is the same as it was in 1947. Remember, the null hypothesis is always a statement of “no difference.” We want to determine if the sample evidence suggests that a *different* proportion of adult Americans today believe it is easier to raise a girl. This is a two-tailed hypothesis test. Symbolically, the null and alternative hypotheses are

$$H_0: p = 0.24 \text{ versus } H_1: p \neq 0.24$$

Step 2 The level of significance is $\alpha = 0.05$.

- The data result from a simple random sample.
- $np_0(1 - p_0) = 1500(0.24)(1 - 0.24) = 273.6 \geq 10$
- There are over 240,000,000 adult Americans, so the sample size is less than 5% of the population size.

Classical Approach

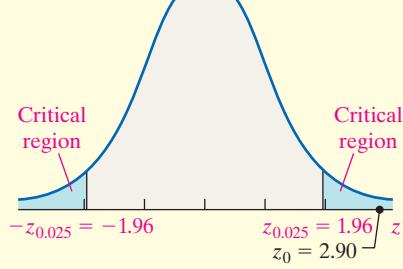
Step 3 Assume the statement in the null hypothesis is true so that $p_0 = 0.24$. The sample proportion of adult Americans who believe it is easier to raise a girl is $\hat{p} = \frac{x}{n} = \frac{408}{1500} = 0.272$. We want to know if it is unusual to obtain a sample proportion of 0.272 from a population whose proportion is 0.24.

The test statistic is

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.272 - 0.24}{\sqrt{\frac{0.24(1 - 0.24)}{1500}}} = 2.90$$

Because this is a two-tailed test, the critical values at the $\alpha = 0.05$ level of significance are $-z_{0.05/2} = -z_{0.025} = -1.96$ and $z_{0.05/2} = z_{0.025} = 1.96$. The critical regions are shown in Figure 10.

Figure 10



Step 4 The test statistic, $z_0 = 2.90$, is labeled in Figure 10. Because the test statistic lies in the critical region ($2.90 > 1.96$), reject the statement in the null hypothesis.

P-Value Approach

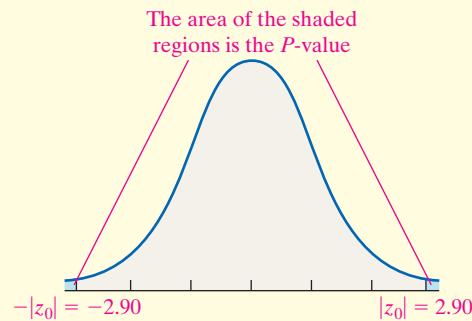
By Hand Step 3 Assume the statement in the null hypothesis is true so that $p_0 = 0.24$. The sample proportion of adult Americans who believe it is easier to raise a girl is $\hat{p} = \frac{x}{n} = \frac{408}{1500} = 0.272$. We want to know if it is unusual to obtain a sample proportion of 0.272 from a population whose proportion is 0.24.

The test statistic is

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} = \frac{0.272 - 0.24}{\sqrt{\frac{0.24(1 - 0.24)}{1500}}} = 2.90$$

Because this is a two-tailed test, the *P*-value is the area under the standard normal distribution to the left of $-|z_0| = -2.90$ and to the right of $|z_0| = 2.90$, as shown in Figure 11.

Figure 11



$$\begin{aligned} \text{P-value} &= P(Z < -|z_0|) + P(Z > |z_0|) \\ &= 2P(Z > 2.90) \quad \text{Use symmetry} \\ &= 2(0.0019) \\ &= 0.0038 \end{aligned}$$

Technology Step 3 Using StatCrunch, we find the *P*-value is 0.0037. See Figure 12.

Figure 12

One sample proportion summary hypothesis test:

p : Proportion of successes
 $H_0 : p = 0.24$
 $H_A : p \neq 0.24$

Hypothesis test results:

Proportion	Count	Total	Sample Prop.	Std. Err.	Z-Stat	P-value
p	408	1500	0.272	0.011027239	2.901905	0.0037

Step 4 The P -value of 0.0038 [Tech: 0.0037] means that if the statement in the null hypothesis that $p = 0.24$ is true, we would expect the type of results observed (or more extreme results) in about 4 out of every 1000 repetitions of this study. The observed results are unusual. Because the P -value is less than the level of significance $\alpha = 0.05$ ($0.0037 < 0.05$), reject the statement in the null hypothesis.

Step 5 There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude the proportion of adult Americans who believe it is easier to raise a girl has changed since 1947. In fact, we are 95% confident the proportion of adult Americans who believe it is easier to raise a girl is between 0.249 and 0.295. 

CAUTION!

Two words of caution. First, the P -values in Examples 1 and 2 are not “the P -values for the hypothesis test”. Instead, they are “the P -values for the hypothesis test based on the sample data obtained.” A different experiment in Example 1 or a different sample in Example 2 will likely lead to different P -values (because we have different individuals in the study).

Second, the P -values obtained in Examples 1 and 2 are found using the normal model. If it is not appropriate to use the normal model to estimate P -values, then the P -value will be inaccurate (and possibly lead to an incorrect conclusion). It is very important that model requirements be verified (and satisfied) before using the normal model to estimate the P -value.

Because P -values may vary from sample to sample, and because P -values are found using a model, it is recommended that we don’t believe with 100% certainty that an association or effect exists just because the evidence is found to be statistically significant (we may have committed a Type I Error). It is also important that we don’t believe an association or effect does not exist just because the evidence is not found to be statistically significant (we may have committed a Type II Error). The original intent of hypothesis tests using P -values was that small P -values warrant further investigation to validate the results of the initial study.

NW Now Work Problem 21**Test a Hypothesis Using a Confidence Interval**

Recall, the level of confidence, $(1 - \alpha) \cdot 100\%$, in a confidence interval represents the percentage of intervals that will contain the unknown parameter if repeated samples are obtained.

Two-Tailed Hypothesis Testing Using Confidence Intervals

When testing $H_0: p = p_0$ versus $H_1: p \neq p_0$, if a $(1 - \alpha) \cdot 100\%$ confidence interval contains p_0 , do not reject the null hypothesis. However, if the confidence interval does not contain p_0 , conclude that $p \neq p_0$ at the level of significance, α .

EXAMPLE 3**Testing a Hypothesis Using a Confidence Interval**

Problem A 2009 study by Princeton Survey Research Associates International found that 34% of teenagers text while driving. Does a recent survey conducted by *Consumer Reports*, which found that 353 of 1200 randomly selected teens had texted while driving, suggest that the proportion of teens who text while driving has changed since 2009? Use a 95% confidence interval to answer the question.

Approach Construct a 95% confidence interval about the proportion of teens who text while driving based on the *Consumer Reports* survey. If the interval does not include 0.34, reject the null hypothesis $H_0: p = 0.34$ in favor of $H_1: p \neq 0.34$. (continued)

Solution The 95% confidence interval for p based on the *Consumer Reports* survey has a lower bound of 0.268 and an upper bound of 0.320. Because 0.34 is not within the bounds of the confidence interval, there is sufficient evidence to conclude that the proportion of teens who text while driving has changed since 2009.

NW Now Work Problem 23

Note: The same *Consumer Reports* article cited in Example 3 states that 75% of teens have friends who text while driving. What does this say about the difficulty in finding truthful responses to questions while conducting a survey?

3 Test Hypotheses about a Population Proportion Using the Binomial Probability Distribution

For the sampling distribution of \hat{p} to be approximately normal, we require that $np(1 - p)$ be at least 10. What if this requirement is not satisfied? In Section 6.2, we used the binomial probability formula to identify unusual events. We stated that an event was unusual if the probability of observing the event was less than 0.05. This criterion is based on the P -value approach to testing hypotheses; the probability that we computed was the P -value. We use this same approach to test hypotheses regarding a population proportion for small samples.

EXAMPLE 4 Hypothesis Test for a Population Proportion: Small Sample Size

Problem According to the U.S. Department of Agriculture (USDA), 48.9% of males aged 20 to 39 years consume the recommended daily requirement of calcium. After an aggressive “Got Milk” advertising campaign, the USDA conducts a survey of 35 randomly selected males aged 20 to 39 and finds that 21 of them consume the recommended daily allowance (RDA) of calcium. At the $\alpha = 0.10$ level of significance, is there evidence to conclude that the percentage of males aged 20 to 39 who consume the RDA of calcium has increased?

Approach We use the following steps:

Step 1 Determine the null and alternative hypotheses.

Step 2 Check whether $np_0(1 - p_0)$ is greater than or equal to 10, where p_0 is the proportion stated in the null hypothesis. If it is, then the sampling distribution of \hat{p} is approximately normal and we can use the steps on page 447. Otherwise, we use Steps 3 and 4, presented next.

Step 3 Compute the P -value. For right-tailed tests, the P -value is the probability of obtaining x or more successes. For left-tailed tests, the P -value is the probability of obtaining x or fewer successes.* The P -value is always computed with the proportion given in the null hypothesis. Remember, assume that the statement in the null is true until we have evidence to the contrary.

Step 4 If the P -value is less than the level of significance, α , reject the null hypothesis.

Solution

Step 1 The status quo or no change proportion of 20- to 39-year-old males who consume the recommended daily requirement of calcium is 0.489. We wish to know whether the advertising campaign increased this proportion. Therefore,

$$H_0: p = 0.489 \quad \text{and} \quad H_1: p > 0.489$$

*We will not address P -values for two-tailed hypothesis tests. For those who are interested, the P -value is two times the probability of obtaining x or more successes if $\hat{p} > p$ and two times the probability of obtaining x or fewer successes if $\hat{p} < p$.

Step 2 From the null hypothesis, we have $p_0 = 0.489$. There were $n = 35$ individuals surveyed, so $np_0(1 - p_0) = 35(0.489)(1 - 0.489) = 8.75$. Because $np_0(1 - p_0) < 10$, the sampling distribution of \hat{p} is not approximately normal.

Step 3 Let the random variable X represent the number of individuals who consume the daily requirement of calcium. We have $x = 21$ successes in $n = 35$ trials, so $\hat{p} = \frac{21}{35} = 0.6$. We want to judge whether the larger proportion is due to an increase in the population proportion or to sampling error. We obtained $x = 21$ successes in the survey and this is a right-tailed test, so the P -value is $P(X \geq 21)$.

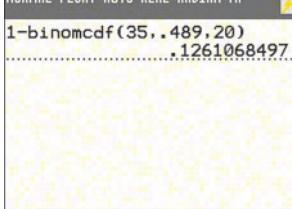
$$P\text{-value} = P(X \geq 21) = 1 - P(X < 21) = 1 - P(X \leq 20)$$

We will compute this P -value using a TI-84 Plus CE graphing calculator, with $n = 35$ and $p = 0.489$. Figure 13 shows the results.

The P -value is 0.1261. Minitab, StatCrunch, and Excel will compute exact P -values using this approach as well.

Step 4 The P -value is greater than the level of significance ($0.1261 > 0.10$), so we do not reject H_0 . There is not sufficient evidence (at the $\alpha = 0.1$ level of significance) to conclude that the proportion of 20- to 39-year-old males who consume the recommended daily allowance of calcium has increased.

Figure 13



NW Now Work Problem 29

CAUTION!

The results of Example 4 raise an interesting point. When hypothesis tests are conducted with small samples, the evidence against the statement in the null hypothesis must be overwhelming. After all, the difference between the sample proportion and proportion in the null hypothesis is $0.6 - 0.489 = 0.111$, which is a large “effect size”. Therefore, watch out for studies where the null hypothesis is not rejected when the sample size is small.

Technology Step-by-Step

Hypothesis Tests Regarding a Population Proportion

TI-83/84 Plus

1. Press STAT, highlight TESTS, and select 5 : 1-PropZTest.
2. For the value of p_0 , enter the value of the population proportion stated in the null hypothesis.
3. Enter the number of successes, x , and the sample size, n .
4. Select the direction of the alternative hypothesis.
5. Highlight Calculate or Draw, and press ENTER.

Minitab

1. If you have raw data, enter them in C1, using 0 for failure and 1 for success.
2. Select the Stat menu, highlight Basic Statistics, then highlight 1-Proportion.
3. If you have raw data, select “One or more samples, each in a column” from the drop-down menu. Place the cursor in the box, highlight the column containing the raw data, and click “Select”. If you have summarized data, select “Summarized data” from the drop-down menu. Enter the number of successes in the “Number of events” box and enter the number of trials. Check the “Perform hypothesis test” box and enter the value of the population proportion stated in the null hypothesis.
4. Click Options. Enter the direction of the alternative hypothesis. Assuming $np_0(1 - p_0) \geq 10$, select “Normal approximation” from the drop-down menu. Click OK twice.

Excel

1. Load the XLSTAT Add-In.
2. Select the XLSTAT menu and select Parametric Tests. From the drop-down menu, select Tests for one proportion.
3. In the cell marked Frequency, enter the number of successes. In the cell marked Sample size, enter the number of trials. In the cell marked Test proportion, enter the proportion stated in the null hypothesis. Check the Frequency radio button. Click Options. Choose the appropriate direction for the alternative hypothesis. Be sure Hypothesize difference (D): is set to zero. Enter the level of significance. Click OK.

StatCrunch

1. If you have raw data, enter them into the spreadsheet. Name the column variable.
2. Select Stat, highlight Proportion Stats, select One Sample, and then choose either With Data or With Summary.
3. If you chose With Data, select the column that has the observations, choose which outcome represents a success. If you chose With Summary, enter the number of successes and the number of trials in the observations box. Choose the hypothesis test radio button. Enter the value of the proportion stated in the null hypothesis and choose the direction of the alternative hypothesis from the pull-down menu. Click Compute!.



10.2 Assess Your Understanding

Vocabulary and Skill Building

1. When observed results are unlikely under the assumption that the null hypothesis is true, we say the result is _____ and we reject the null hypothesis.
2. **True or False:** When testing a hypothesis using the Classical Approach, if the sample proportion is too many standard deviations from the proportion stated in the null hypothesis, we reject the null hypothesis.
3. Put the following P -values in order from weakest to strongest in terms of evidence against the statement in the null hypothesis.
 (a) 0.139 (b) 0.083 (c) 0.091 (d) 0.005 (e) 0.019
4. Determine the critical value for a right-tailed test regarding a population proportion at the $\alpha = 0.01$ level of significance.
5. Determine the critical value for a left-tailed test regarding a population proportion at the $\alpha = 0.1$ level of significance.
6. Determine the critical value for a two-tailed test regarding a population proportion at the $\alpha = 0.05$ level of significance.

In Problems 7–12, test the hypothesis using (a) the classical approach and (b) the P -value approach. Be sure to verify the requirements of the test.

7. $H_0: p = 0.3$ versus $H_1: p > 0.3$

$n = 200; x = 75; \alpha = 0.05$

8. $H_0: p = 0.6$ versus $H_1: p < 0.6$

$n = 250; x = 124; \alpha = 0.01$

9. $H_0: p = 0.55$ versus $H_1: p < 0.55$

$n = 150; x = 78; \alpha = 0.1$

10. $H_0: p = 0.25$ versus $H_1: p < 0.25$

$n = 400; x = 96; \alpha = 0.1$

11. $H_0: p = 0.9$ versus $H_1: p \neq 0.9$

$n = 500; x = 440; \alpha = 0.05$

12. $H_0: p = 0.4$ versus $H_1: p \neq 0.4$

$n = 1000; x = 420; \alpha = 0.01$

13. You Explain It! Stock Analyst Throwing darts at the stock pages to decide which companies to invest in could be a successful stock-picking strategy. Suppose a researcher decides to test this theory and randomly chooses 100 companies to invest in. After 1 year, 53 of the companies were considered winners; that is, they outperformed other companies in the same investment class. To assess whether the dart-picking strategy resulted in a majority of winners, the researcher tested $H_0: p = 0.5$ versus $H_1: p > 0.5$ and obtained a P -value of 0.2743. Explain what this P -value means and write a conclusion for the researcher.

14. You Explain It! ESP Suppose an acquaintance claims to have the ability to determine the birth month of randomly selected individuals. To test such a claim, you randomly select 80 individuals and ask the acquaintance to state the birth month of the individual. If the individual has the ability to determine birth month, then the proportion of correct birth months should exceed $\frac{1}{12} \approx 0.083$, the rate one would expect from simply guessing.

- State the null and alternative hypotheses for this experiment.
- Suppose the individual was able to guess nine correct birth months. The P -value for such results is 0.1726. Explain what this P -value means and write a conclusion for the test.

Applying the Concepts

- 15. Cramer Correct Less Than Half the Time?** The website pundittracker.com keeps track of predictions made by individuals in finance, politics, sports, and entertainment. Jim Cramer is a famous TV financial personality and author. Pundittracker monitored 678 of his stock predictions (such as a recommendation to buy the stock) and found that 320 were correct predictions. Treat these 678 predictions as a random sample of all of Cramer's predictions.
 - Determine the sample proportion of predictions Cramer got correct.
 - Suppose that we want to know whether the evidence suggests Cramer is correct less than half the time. State the null and alternative hypotheses.
 - Verify the normal model may be used to determine the P -value for this hypothesis test.
 - Draw a normal model with area representing the P -value shaded for this hypothesis test.
 - Determine the P -value based on the model from part (d).
 - Interpret the P -value.
 - Based on the P -value, what does the sample evidence suggest? That is, what is the conclusion of the hypothesis test? Assume an $\alpha = 0.05$ level of significance.
- 16. Political Pundits** In his book, *The Signal and the Noise*, Nate Silver analyzed 733 predictions made by experts regarding political events. Of the 733 predictions, 338 were mostly true.
 - Determine the sample proportion of political predictions that were mostly true.
 - Suppose that we want to know whether the evidence suggests the political predictions were mostly true less than half the time. State the null and alternative hypotheses.
 - Verify the normal model may be used to determine the P -value for this hypothesis test.
 - Draw a normal model with the area representing the P -value shaded for this hypothesis test.
 - Determine the P -value based on the model from part (d).
 - Interpret the P -value.
 - Based on the P -value, what does the sample evidence suggest? That is, what is the conclusion of the hypothesis test? Assume an $\alpha = 0.1$ level of significance.
- NW 17. Lipitor** The drug Lipitor is meant to reduce cholesterol and LDL cholesterol. In clinical trials, 19 out of 863 patients taking 10 mg of Lipitor daily complained of flulike symptoms. Suppose that it is known that 1.9% of patients taking competing drugs complain of flulike symptoms. Is there evidence to conclude that more than 1.9% of Lipitor users experience flulike symptoms as a side effect at the $\alpha = 0.01$ level of significance?
- 18. Nexium** Nexium is a drug that can be used to reduce the acid produced by the body and heal damage to the esophagus due to acid reflux. The manufacturer of Nexium claims that more than 94% of patients taking Nexium are healed within 8 weeks. In clinical trials, 213 of 224 patients suffering from acid reflux disease were healed after 8 weeks. Test the manufacturer's claim at the $\alpha = 0.01$ level of significance.
- 19. Fatal Traffic Accidents** According to the National Highway and Traffic Safety Administration, the proportion of fatal traffic accidents in the United States in which the driver had a positive blood alcohol concentration (BAC) is 0.36. Suppose a random sample of 105 traffic fatalities in the state of Hawaii results in 51

that involved a positive BAC. Does the sample evidence suggest that Hawaii has a higher proportion of traffic fatalities involving a positive BAC than the United States at the $\alpha = 0.05$ level of significance?

20. Eating Together In December 2001, 38% of adults with children under the age of 18 reported that their family ate dinner together seven nights a week. In a recent poll, 403 of 1122 adults with children under the age of 18 reported that their family ate dinner together seven nights a week. Has the proportion of families with children under the age of 18 who eat dinner together seven nights a week decreased? Use the $\alpha = 0.05$ significance level.

NW 21. Taught Enough Math? In 1994, 52% of parents with children in high school felt it was a serious problem that high school students were not being taught enough math and science. A recent survey found that 256 of 800 parents with children in high school felt it was a serious problem that high school students were not being taught enough math and science. Do parents feel differently today than they did in 1994? Use the $\alpha = 0.05$ level of significance?
Source: Based on “Reality Check: Are Parents and Students Ready for More Math and Science?” *Public Agenda*, 2006.

22. Living Alone? In 2000, 58% of females aged 15 and older lived alone, according to the U.S. Census Bureau. A sociologist tests whether this percentage is different today by obtaining a random sample of 500 females aged 15 and older and finds that 285 are living alone. Is there sufficient evidence at the $\alpha = 0.1$ level of significance to conclude the proportion has changed since 2000?

NW 23. Quality of Education In August 2002, 47% of parents with children in grades K–12 were satisfied with the quality of education the students receive. A recent Gallup poll found that 437 of 1013 parents with children in grades K–12 were satisfied with the quality of education the students receive. Construct a 95% confidence interval to assess whether this represents evidence that parents’ attitudes toward the quality of education in the United States has changed since August 2002.

24. Infidelity According to menstuff.org, 22% of married men have “strayed” at least once during their married lives.

- (a) Describe how you might go about administering a survey to assess the accuracy of this statement.
- (b) A survey of 500 married men indicated that 122 have “strayed” at least once during their married life. Construct a 95% confidence interval for the population proportion of married men who have strayed. Use this interval to assess the accuracy of the statement made by menstuff.org.

25. Accuracy of the Drive Thru According to QSR Magazine, Chick-fil-A has the best accuracy of drive thru orders with 96.4% of all its drive thru orders filled correctly. The manager of a competing fast food restaurant wants to advertise that her drive thru is more accurate than Chick-fil-A. In a random sample of 350 drive thru orders, how many accurate orders would the manager need out of 350 to be able to claim her drive thru has a statistically significantly better accuracy record than Chick-fil-A at the 0.1 level of significance?

26. Talk to the Animals In an American Animal Hospital Association survey, 37% of respondents stated that they talk to their pets on the telephone. A veterinarian found this result hard to believe, so he randomly selected 150 pet owners and discovered that 54 of them spoke to their pet on the telephone. Does the veterinarian have the right to be skeptical? Use a 0.05 level of significance.

27. Blind Emotion When the area of the brain responsible for vision is destroyed, individuals experience *cortical blindness*. Patients with cortical blindness are unaware of any visual stimulus, including light. In a 52-year-old male patient with cortical blindness (as a result of two strokes within a 38-day timeframe), a series of visual stimuli were presented on a computer screen. The patient was given two choices for each stimulus and asked to report what was on the screen. The patient’s responses were recorded by an individual behind the screen who could not see the contents on the screen.

- (a) An initial baseline test was conducted by presenting black squares and circles on the white background of the screen. The patient correctly “guessed” the item on the screen in 90 of the 200 trials. Researchers concluded the patient’s performance was not statistically different from chance. Explain what this means using the language of hypothesis testing.
- (b) Researchers then presented the patient with a randomized series of pictures depicting angry and happy faces. The patient was asked to report the emotion shown on the face. The patient answered correctly in 118 of 200 trials. Do these results suggest the patient’s performance is statistically different from chance? If the results are statistically different from chance, construct a 95% confidence interval to estimate the proportion of correct answers. **Note:** Similar results were obtained for sad versus happy and fearful versus happy faces.
- (c) The researchers wanted to determine if the patient could identify other facial characteristics. They randomly showed male or female faces and asked the patient to identify the gender. The patient was correct in 89 of 200 trials. What does this suggest?

Source: Pegna, Alan J. et al., “Discriminating Emotional Faces without Primary Visual Cortices Involves the Right Amygdala.” *Nature Neuroscience*, 8(1), 2005.

28. Reproducibility Researchers looked at studies that were reported in newspapers with the goal of determining whether initial studies had results that could be reproduced. Reproducibility of results means that subsequent analysis confirms the conclusion of the original study. Primary studies are studies where the researchers come up with a research objective, clearly state the goals of the study and population, describe the research method, test the research hypotheses, and draw conclusions.

- (a) Among the 156 primary studies reported by newspapers, 76 had results that were validated by subsequent analysis. Does this suggest less than a majority of initial studies reported by newspapers have their results validated by subsequent analysis?
 - (b) In the article, a null effect is defined as any study where the evidence is not in favor of the research (alternative) hypothesis. Put another way, a null effect means the null hypothesis of “no effect” or “no difference” is not rejected. Among 1475 studies reported in newspapers, 75 reported a null effect. Estimate with 95% confidence the proportion of studies in which there is a null effect (such as a particular behavior does not result in a higher risk of disease).
 - (c) What does the result in part (b) suggest about the likelihood of a research study being published with a null effect? Include a discussion of the incentive to have studies that are in favor of the research (alternative) hypothesis.
- NW 29. Small-Sample Hypothesis Test** Professors Honey Kirk and Diane Lerma of Palo Alto College developed a “learning community curriculum that blended the developmental

mathematics and the reading curriculum with a structured emphasis on study skills." In a typical developmental mathematics course at Palo Alto College, 50% of the students complete the course with a letter grade of A, B, or C. In the experimental course, of the 16 students enrolled, 11 completed the course with a letter grade of A, B, or C. Do you believe the experimental course was effective at the $\alpha = 0.05$ level of significance?

- (a) State the appropriate null and alternative hypotheses.
- (b) Verify that the normal model may not be used to estimate the P -value.
- (c) Explain why this is a binomial experiment.
- (d) Determine the P -value using the binomial probability distribution. State your conclusion to the hypothesis test.
- (e) Suppose the course is taught with 48 students and 33 complete the course with a letter grade of A, B, or C. Verify the normal model may now be used to estimate the P -value.
- (f) Use the normal model to obtain and interpret the P -value. State your conclusion to the hypothesis test.
- (g) Explain the role that sample size plays in the ability to reject statements in the null hypothesis.

Source: Kirk, Honey and Lerma, Diane, "Reading Your Way to Success in Mathematics: A Paired Course of Developmental Mathematics and Reading." *MathAMATYC Educator*, Vol. 1. No. 2, 2010.

30. Small-Sample Hypothesis Test In 1997, 4% of mothers smoked more than 21 cigarettes during their pregnancy. An obstetrician believes that the percentage of mothers who smoke 21 cigarettes or more is less than 4% today. She randomly selects 120 pregnant mothers and finds that 3 of them smoked 21 or more cigarettes during pregnancy. Does the sample data support the obstetrician's belief? Use the $\alpha = 0.05$ level of significance.

31. Small Sample Hypothesis Test: Super Bowl Investing From Super Bowl I (1967) through Super Bowl XXXI (1997), the stock market increased if an NFL team won the Super Bowl and decreased if an AFL team won. This condition held 28 out of 31 years.

- (a) Suppose the likelihood of predicting the direction of the stock market (increasing or decreasing) in any given year is 0.50. Decide on the appropriate null and alternative hypotheses to test whether the outcome of the Super Bowl can be used to predict the direction of the stock market.
- (b) Use the binomial probability distribution to determine the P -value for the hypothesis test from part (a).
- (c) Comment on the dangers of using the outcome of the hypothesis test to judge investments. Be sure your comment includes a discussion of circumstances in which associations have a causal relationship.

32. Statistics in the Media A headline read, "More Than Half of Americans Say Federal Taxes Too High." The headline was based on a random sample of 1026 adult Americans in which 534 stated the amount of federal tax they have to pay is too high. Is this an accurate headline?

Source: Gallup Organization, April 14, 2014.

33. Threaded Problem: Tornado The data set "Tornadoes_2017" located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a) Since 1950, the proportion of tornadoes that have been F0 is 0.465. The data in column F0 represent whether a tornado that struck was an F0 (Yes), or not (No). Test whether a different proportion of F0 tornadoes occur in Texas than nationally. (**Note:** Treat the tornadoes that struck in Texas as a simple random sample of all tornadoes that

struck Texas since 1950 using a 0.05 level of significance.) To conduct the hypothesis test in StatCrunch, select Stat > Proportion Stats > One Sample > With Data. In the "Where:" box type "State=TX". The summarized data show that 82 of 168 tornadoes in Texas are F0.

- (b) Test whether a lower proportion of F0 tornadoes occur in Georgia than nationally. The summarized data show that 43 of 118 tornadoes in Georgia are F0.
- (c) Estimate the proportion of F0 tornadoes that occur in Georgia with 95% confidence. Interpret the result.

DATA 34. Gender Income Inequality The Sullivan Statistics Survey II asked, "Do you believe there is an income inequality discrepancy between males and females when each has the same experience and education?" Go to www.pearsonhighered.com/sullivanstats to obtain the data file SullivanStatsSurveyII using the file format of your choice for the version of the text you are using. The data may be found under the column "GenderIncomeInequality." Treat the sample as a random sample of adult Americans. Do the survey results suggest a supermajority (more than 60%) of adult Americans believe there is income inequality among males and females with the same experience and education? Use an $\alpha = 0.05$ level of significance.

DATA 35. Are Spreads Accurate? For every NFL game, there is a team that is expected to win by a certain number of points. In betting parlance, this is called the spread. For example, if the Chicago Bears are expected to beat the Green Bay Packers by three points, a sports bettor would say, "Chicago is minus three." So, if the Bears lose to Green Bay, or do not win by more than three points, a bet on Chicago would be a loser. If point spreads are accurate, we would expect about half of all games played to result in the favored team winning (beating the spread) and about half of all games to result in the team favored to not beat the spread. The following data represent the results of 45 randomly selected games where a 0 indicates the favored team did not beat the spread and a 1 indicates the favored team beat the spread. Do the data suggest that sport books establish accurate spreads?

0	0	0	0	0
1	0	0	1	0
0	0	0	1	0
1	1	0	0	1
0	0	1	1	1
0	0	1	1	0
1	0	0	1	0
0	1	1	0	1
0	0	1	1	1

Source: <http://www.vegasinsider.com>

36. Accept versus Do Not Reject In the United States, historically, 40% of registered voters are Republican. Suppose you obtain a simple random sample of 320 registered voters and find 142 registered Republicans.

- (a) Consider the hypotheses $H_0: p = 0.4$ versus $H_1: p > 0.4$. Explain what the researcher would be testing. Perform the test at the $\alpha = 0.05$ level of significance. Write a conclusion for the test.
- (b) Consider the hypotheses $H_0: p = 0.41$ versus $H_1: p > 0.41$. Explain what the researcher would be testing. Perform the test at the $\alpha = 0.05$ level of significance. Write a conclusion for the test.

- (c) Consider the hypotheses $H_0: p = 0.42$ versus $H_1: p > 0.42$. Explain what the researcher would be testing. Perform the test at the $\alpha = 0.05$ level of significance. Write a conclusion for the test.
- (d) Based on the results of parts (a)–(c), write a few sentences that explain the difference between “accepting” the statement in the null hypothesis versus “not rejecting” the statement in the null hypothesis.

37. Interesting Results Suppose you wish to find out the answer to the age-old question, “Do Americans prefer Coke or Pepsi?” You conduct a blind taste test in which individuals are randomly asked to drink one of the colas first, followed by the other cola, and then asked to disclose which drink they prefer. Results of your taste test indicate that 53 of 100 individuals prefer Pepsi.

- (a) Conduct a hypothesis test (preferably using technology)
 $H_0: p = p_0$ versus $H_1: p \neq p_0$ for $p_0 = 0.42, 0.43, 0.44, \dots, 0.64$ at the $\alpha = 0.05$ level of significance. For which values of p_0 do you not reject the null hypothesis? What do each of the values of p_0 represent?
- (b) Construct a 95% confidence interval for the proportion of individuals who prefer Pepsi.
- (c) Suppose you changed the level of significance in conducting the hypothesis test to $\alpha = 0.01$? What would happen to the range of values of p_0 for which the null hypothesis is not rejected? Why does this make sense?

38. Simulation Simulate drawing 100 simple random samples of size $n = 40$ from a population whose proportion is 0.3.

- (a) Test the null hypothesis $H_0: p = 0.3$ versus $H_1: p \neq 0.3$ for each simulated sample.
- (b) If we test the hypothesis at the $\alpha = 0.1$ level of significance, how many of the 100 samples would you expect to result in a Type I error?
- (c) Count the number of samples that lead to a rejection of the null hypothesis. Is it close to the expected value determined in part (b)?
- (d) How do we know that a rejection of the null hypothesis results in making a Type I error in this situation?

39. Simulation: Predicting the Future Parapsychology (psi) is a field of study that deals with clairvoyance or precognition. Psi made its way back into the news when a professional, refereed journal published an article by Cornell psychologist Daryl Bem, in which he claimed to demonstrate that psi is a real phenomenon. In the article Bem stated that certain individuals behave today as if they already know what is going to happen in the future. That is, individuals adjust current behavior in anticipation of events that are going to happen in the future. Here, we will present a simplified version of Bem’s research.

- (a) Suppose an individual claims to have the ability to predict the color (red or black) of a card from a standard 52-card deck. Of course, simply by guessing we would expect the individual to get half the predictions correct, and half incorrect. What is the statement of no change or no effect in this type of experiment? What statement would we be looking to demonstrate? Based on this, what would be the null and alternative hypotheses?
- (b) Suppose you ask the individual to guess the correct color of a card 40 times, and the alleged savant (wise person) guesses the correct color 24 times. Would you consider this to be convincing evidence that that individual can guess the color of the card at better than a 50/50 rate? To answer this question, we want to determine the likelihood of getting

24 or more colors correct even if the individual is simply guessing. To do this, we assume the individual is guessing so that the probability of a successful guess is 0.5.

Explain how 40 coins flipped independently with heads representing a successful guess can be used to model the card-guessing experiment.

- (c) Now, use a random number generator, or applet such as the Coin-Flip applet in StatCrunch to flip 40 fair coins, 1000 different times. What proportion of time did you observe 24 or more heads due to chance alone? What does this tell you? Do you believe the individual has the ability to guess card color based on the results of the simulation, or could the results simply have occurred due to chance?
- (d) Explain why guessing card color (or flipping coins) 40 times and recording the number of correct guesses (or heads) is a binomial experiment.
- (e) Use the binomial probability function to find the probability of at least 24 correct guesses in 40 trials assuming the probability of success is 0.5.
- (f) Look at the graph of the outcomes of the simulation from part (c). Explain why the normal model might be used to estimate the probability of obtaining at least 24 correct guesses in 40 trials assuming the probability of success is 0.5. Use the model to estimate the P -value.
- (g) Based on the probabilities found in parts (c), (e), and (f), what might you conclude about the alleged savant’s ability to predict card color?

40. Putting It Together: Lupus Based on historical birthing records, the proportion of males born worldwide is 0.51. In other words, the commonly held belief that boys are just as likely as girls is false. Systemic lupus erythematosus (SLE), or lupus for short, is a disease in which one’s immune system attacks healthy cells and tissue by mistake. It is well known that lupus tends to exist more in females than in males. Researchers wondered, however, if families with a child who had lupus had a lower ratio of males to females than the general population. If this were true, it would suggest that something happens during conception that causes males to be conceived at a lower rate when the SLE gene is present. To determine if this hypothesis is true, the researchers obtained records of families with a child who had SLE. A total of 23 males and 79 females were found to have SLE. The 23 males with SLE had a total of 23 male siblings and 22 female siblings. The 79 females with SLE had a total of 69 male siblings and 80 female siblings.

Source: L.N. Moorthy, M.G.E. Peterson, K.B. Onel, and T.J.A. Lehman. “Do Children with Lupus Have Fewer Male Siblings?” *Lupus* 2008 17:128–131, 2008.

- (a) Explain why this is an observational study.
- (b) Is the study retrospective or prospective? Why?
- (c) There are a total of $23 + 69 = 92$ male siblings in the study. How many female siblings are in the study?
- (d) Draw a relative frequency bar graph of gender of the siblings.
- (e) Find a point estimate for the proportion of male siblings in families where one of the children has SLE.
- (f) Does the sample evidence suggest that the proportion of male siblings in families where one of the children has SLE is less than 0.51, the accepted proportion of males born in the general population? Use the $\alpha = 0.05$ level of significance.
- (g) Construct a 95% confidence interval for the proportion of male siblings in a family where one of the children has SLE.

41. Putting It Together: Naughty or Nice? Yale University graduate student J. Kiley Hamlin conducted an experiment in which 16 ten-month-old babies were asked to watch a climber character attempt to ascend a hill. On two occasions, the baby witnesses the character fail to make the climb. On the third attempt, the baby witnesses either a helper toy push the character up the hill or a hinderer toy prevent the character from making the ascent. The helper and hinderer toys were shown to each baby in a random fashion for a fixed amount of time. The baby was then placed in front of each toy and allowed to choose which toy he or she wished to play with. In 14 of the 16 cases, the baby chose the helper toy.

Source: J. Kiley Hamlin et al., "Social Evaluation by Preverbal Infants." *Nature*, Nov. 2007.

- (a) Why is it important to randomly expose the baby to the helper or hinderer toy first?
- (b) What would be the appropriate null and alternative hypotheses if the researcher is attempting to show that babies prefer helpers over hinderers?
- (c) Use the binomial probability formula to determine the *P*-value for this test.
- (d) In testing 12 six-month-old babies, all 12 preferred the helper toy. The *P*-value was reported as 0.0002. Interpret this result.

Explaining the Concepts

- 42. Explain what a *P*-value is. What is the criterion for rejecting the null hypothesis using the *P*-value approach?
- 43. Suppose we are testing the hypothesis $H_0: p = 0.3$ versus $H_1: p > 0.3$ and we find the *P*-value to be 0.23. Explain what this means. Would you reject the null hypothesis? Why?
- 44. Suppose we are testing the hypothesis $H_0: p = 0.65$ versus $H_1: p \neq 0.65$ and we find the *P*-value to be 0.02. Explain what this means. Would you reject the null hypothesis? Why?
- 45. Discuss the advantages and disadvantages of using the Classical Approach to hypothesis testing. Discuss the advantages and disadvantages of using the *P*-value approach to hypothesis testing.
- 46. The headline reporting the results of a poll conducted by the Gallup organization stated "Majority of Americans at Personal Best in the Morning." The results indicated that a survey of 1100 Americans resulted in 55% stating they were at their personal best in the morning. The poll's results were reported with a margin of error of 3%. Explain why the Gallup organization's headline is accurate.
- 47. Explain what "statistical significance" means.

10.3 Hypothesis Tests for a Population Mean



Preparing for This Section Before getting started, review the following:

- Sampling distribution of \bar{x} (Section 8.1, pp. 371–379)
- The *t*-distribution (Section 9.2, pp. 411–414)
- Using probabilities to identify unusual results (Section 5.1, p. 230)
- Confidence intervals for a mean (Section 9.2, pp. 415–418)

Objectives

- ① Test hypotheses about a mean
- ② Understand the difference between statistical significance and practical significance

1 Test Hypotheses about a Mean

In Section 8.1, we learned that the distribution of \bar{x} is approximately normal with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ provided the population from which the sample was drawn is normally distributed or the sample size is sufficiently large (because of the Central Limit Theorem). So $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ follows a standard normal distribution.

However, it is unreasonable to expect to know σ without knowing μ . This problem was resolved by William Gosset, who determined that $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ follows Student's *t*-distribution with $n - 1$ degrees of freedom. We use this distribution to perform hypothesis tests on a mean.

Testing hypotheses about a mean follows the same logic as testing a hypothesis about a population proportion. The only difference is that we use Student's *t*-distribution, rather than the normal distribution.

Testing Hypotheses Regarding a Population Mean

To test hypotheses regarding the population mean, use the following steps, provided that

- the sample is obtained using simple random sampling or from a randomized experiment.
- the sample has no outliers and the population from which the sample is drawn is normally distributed, or the sample size, n , is large ($n \geq 30$).
- the sampled values are independent of each other. That is, the sample size is less than 5% of the population size.

Step 1 Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways:

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$	$H_1: \mu < \mu_0$	$H_1: \mu > \mu_0$

Note: μ_0 is the assumed value of the population mean.

Step 2 Select a level of significance, α , depending on the seriousness of making a Type I error.

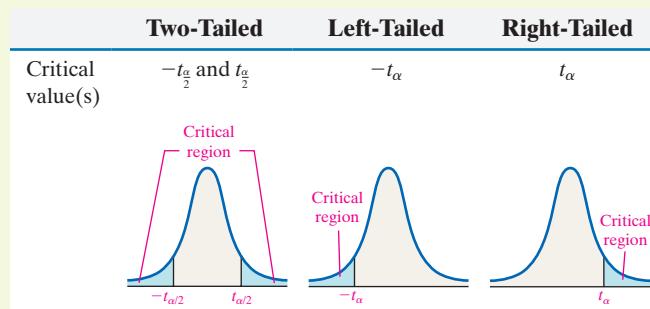
Classical Approach

Step 3 Compute the **test statistic**

$$t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

which follows Student's t -distribution with $n - 1$ degrees of freedom.

Use Table VII to determine the critical value.



Step 4 Compare the critical value to the test statistic.

Two-Tailed	Left-Tailed	Right-Tailed
If $t_0 < -t_{\alpha/2}$ or $t_0 > t_{\alpha/2}$, reject the null hypothesis.	If $t_0 < -t_\alpha$, reject the null hypothesis.	If $t_0 > t_\alpha$, reject the null hypothesis.

Step 5 State the conclusion.

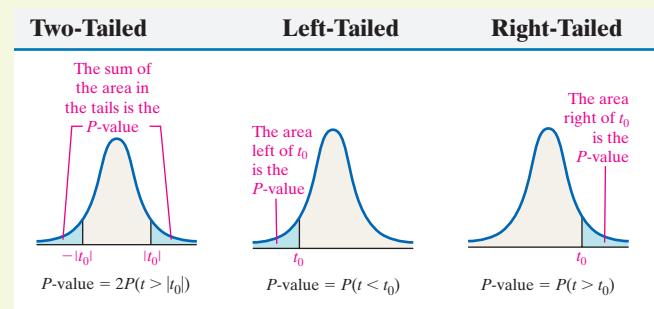
P-Value Approach

By Hand Step 3 Compute the **test statistic**

$$t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

which follows Student's t -distribution with $n - 1$ degrees of freedom.

Use Table VII to approximate the P -value.



Technology Step 3 Use a statistical spreadsheet or calculator with statistical capabilities to obtain the P -value. The directions for obtaining the P -value using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch are in the Technology Step-by-Step on page 464.

Step 4 If the P -value $< \alpha$, reject the null hypothesis.

Notice that the procedure just presented requires either that the population from which the sample was drawn be normal or that the sample size be large ($n \geq 30$). The procedure is robust, so minor departures from normality will not adversely affect the results of the test. However, if the data include outliers, the procedure should not be used.

We will verify these assumptions by constructing normal probability plots (to assess normality) and boxplots (to discover whether there are outliers). This is Option 1 (See page 416). Option 2 for verifying model requirements states the boxplot should be symmetric with no outliers to use Student's t -distribution for inference on the mean. If the model requirements are not satisfied, nonparametric tests should be performed.

Before we look at a couple of examples, it is important to understand that we cannot find exact P -values using the t -distribution table (Table VII) because the table provides t -values only for certain areas. However, we can use the table to calculate lower and

upper bounds on the P -value. To find exact P -values, use statistical software or a graphing calculator with advanced statistical features.

EXAMPLE 1 Testing a Hypothesis about a Population Mean: Large Sample

Problem The mean height of American males is 176.3 cm. The heights of the 44 male U.S. presidents* (Washington through Trump) have a mean 180.1 cm and a standard deviation of 7.1 cm. Treating the 44 presidents as a simple random sample, determine if there is evidence to suggest that U.S. presidents are taller than the typical American male. Use the $\alpha = 0.05$ level of significance.

Approach Assume that all U.S. presidents come from a population whose height is 176.3 cm (that is, there is no difference between heights of U.S. presidents and the general American male population). Then determine the likelihood of obtaining a sample mean of 180.1 cm or higher from a population whose mean is 176.3 cm. If the result is unlikely, reject the assumption stated in the null hypothesis in favor of the more likely notion that the mean height of U.S. presidents is greater than 176.3 cm. However, if obtaining a sample mean of 180.1 cm from a population whose mean is assumed to be 176.3 cm is not unusual, do not reject the null hypothesis (and attribute the difference to sampling error). Assume the population of potential U.S. presidents is large (for independence). Because the sample size is large, the distribution of \bar{x} is approximately normal. Follow Steps 1 through 5.

Solution

Step 1 We want to know if U.S. presidents are taller than the typical American male who is 176.3 cm. We assume there is no difference between the height of a typical American male and U.S. presidents, so

$$H_0: \mu = 176.3 \text{ cm} \quad \text{versus} \quad H_1: \mu > 176.3 \text{ cm}$$

Step 2 The level of significance is $\alpha = 0.05$.

Classical Approach

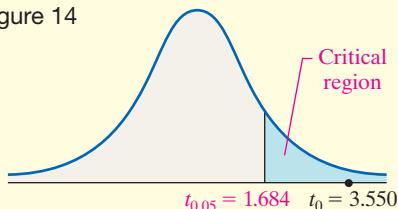
Step 3 The sample mean is $\bar{x} = 180.1$ cm and the sample standard deviation is $s = 7.1$ cm.

The test statistic is

$$t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{180.1 - 176.3}{\frac{7.1}{\sqrt{44}}} = 3.550$$

Because this is a right-tailed test, determine the critical value at the $\alpha = 0.05$ level of significance with $44 - 1 = 43$ degrees of freedom to be $t_{0.05} = 1.684$ (using 40 degrees of freedom since this is closest to 43). The critical region is shown in Figure 14.

Figure 14



P-Value Approach

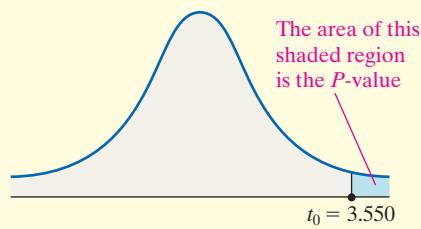
By Hand Step 3 The sample mean is $\bar{x} = 180.1$ cm and the sample standard deviation is $s = 7.1$ cm.

The test statistic is

$$t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{180.1 - 176.3}{\frac{7.1}{\sqrt{44}}} = 3.550$$

Because this is a right-tailed test, the P -value is the area under the t -distribution with 43 degrees of freedom to the right of $t_0 = 3.550$ as shown in Figure 15.

Figure 15



Using Table VII, find the row that corresponds to 40 degrees of freedom (we use 40 degrees of freedom because it is closest to the actual degrees of freedom, $44 - 1 = 43$). The value 3.550 lies between 3.307 and 3.551. The

*Grover Cleveland was elected to two non-consecutive terms, so there have technically been 45 presidents of the United States.

Step 4 The test statistic, $t_0 = 3.550$, is labeled in Figure 14. Because the test statistic lies in the critical region, reject the null hypothesis.

value of 3.307 has an area of 0.001 to the right under the t -distribution with 40 degrees of freedom. The area under the t -distribution with 40 degrees of freedom to the right of 3.551 is 0.0005.

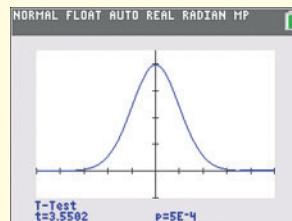
Because 3.550 is between 3.307 and 3.551, the P -value is between 0.0005 and 0.001. So

$$0.0005 < P\text{-value} < 0.001$$

There is an alternate form of Table VII (see Table X in Appendix A) that is useful for finding more accurate P -values. The table is set up similarly to Table V (the standard normal table). To use this alternate version, find the column that corresponds to the degrees of freedom, and the row that corresponds to the test statistic (rounded to the nearest tenth). The intersection of the row and column represents the area under the t -distribution to the right of the test statistic. Using 40 degrees of freedom (because 43 df is not in the table) with a test statistic of 3.6, the P -value is 0.000 (rounded to three decimal places).

Technology Step 3 Using a TI-84 Plus CE graphing calculator, the P -value is 0.0005. See Figure 16.

Figure 16



NOTE

In Figure 16, the P -value is reported in scientific notation.

$$5E-4 = 0.0005$$

The E-4 means move the decimal point four places to the left.

NOTE

If you are using Table XVII to find P -values for a left-tailed test, use the symmetry of the t -distribution. That is,

$$P(t < -t_0) = P(t > t_0)$$

Step 4 The P -value of 0.0005 [by hand: $0.0005 < P\text{-value} < 0.001$] means that, if the null hypothesis that $\mu = 176.3$ cm is true, we expect a sample mean of 180.1 cm or higher in about 5 out of 10,000 samples. The results we obtained do not seem to be consistent with the assumption that the mean height of this population is 176.3 cm. Put another way, because the P -value is less than the level of significance, $\alpha = 0.05$ ($0.0005 < 0.05$), we reject the null hypothesis.

Step 5 There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that U.S. presidents are taller than the typical American male.

NW Now Work Problem 13

EXAMPLE 2

Testing a Hypothesis about a Population Mean: Small Sample

Table 1

19.68	20.66	19.56
19.98	20.65	19.61
20.55	20.36	21.02
21.50	19.74	

Source: Michael Carlisle, student at Joliet Junior College.

Problem The “fun size” of a Snickers bar is supposed to weigh 20 grams. Because the penalty for selling candy bars under their advertised weight is severe, the manufacturer calibrates the machine so the mean weight is 20.1 grams. The quality-control engineer at M&M–Mars, the Snickers manufacturer, is concerned about the calibration. He obtains a random sample of 11 candy bars, weighs them, and obtains the data shown in Table 1. Should the machine be shut down and calibrated? Because shutting down the plant is very expensive, he decides to conduct the test at the $\alpha = 0.01$ level of significance.

Approach Assume that the machine is calibrated correctly. So there is no difference between the actual mean weight and the calibrated weight of the candy. We want to know whether the machine is incorrectly calibrated, which would result in a mean weight that is too high or too low. Therefore, this is a two-tailed test.

(continued)

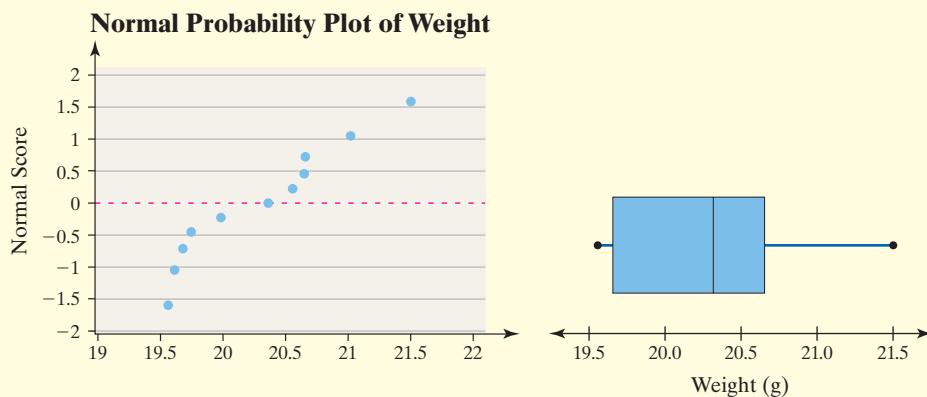
Before performing the hypothesis test, verify that the data come from a population that is normally distributed with no outliers by constructing a normal probability plot and boxplot. Then proceed to follow Steps 1 through 5.

NOTE

If using Option 2, the boxplot in Figure 17 suggests the distribution of weights of Snickers is “symmetric enough” to use Student’s *t*-distribution to conduct the hypothesis test.

Solution Figure 17 displays the normal probability plot and boxplot. The correlation between the weights and expected *z*-scores is 0.967 [Tech: 0.970]. Because $0.967 > 0.923$ (Table VI), the normal probability plot indicates that the data could come from a population that is approximately normal. The boxplot has no outliers. We can proceed with the hypothesis test.

Figure 17



Step 1 The engineer wishes to determine whether the Snickers have a mean weight of 20.1 grams or not. The hypotheses can be written

$$H_0: \mu = 20.1 \text{ grams} \quad \text{versus} \quad H_1: \mu \neq 20.1 \text{ grams}$$

This is a two-tailed test.

Step 2 The level of significance is $\alpha = 0.01$.

Classical Approach

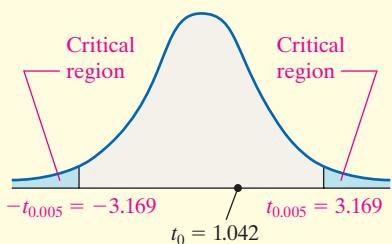
Step 3 From the data in Table 1, the sample mean is $\bar{x} = 20.301$ grams and the sample standard deviation is $s = 0.64$ gram.

The test statistic is

$$t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{20.301 - 20.1}{\frac{0.64}{\sqrt{11}}} = 1.042$$

Because this is a two-tailed test, determine the critical values at the $\alpha = 0.01$ level of significance with $11 - 1 = 10$ degrees of freedom to be $-t_{0.01/2} = -t_{0.005} = -3.169$ and $t_{0.01/2} = t_{0.005} = 3.169$. The critical regions are shown in Figure 18.

Figure 18



P-Value Approach

By Hand Step 3 From the data in Table 1, the sample mean is $\bar{x} = 20.301$ grams and the sample standard deviation is $s = 0.64$ gram.

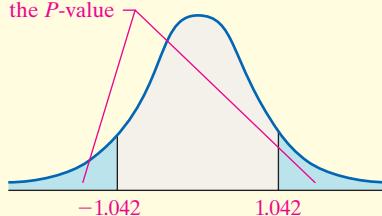
The test statistic is

$$t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{20.301 - 20.1}{\frac{0.64}{\sqrt{11}}} = 1.042$$

Because this is a two-tailed test, the *P*-value is the area under the *t*-distribution with $n - 1 = 11 - 1 = 10$ degrees of freedom to the left of $-t_0 = -1.042$ and to the right of $t_0 = 1.042$, as shown in Figure 19. That is, $P\text{-value} = P(t < -1.042) + P(t > 1.042) = 2P(t > 1.042)$, with 10 degrees of freedom.

Figure 19

The sum of these two areas is the *P*-value



Step 4 Because the test statistic, $t_0 = 1.042$, does not lie in the critical region, do not reject the null hypothesis.

Using Table VII, we find the row that corresponds to 10 degrees of freedom. The value 1.042 lies between 0.879 and 1.093. The value of 0.879 has an area of 0.20 to the right under the t -distribution. The area under the t -distribution to the right of 1.093 is 0.15.

Because 1.042 is between 0.879 and 1.093, the P -value is between $2(0.15)$ and $2(0.20)$. So

$$0.30 < P\text{-value} < 0.40$$

Using Table X we find $P\text{-value} = 2P(t > 1.0) = 2(0.170) = 0.340$ with 10 degrees of freedom.

Technology Step 3 Using Minitab, the exact P -value is 0.323.

Step 4 The P -value of 0.323 [by-hand: $0.30 < P\text{-value} < 0.40$] means that, if the null hypothesis that $\mu = 20.1$ grams is true, we expect about 32 out of 100 samples to result in a sample mean as extreme or more extreme than the one obtained. The result we obtained is not unusual, so we do not reject the null hypothesis.

Step 5 There is not sufficient evidence to conclude that the Snickers have a mean weight different from 20.1 grams at the $\alpha = 0.01$ level of significance. The machine should not be shut down.

NW Now Work Problem 21

IN OTHER WORDS

Results are statistically significant if the difference between the observed result and the statement made in the null hypothesis is unlikely to occur due to chance alone.

②

Understand the Difference between Statistical Significance and Practical Significance

When a large sample size is used in a hypothesis test, the results could be statistically significant even though the difference between the sample statistic and mean stated in the null hypothesis may have no *practical significance*.

Definition

Practical significance refers to the idea that, while small differences between the statistic and parameter stated in the null hypothesis are statistically significant, the difference may not be large enough to cause concern or be considered important.

EXAMPLE 3 Statistical versus Practical Significance

Problem According to the American Community Survey, the mean travel time to work in Collin County, Texas, in 2017 was 29.3 minutes. The Department of Transportation reprogrammed all the traffic lights in Collin County in an attempt to reduce travel time. To determine if there is evidence that travel time has decreased as a result of the reprogramming, the Department of Transportation obtains a random sample of 2500 commuters, records their travel time to work, and finds a sample mean of 29.0 minutes with a standard deviation of 8.5 minutes. Does this result suggest that travel time has decreased at the $\alpha = 0.05$ level of significance?

Approach We will use both the classical and P -value approach to test the hypothesis.

Solution

Step 1 The Department of Transportation wants to know if the mean travel time to work has decreased from 29.3 minutes. From this, we have

$$H_0: \mu = 29.3 \text{ minutes} \quad \text{versus} \quad H_1: \mu < 29.3 \text{ minutes}$$

(continued)

Step 2 The level of significance is $\alpha = 0.05$.

Step 3 The test statistic is

$$t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{29.0 - 29.3}{\frac{8.5}{\sqrt{2500}}} = -1.765$$

Classical Approach

Because this is a left-tailed test, the critical value with $\alpha = 0.05$ and $2500 - 1 = 2499$ degrees of freedom is $-t_{0.05} \approx -1.645$ (use the last row of Table VII when the degrees of freedom is greater than 1000).

Step 4 Because the test statistic is less than the critical value (the test statistic falls in the critical region), we reject the null hypothesis.

Step 5 There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude the mean travel time to work has decreased.

While the difference between 29.0 minutes and 29.3 minutes is statistically significant, it has no practical meaning. After all, is 0.3 minute (18 seconds) really going to make anyone feel better about his or her commute to work? 

The reason that the results from Example 3 were statistically significant had to do with the large sample size. The moral of the story is this:

CAUTION!

Beware of studies with large sample sizes that claim statistical significance because the differences may not have any practical meaning.

Large sample sizes can lead to results that are statistically significant, while the difference between the statistic and parameter in the null hypothesis is not enough to be considered practically significant.

Technology Step-by-Step

Hypothesis Tests Regarding μ

TI-83/84 Plus

1. If necessary, enter raw data in L1.
2. Press STAT, highlight TESTS, and select 2:T-Test.
3. If the data are raw, highlight DATA; make sure that List is set to L1 and Freq is set to 1. If summary statistics are known, highlight STATS and enter the summary statistics. For the value of μ_0 , enter the value of the mean stated in the null hypothesis.
4. Select the direction of the alternative hypothesis.
5. Highlight **Calculate** or **Draw** and press ENTER. The TI-83/84 gives the *P*-value.

Minitab

1. Enter raw data in column C1.
2. Select the Stat menu, highlight **Basic Statistics**, then highlight **1-Sample t . . .**.
3. If you have raw data, select “One or more samples, each in a column” from the drop-down menu. Place the cursor in the box, highlight the column containing the raw data, and click “Select.” If you have summarized data, select “Summarized data” from the drop-down menu. Enter the sample size, sample mean, and sample standard deviation. Check the “Perform hypothesis test” box and enter the value of the population mean stated in the null hypothesis.
4. Click Options. Enter the direction of the alternative hypothesis. Click OK twice.

Excel

1. Enter raw data into Column A.
2. Load the XLSTAT Add-in, if necessary.
3. Select the XLSTAT menu and highlight Parametric tests. Select One-sample *t*-test and *z*-test.
4. Place the cursor in the Data: cell and then highlight the data in the spreadsheet. Check Student’s *t*-test.
5. Click the Options tab. Choose the appropriate direction of the alternative hypothesis. Enter the mean stated in the null hypothesis in the Theoretical mean: cell. Enter the level of significance required for a confidence interval. For example, enter 10 for a 90% confidence interval. Click OK.

StatCrunch

1. If you have raw data, enter them into the spreadsheet. Name the column variable.
2. Select Stat, highlight **T Stats**, select **One Sample**, and then choose either **With Data** or **With Summary**.
3. If you chose **With Data**, select the column that has the observations. If you chose **With Summary**, enter the mean, standard deviation, and sample size. Choose the hypothesis test radio button. Enter the value of the mean stated in the null hypothesis and choose the direction of the alternative hypothesis from the pull-down menu. Click Compute!.



10.3 Assess Your Understanding

Skill Building

1. **(a)** Determine the critical value for a right-tailed test of a population mean at the $\alpha = 0.01$ level of significance with 15 degrees of freedom.
(b) Determine the critical value for a left-tailed test of a population mean at the $\alpha = 0.05$ level of significance based on a sample size of $n = 20$.
(c) Determine the critical values for a two-tailed test of a population mean at the $\alpha = 0.05$ level of significance based on a sample size of $n = 13$.
 2. **(a)** Determine the critical value for a right-tailed test of a population mean at the $\alpha = 0.1$ level of significance with 22 degrees of freedom.
(b) Determine the critical value for a left-tailed test of a population mean at the $\alpha = 0.01$ level of significance based on a sample size of $n = 40$.
(c) Determine the critical values for a two-tailed test of a population mean at the $\alpha = 0.01$ level of significance based on a sample size of $n = 33$.
 3. To test $H_0: \mu = 50$ versus $H_1: \mu < 50$, a simple random sample of size $n = 24$ is obtained from a population that is known to be normally distributed.
(a) If $\bar{x} = 47.1$ and $s = 10.3$, compute the test statistic.
(b) If the researcher decides to test this hypothesis at the $\alpha = 0.05$ level of significance, determine the critical value.
(c) Draw a t -distribution that depicts the critical region.
(d) Will the researcher reject the null hypothesis? Why?
 4. To test $H_0: \mu = 40$ versus $H_1: \mu > 40$, a simple random sample of size $n = 25$ is obtained from a population that is known to be normally distributed.
(a) If $\bar{x} = 42.3$ and $s = 4.3$, compute the test statistic.
(b) If the researcher decides to test this hypothesis at the $\alpha = 0.1$ level of significance, determine the critical value.
(c) Draw a t -distribution that depicts the critical region.
(d) Will the researcher reject the null hypothesis? Why?
 5. To test $H_0: \mu = 100$ versus $H_1: \mu \neq 100$, a simple random sample of size $n = 23$ is obtained from a population that is known to be normally distributed.
(a) If $\bar{x} = 104.8$ and $s = 9.2$, compute the test statistic.
(b) If the researcher decides to test this hypothesis at the $\alpha = 0.01$ level of significance, determine the critical values.
(c) Draw a t -distribution that depicts the critical region.
(d) Will the researcher reject the null hypothesis? Why?
(e) Construct a 99% confidence interval to test the hypothesis.
 6. To test $H_0: \mu = 80$ versus $H_1: \mu < 80$, a simple random sample of size $n = 22$ is obtained from a population that is known to be normally distributed.
(a) If $\bar{x} = 76.9$ and $s = 8.5$, compute the test statistic.
(b) If the researcher decides to test this hypothesis at the $\alpha = 0.02$ level of significance, determine the critical value.
(c) Draw a t -distribution that depicts the critical region.
(d) Will the researcher reject the null hypothesis? Why?
7. To test $H_0: \mu = 20$ versus $H_1: \mu < 20$, a simple random sample of size $n = 18$ is obtained from a population that is known to be normally distributed.
(a) If $\bar{x} = 18.3$ and $s = 4.3$, compute the test statistic.
(b) Draw a t -distribution with the area that represents the P -value shaded.
(c) Approximate and interpret the P -value.
(d) If the researcher decides to test this hypothesis at the $\alpha = 0.05$ level of significance, will the researcher reject the null hypothesis? Why?
 8. To test $H_0: \mu = 4.5$ versus $H_1: \mu > 4.5$, a simple random sample of size $n = 13$ is obtained from a population that is known to be normally distributed.
(a) If $\bar{x} = 4.9$ and $s = 1.3$, compute the test statistic.
(b) Draw a t -distribution with the area that represents the P -value shaded.
(c) Approximate and interpret the P -value.
(d) If the researcher decides to test this hypothesis at the $\alpha = 0.1$ level of significance, will the researcher reject the null hypothesis? Why?
 9. To test $H_0: \mu = 105$ versus $H_1: \mu \neq 105$, a simple random sample of size $n = 35$ is obtained.
(a) Does the population have to be normally distributed to test this hypothesis by using the methods presented in this section? Why?
(b) If $\bar{x} = 101.9$ and $s = 5.9$, compute the test statistic.
(c) Draw a t -distribution with the area that represents the P -value shaded.
(d) Approximate and interpret the P -value.
(e) If the researcher decides to test this hypothesis at the $\alpha = 0.01$ level of significance, will the researcher reject the null hypothesis? Why?
 10. To test $H_0: \mu = 45$ versus $H_1: \mu \neq 45$, a simple random sample of size $n = 40$ is obtained.
(a) Does the population have to be normally distributed to test this hypothesis by using the methods presented in this section? Why?
(b) If $\bar{x} = 48.3$ and $s = 8.5$, compute the test statistic.
(c) Draw a t -distribution with the area that represents the P -value shaded.
(d) Approximate and interpret the P -value.
(e) If the researcher decides to test this hypothesis at the $\alpha = 0.01$ level of significance, will the researcher reject the null hypothesis? Why?
(f) Construct a 99% confidence interval to test the hypothesis.

Applying the Concepts

11. **You Explain It! ATM Withdrawals** According to the Crown ATM Network, the mean ATM withdrawal is \$67. PayEase, Inc., manufactures an ATM that allows one to pay bills (electric, water, parking tickets, and so on), as well as withdraw money. A review of 40 withdrawals shows the mean withdrawal is \$73 from a PayEase ATM machine. Do people withdraw more money from a PayEase ATM machine?
(a) Determine the appropriate null and alternative hypotheses to answer the question.

- (b) Suppose the P -value for this test is 0.02. Explain what this value represents.
 (c) Write a conclusion for this hypothesis test assuming an $\alpha = 0.05$ level of significance.

12. You Explain It! Are Women Getting Taller? In 1990, the mean height of women 20 years of age or older was 63.7 inches based on data obtained from the Centers for Disease Control and Prevention's *Advance Data Report*, No. 347. Suppose that a random sample of 45 women who are 20 years of age or older today results in a mean height of 63.9 inches.

- (a) State the appropriate null and alternative hypotheses to assess whether women are taller today.
 (b) Suppose the P -value for this test is 0.35. Explain what this value represents.
 (c) Write a conclusion for this hypothesis test assuming an $\alpha = 0.10$ level of significance.

NW 13. Ready for College? The ACT is a college entrance exam. ACT has determined that a score of 22 on the mathematics portion of the ACT suggests that a student is ready for college-level mathematics. To achieve this goal, ACT recommends that students take a core curriculum of math courses: Algebra I, Algebra II, and Geometry. Suppose a random sample of 200 students who completed this core set of courses results in a mean ACT math score of 22.6 with a standard deviation of 3.9. Do these results suggest that students who complete the core curriculum are ready for college-level mathematics? That is, are they scoring above 22 on the math portion of the ACT?

- (a) State the appropriate null and alternative hypotheses.
 (b) Verify that the requirements to perform the test using the t -distribution are satisfied.
 (c) Use the classical or P -value approach at the $\alpha = 0.05$ level of significance to test the hypotheses in part (a).
 (d) Write a conclusion based on your results to part (c).

14. SAT Verbal Scores Do students who learned English and another language simultaneously score worse on the SAT Critical Reading exam than the general population of test takers? The mean score among all test takers on the SAT Critical Reading exam is 501. A random sample of 100 test takers who learned English and another language simultaneously had a mean SAT Critical Reading score of 485 with a standard deviation of 116. Do these results suggest that students who learn English as well as another language simultaneously score worse on the SAT Critical Reading exam?

- (a) State the appropriate null and alternative hypotheses.
 (b) Verify that the requirements to perform the test using the t -distribution are satisfied.
 (c) Use the classical or P -value approach at the $\alpha = 0.1$ level of significance to test the hypotheses in part (a).
 (d) Write a conclusion based on your results to part (c).

15. Effects of Alcohol on the Brain In a study published in the *American Journal of Psychiatry* (157:737–744, May 2000), researchers wanted to measure the effect of alcohol on the hippocampal region, the portion of the brain responsible for long-term memory storage, in adolescents. The researchers randomly selected 12 adolescents with alcohol use disorders to determine whether the hippocampal volumes in the alcoholic adolescents were less than the normal volume of 9.02 cubic centimeters (cm^3). An analysis of the sample data revealed that the hippocampal volume is approximately normal with $\bar{x} = 8.10 \text{ cm}^3$ and $s = 0.7 \text{ cm}^3$. Conduct the appropriate test at the $\alpha = 0.01$ level of significance.

16. Effects of Plastic Resin Para-nonylphenol is found in polyvinyl chloride (PVC) used in the food processing and packaging industries. Researchers wanted to determine the effect this substance had on the organ weight of first-generation mice when both parents were exposed to 50 micrograms per liter ($\mu\text{g}/\text{L}$) of para-nonylphenol in drinking water for 4 weeks. After 4 weeks, the mice were bred. After 100 days, the offspring of the exposed parents were sacrificed and the kidney weights were determined. The mean kidney weight of the 12 offspring was found to be 396.9 milligrams (mg), with a standard deviation of 45.4 mg. Is there significant evidence to conclude that the kidney weight of the offspring whose parents were exposed to 50 $\mu\text{g}/\text{L}$ of para-nonylphenol in drinking water for 4 weeks is greater than 355.7 mg, the mean weight of kidneys in normal 100-day-old mice at the $\alpha = 0.05$ level of significance?

Source: Vendula Kyselova et al., "Effects of *p*-nonylphenol and resveratrol on body and organ weight and in vivo fertility of outbred CD-1 mice," *Reproductive Biology and Endocrinology*, 2003.

17. Credit Scores A Fair Isaac Corporation (FICO) score is used by credit agencies (such as mortgage companies and banks) to assess the creditworthiness of individuals. Values range from 300 to 850, with a FICO score over 700 considered to be a quality credit risk. According to Fair Isaac Corporation, the mean FICO score is 703.5. A credit analyst wondered whether high-income individuals (incomes in excess of \$100,000 per year) had higher credit scores. He obtained a random sample of 40 high-income individuals and found the sample mean credit score to be 714.2 with a standard deviation of 83.2. Conduct the appropriate test to determine if high-income individuals have higher FICO scores at the $\alpha = 0.05$ level of significance.

18. TVaholics According to the American Time Use Survey, the typical American spends 154.8 minutes (2.58 hours) per day watching television. A survey of 50 Internet users results in a mean time watching television per day of 128.7 minutes, with a standard deviation of 46.5 minutes. Conduct the appropriate test to determine if Internet users spend less time watching television at the $\alpha = 0.05$ level of significance.

Source: Norman H. Nie and D. Sunshine Hillygus. "Where Does Internet Time Come From? A Reconnaissance." *IT & Society*, 1(2).

19. Age of Death-Row Inmates In 2002, the mean age of an inmate on death row was 40.7 years, according to data from the U.S. Department of Justice. A sociologist wondered whether the mean age of a death-row inmate has changed since then. She randomly selects 32 death-row inmates and finds that their mean age is 38.9, with a standard deviation of 9.6. Construct a 95% confidence interval about the mean age. What does the interval imply?

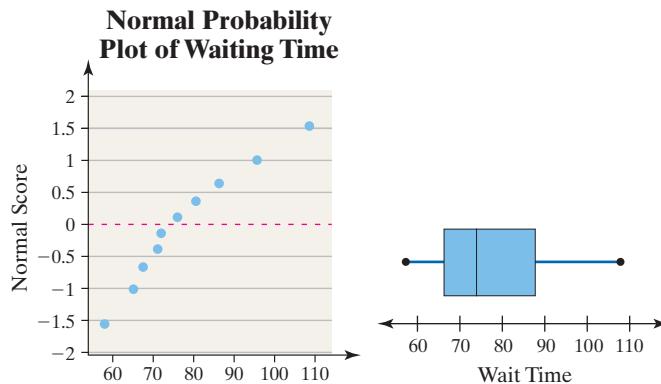
20. Energy Consumption In 2001, the mean household expenditure for energy was \$1493, according to data from the U.S. Energy Information Administration. An economist wanted to know whether this amount has changed significantly from its 2001 level. In a random sample of 35 households, he found the mean expenditure (in 2001 dollars) for energy during the most recent year to be \$1618, with a standard deviation \$321. Construct a 95% confidence interval about the mean energy expenditure. What does the interval imply?

NW 21. Waiting in Line The mean waiting time at the drive-thru of a fast-food restaurant from the time an order is placed to the time the order is received is 84.3 seconds. A manager devises a new drive-thru system that he believes will decrease wait time.

He initiates the new system at his restaurant and measures the wait time for 10 randomly selected orders. The wait times are provided in the table.

108.5	67.4	58.0	75.9	65.1
80.4	95.5	86.3	70.9	72.0

- (a) Because the sample size is small, the manager must verify that wait time is normally distributed and the sample does not contain any outliers. The normal probability plot and boxplot are shown. The correlation between waiting time and expected z -scores is 0.971. Are the conditions for testing the hypothesis satisfied?

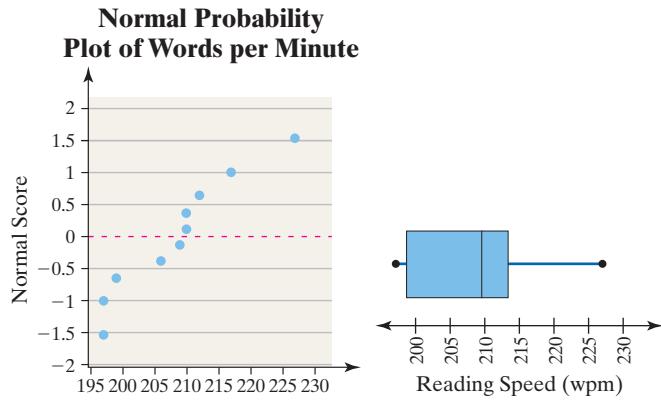


- (b) Is the new system effective? Use the $\alpha = 0.1$ level of significance.

- 22. Reading Rates** Michael Sullivan, son of the author, decided to enroll in a reading course that allegedly increases reading speed and comprehension. Prior to enrolling in the class, Michael read 198 words per minute (wpm). The following data represent the words per minute read for 10 different passages read after the course.

206	217	197	199	210
210	197	212	227	209

- (a) Because the sample size is small, we must verify that reading speed is normally distributed and the sample does not contain any outliers. The normal probability plot and boxplot are shown. The correlation between reading rate and expected z -scores is 0.964. Are the conditions for testing the hypothesis satisfied?



- (b) Was the class effective? Use the $\alpha = 0.10$ level of significance.

- DATA 23. Home Runs** Coors Field is home to the Colorado Rockies baseball team and is located in Denver, Colorado. Denver is approximately one mile above sea level, and the air there is “thinner.” Therefore, baseballs are thought to travel farther in this stadium. Does the evidence support this belief? In 2018, the mean distance of all home runs hit was 397.6 feet. The following data represent the distance, in feet, of a random sample of 12 home runs hit in Coors Field. Does this represent evidence to suggest that the ball travels farther in Coors Field than it does in the other Major League ballparks?

442	395	431	424	391	399
394	413	430	436	413	373

Source: Statcast

- (a) Draw a boxplot of the data. Does the boxplot suggest that it is appropriate to use Student’s t -distribution to determine the P -value?
 (b) Does the sample data suggest that the ball travels farther in Coors Field than it does in the other Major League ballparks? Use the $\alpha = 0.05$ level of significance.

- DATA 24. Four-Seam Fastball** Among all Major League Baseball players, the mean speed of a four-seam fastball is 93.58 miles per hour (mph). The following data represent a random sample of four-seam fastballs thrown by David Price.

93.6	91.4	93.2	93.3	92.3
92.3	94.1	91.6	93.3	90.4
93.4	93.7	90.6	93.9	91.8
92.2	91.5	94.0	92.7	92.5

Source: Statcast.

- (a) Draw a boxplot of the data. Does the boxplot suggest that it is appropriate to use Student’s t -distribution to determine the P -value?
 (b) Does the sample data suggest that David Price has a slower speed on his four-seam fastball? Use the $\alpha = 0.05$ level of significance.

- 25. Calcium in Rainwater** Calcium is essential to tree growth. In 1990, the concentration of calcium in precipitation in Chautauqua, New York, was 0.11 milligram per liter (mg/L). A random sample of 10 precipitation dates in 2018 results in the following data:

0.065	0.087	0.070	0.262	0.126
0.183	0.120	0.234	0.313	0.108

Source: National Atmospheric Deposition Program.

A normal probability plot suggests the data could come from a population that is normally distributed. A boxplot does not show any outliers. Does the sample evidence suggest that calcium concentrations have changed since 1990? Use the $\alpha = 0.05$ level of significance.

- DATA 26. Filling Bottles** A certain brand of apple juice is supposed to have 64 ounces of juice. Because the penalty for underfilling bottles is severe, the target mean amount of juice is 64.05 ounces. However, the filling machine is not precise, and the exact amount of juice varies from bottle to bottle. The quality-control manager wishes to verify that the mean

amount of juice in each bottle is 64.05 ounces so that she can be sure that the machine is not over- or underfilling. She randomly samples 22 bottles of juice, measures the content, and obtains the following data:

64.05	64.05	64.03	63.97	63.95	64.02
64.01	63.99	64.00	64.01	64.06	63.94
63.98	64.05	63.95	64.01	64.08	64.01
63.95	63.97	64.10	63.98		

A normal probability plot suggests the data could come from a population that is normally distributed. A boxplot does not show any outliers.

- (a) Should the assembly line be shut down so that the machine can be recalibrated? Use a 0.01 level of significance.
- (b) Explain why a level of significance of $\alpha = 0.01$ is more reasonable than $\alpha = 0.1$. [Hint: Consider the consequences of incorrectly rejecting the null hypothesis.]

DATA **27. Starbucks Stock** The volume of a stock is the number of shares traded for a given day. In 2011, Starbucks stock had a mean daily volume of 7.52 million shares according to Yahoo!Finance. A random sample of 40 trading days in 2018 was obtained and the volume of shares traded on those days was recorded. Go to www.pearsonhighered.com/sullivanstats to obtain the data file 10_3_27 using the file format of your choice for the version of the text you are using.

- (a) Draw a histogram of the data. Describe the shape of the distribution.
- (b) Draw a boxplot of the data. Are there any outliers?
- (c) Based on the shape of the histogram and boxplot, explain why a large sample size is necessary to perform inference on the mean using Student's t -distribution.
- (d) Does the evidence suggest that the volume of Starbucks stock has changed since 2011? Use an $\alpha = 0.05$ level of significance.

DATA **28. Study Time** Go to www.pearsonhighered.com/sullivanstats to obtain the data file 10_3_28 using the file format of your choice for the version of the text you are using. The data represent the amount of time students in Sullivan's online statistics course spent studying for Section 4.1—Scatter Diagrams and Correlation.

- (a) Draw a histogram of the data. Describe the shape of the distribution.
- (b) Draw a boxplot of the data. Are there any outliers?
- (c) Based on the shape of the histogram and boxplot, explain why a large sample size is necessary to perform inference on the mean using Student's t -distribution.
- (d) According to MyLabStatistics, the mean time students would spend on this assignment nationwide is 95 minutes. Treat the data as a random sample of all Sullivan online statistics students. Do the sample data suggest that Sullivan's students are any different from the country as far as time spent on Section 4.1 goes? Use an $\alpha = 0.05$ level of significance.

Test the hypothesis in the problem given by constructing a 95% confidence interval.

- 29.** Problem 25
31. Problem 27

- 30.** Problem 26
32. Problem 28

33. Statistical Significance versus Practical Significance A math teacher claims that she has developed a review course that increases the scores of students on the math portion of the SAT exam. Based on data from the College Board, SAT scores are normally distributed with $\mu = 515$. The teacher obtains a random sample of 1800 students, puts them through the review class, and finds that the mean SAT math score of the 1800 students is 519 with a standard deviation of 111.

- (a) State the null and alternative hypotheses.
- (b) Test the hypothesis at the $\alpha = 0.10$ level of significance. Is a mean SAT math score of 519 significantly higher than 515?
- (c) Do you think that a mean SAT math score of 519 versus 515 will affect the decision of a school admissions administrator? In other words, does the increase in the score have any practical significance?
- (d) Test the hypothesis at the $\alpha = 0.10$ level of significance with $n = 400$ students. Assume the same sample statistics. Is a sample mean of 519 significantly more than 515? What do you conclude about the impact of large samples on the hypothesis test?

34. Statistical Significance versus Practical Significance The manufacturer of a daily dietary supplement claims that its product will help people lose weight. The company obtains a random sample of 950 adult males aged 20 to 74 who take the supplement and finds their mean weight loss after 8 weeks to be 0.9 pound with standard deviation weight loss of 7.2 pounds.

- (a) State the null and alternative hypotheses.
- (b) Test the hypothesis at the $\alpha = 0.1$ level of significance. Is a mean weight loss of 0.9 pound significant?
- (c) Do you think that a mean weight loss of 0.9 pound is worth the expense and commitment of a daily dietary supplement? In other words, does the weight loss have any practical significance?
- (d) Test the hypothesis at the $\alpha = 0.1$ level of significance with $n = 40$ subjects. Assume the same sample statistics. Is a sample mean weight loss of 0.9 pound significantly more than 0 pound? What do you conclude about the impact of large samples on the hypothesis test?

DATA **35. Threaded Problem: Tornado** The data set "Tornadoes_2017" located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a) Compute the population mean length of a tornado in the United States in 2017.
- (b) The data file 10_3_35b represents the length of a random sample of 40 tornadoes in Texas in 2017. Draw a boxplot of the length data. Are there any outliers? Comment on the shape of the distribution.
- (c) Explain why a large sample size is necessary to use Student's t -distribution to test a hypothesis about a population mean.
- (d) Does the sample data suggest tornadoes have a different length in Texas?
- (e) The data file 10_3_35e represents an independent random sample of 40 tornadoes in Texas in 2017. Does this sample data suggest tornadoes have a different length in Texas?
- (f) Compare the results of parts (d) and (e). Explain why the conclusions may differ.

36. Accept versus Do Not Reject The mean IQ score of humans is 100. Suppose the director of Institutional Research at Joliet Junior College (JJC) obtains a simple random sample of 40 JJC students and finds the mean IQ is 103.4 with a standard deviation of 13.2.

- (a) Consider the hypotheses $H_0: \mu = 100$ versus $H_1: \mu > 100$. Explain what the director of Institutional Research is testing. Perform the test at the $\alpha = 0.05$ level of significance. Write a conclusion for the test.
- (b) Consider the hypotheses $H_0: \mu = 101$ versus $H_1: \mu > 101$. Explain what the director of Institutional Research is testing. Perform the test at the $\alpha = 0.05$ level of significance. Write a conclusion for the test.
- (c) Consider the hypotheses $H_0: \mu = 102$ versus $H_1: \mu > 102$. Explain what the director of Institutional Research is testing. Perform the test at the $\alpha = 0.05$ level of significance. Write a conclusion for the test.
- (d) Based on the results of parts (a)–(c), write a few sentences that explain the difference between “accepting” the statement in the null hypothesis versus “not rejecting” the statement in the null hypothesis.

37. Simulation Simulate drawing 100 simple random samples of size $n = 15$ from a population that is normally distributed with mean 100 and standard deviation 15.

- (a) Test the null hypothesis $H_0: \mu = 100$ versus $H_1: \mu \neq 100$ for each of the 100 simple random samples.
- (b) If we test this hypothesis at the $\alpha = 0.05$ level of significance, how many of the 100 samples would you expect to result in a Type I error?
- (c) Count the number of samples that lead to a rejection of the null hypothesis. Is it close to the expected value determined in part (b)?
- (d) Describe how we know that a rejection of the null hypothesis results in making a Type I error in this situation.

38. Simulation The *exponential probability distribution* can be used to model waiting time in line or the lifetime of electronic components. Its density function is skewed right. Suppose the wait time in a line can be modeled by the exponential distribution with $\mu = \sigma = 5$ minutes.

- (a) Simulate obtaining 100 simple random samples of size $n = 10$ from the population described. That is, simulate obtaining a simple random sample of 10 individuals waiting in a line where the wait time is expected to be 5 minutes.
- (b) Test the null hypothesis $H_0: \mu = 5$ versus the alternative $H_1: \mu \neq 5$ of the $\alpha = 0.05$ level of significance. For each of the 100 simulated simple random samples.
- (c) If we test this hypothesis at the $\alpha = 0.05$ level of significance, how many of the 100 samples would you expect to result in a Type I error?
- (d) Count the number of samples that lead to a rejection of the null hypothesis. Is it close to the expected value determined in part (c)? What might account for any discrepancies?

Retain Your Knowledge

39. Reading at Bedtime It is well-documented that watching TV, working on a computer, or any other activity involving artificial light can be harmful to sleep patterns. Researchers wanted to determine if the artificial light from e-Readers also disrupted sleep. In the study, 12 young adults were given either an iPad or printed book for four hours before bedtime. Then, they switched reading devices. Whether the individual received the iPad or book first was determined randomly. Bedtime was 10 P.M. and the time to fall asleep was measured each evening. It was found that participants took an average of 10 minutes longer to fall asleep after reading on an iPad. The P -value for the test was 0.009.

Source: Anne-Marie Chang, et. al. “Evening Use of Light-Emitting eReaders Negatively Affects Sleep, Circadian Timing, and Next-Morning Alertness” *PNAS* 2015 112(4) 1232–1277. doi:10.1073/pnas.1418490112

- (a) What is the research objective?
- (b) What is the response variable? It is quantitative or qualitative?
- (c) What is the treatment?
- (d) Is this a designed experiment or observational study? What type?
- (e) The null hypothesis for this test would be that there is no difference in time to fall asleep with an e-Reader and printed book. The alternative is that there is a difference. Interpret the P -value.

Explaining the Concepts

40. What’s the Problem? The head of institutional research at a university believes that the mean age of full-time students is declining. In 1995, the mean age of a full-time student was known to be 27.4 years. After looking at the enrollment records of all 4934 full-time students in the current semester, he found that the mean age was 27.1 years, with a standard deviation of 7.3 years. He conducted a hypothesis of $H_0: \mu = 27.4$ years versus $H_1: \mu < 27.4$ years and obtained a P -value of 0.0019. He concluded that the mean age of full-time students did decline. Is there anything wrong with his research?

41. The procedures for testing a hypothesis regarding a population mean are robust. What does this mean?

42. Explain the difference between *statistical significance* and *practical significance*.

43. Wanna Live Longer? Become a Chief Justice The life expectancy of a male during the course of the past 100 years is approximately 27,725 days. Go to Wikipedia.com and download the data that represent the life span of chief justices of Canada for those who have died. Conduct a test to determine whether the evidence suggests that chief justices of Canada live longer than the general population of males. Suggest a reason why the conclusion drawn may be flawed.

10.4 Putting It Together: Which Method Do I Use?

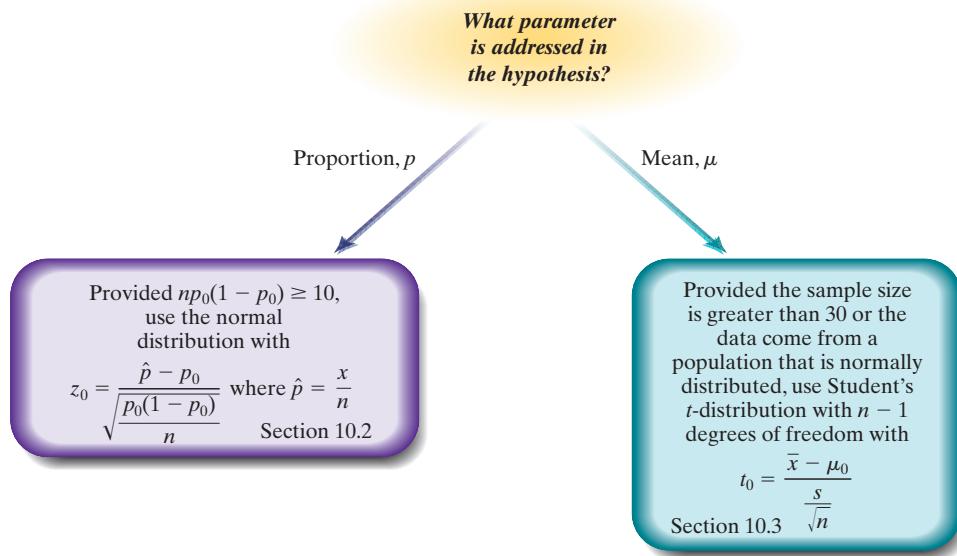


Objective ① Determine the appropriate hypothesis test to perform

1 Determine the Appropriate Hypothesis Test to Perform

Perhaps the most difficult aspect of testing hypotheses is determining which hypothesis test to conduct. To assist in the decision making, we present Figure 20, which shows which approach to take in testing hypotheses for the three parameters discussed in this chapter.

Figure 20



10.4 Assess Your Understanding

Skill Building

1. A simple random sample of size $n = 19$ is drawn from a population that is normally distributed. The sample mean is found to be 0.8, and the sample standard deviation is found to be 0.4. Test whether the population mean is less than 1.0 at the $\alpha = 0.01$ level of significance.
2. A simple random sample of size $n = 200$ individuals with a valid driver's license is asked if they drive an American-made automobile. Of the 200 individuals surveyed, 115 responded that they drive an American-made automobile. Determine if a majority of those with a valid driver's license drive an American-made automobile at the $\alpha = 0.05$ level of significance.
3. A simple random sample of size $n = 15$ is drawn from a population that is normally distributed. The sample mean is found to be 23.8, and the sample standard deviation is found to

be 6.3. Is the population mean different from 25 at the $\alpha = 0.01$ level of significance?

4. A simple random sample of size $n = 65$ is drawn from a population. The sample mean is found to be 583.1, and the sample standard deviation is found to be 114.9. Is the population mean different from 600 at the $\alpha = 0.1$ level of significance?
5. A simple random sample of size $n = 40$ is drawn from a population. The sample mean is found to be 108.5, and the sample standard deviation is found to be 17.9. Is the population mean greater than 100 at the $\alpha = 0.05$ level of significance?
6. A simple random sample of size $n = 320$ adults was asked their favorite ice cream flavor. Of the 320 individuals surveyed, 58 responded that they preferred mint chocolate chip. Do less than 25% of adults prefer mint chocolate chip ice cream? Use the $\alpha = 0.01$ level of significance.

Applying the Concepts

7. Smarter Kids? A psychologist obtains a random sample of 20 mothers in the first trimester of their pregnancy. The mothers are asked to play Mozart in the house at least 30 minutes each day until they give birth. After 5 years, the child is administered an IQ test. We know that IQs are normally distributed with a mean of 100. If the IQs of the 20 children in the study result in a sample mean of 104.2, and a sample standard deviation of 14.7, is there evidence that the children have higher IQs? Use the $\alpha = 0.05$ level of significance.

8. The Atomic Bomb In October 1945, the Gallup organization asked 1487 randomly sampled Americans, “Do you think we can develop a way to protect ourselves from atomic bombs in case other countries tried to use them against us?” with 788 responding yes. Did a majority of Americans feel the United States could develop a way to protect itself from atomic bombs in 1945? Use the $\alpha = 0.05$ level of significance.

9. Course Redesign Pass rates for Intermediate Algebra at a community college are 52.6%. In an effort to improve pass rates in the course, faculty of a community college develop a mastery-based learning model where course content is delivered in a lab through a computer program. The instructor serves as a learning mentor for the students. Of the 480 students who enroll in the mastery-based course, 267 pass.

- (a) What is the variable of interest in this study? What type of variable is it?
- (b) At the 0.01 level of significance, decide whether the sample evidence suggests the mastery-based learning model improved pass rates.
- (c) Explain why a 0.01 level of significance might be used to test this hypothesis.



10. Number of Credit Cards According to the Federal Reserve Bank of Boston, among individuals who had credit cards in 2014, the mean number of cards was 3.5. Treat the individuals who have credit cards in the SullivanStatsSurveyI as a random sample of credit card holders. Go to www.pearsonhighered.com/sullivanstats to obtain the data file SullivanStatsSurveyI using the file format of your choice for the version of the text you are using. The results of the survey are under the column “Number of cards.”

- (a) What is the variable of interest in this study? Is it qualitative or quantitative?
- (b) Do the results of the survey imply that the mean number of cards per individual is less than 3.5? Use the $\alpha = 0.05$ level of significance.

11. Gas Mileage The Environmental Protection Agency (EPA) states that a 2013 Kia Optima should get 28 miles per gallon, on average. The website www.fueleconomy.gov allows users to report the miles per gallon that they get on their vehicle. Treat the following data as a random sample of ten 2013 Kia Optima owners. The data represent the miles per gallon on their vehicle. Is there reason to believe that individuals are getting different gas mileage than the EPA states should be attained? Be sure to verify all conditions necessary to conduct the appropriate test.

30.2	30.3	19.4	26.7	17.6
19.9	23.3	20.2	16.7	27.5

Source: www.fueleconomy.gov

12. Sleepy? According to the National Sleep Foundation, children between the ages of 6 and 11 years should get 10 hours of sleep each night. In a survey of 56 parents of 6 to 11 year olds, it was found that the mean number of hours the children slept was 8.9 with a standard

deviation of 3.2. Does the sample data suggest that 6 to 11 year olds are sleeping less than the required amount of time each night? Use the 0.01 level of significance.

DATA 13. Text while Driving According to the research firm Toluna, the proportion of individuals who text while driving is 0.26. Suppose a random sample of 60 individuals are asked to disclose if they text while driving. Results of the survey are shown next, where 0 indicates no and 1 indicates yes. Do the data contradict the results of Toluna? Use the $\alpha = 0.05$ level of significance.

0	0	0	0	0	0	1	0	0	1
0	0	1	1	1	0	1	1	1	0
1	0	0	0	0	0	0	1	0	1
1	0	1	0	1	0	0	0	0	0
0	1	1	0	0	0	0	0	0	1
0	1	1	0	1	0	0	0	0	1

14. Student Loan Balances Student loan debt has reached record levels in the United States. In a random sample of 100 individuals who have student loan debt, it was found the mean debt was \$23,979 with a standard deviation of \$31,400. Data based on results from the Federal Reserve Bank of New York.

- (a) What do you believe is the shape of the distribution of student loan debt? Explain.
- (b) Use this information to estimate the mean student loan debt among all with such debt at the 95% level of confidence. Interpret this result.
- (c) What could be done to increase the precision of the estimate?

15. Political Decision Politicians often form their positions on various policies through polling. Suppose the U.S. Congress is considering passage of a tax increase to pay down the national debt and national polls suggest the general population is equally split on the matter. A congresswoman wants to poll her constituency on this controversial tax increase. To get a good sense as to how the citizens of her very populous district feel (well over 1 million registered voters), she decides to poll 8250 individuals within the district. Of those surveyed, 4205 indicated they are in favor of the tax increase. Given that politicians are generally leery of voting for tax increases, what level of significance would you recommend the congresswoman use in conducting this hypothesis test? Do the results of the survey represent statistically significant evidence a majority of the district favor the tax increase? What would you recommend to the congresswoman?

16. Quality Control Suppose the mean wait-time for a telephone reservation agent at a large airline is 43 seconds. A manager with the airline is concerned that business may be lost due to customers having to wait too long for an agent. To address this concern, the manager develops new airline reservation policies that are intended to reduce the amount of time an agent needs to spend with each customer. A random sample of 250 customers results in a sample mean wait-time of 42.3 seconds with a standard deviation of 4.2 seconds. Using an $\alpha = 0.05$ level of significance, do you believe the new policies were effective? Do you think the results have any practical significance?

17. Amazon Stock According to Crestmont Research, the stock market has been up in 53.7% of all trading days over the past 50 years. A stock is up if the price per share increases in value day-over-day. An investor wondered whether the proportion of days Amazon stock was up differed from the market. In a random sample of 200 trading days, she found Amazon stock was up 124 days. Does

Does this sample evidence suggest the movement of Amazon stock is different from the stock market? Use the $\alpha = 0.05$ level of significance. *Source: Yahoo!Finance.*

Retain Your Knowledge

- DATA 18. Ideal Number of Children** A survey from the Gallup organization asked, “What do you think is the ideal number of children for a family to have?” Go to www.pearsonhighered.com/sullivanstats to obtain the data file 10_4_18 using the file format of your choice for the version of the text you are using to get the survey results.
- Draw a dot plot of the data. Comment on the shape of the distribution.
 - What is the mode ideal number of children?
 - Determine the mean, median, standard deviation, and interquartile range ideal number of children. Round your answers to the nearest thousandth.
 - Explain why a large sample size is needed to perform any inference regarding this population.
 - In May 1997, the ideal number of children was considered to be 2.64. Do the results of this poll indicate that people’s beliefs as to the ideal number of children have changed? Use the 0.05 level of significance.

19. Confidence Intervals Suppose you wish to determine if the mean IQ of students on your campus is different from the mean IQ in the general population, 100. To conduct this study, you obtain a simple random sample of 50 students on your campus, administer an IQ test, and record the results. The mean IQ of the sample of 50 students is found to be 107.3 with a standard deviation of 13.6.

- Conduct a hypothesis test (preferably using technology) $H_0: \mu = \mu_0$ versus $H_1: \mu \neq \mu_0$ for $\mu_0 = 103, 104, 105, 106, 107, 108, 109, 110, 111, 112$ at the $\alpha = 0.05$ level of significance. For which values of μ_0 do you not reject the null hypothesis?
- Construct a 95% confidence interval for the mean IQ of students on your campus. What might you conclude about how the lower and upper bounds of a confidence interval relate to the values for which the null hypothesis is rejected?

- Suppose you changed the level of significance in conducting the hypothesis test to $\alpha = 0.01$. What would happen to the range of values of μ_0 for which the null hypothesis is not rejected? Why does this make sense?

In Problems 20–25, decide whether the problem requires a confidence interval or hypothesis test, and determine the variable of interest. For any problem requiring a confidence interval, state whether the confidence interval will be for a population proportion or population mean. For any problem requiring a hypothesis test, write the null and alternative hypothesis.

- An investigator with the Food and Drug Administration wanted to determine whether a typical bag of potato chips contained less than the 16 ounces claimed by the manufacturer.
- A researcher wanted to estimate the average length of time mothers who gave birth via Caesarean section spent in a hospital after delivery of the baby.
- An official with the Internal Revenue Service wished to estimate the proportion of high-income (greater than \$100,000 annually) earners who under-reported their net income (and, therefore, their tax liability).
- According to the Pew Research Center, 55% of adult Americans support the death penalty for those convicted of murder. A social scientist wondered whether a higher proportion of adult Americans with at least a bachelor’s degree support the death penalty for those convicted of murder.
- In 2014, of the 37 million borrowers who have outstanding student loan balances, 14% have at least one past due student loan account. A researcher with the United States Department of Education believes this proportion has increased since then.
Source: American Student Assistance.
- Researchers measured regular testosterone levels in a random sample of athletes and then measured testosterone levels prior to an athletic event. They wanted to know whether testosterone levels increase prior to athletic events.



Chapter 10 Review

Summary

In this chapter, we discussed hypothesis testing. Hypothesis testing is the second type of inferential statistics (recall the other type of inference is estimation via confidence intervals).

In hypothesis testing, a statement is made regarding a population parameter, which leads to a null, H_0 , and alternative hypothesis, H_1 . The null hypothesis is a statement of “no change” or “no difference.” We build a probability model under the assumption the statement in the null hypothesis is true, and we use sample data to decide whether to reject or not reject the statement in the null hypothesis. There are three options for structuring the null and alternative hypotheses on a single parameter.

Determining the Null and Alternative Hypotheses

- Equal hypothesis versus not equal hypothesis (**two-tailed test**)

$$H_0: \text{parameter} = \text{some value}$$

$$H_1: \text{parameter} \neq \text{some value}$$

- Equal hypothesis versus less than hypothesis (**left-tailed test**)

$$H_0: \text{parameter} = \text{some value}$$

$$H_1: \text{parameter} < \text{some value}$$

- Equal hypothesis versus greater than hypothesis (**right-tailed test**)

$$H_0: \text{parameter} = \text{some value}$$

$$H_1: \text{parameter} > \text{some value}$$

In performing a hypothesis test, there is always the possibility of making a Type I error (rejecting the null hypothesis when it is true) or making a Type II error (not rejecting the null hypothesis when it is false). The probability of making a Type I error is equal to the level of significance, α , of the test.

The first test introduced was hypothesis tests about a population proportion p . Provided certain requirements were satisfied (simple random sample or randomized experiment, independence, and a large sample size), we were able to use the normal model to assess statements made in the null hypothesis.

Then we discussed hypothesis testing on the mean. Here we also required a simple random sample (or randomized experiment) and independence, but we also required that either the sample come from a population that is normally distributed with no outliers or a large ($n \geq 30$) sample size. When dealing with small sample sizes, we tested the normality requirement with a normal probability plot and the outlier requirement with a

boxplot. These requirements allow us to use Student's t -distribution to test the hypothesis.

Both hypothesis tests were performed using the classical method and the P -value approach. The P -value approach to testing hypotheses has appeal because the rejection rule is always to reject the null hypothesis if the P -value is less than the level of significance, α .

In testing hypotheses, it is important to remember that we never *accept* the null hypothesis, because, without having access to the entire population, we don't know the exact value of the parameter stated in the null hypothesis. Rather, we say that we do not *reject* the null hypothesis.

Remember, statistical significance refers to the idea that the observed results are unlikely to occur if the statement in the null hypothesis is true. Practical significance, on the other hand, refers to the idea that, while small differences between the statistic and parameter stated in the null hypothesis are statistically significant, the difference may not be large enough to cause concern or be considered important.

Vocabulary

Hypothesis (p. 436)

Hypothesis testing (p. 436)

Null hypothesis (p. 436)

Alternative hypothesis (p. 436)

Two-tailed test (p. 436)

Left-tailed test (p. 436)

Right-tailed test (p. 436)

One-tailed test (p. 437)

Type I error (p. 438)

Type II error (p. 438)

Level of significance (p. 440)

Statistically significant (p. 444)

P -value (p. 446)

Test statistic (pp. 447, 459)

Practical significance (p. 463)

Formulas

Test Statistics

- $z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$ follows the standard normal distribution if $np_0(1 - p_0) \geq 10$ and $n \leq 0.05N$

- $t_0 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$ follows Student's t -distribution with $n - 1$ degrees of freedom if the population from which the sample was drawn is normal with no outliers, or if the sample size is large ($n \geq 30$).

Objectives

Section	You should be able to ...	Example(s)	Review Exercises
10.1	1 Determine the null and alternative hypotheses (p. 435)	2	1(a), 2(a)
	2 Explain Type I and Type II errors (p. 438)	3	1(b), 1(c), 2(b), 2(c), 3, 4, 11(e)
	3 State conclusions to hypothesis tests (p. 440)	4	1(d), 1(e), 2(d), 2(e)
10.2	1 Explain the logic of hypothesis testing (p. 443)	pp. 443–447	19, 21, 22
	2 Test hypotheses about a population proportion (p. 447)	1, 2, 3	9, 10, 15, 16
	3 Test hypotheses about a population proportion using the binomial probability distribution (p. 452)	4	17
10.3	1 Test hypotheses about a mean (p. 458)	1, 2	5, 6, 11, 12, 13, 14, 18
	2 Understand the difference between statistical significance and practical significance (p. 463)	3	16, 18
10.4	1 Determine the appropriate hypothesis test to perform (p. 470)		5–18

Review Exercises

For Problems 1 and 2, (a) determine the null and alternative hypotheses, (b) explain what it would mean to make a Type I error, (c) explain what it would mean to make a Type II error, (d) state the conclusion that would be reached if the null hypothesis is not rejected, and (e) state the conclusion that would be reached if the null hypothesis is rejected.

1. Credit-Card Debt According to creditcard.com, the mean outstanding credit-card debt of college undergraduates was \$3173 in 2010. A researcher believes that this amount has decreased since then.

2. More Credit-Card Debt Among all credit cards issued, the proportion of cards that result in default was 0.13 in 2010. A credit analyst with Visa believes this proportion is different today.

3. A test is conducted at the $\alpha = 0.05$ level of significance. What is the probability of a Type I error?

4. β is computed to be 0.113. What is the probability of a Type II error?

5. To test $H_0: \mu = 100$ versus $H_1: \mu > 100$, a simple random sample of size $n = 35$ is obtained from an unknown distribution. The sample mean is 104.3 and the sample standard deviation is 12.4.

- (a) To use the t -distribution, why must the sample size be large?
- (b) Use the classical or P -value approach to decide whether to reject the statement in the null hypothesis at the $\alpha = 0.05$ level of significance.

6. To test $H_0: \mu = 50$ versus $H_1: \mu \neq 50$, a simple random sample of size $n = 15$ is obtained from a population that is normally distributed. The sample mean is 48.1 and the sample standard deviation is 4.1.

- (a) Why must it be the case that the population from which the sample was drawn is normally distributed?
- (b) Use the classical or P -value approach to decide whether to reject the statement in the null hypothesis at the $\alpha = 0.05$ level of significance.

In Problems 7 and 8, test the hypothesis at the $\alpha = 0.05$ level of significance, using (a) the classical approach and (b) the P -value approach. Be sure to verify the requirements of the test.

7. $H_0: p = 0.6$ versus $H_1: p > 0.6$

$n = 250$; $x = 165$

8. $H_0: p = 0.35$ versus $H_1: p \neq 0.35$

$n = 420$; $x = 138$

9. Sneeze According to work done by Nick Wilson of Otago University Wellington, the proportion of individuals who cover their mouth when sneezing is 0.733. As part of a school project, Mary decides to confirm this by observing 100 randomly selected individuals sneeze and finds that 78 covered their mouth when sneezing.

- (a) What are the null and alternative hypotheses for Mary's project?
- (b) Verify the requirements that allow use of the normal model to test the hypothesis are satisfied.
- (c) Does the sample evidence contradict Professor Wilson's findings?

10. Emergency Room The proportion of patients who visit the emergency room (ER) and die within the year is 0.05.

Source: SuperFreakonomics. Suppose a hospital administrator is concerned that his ER has a higher proportion of patients who die within the year. In a random sample of 250 patients who have visited the ER in the past year, 17 have died. Should the administrator be concerned?

11. Linear Rotary Bearing A linear rotary bearing is designed so that the distance between the retaining rings is 0.875 inch. The quality-control manager suspects that the manufacturing process needs to be recalibrated because the mean distance between the retaining rings is greater than 0.875 inch. In a random sample of 36 bearings, he finds the sample mean distance between the retaining rings is 0.876 inch with standard deviation 0.005 inch.

- (a) Are the requirements for conducting a hypothesis test satisfied?
- (b) State the null and alternative hypotheses.
- (c) The quality-control manager decides to use an $\alpha = 0.01$ level of significance. Why do you think this level of significance was chosen?
- (d) Does the evidence suggest the machine be recalibrated?
- (e) What does it mean for the quality-control engineer to make a Type I error? A Type II error?

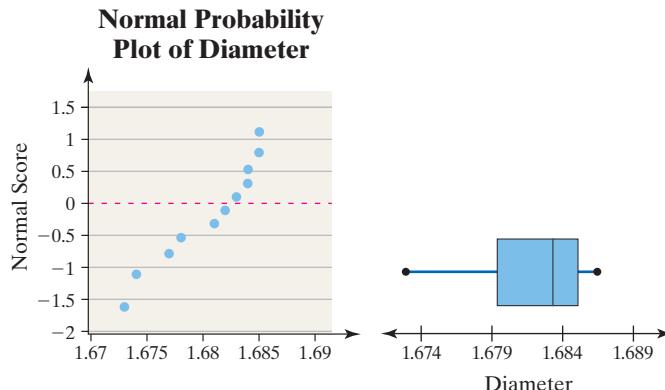
12. Normal Temperature Carl Reinhold August Wunderlich said that the mean temperature of humans is 98.6°F. Researchers Philip Mackowiak, Steven Wasserman, and Myron Levine [JAMA, Sept. 23–30 1992; 268(12):1578–80] thought that the mean temperature of humans is less than 98.6°F. They measured the temperature of 148 healthy adults one to four times daily for 3 days, obtaining 700 measurements. The sample data resulted in a sample mean of 98.2°F and a sample standard deviation of 0.7°F. Test whether the mean temperature of humans is less than 98.6°F at the $\alpha = 0.01$ level of significance.

13. Conforming Golf Balls The U.S. Golf Association (USGA) requires that golf balls have a diameter that is 1.68 inches. To determine if Maxfli XS golf balls conform to USGA standards, a random sample of Maxfli XS golf balls was selected. Their diameters are shown in the table.

1.683	1.684	1.677	1.684	1.681	1.673
1.685	1.685	1.678	1.682	1.686	1.674

Source: Michael McCraith, Joliet Junior College.

- (a) Because the sample size is small, the engineer must verify that the diameter is normally distributed and the sample does not contain any outliers. The normal probability plot and boxplot are shown. The correlation between ball diameter and expected z -scores is 0.951. Are the conditions for testing the hypothesis satisfied?



- (b) Construct a 95% confidence interval to judge whether the golf balls conform to USGA standards. Be sure to state the null and alternative hypotheses and write a conclusion.

14. Studying Enough? College mathematics instructors suggest that students spend 2 hours outside class studying for every hour in class. So, for a 4-credit-hour math class, students should spend at least 8 hours (480 minutes) studying each week. The given data, from Michael Sullivan's College Algebra class, represent the time spent on task recorded in MyLab Math (in minutes) for randomly selected students during the third week of the semester. Determine if the evidence suggests students may not, in fact, be following the advice. That is, does the evidence suggest students are studying less than 480 minutes each week? Use $\alpha = 0.05$ level of significance. **Note:** A normal probability plot and boxplot indicate that the data come from a population that is normally distributed with no outliers.

504	267	220	322	538	542
428	481	413	302	602	

Source: MyLab Math

15. Sleeping Patterns of Pregnant Women A random sample of 150 pregnant women indicated that 81 napped at least twice per week. Do a majority of pregnant women nap at least twice a week? Use the $\alpha = 0.05$ level of significance.

Source: National Sleep Foundation.

16. Grim Report Throughout the country, the proportion of first-time, first-year community college students who return for their second year of studies is 0.52 according to the Community College Survey of Student Engagement. Suppose a community college institutes new policies geared toward increasing student retention. The first year of this policy there were 2843 first-time, first-year students of which 1516 returned for their second year of studies. Treat these students as a random sample of all first-time, first-year students. Do a statistically significant higher proportion of students return for their second year under this policy? Use the $\alpha = 0.1$ level of significance. Would you say the results are practically significant? In other words, do you believe the results are significant enough that other community colleges might consider emulating these same policies?

17. Teen Prayer In 1995, 40% of adolescents stated they prayed daily. A researcher wants to know whether this percentage has risen since then. He surveys 40 adolescents and finds that 18 pray on a daily basis. Is this evidence that the proportion of adolescents who pray daily has increased at the $\alpha = 0.05$ level of significance?

18. A New Teaching Method A large university has a college algebra enrollment of 5000 students each semester. Because of space limitations, the university decides to offer its college algebra courses in a self-study format in which students learn independently, but have access to tutors and other help in a lab setting. Historically, students in traditional college algebra scored 73.2 points on the final exam, and the coordinator of this course is concerned that test scores are going to decrease in the new format. At the end of the first semester using the new delivery system, 3851 students took the final exam and had a mean score of 72.8 and a standard deviation of 12.3. Treating these students as a simple random sample of all students, determine whether or not the scores decreased significantly at the $\alpha = 0.05$ level of significance. Do you think that the decrease in scores has any practical significance?

19. Explain the difference between “accepting” and “not rejecting” a null hypothesis.

20. According to the American Time Use Survey, the mean number of hours each day Americans, aged 15 and older, spend eating and drinking is 1.22. A researcher wanted to know if Americans, aged 15 to 19, spent less time eating and drinking. After surveying 50 Americans, aged 15 to 19, and running the appropriate hypotheses test, she obtained a P -value of 0.0329. State the null and alternative hypothesis this researcher used and interpret the P -value.

21. Explain the procedure for testing a hypothesis using the Classical Approach. What is the criterion for judging whether to reject the null hypothesis?

22. Explain the procedure for testing a hypothesis using the P -value Approach. What is the criterion for judging whether to reject the null hypothesis?



Chapter Test

1. According to the American Time Use Survey, adult Americans spent 42.6 minutes per day on phone calls and answering or writing email in 2006.

- (a) Suppose that we want to judge whether the amount of daily time spent on phone calls and answering or writing email has increased. Write the appropriate null and alternative hypotheses.
- (b) The sample data indicated that the null hypothesis should be rejected. Write a conclusion.
- (c) Explain what it would mean if we made a Type I error in conducting the test from part (a).
- (d) Explain what it would mean if we made a Type II error in conducting the test from part (a).

2. The trade magazine *QSR* routinely examines fast-food drive-thru service times. Their recent research indicates that the mean time a car spends in a McDonald's drive-thru is 167.1 seconds. A McDonald's manager in Salt Lake City feels that she has instituted a drive-thru policy that results in lower drive-thru service times. A random sample of 70 cars results in a mean service time of 163.9 seconds, with a standard deviation of 15.3 seconds. Determine whether the policy is effective in reducing drive-thru service times.

- (a) State the null and alternative hypotheses.
- (b) Because the cost of instituting the policy is quite high, the quality-control researcher at McDonald's chooses

to test the hypothesis using an $\alpha = 0.01$ level of significance. Why is this a good idea?

- (c) Conduct the appropriate test to determine if the policy is effective.

3. “Did you get your 8 hours of sleep last night?” is a common question. In a recent survey of 151 postpartum women, the folks at the National Sleep Foundation found that the mean sleep time was 7.8 hours, with a standard deviation of 1.4 hours. Does the evidence suggest that postpartum women do not get enough sleep? Use $\alpha = 0.05$ level of significance.

4. The outside diameter of a manufactured part must be 1.3825 inches, according to customer specifications. The data shown represent a random sample of ten parts. Use a 95% confidence interval to judge whether the part has been manufactured to specifications.

Note: A normal probability plot and boxplot indicate that the data come from a population that is normally distributed with no outliers.

1.3821	1.3830	1.3823	1.3829	1.3830
1.3829	1.3826	1.3825	1.3823	1.3824

Source: Dennis Johnson, student at Joliet Junior College.

5. In many parliamentary procedures, a supermajority is defined as an excess of 60% of voting members. In a poll conducted by the Gallup organization on May 10, 1939, 1561 adult Americans were asked, “Do you think the United States will have to fight Japan within your lifetime?” Of the

1561 respondents, 954 said no. Does this constitute sufficient evidence that a supermajority of Americans did not feel the United States would have to fight Japan within their lifetimes at the $\alpha = 0.05$ level of significance?

6. A Zone diet is one with a 40%–30%–30% distribution of carbohydrate, protein, and fat, respectively, and is based on the book *Enter the Zone*. In a study conducted by researchers Christopher Gardner and associates, 79 subjects were administered the Zone diet. After 12 months, the mean weight loss was 1.6 kg, with a standard deviation of 5.4 kg. Do these results suggest that the weight loss was statistically significantly greater than zero at the 0.05 level of significance? Do you believe the weight loss has any practical significance? Why?

Source: Christopher D. Gardner, Alexandre Kiazand, and Sofiya Alhassan, et al. “Comparison of the Atkins, Zone, Ornish, and LEARN Diets for Change in Weight and Related Risk Factors among Overweight Premenopausal Women. The A to Z Weight Loss Study: A Randomized Trial.” *Journal of the American Medical Association*, March 2007.

7. According to the Pew Research Center, the proportion of the American population who use only a cellular telephone (no landline) is 0.37. Jason conducts a survey of thirty 20- to 24-year-olds who live on their own and finds that 16 do not have a landline to their home. Does this provide sufficient evidence to conclude that the proportion of 20- to 24-year-olds who live on their own and don’t have a landline is greater than 0.37? Use an $\alpha = 0.10$ level of significance.

Making an Informed Decision

Selecting a Mutual Fund

Suppose you have just received a \$1000 bonus at your job. Rather than waste the money on frivolous items, you decide to invest the money so that you can apply it toward the purchase of a home one day. Many investment options are available. Your family and friends, who have some experience in investing, recommend mutual funds. Which mutual funds should you choose?

Thousands of mutual funds are available to invest in. According to Wikipedia, “a **mutual fund** is a professionally managed type of collective investment scheme that pools money from many investors and invests typically in investment securities (stocks, bonds, and so on).”

- (a) Research the various classifications (Growth, Value, and so on) of mutual funds. Decide on a particular mutual fund classification.
- (b) Go to www.morningstar.com. Morningstar is a mutual fund rating agency that ranks mutual funds according to a

star rating. The 5-star rating system divides the mutual funds into percentile classes. A mutual fund with a five-star rating is in the top 20% of all mutual funds in the category. Choose a mutual fund that has at least a four-star rating and has been around for at least 5 years.



- (c) Research the historical monthly rates of return of the mutual fund you selected in part (b). You will use two criteria in selecting a mutual fund. First, the mean rate of return of the fund over the past 48 months must exceed 7%. Second, the proportion of the past 48 months where the rate of return is positive must exceed 0.7. Treat the selected months as a random sample of rates of return. Conduct the appropriate tests to see if the mutual fund you selected meets your criteria.

11

Inference on Two Population Parameters

Outline

- 11.1** Inference about Two Population Proportions
- 11.2** Inference about Two Means: Dependent Samples
- 11.3** Inference about Two Means: Independent Samples
- 11.4** Putting It Together: Which Method Do I Use?
- 11.5** Inference about Two Population Standard Deviations (eText only)

Making an Informed Decision



You have decided to purchase a car. A couple areas of concern for you are:

1. What kind of gas mileage does the car get?
2. Will the car hold its value?

See the Decisions Project on page 522.

Putting It Together

In Chapters 9 and 10, we discussed inference regarding a single population parameter. The inferential methods presented in those chapters will be modified slightly in this chapter so that we can compare two population parameters.

The first section presents inferential methods for comparing two population proportions. That is, inference when there are two levels of treatment and the response variable is qualitative with two possible outcomes (success or failure). The first order of business is to decide whether the data are obtained from an independent or dependent sample—simply put, we determine if the observations in one sample are somehow related to the observations in the other. We then discuss methods for comparing two proportions from independent samples. Methods for comparing proportions from dependent samples are covered in Section B.2.

Section 11.2 presents inferential methods used to handle dependent samples when the response variable is quantitative. For example, we might want to know whether the reaction time in an individual's dominant hand is different from the reaction time in the nondominant hand.

Section 11.3 presents inferential methods used to handle independent samples when there are two levels of treatment and the response variable is quantitative. For example, we might randomly divide 100 volunteers who have a common cold into two groups. The control group would receive a placebo and the experimental group would receive a specific amount of some experimental drug. The response variable might be time until the cold symptoms go away.

We wrap up the chapter with a “Putting It Together” section. One of the more difficult aspects of inference is determining which inferential method to use. This section helps develop this skill.

11.1 Inference about Two Population Proportions



Preparing for This Section Before getting started, review the following:

- Completely randomized design (Section 1.6, pp. 48–49)
- Matched-pairs design (Section 1.6, pp. 49–50)
- Estimating a population proportion (Section 9.1, pp. 396–404)
- Hypothesis tests about a population proportion (Section 10.2, pp. 443–452)
- Statistical versus practical significance (Section 10.3, pp. 463–464)

Objectives

- ① Distinguish between independent and dependent sampling
- ② Test hypotheses regarding two proportions from independent samples
- ③ Construct and interpret confidence intervals for the difference between two population proportions
- ④ Determine the sample size necessary for estimating the difference between two population proportions

1 Distinguish between Independent and Dependent Sampling

Let's consider two scenarios:

Scenario 1: Among competing acne medications, does one perform better than the other? To answer this question, researchers applied Medication A to one part of the subject's face and Medication B to a different part of the subject's face to determine the proportion of subjects whose acne cleared up for each medication. The part of the face that received Medication A was randomly determined.

Scenario 2: Do individuals who make fast-food purchases with a credit card tend to spend more than those who pay with cash? To answer this question, a marketing manager randomly selects 30 credit-card receipts and 30 cash receipts to determine if the credit-card receipts have a significantly higher dollar amount, on average.

Is there a difference in the approach taken to select the individuals in each study? Yes! In scenario 1, once an individual is selected, one part of his or her face is "matched-up" with a second part of the face. In scenario 2, the receipts selected from the credit-card group have nothing at all to do with the receipts selected from the cash group.

Definitions

A sampling method is **independent** when an individual selected for one sample does not dictate which individual is to be in a second sample. A sampling method is **dependent** when an individual selected to be in one sample is used to determine the individual in the second sample. Dependent samples are often referred to as **matched-pairs** samples. It is possible for an individual to be matched against him- or herself.

So, the sampling method in scenario 1 is dependent (or a matched-pairs sample), while the sampling method in scenario 2 is independent.

EXAMPLE 1

Distinguishing between Independent and Dependent Sampling

Problem Decide whether the sampling method is independent or dependent. Then determine whether the response variable is qualitative or quantitative.

- (a) Joliet Junior College decided to implement a course redesign of its developmental mathematics program. Students either enrolled in a traditional lecture format

course or a lab-based format in which lectures and homework are done using video and the course management system MyLabMath. There were 1200 students enrolled in the traditional lecture format and 300 enrolled in the lab-based format. Once the course ended, the researchers determined whether the student passed the course (with an A, B, or C), or not. The goal of the study was to determine whether the proportion of students who passed the lab-based format exceeded that of the lecture format.

- (b) Are products purchased on Amazon less expensive than those purchased online at Walmart? To answer this question, researchers randomly identified 20 products sold at both stores and determined the selling price at Amazon and the online Walmart store to determine if there was a significant difference in the price of the goods.

Approach Determine whether the individuals in one group were used to determine the individuals in the other group. If so, the sampling method is dependent. If not, the sampling method is independent. Finally, consider the response variable in the study. Is it qualitative with two outcomes? If so, inferential methods based on proportions are appropriate. Is it quantitative? If so, inferential methods based on means may be appropriate.

Solution

- (a) The sampling method is independent because the individuals in the lecture format are not related to the individuals in the lab-based format. The response variable is whether the student passed the course or not. Because there are two outcomes, pass or do not pass, the researchers can compare the proportion of students passing the lecture course to those passing the lab-based course.
- (b) The sampling method is dependent because once a product is selected at Amazon, the same product is selected at the online Walmart store. The response variable is price, which is quantitative.

NW Now Work Problem 3



② Test Hypotheses Regarding Two Proportions from Independent Samples

In Sections 9.1 and 10.2, we discussed inference regarding a single population proportion. We now discuss inference for comparing two proportions. We begin with inference comparing two proportions from independent samples. We will discuss inference comparing two proportions from dependent samples in Chapter 12.

For example, in clinical trials of the drug Nasonex, a drug that is meant to relieve allergy symptoms, 26% of patients receiving 200 micrograms (μg) of Nasonex reported a headache as a side effect, while 22% of patients receiving a placebo reported a headache as a side effect. Researchers want to determine whether the proportion of patients receiving Nasonex and complaining of headaches is significantly higher than the proportion of patients receiving the placebo and complaining of headaches.

To conduct inference about two population proportions from independent samples, we must first determine the sampling distribution of the difference of two proportions. Recall that the point estimate of a population proportion, p , is given by $\hat{p} = \frac{x}{n}$, where x is the number of the n individuals in the sample that have a specific characteristic. In addition, recall that the sampling distribution of \hat{p} is approximately normal with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, provided that $np(1-p) \geq 10$ and $n \leq 0.05N$, so $Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ is approximately normal with mean 0 and standard deviation 1. Using this information along with the idea of independent sampling from two populations, we obtain the sampling distribution of the difference between two proportions.

Sampling Distribution of the Difference between Two Proportions (Independent Sample)

Suppose a simple random sample of size n_1 is taken from a population where x_1 of the individuals have a specified characteristic, and a simple random sample of size n_2 is independently taken from a different population where x_2 of the individuals have a specified characteristic. The sampling distribution of $\hat{p}_1 - \hat{p}_2$, where $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$, is approximately normal, with mean $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$ and standard deviation $\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$, provided that $n_1\hat{p}_1(1-\hat{p}_1) \geq 10$ and $n_2\hat{p}_2(1-\hat{p}_2) \geq 10$ and each sample size is no more than 5% of the population size. The standardized version of $\hat{p}_1 - \hat{p}_2$ is then written as

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

which has an approximate standard normal distribution.

Now that we know the approximate sampling distribution of the difference of two sample proportions, we can introduce a procedure that can be used to test hypotheses regarding two population proportions. We first consider the test statistic. Following the discussion for testing hypotheses on a single proportion, it seems reasonable that the test statistic for the difference of two population proportions would be

$$z_0 = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \quad (1)$$

When comparing two population proportions, the null hypothesis is a statement of “no difference” (as always), so $H_0: p_1 = p_2$. Because the null hypothesis is assumed to be true, the test assumes that $p_1 = p_2$, or $p_1 - p_2 = 0$. Because we assume that both p_1 and p_2 equal p , where p is the common population proportion, substitute p into Equation (1), and obtain

$$z_0 = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2 - 0}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (2)$$

We need a point estimate of p because it is unknown. The best point estimate of p is called the **pooled estimate of p** , denoted \hat{p} , where

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Substituting the pooled estimate of p into Equation (2), we obtain

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sigma_{\hat{p}_1 - \hat{p}_2}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

This test statistic will be used to test hypotheses regarding two population proportions.

Hypothesis Test Regarding the Difference between Two Population Proportions

To test hypotheses regarding two population proportions, p_1 and p_2 , use the steps that follow, provided that

- the samples are independently obtained using simple random sampling or through a completely randomized experiment with two levels of treatment,
- $n_1\hat{p}_1(1-\hat{p}_1) \geq 10$ and $n_2\hat{p}_2(1-\hat{p}_2) \geq 10$, and
- $n_1 \leq 0.05N_1$ and $n_2 \leq 0.05N_2$ (the sample size is no more than 5% of the population size); this requirement ensures the independence necessary for a binomial experiment.

Step 1 Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways:

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: p_1 = p_2$	$H_0: p_1 = p_2$	$H_0: p_1 = p_2$
$H_1: p_1 \neq p_2$	$H_1: p_1 < p_2$	$H_1: p_1 > p_2$

Note: p_1 is the population proportion for population 1, and p_2 is the population proportion for population 2.

Step 2 Select a level of significance, α , depending on the seriousness of making a Type I error.

Classical Approach

Step 3 Compute the test statistic

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ where } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Use Table V to determine the critical value.

	Two-Tailed	Left-Tailed	Right-Tailed
Critical value	$-z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$	$-z_\alpha$	z_α
Critical region			

Step 4 Compare the critical value to the test statistic.

Two-Tailed	Left-Tailed	Right-Tailed
If $z_0 < -z_{\frac{\alpha}{2}}$ or $z_0 > z_{\frac{\alpha}{2}}$, reject the null hypothesis.	If $z_0 < -z_\alpha$, reject the null hypothesis.	If $z_0 > z_\alpha$, reject the null hypothesis.

Step 5 State the conclusion.

If the model requirements are not satisfied, then randomization methods as discussed in Section 11.1A should be performed.

EXAMPLE 2 Testing a Hypothesis Regarding Two Population Proportions

Problem In clinical trials of Nasonex, 3774 adult and adolescent allergy patients (patients 12 years and older) were randomly divided into two groups. The patients in group 1 (experimental group) received 200 µg of Nasonex, while the patients in group 2 (control group) received a placebo. Of the 2103 patients in the experimental group, 547 reported headaches as a side effect. Of the 1671 patients in the control group, 368 reported headaches as a side effect. Is there evidence to conclude that the proportion of Nasonex users who experienced headaches as a side effect is greater than the proportion in the control group at the $\alpha = 0.05$ level of significance?

Approach Note that this is a completely randomized design. The response variable is whether or not the patient reports a headache. The treatments are the drugs: Nasonex or a placebo. The subjects are the 3774 adult and adolescent allergy patients.

(continued)

We want to determine if the evidence suggests that the proportion of patients who report a headache and are receiving Nasonex is greater than the proportion of patients who report a headache and are receiving a placebo. We will call the group that received Nasonex sample 1 and the group that received a placebo sample 2.

Verify the requirements to perform the hypothesis test. That is, the sample must be a simple random sample or the result of a randomized experiment and $n_1 \hat{p}_1(1 - \hat{p}_1) \geq 10$ and $n_2 \hat{p}_2(1 - \hat{p}_2) \geq 10$. In addition, the sample size cannot be more than 5% of the population size. Then follow Steps 1 through 5.

Solution First we verify that the requirements are satisfied.

1. The samples are independent because the subjects were randomly assigned to the treatment.

2. We have $x_1 = 547$, $n_1 = 2103$, $x_2 = 368$, and $n_2 = 1671$, so

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{547}{2103} = 0.260 \text{ and } \hat{p}_2 = \frac{x_2}{n_2} = \frac{368}{1671} = 0.220. \text{ Therefore,}$$

$$n_1 \hat{p}_1(1 - \hat{p}_1) = 2103(0.260)(1 - 0.260) = 404.6 \geq 10$$

$$n_2 \hat{p}_2(1 - \hat{p}_2) = 1671(0.220)(1 - 0.220) = 286.7 \geq 10$$

3. More than 10 million Americans 12 years old or older are allergy sufferers, so both sample sizes are less than 5% of the population size.

All three requirements are satisfied, so we proceed to follow Steps 1 through 5.

Step 1 Is the proportion of patients taking Nasonex who experience a headache greater than the proportion of patients taking the placebo who experience a headache? Letting p_1 represent the population proportion of patients taking Nasonex who experience a headache and p_2 represent the population proportion of patients taking the placebo who experience a headache, we want to know if $p_1 > p_2$. This is a right-tailed hypothesis with

$$H_0: p_1 = p_2 \text{ versus } H_1: p_1 > p_2$$

or, equivalently,

$$H_0: p_1 - p_2 = 0 \text{ versus } H_1: p_1 - p_2 > 0$$

Step 2 The level of significance is $\alpha = 0.05$.

Classical Approach

Step 3 From verifying requirement 2, we have $\hat{p}_1 = 0.260$ and $\hat{p}_2 = 0.220$. To find the test statistic, first compute the pooled estimate of p :

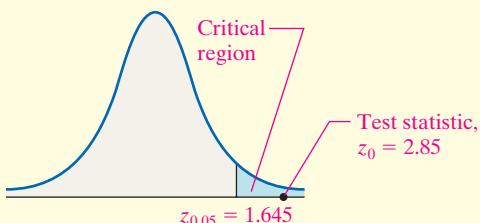
$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{547 + 368}{2103 + 1671} = 0.242$$

The test statistic is

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0.260 - 0.220}{\sqrt{0.242(1 - 0.242)} \sqrt{\frac{1}{2103} + \frac{1}{1671}}} = 2.85$$

Because this is a right-tailed test, we determine the critical value at the $\alpha = 0.05$ level of significance to be $z_{0.05} = 1.645$. The critical region is shown in Figure 1.

Figure 1



P-Value Approach

By Hand Step 3 From verifying requirement 2, we have $\hat{p}_1 = 0.260$ and $\hat{p}_2 = 0.220$. To find the test statistic, first compute the pooled estimate of p :

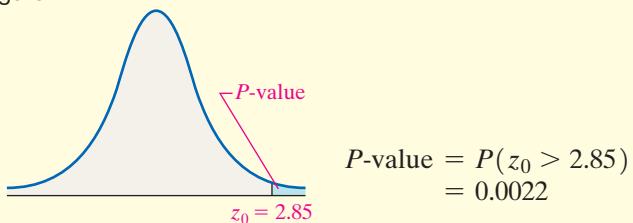
$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{547 + 368}{2103 + 1671} = 0.242$$

The test statistic is

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0.260 - 0.220}{\sqrt{0.242(1 - 0.242)} \sqrt{\frac{1}{2103} + \frac{1}{1671}}} = 2.85$$

Because this is a right-tailed test, the P -value is the area under the standard normal distribution to the right of the test statistic, $z_0 = 2.85$, as shown in Figure 2.

Figure 2



$$P\text{-value} = P(z_0 > 2.85) = 0.0022$$

Step 4 The test statistic is $z_0 = 2.85$. We label this point in Figure 1. Because the test statistic is greater than the critical value ($2.85 > 1.645$), we reject the null hypothesis.

Technology Step 3 Using Minitab, we find the P -value is 0.002. See Figure 3.

Figure 3

Test

Null hypothesis	$H_0: p_1 - p_2 = 0$	
Alternative hypothesis	$H_1: p_1 - p_2 > 0$	
Method	Z-Value	P-Value
Normal approximation	2.84	0.002
Fisher's exact		0.002

The test based on the normal approximation uses the pooled estimate of the proportion (0.242448).

Step 4 The P -value of 0.002 means that if the null hypothesis that $p_1 - p_2 = 0$ (or $p_1 = p_2$) is true, we expect 2 samples in 1000 repetitions of this experiment to yield the results we obtained! The observed results are unusual. Because the P -value is less than the level of significance, $\alpha = 0.05$ ($0.002 < 0.05$), we reject the null hypothesis.

Step 5 There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude the proportion of individuals 12 years and older taking 200 μg of Nasonex who experience headaches is greater than the proportion of individuals 12 years and older taking a placebo who experience headaches.

CAUTION!

In any statistical study, be sure to consider practical significance. Many statistically significant results can be produced simply by increasing the sample size.

Looking back at the results of Example 1, we notice that the proportion of individuals taking 200 μg of Nasonex who experience headaches is *statistically significantly* greater than the proportion of individuals 12 years and older taking a placebo who experience headaches. However, we need to ask ourselves a pressing question. Would you not take an allergy medication because 26% of patients experienced a headache taking the medication versus 22% who experienced a headache taking a placebo? Most people would be willing to accept the additional risk of a headache to relieve their allergy symptoms. While the difference of 4% is statistically significant, it does not have any *practical significance*.

NW Now Work Problem 17

3 Construct and Interpret Confidence Intervals for the Difference between Two Population Proportions

The sampling distribution of the difference of two proportions, $\hat{p}_1 - \hat{p}_2$, from independent samples can also be used to construct confidence intervals for the difference of two proportions.

Constructing a $(1 - \alpha) \cdot 100\%$ Confidence Interval for the Difference between Two Population Proportions (Independent Samples)

To construct a $(1 - \alpha) \cdot 100\%$ confidence interval for the difference between two population proportions from independent samples, the following requirements must be satisfied:

1. The samples are obtained independently, using simple random sampling or from a randomized experiment.
2. $n_1 \hat{p}_1(1 - \hat{p}_1) \geq 10$ and $n_2 \hat{p}_2(1 - \hat{p}_2) \geq 10$.
3. $n_1 \leq 0.05N_1$ and $n_2 \leq 0.05N_2$ (the sample size is no more than 5% of the population size); this ensures the independence necessary for a binomial experiment.

(continued)

Provided that these requirements are met, a $(1 - \alpha) \cdot 100\%$ confidence interval for $p_1 - p_2$ is given by

$$\text{Lower bound: } (\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \quad (3)$$

$$\text{Upper bound: } (\hat{p}_1 - \hat{p}_2) + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Notice that we do not pool the sample proportions. This is because we are not making any assumptions regarding their equality, as we did in hypothesis testing.

EXAMPLE 3

Constructing a Confidence Interval for the Difference between Two Population Proportions

Problem The Gallup organization surveyed 1100 adult Americans on May 6–9, 2002, and conducted an independent survey of 1024 adult Americans on May 1–10, 2018. In both surveys they asked the following: “Right now, do you think the state of moral values in the country as a whole is getting better or getting worse?” On May 1–10, 2018, 784 of the 1024 surveyed responded that the state of moral values is getting worse; on May 6–9, 2002, 737 of the 1100 surveyed responded that the state of moral values is getting worse. Construct and interpret a 90% confidence interval for the difference between the two population proportions.

Approach We can compute a 90% confidence interval for the two population proportions provided that the stated requirements are satisfied. We then construct the interval by hand using Formula (3) or using technology.

Solution We have to verify the requirements for constructing a confidence interval for the difference between two population proportions.

1. The samples were obtained independently through a random sample.
 2. For the May 1–10, 2018, survey (sample 1), we have $n_1 = 1024$ and $x_1 = 784$, so $\hat{p}_1 = \frac{x_1}{n_1} = \frac{784}{1024} = 0.766$. For the May 6–9, 2002, survey (sample 2), we have $n_2 = 1100$ and $x_2 = 737$, so $\hat{p}_2 = \frac{x_2}{n_2} = \frac{737}{1100} = 0.67$. Therefore,
- $$n_1 \hat{p}_1 (1 - \hat{p}_1) = 1024 (0.766)(1 - 0.766) = 183.5 \geq 10$$
- $$n_2 \hat{p}_2 (1 - \hat{p}_2) = 1100 (0.67)(1 - 0.67) = 243.2 \geq 10$$
3. The population of adult Americans exceeded 100 million in 2002 and 2018, so the sample size is definitely less than 5% of the population size.

By-Hand Solution

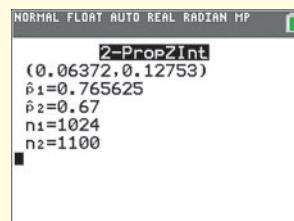
Substituting into Formula (3) with $\hat{p}_1 = 0.766$, $n_1 = 1024$, $\hat{p}_2 = 0.67$, and $n_2 = 1100$, we obtain the lower and upper bounds on the confidence interval:

$$\begin{aligned} \text{Lower bound: } & (\hat{p}_1 - \hat{p}_2) - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \\ &= (0.766 - 0.67) - 1.645 \cdot \sqrt{\frac{0.766(1 - 0.766)}{1024} + \frac{0.67(1 - 0.67)}{1100}} \\ &= 0.096 - 0.032 \\ &= 0.064 \end{aligned}$$

Technology Solution

Figure 4 shows the 90% confidence interval using a TI-84 Plus CE graphing calculator.

Figure 4



In Figure 4, the lower bound is 0.064. The upper bound is 0.128.

$$\begin{aligned}
 \text{Upper bound: } & (\hat{p}_1 - \hat{p}_2) + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \\
 & = (0.766 - 0.67) + 1.645 \cdot \sqrt{\frac{0.766(1 - 0.766)}{1024} + \frac{0.67(1 - 0.67)}{1100}} \\
 & = 0.096 + 0.032 \\
 & = 0.128
 \end{aligned}$$

Interpretation We are 90% confident that the difference between the proportion of adult Americans who believed that the state of moral values in the country as a whole was getting worse from 2002 to 2018 is between 0.064 and 0.128. To put this statement into everyday language, we might say that we are 90% confident that the percentage of adult Americans who believe that the state of moral values in the country as a whole was getting worse increased between 6.4% and 12.8% from 2002 to 2018. Because this interval does not contain 0, we might conclude that a higher proportion of the country believed that the state of moral values was getting worse in the United States in 2018 than in 2002.

NW Now Work Problem 21

④ Determine the Sample Size Necessary for Estimating the Difference between Two Population Proportions

In Section 9.1, we introduced a method for determining the sample size, n , required to estimate a single population proportion within a specified margin of error, E , with a specified level of confidence. This formula was obtained by solving the margin of error,

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \text{ for } n.$$

We follow the same approach to determine the sample size when estimating two population proportions. The margin of error, E , in constructing a confidence interval for the difference between two population proportions is

$$E = z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}. \text{ Assuming that } n_1 = n_2 = n, \text{ we can solve this expression for } n = n_1 = n_2 \text{ and obtain the following result:}$$

Sample Size for Estimating $p_1 - p_2$

The sample size required to obtain a $(1 - \alpha) \cdot 100\%$ confidence interval with a margin of error, E , is given by

$$n = n_1 = n_2 = [\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)] \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \quad (4)$$

rounded up to the next integer, if prior estimates of p_1 and p_2 , \hat{p}_1 and \hat{p}_2 , are available.

If prior estimates of p_1 and p_2 are unavailable, the sample size is

$$n = n_1 = n_2 = 0.5 \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \quad (5)$$

rounded up to the next integer.

The margin of error should always be expressed as a decimal when using Formulas (4) and (5).

CAUTION!

When doing sample size calculations, always round up.

EXAMPLE 4 Determining Sample Size

Problem A nutritionist wants to estimate the difference between the proportion of males and females who consume the USDA's recommended daily intake of calcium. What sample size should be obtained if she wishes the estimate to be within 3 percentage points with 95% confidence, assuming that

- (a) she uses the results of the USDA's 1994–1996 Diet and Health Knowledge Survey, according to which 51.1% of males and 75.2% of females consume the USDA's recommended daily intake of calcium, and
- (b) she does not use any prior estimates?

Approach We have $E = 0.03$ and $z_{\frac{\alpha}{2}} = z_{0.05} = z_{0.025} = 1.96$. To answer part (a), let $\hat{p}_1 = 0.511$ (for males) and $\hat{p}_2 = 0.752$ (for females) in Formula (4). To answer part (b), use Formula (5).

Solution

$$\begin{aligned}
 \text{(a)} \quad n_1 = n_2 &= [\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)] \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \\
 &= [0.511(1 - 0.511) + 0.752(1 - 0.752)] \left(\frac{1.96}{0.03} \right)^2 \\
 &= 1862.6
 \end{aligned}$$

Round this value up to 1863. The nutritionist must survey 1863 randomly selected males and 1863 randomly selected females.

$$\text{(b)} \quad n_1 = n_2 = 0.5 \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 = 0.5 \left(\frac{1.96}{0.03} \right)^2 = 2134.2$$

Round this value up to 2135. The nutritionist must survey 2135 randomly selected males and 2135 randomly selected females.

NW Now Work Problem 33

IN OTHER WORDS

If possible, obtain a prior estimate of \hat{p} when doing sample size computations.

Notice that having prior estimates of the population proportions reduces the number of individuals that need to be surveyed from 2135 to 1863. Because fewer individuals need to be surveyed, the costs of the study (in terms of time and money) are less when prior estimates are available.

Technology Step-by-Step

Inference for Two Population Proportions

TI-83/84 Plus

Hypothesis Tests

1. Press STAT, highlight TESTS, and select 6:2-PropZTest
2. Enter the values of x_1 , n_1 , x_2 , and n_2 .
3. Highlight the appropriate relation between p_1 and p_2 in the alternative hypothesis.
4. Highlight Calculate or Draw and press ENTER. Calculate gives the test statistic and P -value. Draw will draw the Z-distribution with the P -value shaded.

Confidence Intervals

Follow the same steps given for hypothesis tests, except select B:2-PropZInt Also, select a confidence level (such as $95\% = 0.95$).

Minitab

1. Enter the raw data into columns C1 and C2, if necessary.
2. Select the Stat menu, highlight **Basic Statistics**, then highlight **2 Proportions**
3. If you have raw data in different columns, select “Each sample in its own column” and enter C1 for first sample and C2 for second sample. If you have raw data in a single column, select “Both samples are in one column” and enter C1 for the sample IDs and C2 for the Samples. If you have summary statistics, select “Summarized data”. Enter the number of successes in the “events” cell and the sample size in the “trials” cell for each sample.

4. Click Options . . . and enter the level of confidence desired, the “test difference” (usually 0), and the direction of the alternative hypothesis. For confidence intervals, select “Estimate the proportions separately”; for hypothesis tests, select “Use the pooled estimate of the proportion” under Test method. Click OK twice.

Excel

Hypothesis Tests or Confidence Intervals

- Load the XLSTAT Add-in. Select the XLSTAT menu, highlight Parametric tests. From the pull-down menu, select Tests for two proportions.
- Enter the number of successes for sample 1 in the Frequency 1: cell; enter the sample size for sample 1 in the Sample size 1: cell; and so on. Make sure Data format: is set to Frequencies. Click the Options tab. Select the correct alternative hypothesis. Enter the Hypothesized difference (D) (usually 0), and select the Significance level (%). For a 95% confidence interval,

enter 5. For confidence intervals, select the first radio button under Variance; for hypothesis tests, select the second radio button. Click OK.

StatCrunch

Hypothesis Tests or Confidence Intervals

- If you have raw data, enter them into the spreadsheet. Name each column variable.
- Select Stat, highlight Proportion Stats, highlight Two Sample, and then choose either With Data or With Summary.
- If you chose With Data, select the column that has the observations, choose which outcome represents a success for each sample. If you chose With Summary, enter the number of successes and the number of trials for each sample. If you choose the hypothesis test radio button, enter the value of the proportion stated in the null hypothesis and choose the direction of the alternative hypothesis from the pull-down menu. If you choose the confidence interval radio button, enter the level of confidence. Click Calculate.



11.1 Assess Your Understanding

Vocabulary and Skill Building

- A sampling method is _____ when the individuals selected for one sample do not dictate which individuals are selected to be in a second sample.
- A sampling method is _____ when the individuals selected for one sample are used to determine the individuals in the second sample.

In Problems 3–8, determine whether the sampling is dependent or independent. Indicate whether the response variable is qualitative or quantitative.

- NW** 3. A sociologist wishes to compare the annual salaries of married couples in which both spouses work and determines each spouse’s annual salary.
4. A researcher wishes to determine the effects of alcohol on people’s reaction time to a stimulus. She randomly divides 100 people aged 21 or older into two groups. Group 1 is asked to drink 3 ounces of alcohol, while group 2 drinks a placebo. Both drinks taste the same, so the individuals in the study do not know which group they belong to. Thirty minutes after consuming the drink, the subjects in each group perform a series of tests meant to measure reaction time.
5. The Gallup Organization asked 1050 randomly selected adult Americans age 18 or older who consider themselves to be religious, “Do you believe it is morally acceptable or morally wrong [rotated] to conduct medical research using stem cells obtained from human embryos?” The same question was asked to 1050 randomly selected adult Americans age 18 or older who do not consider themselves to be religious. The goal of the study was to determine whether the proportion of religious adult Americans who believe it is morally wrong to conduct medical research using stem cells obtained from human embryos differed from the proportion of non-religious adult Americans who believe it is morally wrong to conduct medical research using stem cells obtained from human embryos.

6. A political scientist wants to know how a random sample of 18- to 25-year-olds feel about Democrats and Republicans in Congress. She obtains a random sample of 1030 registered voters 18 to 25 years of age and asks, “Do you have favorable/unfavorable [rotated] opinion of the Democratic/Republican [rotated] party?” Each individual was asked to disclose his or her opinion about each party.

7. An educator wants to determine whether a new curriculum significantly improves standardized test scores for third grade students. She randomly divides 80 third-graders into two groups. Group 1 is taught using the new curriculum, while group 2 is taught using the traditional curriculum. At the end of the school year, both groups are given the standardized test and the mean scores are compared.

8. A psychologist wants to know whether subjects respond faster to a go/no go stimulus or a choice stimulus. With the go/no go stimulus, subjects must respond to a particular stimulus by pressing a button and disregard other stimuli. In the choice stimulus, the subjects respond differently depending on the stimulus. The psychologist randomly selects 20 subjects, and each subject is presented a series of go/no go stimuli and choice stimuli. The mean reaction time to each stimulus is compared.

In Problems 9–12, conduct each test at the $\alpha = 0.05$ level of significance by determining (a) the null and alternative hypotheses, (b) the test statistic, (c) the critical value, and (d) the P-value. Assume that the samples were obtained independently using simple random sampling.

- Test whether $p_1 > p_2$. Sample data: $x_1 = 368, n_1 = 541, x_2 = 351, n_2 = 593$
- Test whether $p_1 < p_2$. Sample data: $x_1 = 109, n_1 = 475, x_2 = 78, n_2 = 325$
- Test whether $p_1 \neq p_2$. Sample data: $x_1 = 28, n_1 = 254, x_2 = 36, n_2 = 301$
- Test whether $p_1 \neq p_2$. Sample data: $x_1 = 804, n_1 = 874, x_2 = 902, n_2 = 954$

In Problems 13–16, construct a confidence interval for $p_1 - p_2$ at the given level of confidence.

13. $x_1 = 368, n_1 = 541, x_2 = 421, n_2 = 593$, 90% confidence
14. $x_1 = 109, n_1 = 475, x_2 = 78, n_2 = 325$, 99% confidence
15. $x_1 = 28, n_1 = 254, x_2 = 36, n_2 = 301$, 95% confidence
16. $x_1 = 804, n_1 = 874, x_2 = 892, n_2 = 954$, 95% confidence

Applying the Concepts

NW 17. Prevnar The drug Prevnar is a vaccine meant to prevent certain types of bacterial meningitis, typically administered to infants around 2 months. In randomized, double-blind clinical trials of Prevnar, infants were randomly divided into two groups. Subjects in group 1 received Prevnar, while subjects in group 2 received a control vaccine. After the first dose, 107 of 710 subjects in the experimental group (group 1) experienced fever as a side effect. After the first dose, 67 of 611 of the subjects in the control group (group 2) experienced fever as a side effect. Does the evidence suggest that a higher proportion of subjects in group 1 experienced fever as a side effect than subjects in group 2 at the $\alpha = 0.05$ level of significance?

18. Prevnar Part 2 In randomized, double-blind clinical trials of Prevnar, infants were randomly divided into two groups. Subjects in group 1 received Prevnar, while subjects in group 2 received a control vaccine. After the second dose, 137 of 452 subjects in the experimental group (group 1) experienced drowsiness as a side effect. After the second dose, 31 of 99 subjects in the control group (group 2) experienced drowsiness as a side effect. Does the evidence suggest that a lower proportion of subjects in group 1 experienced drowsiness as a side effect than subjects in group 2 at the $\alpha = 0.05$ level of significance?

19. Abstain from Alcohol In October 1947, the Gallup organization surveyed 1100 adult Americans and asked, “Are you a total abstainer from, or do you on occasion consume, alcoholic beverages?” Of the 1100 adults surveyed, 407 indicated that they were total abstainers. In a recent survey, the same question was asked of 1100 adult Americans and 333 indicated that they were total abstainers. Has the proportion of adult Americans who totally abstain from alcohol changed? Use the $\alpha = 0.05$ level of significance.

20. Views on the Death Penalty The Pew Research Group conducted a poll in which they asked, “Are you in favor of, or opposed to, executing persons as a general policy when the crime was committed while under the age of 18?” Of the 580 Catholics surveyed, 180 indicated they favored capital punishment; of the 600 seculars (those who do not associate with a religion) surveyed, 238 favored capital punishment. Is there a significant difference in the proportion of individuals in these groups in favor of capital punishment for persons under the age of 18? Use the $\alpha = 0.01$ level of significance.

NW 21. Tattoos The Harris Poll conducted a survey in which they asked, “How many tattoos do you currently have on your body?” Of the 1205 males surveyed, 181 responded that they had at least one tattoo. Of the 1097 females surveyed, 143 responded that they had at least one tattoo. Construct a 95% confidence interval to judge whether the proportion of males that have at least one tattoo differs significantly from the proportion of females that have at least one tattoo. Interpret the interval.

22. Body Mass Index The body mass index (BMI) of an individual is a measure used to judge whether an individual is overweight or not. A BMI between 20 and 25 indicates a normal

weight. In a survey of 750 men and 750 women, the Gallup organization found that 203 men and 270 women were normal weight. Construct a 90% confidence interval to gauge whether there is a difference in the proportion of men and women who are normal weight. Interpret the interval.

23. Public Cell Phone Conversations Researchers at Harris Interactive wondered if there was a difference between males and females in regard to some common annoyances. They asked a random sample of males and females, the following question: “Are you annoyed by people who repeatedly check their mobile phones while having an in-person conversation?” Among the 540 males surveyed, 178 responded “Yes”; among the 560 females surveyed, 206 responded “Yes.” Does the evidence suggest a higher proportion of females are annoyed by this behavior?

- (a) Explain why this study can be analyzed using the methods for conducting a hypothesis test regarding two independent proportions.
- (b) What are the null and alternative hypotheses?
- (c) Describe the sampling distribution of $\hat{p}_{\text{female}} - \hat{p}_{\text{male}}$. Draw a normal model with the area representing the P -value shaded for this hypothesis test.
- (d) Determine the P -value based on the model from part (c).
- (e) Interpret the P -value.
- (f) Based on the P -value, what does the sample evidence suggest? That is, what is the conclusion of the hypothesis test? Assume an $\alpha = 0.05$ level of significance.

24. Name Brand Researchers at Harris Interactive wondered if there was a difference between males and females in regard to whether they typically buy name-brand or store-brand products. They asked a random sample of males and females the following question: “For each of the following types of products, please indicate whether you typically buy name-brand products or store-brand products?” Among the 1104 males surveyed, 343 indicated they buy name-brand over-the-counter drugs; among the 1172 females surveyed, 295 indicated they buy name-brand over-the-counter drugs. Does the evidence suggest a lower proportion of females buy name-brand over-the-counter drugs?

- (a) Explain why this study can be analyzed using the methods for conducting a hypothesis test regarding two independent proportions.
- (b) What are the null and alternative hypotheses?
- (c) Describe the sampling distribution of $\hat{p}_{\text{female}} - \hat{p}_{\text{male}}$. Draw a normal model with the area representing the P -value shaded for this hypothesis test.
- (d) Determine the P -value based on the model from part (c).
- (e) Interpret the P -value.
- (f) Based on the P -value, what does the sample evidence suggest? That is, what is the conclusion of the hypothesis test? Assume an $\alpha = 0.05$ level of significance.

25. The Process of Statistics: Side Effects In clinical trials of the allergy medicine Clarinex (5 mg), 3307 allergy sufferers were randomly assigned to either a Clarinex group or a placebo group. It was reported that 50 out of 1655 individuals in the Clarinex group and 31 out of 1652 individuals in the placebo group experienced dry mouth as a side effect of their respective treatments. Source: www.clarinex.com

- (a) What type of experimental design is this?
- (b) What is the response variable? Is it qualitative or quantitative?
- (c) What is the explanatory variable? How many levels does the treatment have?

- (d) The clinical trial was double-blind. What does this mean?
- (e) Why is it important to have a placebo group?
- (f) Does the sample evidence suggest that the proportion of individuals experiencing dry mouth is greater for those taking Clarinex than for those taking a placebo at the $\alpha = 0.05$ level of significance?
- (g) Do you believe the difference between the groups is practically significant?

26. Practical versus Statistical Significance In clinical trials for treatment of a skin disorder, 642 of 2105 patients receiving the current standard treatment were cured of the disorder and 697 of 2115 patients receiving a new proposed treatment were cured of the disorder.

- (a) Does the new procedure cure a higher proportion of patients at the $\alpha = 0.05$ level of significance?
- (b) Do you think that the difference in success rates is practically significant? What factors might influence your decision?

27. The Process of Statistics: Metastatic Melanoma OPDIVA is a drug developed by Bristol-Meyers Squib that is meant to treat metastatic melanoma, which is the worst form of skin cancer. Survival rates for patients with this cancer are 6 to 10 months.

Historically, patients with metastatic melanoma were treated with a combination of chemotherapy and dacarbazine. In clinical trials, 418 patients with metastatic melanoma were randomly assigned to either OPDIVO ($n = 210$) by receiving 3 mg/kg by intravenous infusion every 2 weeks or dacarbazine ($n = 208$) by receiving 1000 mg/m² by intravenous infusion every 3 weeks. Of the 210 patients in the OPDIVO group, 45 had survived 12 months; of the 208 patients in the dacarbazine group, 22 had survived 12 months.

- (a) What type of experimental design is this study?
- (b) What is the response variable? Is it qualitative or quantitative?
- (c) What is the factor that is controlled and set at various levels? How many levels are there?
- (d) Estimate the difference in proportion of patients who had survived 12 months in the OPDIVO group versus the dacarbazine group with 95% confidence. Interpret the interval.

28. Iraq War In March 2003, the Pew Research Group surveyed 1508 adult Americans and asked, “Do you believe the United States made the right or wrong decision to use military force in Iraq?” Of the 1508 adult Americans surveyed, 1086 stated the United States made the right decision. In August 2010, the Pew Research Group asked the same question of 1508 adult Americans and found that 618 believed the United States made the right decision.

- (a) In the survey question, the choices “right” and “wrong” were randomly rotated. Why?
- (b) Construct and interpret a 90% confidence interval for the difference between the two population proportions, $p_{2003} - p_{2010}$.

29. Threaded Problem: Tornado The data set “Tornadoes_2017” located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a) Is there a difference in the proportion of F0 tornadoes in Texas versus Georgia? The data in column F Scale represent the Fujita F scale of the tornado (on a scale from 0 to 5). An F0 tornado is one where the wind speeds are less than 73 miles per hour. Test whether a different proportion of F0 tornadoes occurs in Texas than in Georgia. **Note:** Treat the tornadoes that struck in Texas and Georgia as a simple random sample of all tornadoes that struck Texas since 1950 using a 0.05 level

of significance. To conduct the hypothesis test in StatCrunch, select Stat > Proportion Stats > Two Sample > With Data. Let 0 from the F scale column be a success for both states. In the “Where:” box type “State = TX” or “State = GA”. The summarized data show that 82 of 168 tornadoes in Texas and 43 of 118 tornadoes in Georgia are F0.

- (b) Estimate the difference in the proportion of F0 tornadoes that occur in Texas versus Georgia (compute the difference in proportions as Texas – Georgia) with 95% confidence. Interpret the result.

DATA **30. Deficit Reduction** In the Sullivan Statistics Survey I, respondents were asked, “Would you be willing to pay higher taxes if the tax revenue went directly toward deficit reduction?” Treat the respondents as a simple random sample of adult Americans. The results of the survey may be downloaded at www.pearsonhighered.com/sullivanstats.

- (a) What proportion of the males who took the survey is willing to pay higher taxes to reduce the deficit? What proportion of the females who took the survey is willing to pay higher taxes to reduce the deficit?
- (b) Is there significant evidence to suggest the proportions of males and females who are willing to pay higher taxes to reduce the deficit differs at the $\alpha = 0.05$ level of significance?

31. Political Grammar Psychologists asked students to read two sentences about hypothetical politicians. One group of students read, “Last year, Mark was having an affair with his assistant and was taking hush money from a prominent constituent.” Let’s call this sentence A. Students in a second group read, “Last year, Mark had an affair with his assistant and took hush money from a prominent constituent.” We will call this sentence B. For each sentence, the students were asked if they felt the politician would be re-elected. The results are presented in the contingency table below.

	Sentence A	Sentence B
Re-elected	27	49
Not Re-elected	71	49

Source: Fausey, C.M., and Matlock, T., “Can Grammar Win Elections?” *Political Psychology*, no. doi: 10.1111/j.1467-9221.2010.00802.x

- (a) What are the specific differences in the way the two sentences are phrased?
- (b) How many students were given sentence A? How many students were given sentence B?
- (c) What proportion of the students given sentence A felt the politician would be re-elected? What proportion of the students given sentence B felt the politician would be re-elected?
- (d) Do the results presented in the contingency table suggest that the sentence structure makes a difference in deciding whether the politician would be re-elected?
- (e) Research “imperfect aspect” and “perfect aspect.” Does this help explain any differences in the results of the survey?

32. Relationship Deal-breakers A random sample of single males and single females (aged 21 to 76) was obtained by a group called Singles in America. The sample includes those currently separated, divorced, or widowed. The subjects were given a trait and asked to respond whether the trait was a deal-breaker in terms of maintaining a relationship. For each trait given in parts (a) through (c), determine if there is a statistically significant difference in the proportion of individuals responding yes. If there is a significant difference, report a 95% confidence interval for the difference. Source: Jonason, Peter K., et al,

"Relationship Deal-breakers: Traits People Avoid in Potential Mates." *Personality and Social Psychology Bulletin*, 2015, 41(12): 1697–1711.

(a) Trait: Lazy

	Males	Females
Yes	1646	2014
No	1098	783

(b) Trait: Stubborn

	Males	Females
Yes	878	951
No	1866	1846

(c) Trait: Talks Too Much

	Males	Females
Yes	713	559
No	2031	2238

NW 33. Determining Sample Size A physical therapist wants to determine the difference in the proportion of men and women who participate in regular, sustained physical activity. What sample size should be obtained if she wishes the estimate to be within 3 percentage points with 95% confidence, assuming that

- (a) she uses the 1998 estimates of 21.9% male and 19.7% female from the U.S. National Center for Chronic Disease Prevention and Health Promotion?
- (b) she does not use any prior estimates?

34. Determining Sample Size An educator wants to determine the difference between the proportion of males and females who have completed four or more years of college. What sample size should be obtained if she wishes the estimate to be within 2 percentage points with 90% confidence, assuming that

- (a) she uses the 1999 estimates of 27.5% male and 23.1% female from the U.S. Census Bureau?
- (b) she does not use any prior estimates?

35. Designing a Study Stock fund managers are investment professionals who decide which stocks should be part of a portfolio. In an article in the *Wall Street Journal* ("Not a Stock-Picker's Market," WSJ, January 25, 2014), the performance of stock fund managers was considered based on dispersion in the market. In the stock market, risk is measured by the standard deviation rate of return of stock (dispersion). When dispersion is low, then the rate of return of the stocks that make up the market are not as spread out. That is, the return on Company X is close to that of Y is close to

that of Z, and so on. When dispersion is high, then the rate of return of stocks is more spread out; meaning some stocks outperform others by a substantial amount. Since 1991, the dispersion of stocks has been about 7.1%. In some years, the dispersion is higher (such as 2001 when dispersion was 10%), and in some years it is lower (such as 2013 when dispersion was 5%). So, in 2001, stock fund managers would argue, one needed to have more investment advice in order to identify the stock market winners, whereas in 2013, since dispersion was low, virtually all stocks ended up with returns near the mean, so investment advice was not as valuable.

- (a) Suppose you want to design a study to determine whether the proportion of fund managers who outperform the market in low-dispersion years is less than the proportion of fund managers who outperform the market in high-dispersion years. What would be the response variable in this study? What is the explanatory variable in this study?
- (b) What or who are the individuals in this study?
- (c) To what population does this study apply?
- (d) What would be the null and alternative hypothesis?
- (e) Suppose this study was conducted and the data yielded a *P*-value of 0.083. Explain what this result suggests.

36. Putting It Together: Salk Vaccine On April 12, 1955, Dr. Jonas Salk released the results of clinical trials for his vaccine to prevent polio. In these clinical trials, 400,000 children were randomly divided in two groups. The subjects in group 1 (the experimental group) were given the vaccine, while the subjects in group 2 (the control group) were given a placebo. Of the 200,000 children in the experimental group, 33 developed polio. Of the 200,000 children in the control group, 115 developed polio.

- (a) What type of experimental design is this?
- (b) What is the response variable?
- (c) What are the treatments?
- (d) What is a placebo?
- (e) Why is such a large number of subjects needed for this study?
- (f) Does it appear to be the case that the vaccine was effective?

Explaining the Concepts

37. Why do we use a pooled estimate of the population proportion when testing a hypothesis about two proportions? Why do we not use a pooled estimate of the population proportion when constructing a confidence interval for the difference of two proportions?

38. Explain the difference between an independent and dependent sample.

11.2 Inference about Two Means: Dependent Samples



Preparing for This Section Before getting started, review the following:

- Matched-pairs design (Section 1.6, pp. 49–50)
- Confidence intervals about μ (Section 9.2, pp. 415–418)
- Hypothesis tests about μ (Section 10.3, pp. 458–463)
- Type I and Type II errors (Section 10.1, pp. 438–440)

Objectives

- ① Test hypotheses for a population mean from matched-pairs data
- ② Construct and interpret confidence intervals about the population mean difference of matched-pairs data

In this section, we discuss inference on the difference of two means for dependent sampling. We will address inference when the sampling is independent in Section 11.3.

1 Test Hypotheses for a Population Mean from Matched-Pairs Data

Inference on matched-pairs data is similar to inference regarding a population mean. Recall that if the population from which the sample was drawn is normally distributed or the sample size is large ($n \geq 30$), we said that

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

follows Student's t -distribution with $n - 1$ degrees of freedom.

When analyzing matched-pairs data, compute the difference in each matched pair and then perform inference on the differenced data using the methods of Section 9.2 or 10.3.

IN OTHER WORDS

Statistical inference methods on matched-pairs data use the same methods as inference on a single population mean, except that the **differences** are analyzed.

Testing Hypotheses Regarding the Difference of Two Means Using a Matched-Pairs Design

To test hypotheses regarding the mean difference of data obtained from a dependent sample (matched-pairs data), use the following steps, provided that

- the sample is obtained by simple random sampling or the data result from a matched-pairs design experiment.
- the sample data are dependent (matched pairs).
- the differences are normally distributed with no outliers or the sample size, n , is large ($n \geq 30$).
- the sampled values are independent (sample size is no more than 5% of population size).

Step 1 Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways, where μ_d is the population mean difference of the matched-pairs data.

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: \mu_d = 0$	$H_0: \mu_d = 0$	$H_0: \mu_d = 0$
$H_1: \mu_d \neq 0$	$H_1: \mu_d < 0$	$H_1: \mu_d > 0$

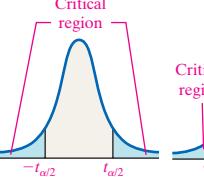
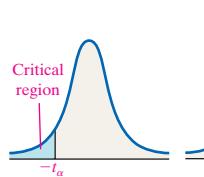
Step 2 Select a level of significance, α , depending on the seriousness of making a Type I error.

Classical Approach

Step 3 Compute the **test statistic**

$$t_0 = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

which follows Student's t -distribution with $n - 1$ degrees of freedom. The values of \bar{d} and s_d are the mean and standard deviation of the differenced data. Use Table VII to determine the critical value.

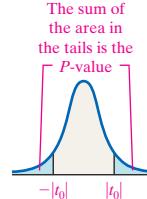
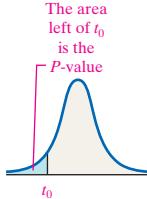
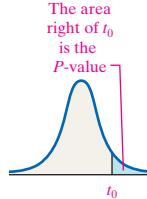
	Two-Tailed	Left-Tailed	Right-Tailed
Critical value	$-t_{\alpha/2}$ and $t_{\alpha/2}$	$-t_\alpha$	t_α
Critical region			

P-Value Approach

By Hand Step 3 Compute the **test statistic**

$$t_0 = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

which follows Student's t -distribution with $n - 1$ degrees of freedom. The values of \bar{d} and s_d are the mean and standard deviation of the differenced data. Use Table VII to approximate the P -value.

	Two-Tailed	Left-Tailed	Right-Tailed
	The sum of the area in the tails is the P -value	The area left of t_0 is the P -value	The area right of t_0 is the P -value
			
	$P\text{-value} = 2P(t > t_0)$	$P\text{-value} = P(t < t_0)$	$P\text{-value} = P(t > t_0)$

(continued)

Step 4 Compare the critical value to the test statistic.

Two-Tailed	Left-Tailed	Right-Tailed
If $t_0 < -t_{\alpha/2}$ or $t_0 > t_{\alpha/2}$, reject the null hypothesis.	If $t_0 < -t_{\alpha}$, reject the null hypothesis.	If $t_0 > t_{\alpha}$, reject the null hypothesis.

Technology Step 3 Use a statistical spreadsheet or calculator with statistical capabilities to obtain the P -value. The directions for obtaining the P -value using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on page 496.

Step 4 If P -value $< \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

The procedures just presented are **robust**, which means that minor departures from normality will not adversely affect the results of the test. If the data have outliers, however, the procedure should not be used.

Verify the assumption that the differenced data come from a population that is normally distributed by constructing a normal probability plot. Use a boxplot to determine whether there are outliers. If the normal probability plot indicates that the differenced data are not normally distributed or the boxplot reveals outliers, then bootstrap hypothesis tests (Section 11.2A) or nonparametric tests should be performed.

EXAMPLE 1 Testing Hypotheses Regarding Matched-Pairs Data

Problem Professor Andy Neill measured the time (in seconds) required to catch a falling meter stick for 12 randomly selected students' dominant hand and nondominant hand. Professor Neill wants to know if the reaction time in an individual's dominant hand is less than the reaction time in his or her nondominant hand. A coin flip is used to determine whether reaction time is measured using the dominant or nondominant hand first. Conduct the test at the $\alpha = 0.05$ level of significance. The data obtained are presented in Table 1.

Table 1

Student	Dominant Hand, X_i	Nondominant Hand, Y_i
1	0.177	0.179
2	0.210	0.202
3	0.186	0.208
4	0.189	0.184
5	0.198	0.215
6	0.194	0.193
7	0.160	0.194
8	0.163	0.160
9	0.166	0.209
10	0.152	0.164
11	0.190	0.210
12	0.172	0.197

Source: Professor Andy Neill, Joliet Junior College.

Approach This is a matched-pairs design because the variable is measured on the same subject for both the dominant and nondominant hand, the treatment in this experiment. Compute the difference between the dominant time and the nondominant time. So, for the first student compute $X_1 - Y_1$, for the second student compute $X_2 - Y_2$, and so on. If the reaction time in the dominant hand is less than the reaction time in the nondominant hand, we would expect the values of $X_i - Y_i$ to be negative. We assume that there is no difference and seek evidence that leads us to believe that there is a difference.

Before performing the hypothesis test, verify that the differences come from a population that is approximately normally distributed with no outliers because the sample size is small. Construct a normal probability plot and boxplot of the differenced data to verify these requirements. Then proceed to follow Steps 1 through 5.

Solution We compute the differences as $d_i = X_i - Y_i$ = time of dominant hand for i th student minus time of nondominant hand for i th student. We expect these differences to be negative, so we want to determine if $\mu_d < 0$. Table 2 displays the differences.

CAUTION!

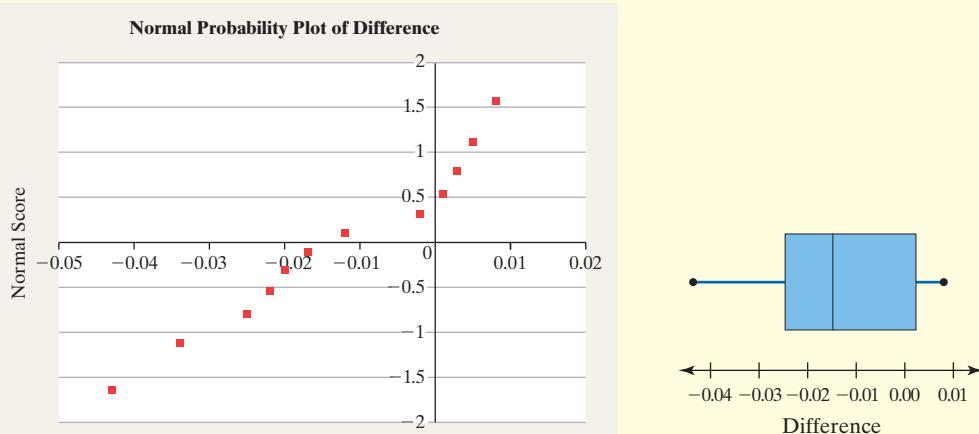
The way that we define the difference determines the direction of the alternative hypothesis in one-tailed tests. In Example 1, we expect $X_i < Y_i$, so the difference $X_i - Y_i$ is expected to be negative. Therefore, the alternative hypothesis is $H_1: \mu_d < 0$, and we have a left-tailed test. However, if we computed the differences as $Y_i - X_i$, we would expect the differences to be positive, and we have a right-tailed test!

Table 2

Student	Dominant Hand, X_i	Nondominant Hand, Y_i	Difference, d_i
1	0.177	0.179	$0.177 - 0.179 = -0.002$
2	0.210	0.202	$0.210 - 0.202 = 0.008$
3	0.186	0.208	-0.022
4	0.189	0.184	0.005
5	0.198	0.215	-0.017
6	0.194	0.193	0.001
7	0.160	0.194	-0.034
8	0.163	0.160	0.003
9	0.166	0.209	-0.043
10	0.152	0.164	-0.012
11	0.190	0.210	-0.020
12	0.172	0.197	-0.025
			$\sum d_i = -0.158$

Compute the mean and standard deviation of the differences and obtain $\bar{d} = -0.0132$ and $s_d = 0.0164$, each rounded to four decimal places. Verify that the data come from a population that is approximately normal with no outliers. Figure 5 shows the normal probability plot and boxplot of the differenced data.

Figure 5



NOTE

For those using Option 2 to assess model requirements, the boxplot indicates that the distribution of the differenced data is “symmetric enough” to use Student’s t -distribution.

The correlation between the differenced data and expected z -scores is 0.978. Because $0.978 > 0.928$ (Table VI), it is reasonable to conclude the data come from a population that is normally distributed.

Step 1 Professor Neill wants to know if the reaction time in the dominant hand is less than the reaction time in the nondominant hand. Express this as $\mu_d < 0$. We have

$$H_0: \mu_d = 0 \text{ second} \quad \text{versus} \quad H_1: \mu_d < 0 \text{ second}$$

This test is left-tailed.

Step 2 The level of significance is $\alpha = 0.05$.

(continued)

Classical Approach

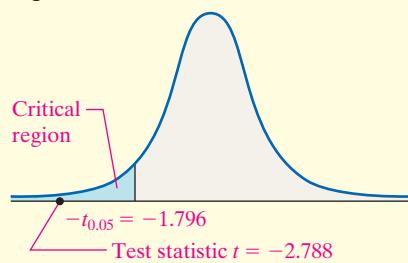
Step 3 The sample mean is $\bar{d} = -0.0132$ second and the sample standard deviation is $s_d = 0.0164$ second.

The test statistic is

$$t_0 = \frac{\bar{d}_0}{\frac{s_d}{\sqrt{n}}} = \frac{-0.0132}{\frac{0.0164}{\sqrt{12}}} = -2.788$$

Because this is a left-tailed test, determine the critical value at the $\alpha = 0.05$ level of significance with $12 - 1 = 11$ degrees of freedom to be $-t_{0.05} = -1.796$. The critical region is shown in Figure 6.

Figure 6



Step 4 The test statistic, $t_0 = -2.788$, is labeled in Figure 6. Because the test statistic lies in the critical region, we reject the null hypothesis.

P-Value Approach

By Hand Step 3 The sample mean is $\bar{d} = -0.0132$ second and the sample standard deviation is $s_d = 0.0164$ second.

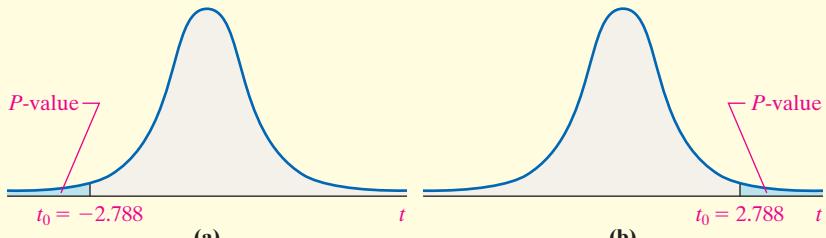
The test statistic is

$$t_0 = \frac{\bar{d}_0}{\frac{s_d}{\sqrt{n}}} = \frac{-0.0132}{\frac{0.0164}{\sqrt{12}}} = -2.788$$

Because this is a left-tailed test, the P -value is the area under the t -distribution with $12 - 1 = 11$ degrees of freedom to the left of the test statistic, $t_0 = -2.788$, as shown in Figure 7(a). That is, P -value = $P(t < t_0) = P(t < -2.788)$ with 11 degrees of freedom.

Because of the symmetry of the t -distribution, the area under the t -distribution to the left of -2.788 equals the area under the t -distribution to the right of 2.788 . So the P -value = $P(t_0 < -2.788) = P(t_0 > 2.788)$. See Figure 7(b).

Figure 7



Using Table VII, find the row that corresponds to 11 degrees of freedom. The value 2.788 lies between 2.718 and 3.106. The value of 2.718 has an area of 0.01 to the right under the t -distribution with 11 degrees of freedom. The value of 3.106 has an area of 0.005 to the right under the t -distribution with 11 degrees of freedom.

Because 2.788 is between 2.718 and 3.106, the P -value is between 0.005 and 0.01. So $0.005 < P\text{-value} < 0.01$.

Technology Step 3 Using StatCrunch, we find the P -value is 0.009. See Figure 8.

Figure 8

Hypothesis test results:

$\mu_1 - \mu_2$: mean of the paired difference between Dominant and Nondominant

$H_0: \mu_1 - \mu_2 = 0$

$H_A: \mu_1 - \mu_2 < 0$

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
Dominant - Nondominant	-0.013166667	0.0047431504	11	-2.7759328	0.009

Differences stored in column, Differences.

Step 4 The P -value of 0.009 [by hand: $0.005 < P\text{-value} < 0.01$] means that if the null hypothesis that the mean difference is zero is true, we expect a sample mean of -0.0132 second or lower in about 9 out of 1000 repetitions of this experiment. The results we obtained are not consistent with the assumption the mean difference in reaction time between the dominant and nondominant hand is 0 second. Simply put, because the P -value is less than the level of significance, $\alpha = 0.05$ ($0.009 < 0.05$), Professor Neill rejects the null hypothesis.

Step 5 The sample data provide sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the reaction time in the dominant hand is less than the reaction time in the nondominant hand.

NW Now Work Problem 7 (a) and (b)



② Construct and Interpret Confidence Intervals about the Population Mean Difference of Matched-Pairs Data

We can also obtain a confidence interval for the population mean difference, μ_d , using the sample mean difference, \bar{d} , the sample standard deviation difference, s_d , the sample size, and $t_{\frac{\alpha}{2}}$. Remember, a confidence interval about a population mean is given in the following form:

$$\text{Point estimate} \pm \text{margin of error}$$

Based on this, we compute the confidence interval for μ_d as follows:

Confidence Interval for Matched-Pairs Data

A $(1 - \alpha) \cdot 100\%$ confidence interval for μ_d is given by

$$\text{Lower bound: } \bar{d} - t_{\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}} \quad \text{Upper bound: } \bar{d} + t_{\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}} \quad (1)$$

The critical value $t_{\alpha/2}$ is determined using $n - 1$ degrees of freedom. The values of \bar{d} and s_d are the mean and standard deviation of the differenced data.

Note: The interval is exact when the population is normally distributed and approximately correct for nonnormal populations, provided that n is large.

EXAMPLE 2 Constructing a Confidence Interval for Matched-Pairs Data

Problem Using the data from Table 2 on page 493, construct a 95% confidence interval estimate of the mean difference, μ_d .

By Hand Approach

Step 1 Compute the differenced data. Because the sample size is small, verify that the differenced data come from a population that is approximately normal with no outliers.

Step 2 Compute the sample mean difference, \bar{d} , and the sample standard deviation difference, s_d .

Step 3 Determine the critical value, $t_{\frac{\alpha}{2}}$, with $\alpha = 0.05$ and $n - 1$ degrees of freedom.

Step 4 Use Formula (1) to determine the lower and upper bounds.

Step 5 Interpret the results.

By-Hand Solution

Step 1 We computed the differenced data and verified that they come from a population that is approximately normally distributed with no outliers in Example 1.

Step 2 We computed the sample mean difference, \bar{d} , to be -0.0132 and the sample standard deviation of the difference, s_d , to be 0.0164 in Example 1.

Step 3 Using Table VII with $\alpha = 0.05$ and $12 - 1 = 11$ degrees of freedom, we find $t_{\frac{\alpha}{2}} = t_{0.025} = 2.201$.

Technology Approach

Step 1 Compute the differenced data. Because the sample size is small, verify the differenced data come from a population that is approximately normal with no outliers.

Step 2 Use a statistical spreadsheet or graphing calculator with advanced statistical features to obtain the confidence interval. We will use a TI-84 Plus CE to construct the confidence interval. The steps for constructing confidence intervals using StatCrunch, Minitab, Excel, and the TI-83/84 graphing calculators are given in the Technology Step-by-Step on page 496.

Step 3 Interpret the result.

Technology Solution

Step 1 This was done in Example 1 on page 493.

Step 2 Figure 9 shows the results from a TI-84 Plus CE graphing calculator.

Figure 9



(continued)

Step 4 Substituting into Formula (1), we find

$$\begin{aligned}\text{Lower bound: } \bar{d} - t_{\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}} &= -0.0132 - 2.201 \cdot \frac{0.0164}{\sqrt{12}} \\ &= -0.0236\end{aligned}$$

$$\begin{aligned}\text{Upper bound: } \bar{d} + t_{\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}} &= -0.0132 + 2.201 \cdot \frac{0.0164}{\sqrt{12}} \\ &= -0.0028\end{aligned}$$

Step 5 We are 95% confident the mean difference between the dominant hand's reaction time and the nondominant hand's reaction time is between -0.0236 and -0.0028 second. In other words, we are 95% confident the dominant hand has a mean reaction time that is somewhere between 0.0028 and 0.0236 second faster than the nondominant hand. Because the confidence interval does not contain zero, the evidence suggests the reaction time of a person's dominant hand is different from the reaction time of the nondominant hand.

Step 3 We are 95% confident the mean difference between the dominant hand's reaction time and the nondominant hand's reaction time is between -0.0236 and -0.0027 second. In other words, we are 95% confident the dominant hand has a mean reaction time that is somewhere between 0.0027 and 0.0236 second faster than the nondominant hand. Because the confidence interval does not contain zero, the evidence suggests the reaction time of the dominant hand is different from the reaction time of the nondominant hand.

NW Now Work Problem 7 (c)

Technology Step-by-Step

Two-Sample *t*-Tests, Dependent Sampling

TI-83/84 Plus

Hypothesis Tests

- If necessary, enter raw data in L1 and L2. Let $L3 = L1 - L2$ (or $L2 - L1$), depending on how the alternative hypothesis is defined.
- Press STAT, highlight TESTS, and select 2:T-Test
- If the data are raw, highlight DATA, making sure that List is set to L3 with frequency set to 1. If summary statistics are known, highlight STATS and enter the summary statistics.
- Highlight the appropriate relation in the alternative hypothesis.
- Highlight Calculate or Draw and press ENTER. Calculate gives the test statistic and *P*-value. Draw will draw the *t*-distribution with the *P*-value shaded.

Confidence Intervals

Follow the same steps given for hypothesis tests, except select 8: TInterval. Also, select a confidence level (such as $95\% = 0.95$).

Minitab

- If necessary, enter raw data in columns C1 and C2.
- Select the Stat menu, highlight **Basic Statistics**, and then select **Paired t**
- If you have raw data, select "Each sample is in a column" from the drop-down menu. Enter C1 in the cell marked "Sample 1" and enter C2 in the cell marked "Sample 2." If you have summarized data, select "Summarized data (differences)" from the drop-down menu and enter the summary statistics. Click Options . . . , select the direction of the alternative hypothesis and select a confidence level. Click OK twice.

Excel

- Enter the raw data into Columns A and B.
- Select the Formulas menu. Select More Functions. Highlight Statistical, and select T.TEST from the drop-down menu.
- Place the cursor in Array1. Highlight the data in Column A. Place the cursor in Array2. Highlight the data in Column B. Place the cursor in the Tails cell. Enter the number corresponding to the test you desire (1 for a one-tailed distribution; 2 for a two-tailed distribution). Place the cursor in the Type cell. Enter 1 for a paired *t*-test. Click OK.

StatCrunch

- If necessary, enter the raw data into the first two columns of the spreadsheet. Name each column variable.
- Select Stat, highlight **T Stats**, select **Paired**.
- Select the column that contains the data for Sample 1. Select the column that contains the data for Sample 2. Note that the differences are computed Sample 1 – Sample 2. If you choose the hypothesis test radio button, enter the value of the mean stated in the null hypothesis and choose the direction of the alternative hypothesis from the pull-down menu. If you choose the confidence interval radio button, enter the level of confidence. Click Compute!.

Note: If you have summarized data, then follow the steps for performing a single sample *t*-test.



11.2 Assess Your Understanding

Skill Building

1. A researcher wants to show the mean from population 1 is less than the mean from population 2 in matched-pairs data. If the observations from sample 1 are X_i and the observations from sample 2 are Y_i , and $d_i = X_i - Y_i$, then the null hypothesis is $H_0: \mu_d = 0$ and the alternative hypothesis is $H_1: \mu_d \underline{\hspace{2cm}} 0$.
2. A researcher wants to show the mean from population 1 is less than the mean from population 2 in matched-pairs data. If the observations from sample 1 are X_i and the observations from sample 2 are Y_i and $d_i = Y_i - X_i$, then the null hypothesis is $H_0: \mu_d = 0$ and the alternative hypothesis is $H_1: \mu_d \underline{\hspace{2cm}} 0$.

In Problems 3 and 4, assume that the differences are normally distributed.



3.

Observation	1	2	3	4	5	6	7
X_i	7.6	7.6	7.4	5.7	8.3	6.6	5.6
Y_i	8.1	6.6	10.7	9.4	7.8	9.0	8.5

- (a) Determine $d_i = X_i - Y_i$ for each pair of data.
 (b) Compute \bar{d} and s_d .
 (c) Test if $\mu_d < 0$ at the $\alpha = 0.05$ level of significance.
 (d) Construct a 95% confidence interval about the population mean difference μ_d .



4.

Observation	1	2	3	4	5	6	7	8
X_i	19.4	18.3	22.1	20.7	19.2	11.8	20.1	18.6
Y_i	19.8	16.8	21.1	22.0	21.5	18.7	15.0	23.9

- (a) Determine $d_i = X_i - Y_i$ for each pair of data.
 (b) Compute \bar{d} and s_d .
 (c) Test if $\mu_d \neq 0$ at the $\alpha = 0.01$ level of significance.
 (d) Construct a 99% confidence interval about the population mean difference μ_d .

Applying the Concepts

5. **Naughty or Nice?** An experiment was conducted in which 16 ten-month-old babies were asked to watch a climber character attempt to ascend a hill. On two occasions, the baby witnesses the character fail to make the climb. On the third attempt, the baby witnesses either a helper toy push the character up the hill, or a hinderer toy preventing the character from making the ascent. The helper and hinderer toys were shown to each baby in a random fashion for a fixed amount of time. In Problem 41 from Section 10.2, we learned that, after watching both the helper and hinderer toy in action, 14 of 16 ten-month-old babies preferred to play with the helper toy when given a choice as to which toy to play with. A second part of this experiment showed the climber approach the helper toy, which is not a surprising action, and then alternatively the climber approached the hinderer toy, which is a surprising action. The amount of time the ten-month-old watched the event was recorded. The mean difference in time spent watching the climber approach the hinderer toy versus watching the climber approach the helper toy was 1.14 seconds with a standard deviation of 1.75 second. *Source: J. Kiley Hamlin et al., "Social Evaluation by Preverbal Infants," Nature, Nov. 2007.*

- (a) State the null and alternative hypothesis to determine if babies tend to look at the hinderer toy longer than the helper toy.

- (b) Assuming the differences are normally distributed with no outliers, test if the difference in the amount of time the baby will watch the hinderer toy versus the helper toy is greater than 0 at the 0.05 level of significance.

- (c) What do you think the results of this experiment imply about 10-month-olds' ability to assess surprising behavior?

6. **Platelet-Rich Plasma** In a prospective cohort study, 20 patients with alopecia (hair loss) had platelet-rich plasma (PRP) injected in their scalps. After three months, the mean difference in hair density (after – before) was 170.70 hairs per square centimeter with a standard deviation of 37.81 hairs/cm². *Source: Gkini MA, Kouskoukis AE, Tripsianis G, Rigopoulos D, Kouskoukis K., "Study of Platelet-Rich Plasma Injections in the Treatment of Androgenetic Alopecia through a One-Year Period". J Cutan Aesthet Surg, 2014; 7:213–219.*

- (a) What does it mean for this study to be a prospective cohort study?

- (b) What is the variable of interest in this study? Is it qualitative or quantitative?

- (c) State the null and alternative hypotheses to determine if hair density increased.

- (d) The researchers indicated that a test for normality indicated it was reasonable to conclude the change in hair density is approximately normal with no outliers. Does the evidence suggest that PRP injected in the scalp increases hair density? Use an $\alpha = 0.05$ level of significance.

- (e) How would the alternative hypothesis change if the differences were computed "before – after"?

NW 7. Muzzle Velocity The following data represent the muzzle

DATA velocity (in feet per second) of rounds fired from a 155-mm gun. For each round, two measurements of the velocity were recorded using two different measuring devices, with the following data obtained:

Observation	1	2	3	4	5	6
A	793.8	793.1	792.4	794.0	791.4	792.4
B	793.2	793.3	792.6	793.8	791.6	791.6

Observation	7	8	9	10	11	12
A	791.7	792.3	789.6	794.4	790.9	793.5
B	791.6	792.4	788.5	794.7	791.3	793.5

Source: Ronald Christenson and Larry Blackwood. "Tests for Precision and Accuracy of Multiple Measuring Devices." Technometrics, 35(4):411–421, 1993.

- (a) Why are these matched-pairs data?

- (b) Is there a difference in the measurement of the muzzle velocity between device A and device B at the $\alpha = 0.01$ level of significance? **Note:** A normal probability plot and boxplot of the data indicate that the differences are approximately normally distributed with no outliers.

- (c) Construct a 99% confidence interval about the population mean difference. Interpret your results.

- (d) Draw a boxplot of the differenced data. Does this visual evidence support the results obtained in part (b)?

8. **Reaction Time** In an experiment conducted online at the University of Mississippi, study participants are asked to react to a stimulus. In one experiment, the participant must press a key on seeing a blue screen and reaction time (in seconds) to press the key is measured. The same person is then asked to press a key on

seeing a red screen, again with reaction time measured. The results for six randomly sampled study participants are as follows:

Participant	1	2	3	4	5	6
Blue	0.582	0.481	0.841	0.267	0.685	0.450
Red	0.408	0.407	0.542	0.402	0.456	0.533

Source: PsychExperiments at the University of Mississippi.

- (a) Why are these matched-pairs data?
- (b) In this study, the color that the participant was first asked to react to was randomly selected. Why is this a good idea in this experiment?
- (c) Is the reaction time to the blue stimulus different from the reaction time to the red stimulus at the $\alpha = 0.01$ level of significance? **Note:** A normal probability plot and boxplot of the data indicate that the differences are approximately normally distributed with no outliers.
- (d) Construct a 99% confidence interval about the population mean difference. Interpret your results.
- (e) Draw a boxplot of the differenced data. Does this visual evidence support the results obtained in part (c)?

9. SUV versus Car It is a commonly held belief that SUVs are safer than cars. If an SUV and car are in a collision, does the SUV sustain less damage (as suggested by the cost of repair)? The Insurance Institute for Highway Safety crashed SUVs into cars, with the SUV moving 10 miles per hour and the front of the SUV crashing into the rear of the car.

SUV into Car	SUV Damage	Car Damage
Honda CR-V into Honda Civic	1721	1274
Toyota RAV4 into Toyota Corolla	1434	2327
Hyundai Tucson into Kia Forte	850	3223
Volkswagen Tiguan into VW Golf	2329	2058
Jeep Patriot into Dodge Caliber	1415	3095
Ford Escape into Ford Focus	1470	3386
Nissan Rogue into Nissan Sentra	2884	4560

Source: Insurance Institute for Highway Safety.

- (a) Why are these matched-pairs data?
- (b) Draw a boxplot of the differenced data. Does the visual evidence support the belief that SUVs have a lower repair cost?
- (c) Do the data suggest the repair cost for the car is higher? Use an $\alpha = 0.05$ level of significance.

Note: A normal probability plot indicates the differenced data are approximately normal with no outliers.

10. Secchi Disk A Secchi disk is an 8-inch-diameter weighted disk that is painted black and white and attached to a rope. The disk is lowered into water and the depth (in inches) at which it is no longer visible is recorded. The measurement is an indication of water clarity. An environmental biologist interested in determining whether the water clarity of the lake at Joliet Junior College is improving takes measurements at the same location on eight dates during the course of a year and repeats the measurements on the same dates five years later. She obtains the following results:

Observation	1	2	3	4	5	6	7	8
Date	5/11	6/7	6/24	7/8	7/27	8/31	9/30	10/12
Initial depth, X_i	38	58	65	74	56	36	56	52
Depth five years later, Y_i	52	60	72	72	54	48	58	60

Source: Virginia Piekarski, Joliet Junior College.

- (a) Why is it important to take the measurements on the same date?
- (b) Does the evidence suggest that the clarity of the lake is improving at the $\alpha = 0.05$ level of significance? **Note:** A normal probability plot and boxplot of the data indicate that the differences are approximately normally distributed with no outliers.
- (c) Draw a boxplot of the differenced data. Does this visual evidence support the results obtained in part (b)?

11. Getting Taller? To test the belief that sons are taller than their fathers, a student randomly selects 13 fathers who have adult male children. She records the height of both the father and son in inches and obtains the following data. Does the evidence suggest that sons are taller than their fathers? Use the $\alpha = 0.1$ level of significance. **Note:** A normal probability plot and boxplot of the data indicate that the differences are approximately normally distributed with no outliers.

	1	2	3	4	5	6	7
Height of father, X_i	70.3	67.1	70.9	66.8	72.8	70.4	71.8
Height of son, Y_i	74.1	69.2	66.9	69.2	68.9	70.2	70.4
	8	9	10	11	12	13	
Height of father, X_i	70.1	69.9	70.8	70.2	70.4	72.4	
Height of son, Y_i	69.3	75.8	72.3	69.2	68.6	73.9	

Source: Anna Behounek, student at Joliet Junior College.

12. Waiting in Line A quality-control manager at an amusement park feels that the amount of time that people spend waiting in line for the American Eagle roller coaster is too long. To determine if a new loading/unloading procedure is effective in reducing wait time in line, he measures the amount of time (in minutes) people are waiting in line on 7 days. After implementing the new procedure, he again measures the amount of time (in minutes) people are waiting in line on 7 days and obtains the following data. To make a reasonable comparison, he chooses days when the weather conditions are similar. Is the new loading/unloading procedure effective in reducing wait time at the $\alpha = 0.05$ level of significance? **Note:** A normal probability plot and boxplot of the data indicate that the differences are approximately normally distributed with no outliers.

Day	Mon (2 P.M.)	Tues (2 P.M.)	Wed (2 P.M.)	Thurs (2 P.M.)	Fri (2 P.M.)
Wait time before, X_i	11.6	25.9	20.0	38.2	57.3
Wait time after, Y_i	10.7	28.3	19.2	35.9	59.2
Day	Sat (11 A.M.)	Sat (4 P.M.)	Sun (12 noon)	Sun (4 P.M.)	
Wait time before, X_i	32.1	81.8	57.1	62.8	
Wait time after, Y_i	31.8	77.3	54.9	62.0	

13. Hardness Testing The manufacturer of hardness testing equipment uses steel-ball indenters to penetrate a metal that is being tested. However, the manufacturer thinks it would be better to use a diamond indenter so that all types of metal can be tested. Because of differences between the two types of indenters, it is suspected that the two methods will produce different hardness readings. The metal specimens to be tested are large enough so

that two indentations can be made. Therefore, the manufacturer uses both indenters on each specimen and compares the hardness readings. Construct a 95% confidence interval to judge whether the two indenters result in different measurements.

Specimen	1	2	3	4	5	6	7	8	9
Steel ball	50	57	61	71	68	54	65	51	53
Diamond	52	56	61	74	69	55	68	51	56

Note: A normal probability plot and boxplot of the data indicate that the differences are approximately normally distributed with no outliers.

- DATA 14. Car Rentals** The following data represent the daily rental for a compact automobile charged by two car rental companies, Thrifty and Hertz, in ten locations. Test whether Thrifty is less expensive than Hertz at the $\alpha = 0.1$ level of significance. **Note:** A normal probability plot and boxplot of the data indicate that the differences are approximately normally distributed with no outliers.

City	Thrifty	Hertz
Chicago	21.81	18.99
Los Angeles	29.89	37.99
Houston	17.90	19.99
Orlando	27.98	35.99
Boston	24.61	25.60
Seattle	21.96	22.99
Pittsburgh	20.90	19.99
Phoenix	37.75	36.99
New Orleans	33.81	26.99
Minneapolis	33.49	30.99

Source: Yahoo!Travel.

- DATA 15. DUI Simulator** To illustrate the effects of driving under the influence (DUI) of alcohol, a police officer brought a DUI simulator to a local high school. Student reaction time in an emergency was measured with unimpaired vision and also while wearing a pair of special goggles to simulate the effects of alcohol on vision. For a random sample of nine teenagers, the time (in seconds) required to bring the vehicle to a stop from a speed of 60 miles per hour was recorded.

Subject	1	2	3	4	5	6	7	8	9
Normal, X_i	4.47	4.24	4.58	4.65	4.31	4.80	4.55	5.00	4.79
Impaired, Y_i	5.77	5.67	5.51	5.32	5.83	5.49	5.23	5.61	5.63

- (a) Whether the student had unimpaired vision or wore goggles first was randomly selected. Why is this a good idea in designing the experiment?
(b) Use a 95% confidence interval to test if there is a difference in braking time with impaired vision and normal vision where the differences are computed as “impaired minus normal.” **Note:** A normal probability plot and boxplot of the data indicate that the differences are approximately normally distributed with no outliers.

- DATA 16. Braking Distance** An automotive researcher wanted to estimate the difference in distance required to come to a complete stop while traveling 40 miles per hour on wet versus dry pavement. Because car type plays a role, the researcher used eight different cars with the same driver and tires. The braking distance (in feet) on

both wet and dry pavement is shown in the table below. Construct a 95% confidence interval for the mean difference in braking distance on wet versus dry pavement where the differences are computed as “wet minus dry.” Interpret the interval. **Note:** A normal probability plot and boxplot of the data indicate that the differences are approximately normally distributed with no outliers.

Car	1	2	3	4	5	6	7	8
Wet	106.9	100.9	108.8	111.8	105.0	105.6	110.6	107.9
Dry	71.8	68.8	74.1	73.4	75.9	75.2	75.7	81.0

- DATA 17. Red Light Cameras** Chicago has installed cameras at various intersections through the city. The cameras photograph the license plate of any car engaging in a moving violation (such as driving through a red light or failing to stop completely prior to turning on red). Researchers with the city wanted to know whether there is a difference between the number of violations on Wednesdays and Saturdays. They randomly selected 30 cameras on a Wednesday and Saturday of the same week. The following data represent the number of violations recorded.

Camera ID	Sat	Wed	Camera ID	Sat	Wed
2244	2	19	1031	3	3
1671	14	6	2412	9	1
2081	10	3	2013	13	0
1964	5	6	1612	7	5
1743	4	26	1142	2	47
1151	1	2	1982	24	2
2321	11	4	2023	5	2
1142	30	8	2523	2	16
1652	3	11	1011	5	3
1514	3	17	1502	1	4
2054	3	24	2154	13	1
1934	6	9	1274	2	0
1234	7	3	1282	5	2
3082	1	4	1123	2	3
2672	12	3	1721	19	40

Source: City of Chicago Data Portal.

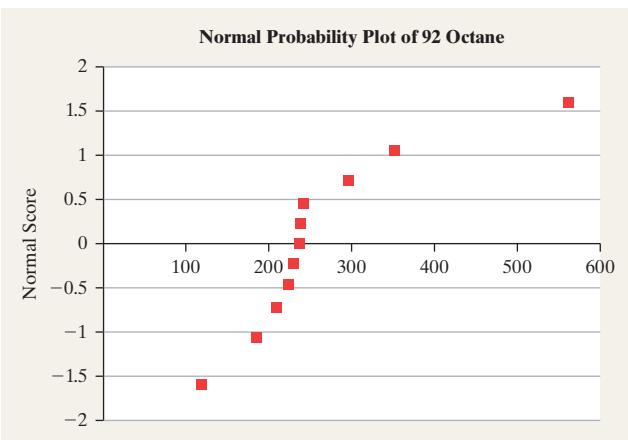
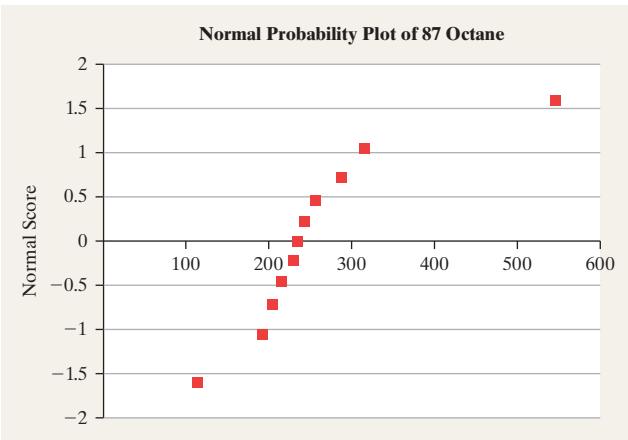
- (a) Why does it make sense to pair by the camera?
(b) Determine the difference in the data as “Sat – Wed.” Draw a boxplot of the differenced data. Explain why a large sample size is needed to analyze this data using Student’s t -distribution.
(c) Does the sample evidence suggest there is a difference in the number of violations on a Wednesday versus Saturday?
(d) Can you think of any variables that may confound the results of the study?

- DATA 18. Does Octane Matter?** Octane is a measure of how much fuel can be compressed before it spontaneously ignites. Some people believe that higher-octane fuels result in better gas mileage for their cars. To test this claim, a researcher randomly selected 11 individuals (and their cars) to participate in the study. Each participant received 10 gallons of gas and drove his or her car on a closed course that simulated both city and highway driving. The number of miles driven until the car ran out of gas was recorded. A coin flip was used to determine

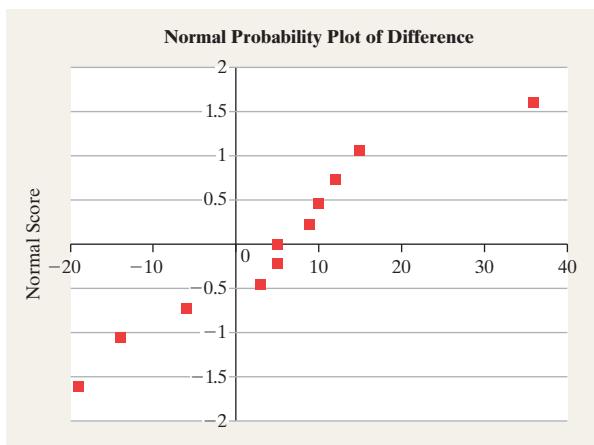
whether the car was filled up with 87-octane or 92-octane fuel first, and the driver did not know which type of fuel was in the tank. The results are in the following table:

Driver	1	2	3	4	5	6
Miles on 87 octane	234	257	243	215	114	287
Driver	7	8	9	10	11	
Miles on 87 octane	315	229	192	204	547	
Miles on 92 octane	351	241	186	209	562	

- (a) Why is it important that the matching be done by driver and car?
- (b) Why is it important to conduct the study on a closed track?
- (c) The normal probability plots for miles on 87 octane and miles on 92 octane are shown. The correlation between 87 octane and the expected z -scores is 0.877. The correlation between 92 octane and the expected z -scores is 0.879. Are either of these variables approximately normally distributed?



- (d) The differences are computed as “92 octane minus 87 octane.” The normal probability plot of the differences is shown in the next column. The correlation between the differenced data and the expected z -scores is 0.966. Is there reason to believe that the differences are normally distributed? Conclude that the differences can be normally distributed even though the original data are not.



- (e) The researchers used Minitab to determine whether the mileage from 92 octane is greater than the mileage from 87 octane. The results are as follows:

Test

Null hypothesis $H_0: \mu_{\text{difference}} = 0$

Alternative hypothesis $H_1: \mu_{\text{difference}} > 0$

T-Value	P-Value
1.14	0.141

What do you conclude? Why?

- DATA** 19. **Putting It Together: Glide Testing** You are a passenger in a single-propeller-driven aircraft that experiences engine failure in the middle of a flight. The pilot wants to maximize the distance that the plane can glide to increase the likelihood of finding a safe place to land. To accomplish this goal, should the pilot allow the propeller to “windmill” or should the pilot force the propeller to stop?

To obtain the data needed to answer the research question, a pilot climbed to 8000 feet at a speed of 60 knots and then killed the engine with the propeller either windmilling or stopped. Because the time to descend is directly proportional to glide distance, the time to descend to 7200 feet was recorded in seconds and used as a proxy for glide distance. The design called for randomly choosing the order in which the propeller would windmill or be stopped. The data in the table represent the time to descend 800 feet for each of 27 trials. **Note:** Visit www.aceaerobaticschool.com to see footage of this scenario.

Trial	Windmilling	Stopped	Trial	Windmilling	Stopped
1	73.4	82.3	15	64.2	82.5
2	68.9	75.8	16	67.5	81.1
3	74.1	75.7	17	71.2	72.3
4	71.7	71.7	18	75.6	77.7
5	74.2	68.8	19	73.1	82.6
6	63.5	74.2	20	77.4	79.5
7	64.4	78.0	21	77.0	82.3
8	60.9	68.5	22	77.8	79.5
9	79.5	90.6	23	77.0	79.7
10	74.5	81.9	24	72.3	73.4
11	76.5	72.9	25	69.2	76.0
12	70.3	75.7	26	63.9	74.2
13	71.3	77.6	27	70.3	79.0
14	72.7	174.3			

Source: Catherine Elizabeth Cavagnaro. “Glide Testing: A Paired Samples Experiment.” *Stats* 46, Fall 2006.

- (a) The trials took place over the course of a few days. However, for each trial, the pilot conducted both windmilling and stopped propeller one right after the other to minimize any impact of a change in weather conditions. Knowing this, explain why this is matched-pair data.
- (b) Why does the researcher randomly determine whether to windmill or stop the propeller first for each trial?
- (c) Explain why blinding is not possible for this experiment.
- (d) What is the response variable in the study? What are the treatments?
- (e) Compute the difference as “stopped – windmilling.” Draw a boxplot of the differenced data. What do you notice?
- (f) From part (e), you should notice that trial 14 results in an outlier. Because our sample size is small, this outlier will

have a major effect on any results. The author of the article indicated that it was possibly a situation in which there was an updraft of wind, causing the plane to take quite a bit longer than normal to fall 800 feet. Explain why this explanation makes it reasonable to eliminate trial 14 from the analysis.

- (g) Redraw a boxplot of the data with trial 14 eliminated. Based on the shape of the boxplot, do you believe it is reasonable to proceed with a matched-pair t -test?
- (h) The researchers wanted to determine if stopping the propeller resulted in a longer glide distance. Based on this goal, determine the null and alternative hypotheses.
- (i) Conduct the appropriate test to answer the researcher’s question.
- (j) Write a few sentences outlining your recommendations to pilots who experience engine failure.

11.3 Inference about Two Means: Independent Samples



Preparing for This Section Before getting started, review the following:

- The Completely Randomized Design (Section 1.6, pp. 48–49)
- Confidence intervals about μ (Section 9.2, pp. 415–418)

- Hypothesis tests about μ (Section 10.3, pp. 458–463)
- Type I and Type II errors (Section 10.1, pp. 438–440)

Objectives

- ① Test hypotheses regarding the difference of two independent means
- ② Construct and interpret confidence intervals regarding the difference of two independent means

We now turn our attention to inferential methods for comparing means from two independent samples. For example, suppose we want to know whether a new experimental drug relieves symptoms attributable to the common cold. The response variable might be the time until the cold symptoms go away—a quantitative variable. If the drug is effective, the mean time until the cold symptoms go away should be less for individuals taking the drug than for those not taking the drug. If we let μ_1 represent the mean time until cold symptoms go away for the individuals taking the drug and μ_2 represent the mean time until cold symptoms go away for individuals taking a placebo, the null and alternative hypotheses will be

$$H_0: \mu_1 = \mu_2 \quad \text{versus} \quad H_1: \mu_1 < \mu_2$$

or, equivalently,

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{versus} \quad H_1: \mu_1 - \mu_2 < 0$$

To conduct this test, we might randomly divide 500 volunteers who have a common cold into two groups: an experimental group (group 1) and a control group (group 2). The control group will receive a placebo and the experimental group will receive a predetermined amount of the experimental drug. Next, determine the time until the cold symptoms go away. Compute \bar{x}_1 , the sample mean time until cold symptoms go away in the experimental group, and \bar{x}_2 , the sample mean time until cold symptoms go away in the control group. Now we determine whether the difference in the sample means, $\bar{x}_1 - \bar{x}_2$, is significantly less than 0, the assumed difference stated in the null hypothesis. To do this, we need to know the sampling distribution of $\bar{x}_1 - \bar{x}_2$.

It is unreasonable to expect to know information regarding σ_1 and σ_2 without knowing information regarding the population means. Therefore, we must develop a sampling distribution for the difference of two means when the population standard deviations are unknown.

The comparison of two means with unequal (and unknown) population variances is called the Behrens–Fisher problem. While an exact method for performing inference on the equality of two means with unequal population standard deviations does not exist, an approximate solution is available. The approach that we use is known as **Welch's approximate *t***, in honor of English statistician Bernard Lewis Welch (1911–1989).

Sampling Distribution of the Difference of Two Means: Independent Samples with Population Standard Deviations Unknown (Welch's *t*)

Suppose that a simple random sample of size n_1 is taken from a population with unknown mean μ_1 and unknown standard deviation σ_1 . In addition, a simple random sample of size n_2 is taken from a second population with unknown mean μ_2 and unknown standard deviation σ_2 . If the two populations are normally distributed or the sample sizes are sufficiently large ($n_1 \geq 30$ and $n_2 \geq 30$), then

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (1)$$

approximately follows Student's *t*-distribution with the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom, where \bar{x}_1 is the sample mean and s_1 is the sample standard deviation from population 1, and \bar{x}_2 is the sample mean and s_2 is the sample standard deviation from population 2.

① Test Hypotheses Regarding the Difference of Two Independent Means

Now that we know the approximate sampling distribution of $\bar{x}_1 - \bar{x}_2$, we can introduce a procedure that can be used to test hypotheses regarding two population means.

Testing Hypotheses Regarding the Difference of Two Means

To test hypotheses regarding two population means, μ_1 and μ_2 , with unknown population standard deviations, use the following steps, provided that

- the samples are obtained using simple random sampling, or through a completely randomized experiment with two levels of treatment.
- the samples are independent.
- the populations from which the samples are drawn are normally distributed or the sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$).
- for each sample, the sample size is no more than 5% of the population size.

Step 1 Determine the null and alternative hypotheses.

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: \mu_1 = \mu_2$	$H_0: \mu_1 = \mu_2$	$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$	$H_1: \mu_1 < \mu_2$	$H_1: \mu_1 > \mu_2$

Note: μ_1 is the population mean for population 1, and μ_2 is the population mean for population 2.

Step 2 Select a level of significance α , depending on the seriousness of making a Type I error.

Classical Approach

Step 3 Compute the **test statistic**

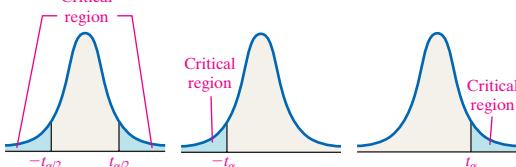
$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

P-Value Approach

By Hand Step 3 Compute the **test statistic**

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

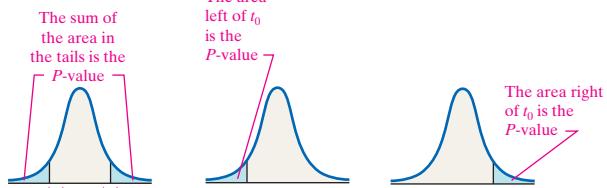
which approximately follows Student's t -distribution. Use Table VII to determine the critical value using the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom.

	Two-Tailed	Left-Tailed	Right-Tailed
Critical value(s)	$-t_{\frac{\alpha}{2}}$ and $t_{\frac{\alpha}{2}}$	$-t_{\alpha}$	t_{α}
Critical region			

Step 4 Compare the critical value to the test statistic.

Two-Tailed	Left-Tailed	Right-Tailed
If $t_0 < -t_{\frac{\alpha}{2}}$ or $t_0 > t_{\frac{\alpha}{2}}$, reject the null hypothesis.	If $t_0 < -t_{\alpha}$, reject the null hypothesis.	If $t_0 > t_{\alpha}$, reject the null hypothesis.

which approximately follows Student's t -distribution. Use Table VII to approximate the P -value using the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom.

Two-Tailed	Left-Tailed	Right-Tailed
		

Technology Step 3 Use a statistical spreadsheet or calculator with statistical capabilities to obtain the P -value. The directions for obtaining the P -value using the TI-83/84 Plus graphing calculators, Excel, Minitab, and StatCrunch are in the Technology Step-by-Step on page 508.

Step 4 If P -value $< \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

The procedure just presented is robust, so minor departures from normality will not adversely affect the results of the test. If the data have outliers, however, the procedure should not be used.

We verify these requirements by constructing normal probability plots (to assess normality) and boxplots (to determine whether there are outliers). If the normal probability plots indicate that the data come from populations that are not normally distributed or the boxplots reveal outliers, then randomization tests (Section 11.3A) or nonparametric tests (Section 15.4) should be performed.

EXAMPLE 1 Testing Hypotheses Regarding Two Means

Problem In the Spacelab Life Sciences 2 payload, 14 male rats were sent to space. Upon their return, the red blood cell mass (in milliliters) of the rats was determined. A control group of 14 male rats was held under the same conditions (except for space flight) as the space rats, and their red blood cell mass was also determined when the space rats returned. The project resulted in the data listed in Table 3. Does the evidence suggest that the flight animals have a different red blood cell mass from the control animals at the $\alpha = 0.05$ level of significance?

Table 3

Flight					Control				
8.59	8.64	7.43	7.21	6.39	8.65	6.99	8.40	9.66	7.14
6.87	7.89	9.79	6.85	7.54	7.62	7.44	8.55	8.70	9.14
7.00	8.80	9.30	8.03		7.33	8.58	9.88	9.94	

Source: NASA Life Sciences Data Archive.

Approach This experiment is a completely randomized design with response variable red blood cell mass. The treatment is space flight, which is set at two levels: space flight or no space flight. The experimental units are the 28 rats. The expectation is that all other variables that may affect red blood cell mass are accounted for by holding both groups of rats in the same condition (other than space flight) and through random assignment.

(continued)

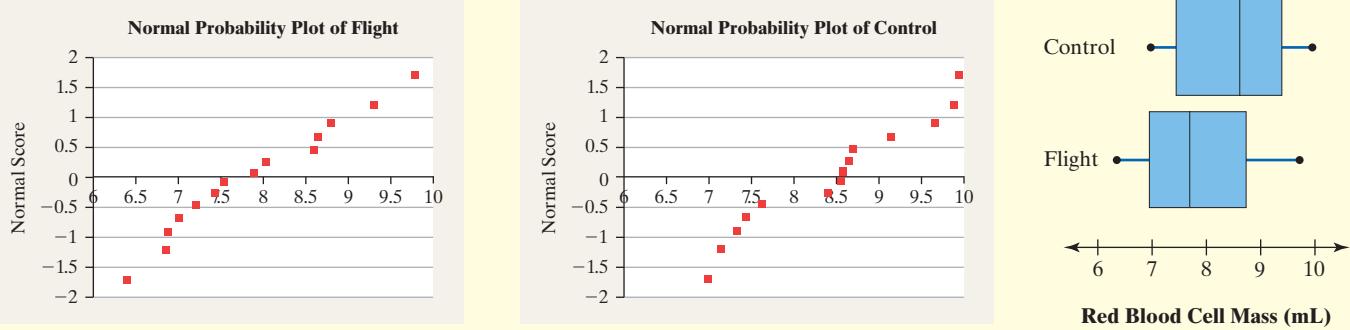
We are attempting to determine if the evidence suggests that space flight affects red blood cell mass. We assume no difference, so we assume the mean red blood cell mass of the flight group equals that of the control group. We want to show that the mean of the flight group is different from the mean of the control group.

Verify that each sample comes from a population that is approximately normal with no outliers by drawing normal probability plots and boxplots. The boxplots will be drawn on the same graph so that we can visually compare the two samples. Then follow Steps 1 through 5, listed on pages 502–503.

NOTE

The boxplots shown in Figure 10 are roughly symmetric. The robustness of Welch's *t*-test suggests the model requirements are satisfied.

Figure 10



Step 1 We want to know whether the flight animals have a different red blood cell mass from the control animals. Let μ_1 represent the mean red blood cell mass of the flight animals and μ_2 represent the mean red blood cell mass of the control animals. We are attempting to gather evidence that shows $\mu_1 \neq \mu_2$, and we have the hypotheses

$$\begin{array}{ll} H_0: \mu_1 = \mu_2 & H_0: \mu_1 - \mu_2 = 0 \\ \text{versus} & \text{versus} \\ H_1: \mu_1 \neq \mu_2 & H_1: \mu_1 - \mu_2 \neq 0 \end{array}$$

Step 2 The level of significance $\alpha = 0.05$.

Classical Approach

Step 3 The statistics for sample 1 (the flight rats) are $n_1 = 14$, $\bar{x}_1 = 7.881$, and $s_1 = 1.017$.

The statistics for sample 2 (the control rats) are $n_2 = 14$, $\bar{x}_2 = 8.430$, and $s_2 = 1.005$.

The test statistic is

$$\begin{aligned} t_0 &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{(7.881 - 8.430) - 0}{\sqrt{\frac{1.017^2}{14} + \frac{1.005^2}{14}}} \\ &= \frac{-0.549}{\sqrt{\frac{1.017^2}{14} + \frac{1.005^2}{14}}} \\ &= \frac{-0.549}{0.3821288115} \\ &= -1.437 \end{aligned}$$

P-Value Approach

By Hand Step 3 The statistics for sample 1 (the flight rats) are $n_1 = 14$, $\bar{x}_1 = 7.881$, and $s_1 = 1.017$. The statistics for sample 2 (the control rats) are $n_2 = 14$, $\bar{x}_2 = 8.430$, and $s_2 = 1.005$.

The test statistic is

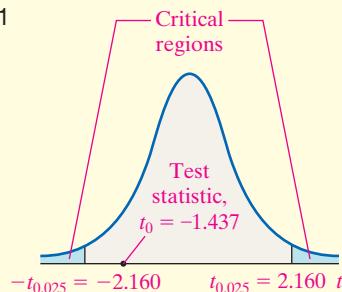
$$\begin{aligned} t_0 &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(7.881 - 8.430) - 0}{\sqrt{\frac{1.017^2}{14} + \frac{1.005^2}{14}}} \\ &= \frac{-0.549}{\sqrt{\frac{1.017^2}{14} + \frac{1.005^2}{14}}} \\ &= \frac{-0.549}{\sqrt{0.3821288115}} \\ &= -1.437 \end{aligned}$$

Because this is a two-tailed test, the *P*-value is the area under the *t*-distribution to the left of $t_0 = -1.437$ plus the area under the *t*-distribution to the right of $t_0 = 1.437$. See Figure 12.

Step 4 The test statistic is $t_0 = -1.437$. We label this point in Figure 11.

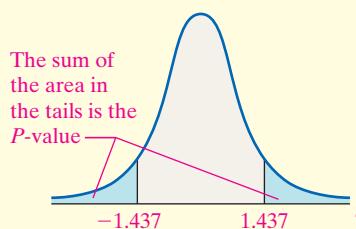
This is a two-tailed test with $\alpha = 0.05$. Since the sample sizes of the experimental group and control group are both 14, we have $n_1 - 1 = 14 - 1 = 13$ degrees of freedom. The critical values are $t_{\frac{\alpha}{2}} = t_{0.025} = 2.160$ and $-t_{0.025} = -2.160$. The critical region is displayed in Figure 11.

Figure 11



Because the test statistic does not lie within a critical region, we do not reject the null hypothesis.

Figure 12



Since the sample size of the experimental group and control group are both 14, we have $n_1 - 1 = 14 - 1 = 13$ degrees of freedom. Because of symmetry, use Table VII to estimate the area under the t -distribution to the right of $t_0 = 1.437$ and double it.

$$P\text{-value} = P(t_0 < -1.437 \text{ or } t_0 > 1.437) = 2P(t_0 > 1.437)$$

Using Table VII, find the row that corresponds to 13 degrees of freedom. The value 1.437 lies between 1.350 and 1.771. The area under the t -distribution with 13 degrees of freedom to the right of 1.350 is 0.10. The area under the t -distribution with 13 degrees of freedom to the right of 1.771 is 0.05. After doubling these values, we have

$$0.10 < P\text{-value} < 0.20$$

Technology Step 3 Using StatCrunch, we find the P -value is 0.1627. See Figure 13.

Figure 13

Hypothesis test results:
 μ_1 : Mean of Flight
 μ_2 : Mean of Control
 $\mu_1 - \mu_2$: Difference between two means
 H_0 : $\mu_1 - \mu_2 = 0$
 H_A : $\mu_1 - \mu_2 \neq 0$
 (without pooled variances)

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	-0.54928571	0.38230283	25.996352	-1.4367817	0.1627

Step 4 The P -value of 0.1627 [by hand: $0.10 < P\text{-value} < 0.20$] means that if the null hypothesis (the mean difference of red blood cell mass in the flight group and control group is zero) is true, we expect a sample difference at least as extreme as the one obtained in about 16 out of 100 repetitions of this experiment. The results obtained are not unusual if the statement in null hypothesis is true. Simply put, because the P -value is greater than the level of significance ($0.1627 > 0.05$), we do not reject the null hypothesis.

Step 5 The sample data does not provide sufficient evidence to conclude that the flight animals have a different red blood cell mass from the control animals at the $\alpha = 0.05$ level of significance.

The degrees of freedom used to determine the critical value(s) in the classical approach and by-hand P -value presented in Example 1 are conservative (this means we would require more evidence against the statement in the null hypothesis than if we use the formula for degrees of freedom given below). Results that are more accurate can be obtained by using the following formula for degrees of freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(s_1^2 \right)^2}{n_1 - 1} + \frac{\left(s_2^2 \right)^2}{n_2 - 1}} \quad (2)$$

CAUTION!

The degrees of freedom in by-hand solutions will not equal the degrees of freedom in technology solutions unless you use Formula (2) to compute degrees of freedom.

When using Formula (2) to compute degrees of freedom, round down to the nearest integer to use Table VII. For by-hand inference, it is recommended that you use the smaller of $n_1 - 1$ or $n_2 - 1$ as the degrees of freedom to ease computation. However, for increased precision in determining the P -value, computer software will use Formula (2) when computing the degrees of freedom.

Notice that the degrees of freedom in the technology solution are 25.996352 versus 13 in the conservative solution done by hand in Example 1. With the lower degrees of freedom, the critical t is larger (2.160 with 13 degrees of freedom versus 2.056 with approximately 26 degrees of freedom). The larger critical value increases the number of standard deviations the difference in the sample means must be from the hypothesized mean difference before the null hypothesis is rejected. Therefore, in using the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom, we need more substantial evidence to reject the null hypothesis. This requirement decreases the probability of a Type I error (rejecting the null hypothesis when the null hypothesis is true) below the actual level of α chosen by the researcher. This is what we mean when we say that the method of using the lesser of $n_1 - 1$ and $n_2 - 1$ as a proxy for degrees of freedom is conservative compared with using Formula (2).

NW Now Work Problem 13

② Construct and Interpret Confidence Intervals Regarding the Difference of Two Independent Means

Constructing a confidence interval for the difference of two means is an extension of the results presented in Section 9.2.

Constructing a $(1 - \alpha) \cdot 100\%$ Confidence Interval for the Difference of Two Means

A simple random sample of size n_1 is taken from a population with unknown mean μ_1 and unknown standard deviation σ_1 . Also, a simple random sample of size n_2 is taken from a population with unknown mean μ_2 and unknown standard deviation σ_2 . If the two populations are normally distributed or the sample sizes are sufficiently large ($n_1 \geq 30$ and $n_2 \geq 30$), a $(1 - \alpha) \cdot 100\%$ confidence interval about $\mu_1 - \mu_2$ is given by

$$\text{Lower bound: } (\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and

$$\text{Upper bound: } (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (3)$$

where $t_{\frac{\alpha}{2}}$ is computed using the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom or Formula (2).

EXAMPLE 2

Constructing a Confidence Interval for the Difference of Two Means

Problem A Gallup poll of 513 national adults aged 18 years or older who consider themselves to be Republican asked, “Of every tax dollar that goes to the federal government in Washington, D.C., how many cents of each dollar would you say are wasted?” The mean amount wasted was found to be 54 cents with a standard deviation of 2.9 cents. The same question was asked of 513 national adults aged 18 years or older who consider themselves to be Democrat. The mean amount wasted was found to be 41 cents with a standard deviation of 2.6 cents. Construct a 95% confidence interval for the mean difference in government waste, $\mu_R - \mu_D$. Interpret the interval.

Approach Verify the model requirements. If satisfied, construct a confidence interval for the difference of two independent means using Formula (3) or technology.

Solution The samples are independent and obtained through a simple random sample. The sample sizes for both Republicans and Democrats are large, and the sample size is small relative to the size of the population ($n \leq 0.05N$). The requirements for constructing a confidence interval for $\mu_R - \mu_D$ using Student's t -distribution are satisfied.

By Hand The sample statistics are provided in the problem and summarized below:

$$\bar{x}_R = 54, s_R = 2.9, \bar{x}_D = 41, s_D = 2.6$$

There were $n_R = n_D = 513$ Republicans and Democrats surveyed. Using 100 degrees of freedom (the highest available without going over), the critical t -value is $t_{0.025} = 1.984$.

$$\begin{aligned}\text{Lower Bound: } & (\bar{x}_R - \bar{x}_D) - t_{\alpha/2} \cdot \sqrt{\frac{s_R^2}{n_R} + \frac{s_D^2}{n_D}} \\ &= (54 - 41) - 1.984 \cdot \sqrt{\frac{2.9^2}{513} + \frac{2.6^2}{513}} \\ &= 13 - 0.3 \\ &= 12.7\end{aligned}$$

$$\begin{aligned}\text{Upper Bound: } & (\bar{x}_R - \bar{x}_D) + t_{\alpha/2} \cdot \sqrt{\frac{s_R^2}{n_R} + \frac{s_D^2}{n_D}} \\ &= (54 - 41) + 1.984 \cdot \sqrt{\frac{2.9^2}{513} + \frac{2.6^2}{513}} \\ &= 13 + 0.3 \\ &= 13.3\end{aligned}$$

Technology The sample statistics are provided in the problem and summarized below:

$$\bar{x}_R = 54, s_R = 2.9, \bar{x}_D = 41, s_D = 2.6$$

There were $n_R = n_D = 513$ Republicans and Democrats surveyed. Enter these values into the statistical software you are using to obtain the 95% confidence interval. Figure 14 shows the results from Minitab.

Figure 14

Descriptive Statistics				
Sample	N	Mean	StDev	SE Mean
Sample 1	513	54.00	2.90	0.13
Sample 2	513	41.00	2.60	0.11

Estimation for Difference	
Difference	95% CI for Difference
13.000	(12.663, 13.337)

We are 95% confident the Republicans believe the mean amount of each dollar collected by the federal government that is wasted is between 12.7 cents and 13.3 cents higher than the mean amount Democrats believe is wasted.

CAUTION!

We would use the pooled two-sample t -test when the two samples come from populations that have the same variance. *Pooling* refers to finding a weighted average of the two sample variances from the independent samples. It is difficult to verify that two population variances might be equal based on sample data, so we will always use Welch's t when comparing two means.

What about the Pooled Two-Sample t -Tests?

Perhaps you have noticed that statistical software and graphing calculators with advanced statistical features provide an option for two types of two-sample t -tests: one that assumes equal population variances (pooling) and one that does not assume equal population variances. Welch's t -statistic does not assume that the population variances are equal and can be used whether the population variances are equal or not. The test that assumes equal population variances is referred to as the *pooled t -statistic*.

The **pooled t -statistic** is computed by finding a weighted average of the sample variances and uses this average in the computation of the test statistic. The advantage of this test statistic is that it exactly follows Student's t -distribution with $n_1 + n_2 - 2$ degrees of freedom.

The disadvantage of the test statistic is that it requires that the population variances be equal. How is this requirement to be verified? While a test for determining the equality of variances does exist (F -test, Section 11.5 in the eText), the test *requires* that each population be normally distributed. So, the F -test is not robust. Any minor departures from normality will make the results of the F -test unreliable. It has been recommended by many statisticians* that a preliminary F -test to check the requirement of equality of variance not be performed. In fact, George Box once said, "To make preliminary tests on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!"

Because the formal F -test for testing the equality of variances is so volatile, we are content to use Welch's t . Welch's t -test is more conservative than the pooled t . The price

*Moser and Stevens, "Homogeneity of Variance in the Two-Sample Means Test." *American Statistician* 46(1).

that must be paid for the conservative approach is that the probability of a Type II error is higher with Welch's t than with the pooled t when the population variances are equal. However, the two tests typically provide the same conclusion, even if the assumption of equal population standard deviations seems reasonable.

Technology Step-by-Step

Two-Sample t -Tests, Independent Sampling

TI-83/84 Plus

Hypothesis Tests

1. If necessary, enter raw data in L1 and L2.
2. Press STAT, highlight TESTS, and select 4:2-SampTTest
3. If the data are raw, highlight Data, making sure that List1 is set to L1 and List2 is set to L2, with frequencies set to 1. If summary statistics are known, highlight Stats and enter the summary statistics.
4. Highlight the appropriate relation between μ_1 and μ_2 in the alternative hypothesis. Set Pooled to No.
5. Highlight Calculate or Draw and press ENTER. Calculate gives the test statistic and P -value. Draw will draw the t -distribution with the P -value shaded.

Confidence Intervals

Follow the steps given for hypothesis tests, except select Ø: 2-SampTInt. Also, select a confidence level (such as 95% = 0.95).

Minitab

1. If necessary, enter raw data in columns C1 and C2.
2. Select the Stat menu, highlight **Basic Statistics**, then highlight **2-Sample t**
3. If you have raw data, select "Each sample is in its own column." Enter C1 in the cell marked "Sample 1" and enter C2 in the cell marked "Sample 2." If you have summarized data, select "Summarized data." Enter the summary statistics in the appropriate cell. Click Options Select the direction of the alternative

hypothesis, enter the "test difference" (usually zero), and select a confidence level. Click OK twice.

Excel

1. Enter the raw data into Columns A and B.
2. Select the Formulas menu. Select More Functions. Highlight Statistical, and select T.TEST from the drop-down menu.
3. Place the cursor in Array1. Highlight the data in Column A. Place the cursor in Array2. Highlight the data in Column B. Place the cursor in the Tails cell. Enter the number corresponding to the test you desire (1 for a one-tailed distribution; 2 for a two-tailed distribution). Place the cursor in the Type cell. Enter 3 for a two-sample unequal variance test.

StatCrunch

1. If necessary, enter the raw data into the first two columns of the spreadsheet. Name the column variables.
2. Select **Stat**, highlight **T Stats**, select **Two Sample**, and then choose either **With Data** or **With Summary**.
3. Select the column that contains the data for Sample 1. Select the column that contains the data for Sample 2. Note that the differences are computed Sample 1 – Sample 2. Leave the box "Pool variances" unchecked. If you choose the hypothesis test radio button, enter the value of the mean stated in the null hypothesis and choose the direction of the alternative hypothesis from the pull-down menu. If you choose the confidence interval radio button, enter the level of confidence. Click Compute!.



11.3 Assess Your Understanding

Skill Building*

In Problems 1–6, assume that the populations are normally distributed.

1. (a) Test whether $\mu_1 \neq \mu_2$ at the $\alpha = 0.05$ level of significance for the given sample data.
(b) Construct a 95% confidence interval about $\mu_1 - \mu_2$.

	Sample 1	Sample 2
n	15	15
\bar{x}	15.3	14.2
s	3.2	3.5

2. (a) Test whether $\mu_1 \neq \mu_2$ at the $\alpha = 0.05$ level of significance for the given sample data.
(b) Construct a 95% confidence interval about $\mu_1 - \mu_2$.

	Sample 1	Sample 2
n	20	20
\bar{x}	111	104
s	8.6	9.2

3. (a) Test whether $\mu_1 > \mu_2$ at the $\alpha = 0.1$ level of significance for the given sample data.
(b) Construct a 90% confidence interval about $\mu_1 - \mu_2$.

	Sample 1	Sample 2
n	25	18
\bar{x}	50.2	42.0
s	6.4	9.9

*The by-hand confidence intervals in the back of the text were computed using the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom.

4. (a) Test whether $\mu_1 < \mu_2$ at the $\alpha = 0.05$ level of significance for the given sample data.
 (b) Construct a 95% confidence interval about $\mu_1 - \mu_2$.

Sample 1	Sample 2	
n	40	32
\bar{x}	94.2	115.2
s	15.9	23.0

5. Test whether $\mu_1 < \mu_2$ at the $\alpha = 0.02$ level of significance for the given sample data.

Sample 1	Sample 2	
n	32	25
\bar{x}	103.4	114.2
s	12.3	13.2

6. Test whether $\mu_1 > \mu_2$ at the $\alpha = 0.05$ level of significance for the given sample data.

Sample 1	Sample 2	
n	23	13
\bar{x}	43.1	41.0
s	4.5	5.1

Applying the Concepts

7. **Elapsed Time to Earn a Bachelor's Degree** A researcher with the Department of Education followed a cohort of students who graduated from high school in 1992, monitoring the progress the students made toward completing a bachelor's degree. One aspect of his research was to determine whether students who first attended community college took longer to attain a bachelor's degree than those who immediately attended and remained at a 4-year institution. The data in the table summarize the results of his study.

	Community College to Four-Year Transfer	No Transfer
n	268	1145
Sample mean time to graduate, in years	5.43	4.43
Sample standard deviation time to graduate, in years	1.162	1.015

Source: Clifford Adelman, *The Toolbox Revisited*. United States Department of Education, 2006.

- (a) What is the response variable in this study? What is the explanatory variable?
 (b) Explain why this study can be analyzed using the methods of this section.
 (c) Does the evidence suggest that community college transfer students take longer to attain a bachelor's degree? Use an $\alpha = 0.01$ level of significance.
 (d) Construct a 95% confidence interval for $\mu_{\text{community college}} - \mu_{\text{no transfer}}$ to approximate the mean additional time it takes to complete a bachelor's degree if you begin in community college.
 (e) Do the results of parts (c) and (d) imply that community college causes you to take extra time to earn a bachelor's degree? Cite some reasons that you think might contribute to the extra time to graduate.

8. **Influence of Marriage on Testosterone Levels** Testosterone is a steroid in men that affects sex drive, bone, and muscle mass. Researchers wanted to determine the impact of marriage on testosterone levels in males. Source: Stine A. Hogmoe et al., "Influence of Marital Status on Testosterone Levels—A Ten Year Follow-Up of 1113 Men" *Psychoneuroendocrinology* 80(2017):155–161.

- (a) The researchers obtained baseline measures of married and unmarried males in the study. The testosterone levels of both groups are shown in the table below. Does the sample evidence suggest the testosterone levels of the married men are lower than the testosterone levels of the unmarried men at the $\alpha = 0.05$ level of significance?

Married	Unmarried	
\bar{x}	20.1 nmol/L	22.1 nmol/L
s	7.2	7.5 nmol/L
n	1454	245

- (b) Testosterone levels are known to decline with age. Does the sample evidence suggest the married males are older than the unmarried males at the $\alpha = 0.05$ level of significance? What do these results suggest about any conclusions drawn in part (a)?

Married	Unmarried	
\bar{x}	46.9	36.5
s	10.6	9.8
n	1454	245

- (c) The results of parts (a) and (b) suggest that we cannot simply compare testosterone levels of married and unmarried males. To determine the role that marriage may play in testosterone levels, the researchers computed the decline in testosterone over a 10-year period. The table below shows the amount testosterone levels declined among males who transitioned from married to unmarried and the decline in testosterone among males who transitioned from unmarried to married within the 10-year period. Do the results suggest that males who transitioned from unmarried to married experience a greater decline in testosterone levels than those who transitioned from married to unmarried at the $\alpha = 0.05$ level of significance? Note: The males who transitioned from unmarried to married were significantly younger than those who transitioned from married to unmarried.

Married to Unmarried	Unmarried to Married	
\bar{x}	2.3	6.6
s	7.3	5.2
n	67	81

9. **Walking in the Airport, Part I** Do people walk faster in the airport when they are departing (getting on a plane) or when they are arriving (getting off a plane)? Researcher Seth B. Young measured the walking speed of travelers in San Francisco International Airport and Cleveland Hopkins International Airport. His findings are summarized in the table.

Direction of Travel	Departure	Arrival
Mean speed (feet per minute)	260	269
Standard deviation (feet per minute)	53	34
Sample size	35	35

Source: Seth B. Young. "Evaluation of Pedestrian Walking Speeds in Airport Terminals." *Transportation Research Record*. Paper 99-0824.

- (a) Is this an observational study or a designed experiment? Why?
 (b) Explain why it is reasonable to use Welch's t -test.
 (c) Do individuals walk at different speeds depending on whether they are departing or arriving at the $\alpha = 0.05$ level of significance?

10. Walking in the Airport, Part II Do business travelers walk at a different pace than leisure travelers? Researcher Seth B. Young measured the walking speed of business and leisure travelers in San Francisco International Airport and Cleveland Hopkins International Airport. His findings are summarized in the table.

Type of Traveler	Business	Leisure
Mean speed (feet per minute)	272	261
Standard deviation (feet per minute)	43	47
Sample size	20	20

Source: Seth B. Young. "Evaluation of Pedestrian Walking Speeds in Airport Terminals." *Transportation Research Record*. Paper 99-0824.

- (a) Is this an observational study or a designed experiment? Why?
 (b) What must be true regarding the populations to use Welch's t -test to compare the means?
 (c) Assuming that the requirements listed in part (b) are satisfied, determine whether business travelers walk at a different speed from leisure travelers at the $\alpha = 0.05$ level of significance.

11. Priming Two Dutch researchers conducted a study in which two groups of students were asked to answer 42 questions from Trivial Pursuit. The students in group 1 were asked to spend 5 minutes thinking about what it would mean to be a professor, while the students in group 2 were asked to think about soccer hooligans. The 200 students in group 1 had a mean score of 23.4 with a standard deviation of 4.1, while the 200 students in group 2 had a mean score of 17.9 with a standard deviation of 3.9.

Source: Based on the research of Dijksterhuis, Ap, and Ad van Knippenberg. "The relation between perception and behavior, or how to win a game of Trivial Pursuit." *Journal of Personality and Social Psychology* 74.4 (1998): 865+. Academic OneFile. Web. 5 July 2010.

- (a) Determine the 95% confidence interval for the difference in scores, $\mu_1 - \mu_2$. Interpret the interval.
 (b) What does this say about priming?

12. Popcorn Researchers Brian Wansink and Junyong Kim randomly gave 157 moviegoers a free medium (120 grams) or large (250 gram) bucket of popcorn before entering a movie. After the show, the researchers measured how much popcorn the moviegoers consumed. The 77 individuals randomly assigned the medium bucket had a mean consumption of 58.9 grams with a standard deviation of 16.7 grams. The 80 individuals randomly assigned the large bucket had a mean consumption of 85.6 grams with a standard deviation of 14.1 grams. With 95% confidence, determine how much more popcorn was consumed by individuals given the large bucket of popcorn. What is the implication? Source: Wansink, B. Junyong, K. "Bad Popcorn in Big Buckets: Portion Size Can Influence Intake as Much as Taste." *Journal of Nutrition Education & Behavior*, September 2005; 35(5):242–245.

NW 13. Ramp Metering Ramp metering is a traffic engineering idea that requires cars entering a freeway to stop for a certain period of time before joining the traffic flow. The theory is that ramp metering controls the number of cars on the freeway and the number of cars accessing the freeway, resulting in a freer flow of cars, which ultimately results in faster travel times. To test whether ramp metering is effective in reducing travel times, engineers in Minneapolis, Minnesota, conducted an experiment

in which a section of freeway had ramp meters installed on the on-ramps. The response variable for the study was speed of the vehicles. A random sample of 15 cars on the highway for a Monday at 6 P.M. with the ramp meters on and a second random sample of 15 cars on a different Monday at 6 P.M. with the meters off resulted in the following speeds (in miles per hour).

Ramp Meters On	Ramp Meters Off	
28	48	56
38	31	25
43	46	50
35	55	40
42	26	47
	24	26
	34	37
	47	38
	29	23
	37	52
		41

- (a) Draw side-by-side boxplots of each data set. Does there appear to be a difference in the speeds? Are there any outliers?
 (b) Are the ramp meters effective in maintaining a higher speed on the freeway? Use the $\alpha = 0.10$ level of significance.

Note: Normal probability plots indicate the data could come from a population that is normally distributed.

14. Measuring Reaction Time Researchers wanted to determine whether the reaction time (in seconds) of males differed from that of females to a go/no go stimulus. The researchers randomly selected 20 females and 15 males to participate in the study. The go/no go stimulus required the student to respond to a particular stimulus and not to respond to other stimuli. The results are as follows:

Female Students				Male Students		
0.588	0.652	0.442	0.293	0.375	0.256	0.427
0.340	0.636	0.391	0.367	0.654	0.563	0.405
0.377	0.646	0.403	0.377	0.374	0.465	0.402
0.380	0.403	0.617	0.434	0.373	0.488	0.337
0.443	0.481	0.613	0.274	0.224	0.477	0.655

Source: PsychExperiments at the University of Mississippi.

- (a) Is it reasonable to use Welch's t -test? Why?
 (b) Test whether there is a difference in the reaction times of males and females at the $\alpha = 0.05$ level of significance.
 (c) Draw boxplots of each data set using the same scale. Does this visual evidence support the results obtained in part (b)?

15. Bacteria in Hospital Carpeting Researchers wanted to determine if carpeted rooms contained more bacteria than uncarpeted rooms. To determine the amount of bacteria in a room, researchers pumped the air from the room over a Petri dish at the rate of 1 cubic foot per minute for eight carpeted rooms and eight uncarpeted rooms. Colonies of bacteria were allowed to form in the 16 Petri dishes. The results are presented in the table. A normal probability plot and boxplot indicate that the data are approximately normally distributed with no outliers. Do carpeted rooms have more bacteria than uncarpeted rooms at the $\alpha = 0.05$ level of significance?

Carpeted Rooms (bacteria/cubic foot)				Uncarpeted Rooms (bacteria/cubic foot)			
11.8	10.8	7.1	14.6	12.1	12.0	3.8	10.1
8.2	10.1	13.0	14.0	8.3	11.1	7.2	13.7

Source: William G. Walter and Angie Stober. "Microbial Air Sampling in a Carpeted Hospital." *Journal of Environmental Health*, 30 (1968), p. 405.

- DATA 16. Visual versus Textual Learners** Researchers wanted to know whether there was a difference in comprehension among students learning a computer program based on the style of the text. They randomly divided 36 students of similar educational level, age, and so on, into two groups of 18 each. Group 1 individuals learned the software using a visual manual (*multimodal instruction*), while group 2 individuals learned the software using a textual manual (*unimodal instruction*). The following data represent scores that the students received on an exam given to them after they studied from the manuals.

Visual Manual	Textual Manual
51.08	60.35
57.03	76.60
44.85	70.77
75.21	70.15
56.87	47.60
75.28	46.59
57.07	81.23
80.30	67.30
52.20	60.82
	49.57
	61.16

Source: Mark Gellevij et al. "Multimodal versus Unimodal Instruction in a Complex Learning Context," *Journal of Experimental Education* 70(3):215–239, 2002.

- (a) What type of experimental design is this?
- (b) What are the treatments?
- (c) A normal probability plot and boxplot indicate it is reasonable to use Welch's *t*-test. Is there a difference in test scores at the $\alpha = 0.05$ level of significance?

- DATA 17. Threaded Problem: Tornado** The data set "Tornadoes_2017" located at www.pearsonhighered.com/sullivanstats contains a variety of variables that were measured for all tornadoes in the United States in 2017.

- (a) Is there a difference in the length of tornadoes in Texas versus Georgia? The data in column "Length" represent the length of the tornado (in miles). Test whether the length of tornadoes that occur in Texas is different from the length of tornadoes that occur in Georgia. **Note:** Treat the tornadoes that struck in Texas and Georgia as a simple random sample of all tornadoes that struck Texas and Georgia since 1950 using a 0.05 level of significance. To conduct the hypothesis test in StatCrunch, select Stat > T Stats > Two Sample > With Data. In the "Where:" box type "State = TX" or "State = GA." The summarized data show that $\bar{x}_T = 2.721$ miles; $\bar{x}_G = 5.346$ miles; $s_T = 4.362$ miles; $s_G = 7.726$ miles; $n_T = 168$; $n_G = 118$.
- (b) Estimate the difference in the length of tornadoes that occur in Texas and Georgia (compute the difference in means as Georgia – Texas) with 95% confidence. Interpret the result.

- 18. Putting It Together: Wait Times at Disney** The website touringplans.com records actual wait times (in minutes) for the Pirates of the Caribbean ride and Splash Mountain ride at Walt Disney World. Do the wait times at these two rides differ?

- DATA 19. Rates of Returns of Stocks** Stocks may be categorized by sectors. Go to www.pearsonhighered.com/sullivanstats to obtain the data file 11_3_19 using the file format of your choice for the version of the text you are using. The data represent the one-year rate of return (in percent) for a sample of consumer cyclical stocks and industrial stocks for the period December, 2013, through November, 2014. **Note:** Consumer cyclical stocks include names such as Starbucks and Home Depot. Industrial stocks include names such as 3M and FedEx.

- (b) A flaw in the analysis from part (a) is that it did not consider the date and time the wait time was measured. Explain how date and time might impact wait times at the two rides.
- (c) The data set 11_3_18c located at www.pearsonhighered.com/sullivanstats represents a random sample of wait times at each ride on the same date and at the same time of day. Explain how this is matched-pairs data.
- (d) Test whether the wait times at the two rides differ. **Note:** A normal probability plot suggests the differenced data is approximately normal and a boxplot shows no outliers.
- (e) Answer the following for the data in 11_3_18c. Compute the sample mean difference and standard error of the mean difference treating the data as an independent sample. Compute the sample mean difference and standard error of the mean difference treating the data as a dependent sample. What do the results suggest?

- DATA 19. Rates of Returns of Stocks** Stocks may be categorized by sectors. Go to www.pearsonhighered.com/sullivanstats to obtain the data file 11_3_19 using the file format of your choice for the version of the text you are using. The data represent the one-year rate of return (in percent) for a sample of consumer cyclical stocks and industrial stocks for the period December, 2013, through November, 2014. **Note:** Consumer cyclical stocks include names such as Starbucks and Home Depot. Industrial stocks include names such as 3M and FedEx.

- (a) Draw side-by-side boxplots of one-year rate of return by sector. Does there appear to be a difference in the one-year rate of return for these two sectors?
- (b) Explain why the methods of this section may be used to test whether the mean rate of return for the two sectors differ.
- (c) Test whether the mean one-year rate of return for consumer cyclical stocks is different from that of industrial stocks at the $\alpha = 0.05$ level of significance.
- (d) Construct a 95% confidence interval for the mean difference in rate of return between industrial stocks and consumer cyclical stocks. Interpret the interval.

- DATA 20. Tax Rates** Do women feel differently from men when it comes to federal tax rates? One question on the Sullivan Statistics Survey II was, "What percent of income do you believe individuals should pay in federal income tax?" Results of the survey may be found at www.pearsonhighered.com/sullivanstats. Select the data file SullivanSurveyII using the file format of your choice for the version of the text you are using. The Tax Rate column contains the response.

- (a) Draw side-by-side boxplots of tax rates by gender. Does there appear to be a difference in the tax rates for the genders?
- (b) Explain why the methods of this section may be used to test whether the mean tax rates for the two genders differ.
- (c) Test whether the mean tax rate for females differs from that of males at the $\alpha = 0.05$ level of significance.

- NW 21. Kids and Leisure** Young children require a lot of time and this time commitment cuts into a parent's leisure time. A sociologist wanted to estimate the difference in the amount of daily leisure time (in hours) of adults who do not have children under the age of 18 years and adults who have children under the age of 18 years. A random sample of 40 adults with no children under the age of 18 years results in a mean daily leisure time of 5.62 hours, with a standard deviation of 2.43 hours. A random sample of 40 adults with children under the age of 18 years results in a mean daily leisure time of 4.10 hours, with a standard deviation of 1.82 hours. Construct and interpret a 90% confidence interval for the mean difference in leisure time between adults with no children and adults with children. **Source:** American Time Use Survey.

22. Aluminum Bottles The aluminum bottle, first introduced in 1991 by CCL Container for mainly personal and household items such as lotions, has become popular with beverage manufacturers. Besides being lightweight and requiring less packaging, the aluminum bottle is reported to cool faster and stay cold longer than typical glass bottles. A small brewery tests this claim and obtains the following information regarding the time (in minutes) required to chill a bottle of beer from room temperature (75°F) to serving temperature (45°F). Construct and interpret a 90% confidence interval for the mean difference in cooling time for clear glass versus aluminum.

	Clear Glass	Aluminum
Sample size	42	35
Mean time to chill (minutes)	133.8	92.4
Sample standard deviation (minutes)	9.9	7.3

23. Putting It Together: Online Homework Professor Stephen Zuro of Joliet Junior College wanted to determine whether an online homework system (meaning students did homework online and received instant feedback with helpful guidance about their answers) improved scores on a final exam. In the fall semester, he taught a precalculus class using an online homework system. In the spring semester, he taught a precalculus class without the homework system (which meant students were responsible for doing their homework the old-fashioned way—paper and pencil). Professor Zuro made sure to teach the two courses identically (same text, syllabus, tests, meeting time, meeting location, and so on). The table summarizes the results of the two classes on their final exam.

	Fall Semester	Spring Semester
Number of students	27	25
Mean final exam score	73.6	67.9
Standard deviation final exam score	10.3	12.4

- (a) What type of experimental design is this?
- (b) What is the response variable? What are the treatments in the study?
- (c) What factors are controlled in the experiment?
- (d) In many experiments, the researcher will recruit volunteers and randomly assign the individuals to a treatment group. In what regard was this done for this experiment?
- (e) Did the students perform better on the final exam in the fall semester? Use an $\alpha = 0.05$ level of significance.
- (f) Can you think of any factors that may confound the results? Could Professor Zuro have done anything about these confounding factors?

Retain Your Knowledge



- 24. Graduation Rates** PayScale reports statistics on colleges and universities. Go to www.pearsonhighered.com/sullivanstats to obtain the data file 11_3_24 using the file format of your choice for the version of the text you are using. The data contain the

four-year cost and graduation rate for over 1300 colleges and universities. Do schools that charge more have higher graduation rates? The variable “4 Year Cost” represents the four-year cost of attending the college or university. The variable “Grad Rate” represents the percentage of incoming freshman who graduate within six years.

- (a) Draw a scatter diagram treating “4 Year Cost” as the explanatory variable and “Grad Rate” as the response variable.
- (b) Determine the correlation coefficient between “4 Year Cost” and “Grad Rate.”
- (c) Is there a linear relation between “4 Year Cost” and “Grad Rate”?
- (d) Find the least-squares regression line.
- (e) Is the graduation rate of Harvey Mudd College higher than would be expected among all schools that charge \$272,000? Explain.
- (f) What proportion of the variability in graduation rates is explained by the cost of attending?

Explaining the Concepts

25. College Skills The Collegiate Learning Assessment Plus (CLA+) is an exam that is meant to assess the intellectual gains made between one’s freshman and senior year of college. The exam, graded on a scale of 400 to 1600, assesses critical thinking, analytical reasoning, document literacy, writing, and communication. The exam was administered to 135 freshman in Fall 2012 at California State University Long Beach (CSULB). The mean score on the exam was 1191 with a standard deviation of 187. The exam was also administered to graduating seniors of CSULB in Spring 2013. The mean score was 1252 with a standard deviation of 182. Explain the type of analysis that could be applied to these data to assess whether CLA+ scores increase while at CSULB. Explain the shortcomings in the data available and provide a better data collection technique.

26. MythBusters In a MythBusters episode, the question was asked, “Which is better? A four-way stop or a roundabout?” “Better” was determined based on determining the number of vehicles that travel through the four-way stop over a 5-minute interval of time. Suppose the folks at MythBusters conducted this experiment 15 times for each intersection design.

- (a) What is the variable of interest in this study? Is it qualitative or quantitative?
- (b) Explain why the data might be analyzed by comparing two independent means. Include in this explanation what the null and alternative hypotheses would be and what each mean represents.
- (c) A potential improvement on the experimental design might be to identify 15 groups of different drivers and ask each group to drive through each intersection design for the 5-minute time interval. Explain why this is a better design and explain the role randomization would play. What would be the null and alternative hypotheses here and how would the mean be computed?

- 27.** Explain why using the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom to determine the critical t instead of Formula (2) is conservative.

11.4 Putting It Together: Which Method Do I Use?

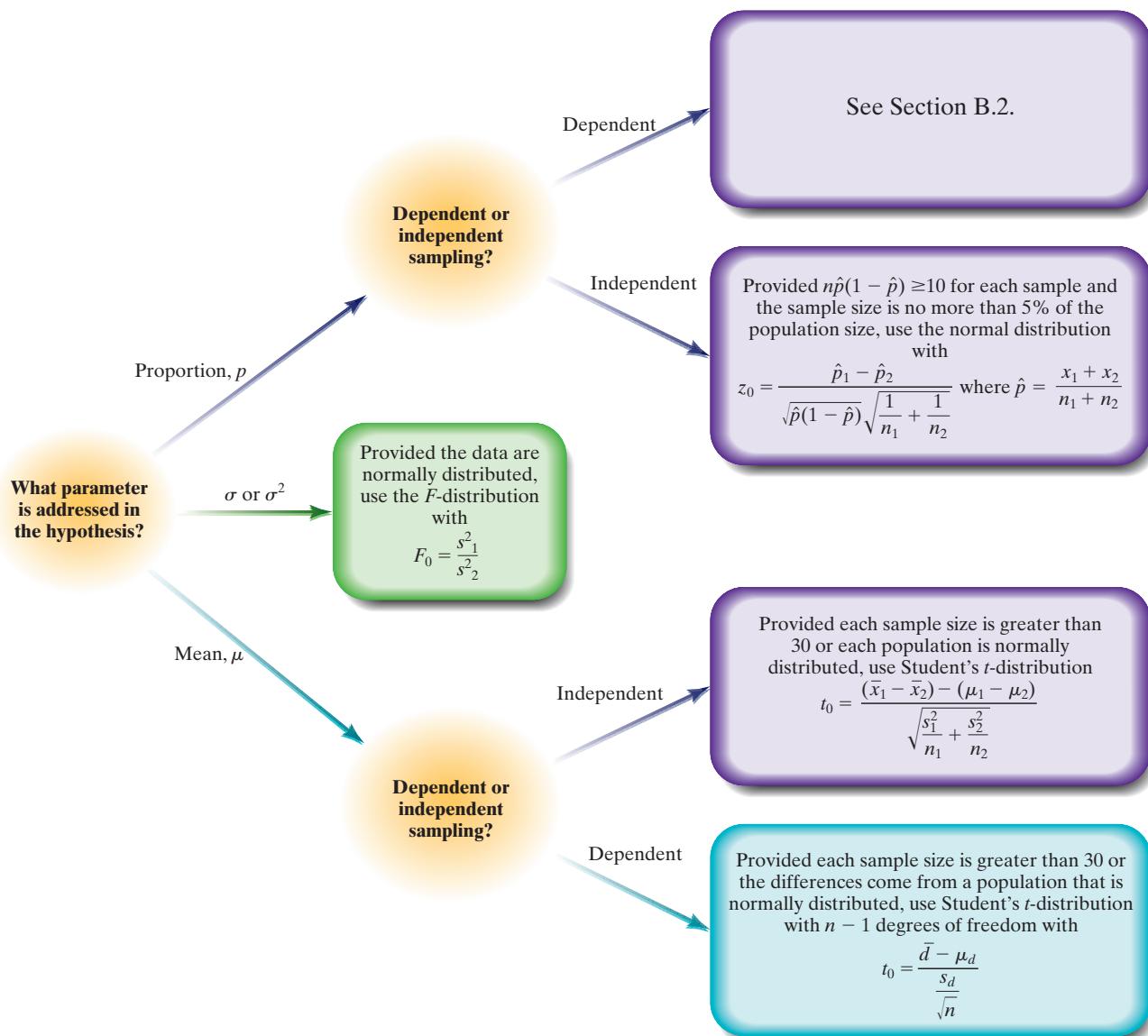


Objective ① Determine the appropriate hypothesis test to perform

① Determine the Appropriate Hypothesis Test to Perform

The ability to recognize the type of test to perform is one of the most important aspects of statistical analysis. To help decide which test to use when conducting inference on two samples, we provide the flowchart in Figure 15. Use it to assist you in the problems that follow.

Figure 15





11.4 Assess Your Understanding

Skill Building

In Problems 1–8, perform the appropriate hypothesis test.

1. A random sample of $n_1 = 120$ individuals results in $x_1 = 43$ successes. An independent sample of $n_2 = 130$ individuals results in $x_2 = 56$ successes. Does this represent sufficient evidence to conclude that $p_1 \neq p_2$ at the $\alpha = 0.01$ level of significance?
2. A random sample of size $n = 13$ obtained from a population that is normally distributed results in a sample mean of 45.3 and sample standard deviation of 12.4. An independent sample of size $n = 18$ obtained from a population that is normally distributed results in a sample mean of 52.1 and sample standard deviation of 14.7. Does this constitute sufficient evidence to conclude that the population means differ at the $\alpha = 0.05$ level of significance?
3. A random sample of $n_1 = 135$ individuals results in $x_1 = 40$ successes. An independent sample of $n_2 = 150$ individuals results in $x_2 = 60$ successes. Does this represent sufficient evidence to conclude that $p_1 < p_2$ at the $\alpha = 0.05$ level of significance?
4. A random sample of size $n = 41$ results in a sample mean of 125.3 and a sample standard deviation of 8.5. An independent sample of size $n = 50$ results in a sample mean of 130.8 and sample standard deviation of 7.3. Does this constitute sufficient evidence to conclude that the population means differ at the $\alpha = 0.01$ level of significance?
5. The following data represent the measure of a variable before and after a treatment.

Individual	1	2	3	4	5
Before, X_i	93	102	90	112	107
After, Y_i	95	100	95	115	107

Does the sample evidence suggest that the treatment is effective in increasing the value of the response variable? Use the $\alpha = 0.05$ level of significance.

Note: Assume that the differenced data come from a population that is normally distributed with no outliers.

6. The following data represent the measure of a variable before and after a treatment.

Individual	1	2	3	4	5
Before, X_i	40	32	53	48	38
After, Y_i	38	33	49	48	33

Does the sample evidence suggest that the treatment is effective in decreasing the value of the response variable? Use the $\alpha = 0.10$ level of significance.

Note: Assume that the differenced data come from a population that is normally distributed with no outliers.

Applying the Concepts

7. **Collision Claims** Automobile collision insurance is used to pay for any claims made against the driver in the event of an accident. This type of insurance will typically pay to repair any assets that your vehicle damages.

- (a) Collision claims tend to be skewed right. Why do you think this is the case?

- (b) A random sample of 40 collision claims of 30- to 59-year-old drivers results in a mean claim of \$3669 with a standard deviation of \$2029. An independent random sample of 40 collision claims of 20- to 24-year-old drivers results in a mean claim of \$4586 with a standard deviation of \$2302. Using the concept of hypothesis testing, provide an argument that justifies charging a higher insurance premium to 20- to 24-year-old drivers. *Source:* Based on data obtained from the Insurance Institute for Highway Safety.

8. **TIMS Report and Kumon** TIMS is an acronym for the Third International Mathematics and Science Study. Kumon promotes a method of studying mathematics that it claims develops mathematical ability. Do data support this claim? In one particular question on the TIMS exam, a random sample of 400 non-Kumon students resulted in 73 getting the correct answer. For the same question, a random sample of 400 Kumon students resulted in 130 getting the correct answer. Perform the appropriate test to substantiate Kumon's claims. Are there any confounding factors regarding the claim? *Source:* TIMS and PROM/SE Assessment of Kumon Students: What the Results Indicate, Kumon North America.

9. **Age Difference in Married Couples** What is the typical age difference between husband and wife? The following data represent the ages of husbands and wives, based on results from the Current Population Survey.

	Couple 1	Couple 2	Couple 3	Couple 4
Husband	47	53	45	50
Wife	43	43	45	48
	Couple 5	Couple 6	Couple 7	Couple 8
Husband	28	65	25	56
Wife	29	61	27	51

- (a) What is the response variable in this study?
(b) Is the sampling method dependent or independent? Explain.
(c) Use the data to estimate the mean difference in age of husband and wives with 95% confidence. Explain the technique that you used.

10. **Cash or Credit?** Do people tend to spend more money on fast-food when they use a credit card? The following data represent a random sample of credit-card and cash purchases.

Credit				
16.78	23.89	13.89	15.54	10.35
12.76	18.32	20.67	18.36	19.16
Cash				
10.76	6.26	18.98	11.36	6.78
21.76	8.90	15.64	13.78	9.21

Source: Brian Ortiz, student at Joliet Junior College.

- (a) Draw boxplots of each data set using the same scale. What do the boxplots imply for cash versus credit?

- (b) Test whether the sample evidence suggests that people spend more when using a credit card. Use the $\alpha = 0.01$ level of significance. **Note:** Normal probability plots indicate that each sample could come from a population that is normally distributed.
- (c) Suppose that you were looking to gather evidence to convince your manager that an effort needs to be made to boost credit-card sales. Would it be legitimate to change the level to $\alpha = 0.05$ after seeing the results from part (b)? Why?

DATA 11. Major League Fastball The average major league fastball is 92.0 miles per hour (mph). While there are many other factors other than velocity that are used to judge the quality of this pitch (location and movement, for example), velocity is a major factor in deciding whether a pitcher has “big league stuff.” Suppose a scout is judging a college player and records the velocity of a random sample of 14 fastballs. Does the evidence imply that this pitcher has better than average major league “stuff” in regards to velocity?

91.3	93.5	93.4	91.9	91.8	90.3	92.3
92.8	91.6	93.0	91.9	92.6	92.6	92.5

DATA 12. Wet Suits Do wet suits allow a swimmer to swim faster? Researchers measured the speed (in meters per second) of swimmers both with and without a wetsuit. The results of the study are shown in the table. Conduct the appropriate test to determine whether the data suggest that wet suits allow a swimmer to swim faster. Be sure to check all model requirements. Use an $\alpha = 0.05$ level of significance. If you reject the null hypothesis, estimate the average difference in speed of the swimmer with the wet suit with 95% confidence.

Swimmer	1	2	3	4	5	6
Without	1.49	1.37	1.35	1.27	1.12	1.64
With	1.57	1.47	1.42	1.35	1.22	1.75
Swimmer	7	8	9	10	11	12
Without	1.59	1.52	1.50	1.45	1.44	1.41
With	1.64	1.57	1.56	1.53	1.49	1.51

Source: Data from de Lucas, R. D., Balildan, P., Neiva, C. M., Greco, C. C., & Denadai, B. S. (2000). “The Effects of Wet Suits on Physiological and Biomechanical Indices during Swimming.” *Journal of Science and Medicine in Sport* 3(1):1–8.

13. Predicting Election Outcomes Researchers conducted an experiment in which 695 individuals were shown black and white photographs of individuals running for Congress (either the U.S. Senate or House of Representatives). In each instance, the individuals were exposed to the photograph of both the winner and runner-up (in random order) for 1 second. The individuals were then asked to decide who they believed was more competent (and, therefore, more likely to receive their vote). Of the 695 individuals exposed to the photos, 469 correctly predicted the winner of the race. Do the results suggest that a quick 1-second view of a black and white photo represents enough information to judge the winner of an election (based on perceived level of competence of the individual) more often than not? Use the $\alpha = 0.05$ level of significance. Source: Todorov, Mandisodza, Goren, Hall, “Inferences of Competence from Faces Predict Election Outcomes.” *Science* Vol. 308.

14. Bribe ‘em with Chocolate In a study published in the journal *Teaching of Psychology*, the article “Fudging the Numbers: Distributing Chocolate Influences Student Evaluations of an Undergraduate Course” states that distributing chocolate to students prior to teacher evaluations increases results. The authors randomly divided three sections of a course taught by the same instructor into two groups. Fifty of the students were given chocolate by an individual not associated with the course and 50 of the students were not given chocolate. The mean score from students who received chocolate was 4.2, while the mean score for the nonchocolate groups was 3.9. Suppose that the sample standard deviation of both the chocolate and nonchocolate groups was 0.8. Does chocolate appear to improve teacher evaluations? Use the $\alpha = 0.1$ level of significance.

15. Unwed Women Having Children The Pew Research Group asked the following question of individuals who earned in excess of \$100,000 per year and those who earned less than \$100,000 per year: “Do you believe that it is morally wrong for unwed women to have children?” Of the 1205 individuals who earned in excess of \$100,000 per year, 710 said yes; of the 1310 individuals who earned less than \$100,000 per year, 695 said yes. Construct a 95% confidence interval to determine if there is a difference in the proportion of individuals who believe it is morally wrong for unwed women to have children.

16. Extramarital Affairs Is there a difference in the attitude toward extramarital affairs in the United States versus Canada? Pew Research surveyed a random sample of adults from each country and asked, “Do you personally believe that married people having an affair is morally acceptable?” Results of the surveys are below:

	Number Surveyed	Number Responding “Morally Acceptable”
Canada	701	533
United States	1002	841

- (a) What is the response variable in this study?
 (b) Is the sampling method dependent or independent? Explain.
 (c) What approach should be used to determine whether the attitudes in Canada differ from the United States when it comes to extramarital affairs?
 (d) Conduct the appropriate test to determine whether the attitudes in Canada differ from the United States when it comes to extramarital affairs. Use the $\alpha = 0.05$ level of significance.
 (e) Why do you think fewer individuals were surveyed in Canada?

17. Multitasking in School In today’s “wired” society, students believe that they can multitask. Research suggests that 5% of individuals truly have the ability to multitask. Can students multitask, or do they perform worse while multitasking? In an introductory accounting class, students were randomly given an instruction sheet at the beginning of class. Half the instruction sheets required students to send three text messages to the instructor during the lecture. The other half of class was told to turn off their cell phones. At the end of class, a quiz was administered based on information

(continued)

shared during class. The results of the quiz are shown in the table below.

	n	Sample Mean	Sample Standard Deviation
Texting Group	31	42.81	9.91
Cell-phone Off Group	31	58.67	10.42

Source: Ellis, Y., Daniels, W. and Jauregui, A. (2010) "The Effect of Multitasking on the Grade Performance of Business Students." *Research in Higher Education Journal*, 8.

- (a) What is the response variable? What is the explanatory variable?
- (b) How is randomization used in this study?
- (c) Is the sampling method dependent or independent? Explain.
- (d) Conduct the appropriate test to determine whether the data suggest students score worse in the texting group. Use an $\alpha = 0.05$ level of significance.
- (e) Estimate the mean difference in scores between the texting group and cell-phone off group with 90% confidence.

- DATA 18. Comparing Stock Sectors** The data below represent the 5-year rate of return (in percent) for a random sample of stocks in financial services and an independent random sample of stocks in health care. Has there been a difference in the rate of return of companies in the Financial Services sector versus those in the Health Care sector over the past 5 years? Use an $\alpha = 0.05$ level of significance.

Source: morningstar.com

Financial Services				
17.34	16.13	10.26	15.62	15.91
10.16	21.25	22.32	28.30	26.31
23.90	17.51	16.93	24.48	21.76
27.53	21.63	15.41	20.86	17.19

Health Care				
22.08	21.81	31.07	15.96	24.30
24.42	12.96	10.35	30.71	9.18
11.07	21.66	23.08	14.70	15.08
17.76	9.83	33.67	28.28	

- 19. Hospital Readmission** As of October 1, 2012, hospitals in the United States with excessive numbers of readmissions based on the Centers for Medicare and Medicaid Services (CMS) data were penalized. Therefore, it is important for hospitals to identify the risk factors associated with readmission. The following data represent statistics on patients readmitted within 30 days and those not readmitted within 30 days of being discharged at a community hospital. Conduct the appropriate test for each variable that might be associated with readmission (age, length of stay, and so on). Researchers wonder if:

- the readmits tend to be older
- is length of stay longer for readmits
- were a higher proportion of readmits admitted the previous calendar year
- were a higher proportion of readmits discharged in winter
- were a higher proportion of readmits on the cardiac floor

Write a short report detailing your findings.

Source: Lee Park, Danielle Andrade, Andrew Mastey, James Sun, LeRoi Hicks "Institution Specific Risk Factors for 30 Day Readmission at a Community Hospital: A Retrospective Observational Study" *BioMedCentral* 2014 14:40.

	Readmit	Non-Readmit
n	637	3137
Age	Mean: 78.0 Standard deviation: 13.7	Mean: 76.0 Standard deviation: 15.8
Length of stay on previous visit	Mean: 4.8 Standard deviation: 4.3	Mean: 3.9 Standard deviation: 3.1
Admission in previous calendar year	139 out of 637	445 out of 3137
Season	199 out of 637 readmits were discharged in winter	765 out of 3137 non-readmits were discharged in winter
Floor	332 out of 637 readmits were on the cardiac floor	1617 out of 3137 non-readmits were on the cardiac floor

20. Gender Bias in Grades In the fall of 1998, the National Center for Education Statistics (NCES) established a cohort of kindergarten students based on a random sample of all kindergarten students throughout the United States. Objective reading and math assessments were administered to the students in kindergarten, first grade, third grade, and fifth grade. **Note:** Science assessments were also administered, but we do not present those results here. Subjective assessments were also obtained by teacher ratings of each student. The teacher rating was translated into an Academic Rating Scale (ARS) from 0 to 4, where 0 indicates no understanding of the content and 4 indicates complete mastery of the content. The ARS serves as a grade the teacher would assign the student. Teachers were unaware of the students' objective test scores. Source: Christopher Cornwell, David B. Mustard, and Jessica Van Parys, "Noncognitive Skills and Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School," *J. Human Resources*, Winter 2013, 48(1):236–264. doi: 10.3386/jhr.48.1.236

- (a) The goal of the research is to demonstrate whether the objective scores correspond with the subjective grades received by the teacher in the class. For example, if females score higher in Reading, do females also earn better grades (ARS scores) in Reading? The table below summarizes the results for objective test scores and ARS score. Perform the appropriate test to answer each research question.

	Female	Male	Research Question
Reading Score	$\bar{x} = 144.21$ $s = 20.86$ $n = 3008$	$\bar{x} = 141.23$ $s = 23.07$ $n = 2833$	Do females score higher in reading than males?
Reading Grade	$\bar{x} = 3.62$ $s = 0.80$ $n = 3008$	$\bar{x} = 3.37$ $s = 0.82$ $n = 2833$	Do females earn higher grades than males in reading?
Math Score	$\bar{x} = 114.49$ $s = 19.61$ $n = 1452$	$\bar{x} = 118.88$ $s = 19.72$ $n = 1368$	Do males score higher in math than females?
Math Grade	$\bar{x} = 3.44$ $s = 0.65$ $n = 1452$	$\bar{x} = 3.45$ $s = 0.72$ $n = 1368$	Do males earn higher grades than females in math?

(b) Teachers were also asked to rate each student's classroom behavior. This included engagement in the classroom, how often the student lost control, and the student's interpersonal skills. The NCES used the results of these questions to establish a Social Rating Scale (SRS) as a way to measure the student's approach to learning. The SRS scale ranges from 0 to 3, with a higher SRS representing a higher approach to learning. For fifth-grade females the mean SRS is 2.30 with a standard deviation of 0.60 ($n = 2995$), and for fifth-grade males the mean SRS is 1.94 with a standard deviation of 0.67 ($n = 2820$). Does the evidence suggest females received a higher SRS score than males?

(c) Write some overall conclusions based on your data analysis in parts (a) and (b).

21. Web Page Design John has an online company that sells custom rims for cars. A web-design firm hired by John designed two different web pages to be used to sell his rims online. However, he cannot decide which page to go with, so he decides to collect some data. John hires a second firm that has the ability to randomly assign one of the two web page designs to potential customers. John records whether a sale was made on the site, or not. The data are displayed in the table below.

	Design I	Design II
Sale	54	62
No Sale	469	450

(a) What is the response variable in this study? What is the explanatory variable?

(b) Based on the results, which web design, if any, should John go with? **Note:** This problem is based on the type of research done by Adobe Test & Target.

22. Gender Wage Gap It has long been a concern that there is a wage gap between men and women in the United States with some reports suggesting that women only make \$0.77 for every dollar earned by a man. Design a study that would allow you to confirm whether a wage gap does actually exist.

23. Wet Suits Revisited Refer to the data in Problem 12. Treat the data as an independent sample. Compute the sample mean difference and standard error of this data. Compute the sample mean difference and standard error treating the data as a dependent sample. What do the results suggest?

Explaining the Concepts

In Problems 24–33, for each study, explain which statistical procedure (estimating a single proportion; estimating a single mean; hypothesis test for a single proportion; hypothesis test for a single mean; hypothesis test or estimation of two proportions; hypothesis test or estimation of two means, dependent or independent) would most likely be used for the research objective given. Assume all model requirements for conducting the appropriate procedure have been satisfied.

24. Is the mean IQ of the students in Professor Dang's statistics class higher than that of the general population, 100?

25. Do adult males who take a single aspirin daily experience a lower rate of heart attacks than adult males who do not take aspirin daily?

26. Does Marriott Courtyard charge more than Holiday Inn Express for a one-night stay?

27. What is the typical amount of time 20- to 24-year-old males spend brushing their teeth (each time they brush)?

28. What proportion of registered voters is in favor of a tax increase to reduce the federal debt?

29. Does drinking two cups of water before a meal assist with weight loss?

30. Does turmeric (an antioxidant that can be added to foods) help with depression? Researchers randomly assigned 200 adult women who were clinically depressed to two groups. Group 1 had turmeric added to their regular diet for one week; group 2 had no additives in their diet. At the end of one week, the change in their scores on the Beck Depression Inventory was compared.

31. While exercising by climbing stairs, is it better to take one stair, or two stairs, at a time? Researchers identified 30 volunteers who were asked to climb stairs for two different 15-minute intervals taking both one stair and two stairs at a time. Whether the volunteer did one stair or two stairs first was determined randomly. The goal of the research was to determine if energy expenditure for each exercise routine was different.

32. By how much does adiposity (a measure of body fat) differ between adult women who maintain a regular sleep schedule versus women whose sleep schedule fluctuates by 90 minutes or more?

33. Do recent graduates from college who have no debt start their own business at a higher rate than recent graduates who have debt between \$20,000 and \$40,000?



Chapter 11 Review

Summary

This chapter continues the presentation of inferential statistics. Here, we discussed performing statistical inference by comparing two population parameters. In particular, we compared two population proportions, two population means, and two population standard deviations. The first step in performing this inference is to decide whether the sampling method is independent or dependent. A sampling method is independent when the choice of individuals for one sample does not dictate which individuals will be in a second sample. Data obtained from a completely randomized design with

two treatments can be analyzed using the methods based on independent sampling presented in this chapter. A sampling method is dependent when the individuals selected for one sample are used to determine the individuals in the second sample. Data obtained from a matched-pairs experimental design can be analyzed using the dependent sampling inferential techniques presented in this chapter.

Section 11.1 presented statistical inference for comparing two population proportions from independent samples. To perform these tests, the samples must be

obtained independently using simple random sampling or the data result from a completely randomized experiment with two levels of treatment. The response variable is qualitative with two outcomes.

In addition, $n\hat{p}(1 - \hat{p})$ must be greater than or equal to 10 for each sample, and each sample size can be no more than 5% of the population size. The distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal, with mean $p_1 - p_2$

and standard deviation $\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$. We

use this distribution to perform hypothesis tests using either the classical approach or P -value approach. Follow the same five steps for conducting a hypothesis test that we introduced in Chapter 10. We also use the model to construct confidence intervals about the difference of two independent proportions. We present inference on two dependent proportions in Section B.2, which is online at www.pearsonhighered.com/sullivanstats.

Section 11.2 presented inference on the difference of two means from dependent sampling. We use Student's t -distribution to perform hypothesis testing or construct confidence intervals. However, the following conditions must be satisfied:

1. Data must be obtained through a simple random sample or result from a randomized experiment.

Data must be matched-pairs (dependent). Response variable is quantitative.

2. Differenced data come from a population that is normally distributed or the sample size is large.
3. Sample size may be no more than 5% of the population size (independence requirement).

Again, use the same five steps for conducting a hypothesis test introduced in Section 10.3.

In Section 11.3, we presented inference on the difference of two means from independent sampling. Here, we use Welch's approximate t provided:

1. The samples are independently obtained using simple random sampling or the data result from a completely randomized experiment with two levels of treatment. Response variable is quantitative.
2. The sample data come from a population that is normally distributed with no outliers, or the sample sizes are large.
3. For each sample, the sample size is no more than 5% of the population size.

The steps for conducting a hypothesis test are the same as they have been throughout the chapters.

To help determine which test to use, we included the flow chart in Figure 15 within Section 11.4.

Vocabulary

Independent (p. 478)
Dependent (p. 478)
Matched pairs (p. 478)

Pooled estimate of p (p. 480)
Robust (p. 492)

Welch's approximate t (p. 502)
Pooled t -statistic (p. 507)

Formulas

- Test statistic comparing two population proportions (independent sampling):

$$z_0 = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$

- Confidence interval for the difference of two proportions (independent sampling):

Lower bound: $(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$

Upper bound: $(\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$

- Sample size for estimating $p_1 - p_2$:

$$n = n_1 = n_2 = [\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)] \left(\frac{z_{\alpha/2}}{E} \right)^2$$

$$n = n_1 = n_2 = 0.5 \left(\frac{z_{\alpha/2}}{E} \right)^2$$

- Test statistic for matched-pairs data:

$$t_0 = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

where \bar{d} is the mean and s_d is the standard deviation of the differenced data

- Confidence interval for matched-pairs data:

Lower bound: $\bar{d} - t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}}$

Upper bound: $\bar{d} + t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}}$

- Test statistic comparing two means (independent sampling):

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Confidence interval for the difference of two means (independent samples):

$$\text{Lower bound: } (\bar{x}_1 - \bar{x}_2) - t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{Upper bound: } (\bar{x}_1 - \bar{x}_2) + t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Objectives

Section	You should be able to ...	Example(s)	Review Exercises
11.1	1 Distinguish between independent and dependent sampling (p. 478) 2 Test hypotheses regarding two proportions from independent samples (p. 479) 3 Construct and interpret confidence intervals for the difference between two population proportions (p. 483) 4 Determine the sample size necessary for estimating the difference between two population proportions (p. 485)	1 2 3 4	1, 2 6, 9(c) 9(d) 10
11.2	1 Test hypotheses for a population mean from matched-pairs data (p. 491) 2 Construct and interpret confidence intervals about the population mean difference of matched-pairs data (p. 495)	1 2	3(c), 7(b) 3(d), 11
11.3	1 Test hypotheses regarding the difference of two independent means (p. 502) 2 Construct and interpret confidence intervals regarding the difference of two independent means (p. 506)	1 2	4(a), 5(a), 8(c) 4(b), 12
11.4	1 Determine the appropriate hypothesis test to perform (p. 513)		3–9, 11–12

Review Exercises

In Problems 1 and 2, determine if the sampling is dependent or independent. Indicate whether the response variable is qualitative or quantitative.

- A researcher wants to know if the mean length of stay in for-profit hospitals is different from that in not-for-profit hospitals. He randomly selected 20 individuals in the for-profit hospital and matched them with 20 individuals in the not-for-profit hospital by diagnosis.
- An urban economist believes that commute times to work in the South are less than commute times to work in the Midwest. He randomly selects 40 employed individuals in the South and 40 employed individuals in the Midwest and determines their commute times.

In Problem 3, assume that the paired differences come from a population that is normally distributed.

3. DATA	Observation	1	2	3	4	5	6
	X_i	34.2	32.1	39.5	41.8	45.1	38.4
	Y_i	34.9	31.5	39.5	41.9	45.5	38.8

- Compute $d_i = X_i - Y_i$ for each pair of data.
- Compute \bar{d} and s_d .
- Test the hypothesis that $\mu_d < 0$ at the $\alpha = 0.05$ level of significance.
- Compute a 98% confidence interval for the population mean difference μ_d .

In Problems 4 and 5, assume that the populations are normally distributed and that independent sampling occurred.

4.	Sample 1	Sample 2
n	13	8
\bar{x}	32.4	28.2
s	4.5	3.8

- Test the hypothesis that $\mu_1 \neq \mu_2$ at the $\alpha = 0.1$ level of significance for the given sample data.
- Construct a 90% confidence interval for $\mu_1 - \mu_2$.

5.	Sample 1	Sample 2
n	45	41
\bar{x}	48.2	45.2
s	8.4	10.3

Test the hypothesis that $\mu_1 > \mu_2$ at the $\alpha = 0.01$ level of significance for the given sample data.

- A random sample of $n_1 = 555$ individuals results in $x_1 = 451$ successes. An independent sample of $n_2 = 600$ individuals results in $x_2 = 510$ successes. Does this represent sufficient evidence to conclude that $p_1 \neq p_2$ at the $\alpha = 0.05$ level of significance?

- DATA 7. Height versus Arm Span A statistics student heard that an individual's arm span is equal to the individual's height. To test this hypothesis, the student used a random sample of ten students and obtained the data on the next page.

Student:	1	2	3	4	5
Height (inches)	59.5	69	77	59.5	74.5
Arm span (inches)	62	65.5	76	63	74
Student:	6	7	8	9	10
Height (inches)	63	61.5	67.5	73	69
Arm span (inches)	66	61	69	70	71

Source: John Climent, Cecil Community College.

- (a) Is the sampling method dependent or independent? Why?
- (b) Does the sample evidence contradict the belief that an individual's height and arm span are the same at the $\alpha = 0.05$ level of significance?

Note: A normal probability plot indicates that the data and differenced data are normally distributed. A boxplot indicates that the data and differenced data have no outliers.

DATA 8. McDonald's versus Wendy's A student wanted to determine whether the wait time in the drive-thru at McDonald's differed from that at Wendy's. She used a random sample of 30 cars at McDonald's and 27 cars at Wendy's and obtained these results:

Wait Time at McDonald's Drive-Thru (seconds)				
151.09	227.38	111.84	131.21	128.75
191.60	126.91	137.90	195.44	246.59
141.78	127.35	121.21	101.03	95.09
122.06	122.62	100.04	71.37	153.34
140.44	126.62	116.72	131.69	100.94
115.66	147.28	81.43	86.31	156.34
Wait Time at Wendy's Drive-Thru (seconds)				
281.90	71.02	204.29	128.59	133.56
187.53	199.86	190.91	110.55	110.64
196.84	233.65	171.01	182.54	183.79
284.48	363.34	270.82	390.50	471.62
123.66	174.43	385.90	386.71	155.53
203.62	119.61			

Source: Catherine M. Simmons, student at Joliet Junior College.

Note: The sample size for Wendy's is less than 30. However, the data do not contain any outliers, so the Central Limit Theorem can be used.

- (a) Is the sampling method dependent or independent?
- (b) What is the variable of interest? Is it qualitative or quantitative?

- (c) Is there a difference in wait times at each restaurant's drive-thru? Use the $\alpha = 0.1$ level of significance.
- (d) Draw boxplots of each data set using the same scale. Does this visual evidence support the results obtained in part (b)?

9. Treatment for Osteoporosis Osteoporosis is a condition in which people experience decreased bone mass and an increase in the risk of bone fracture. Actonel is a drug that helps combat osteoporosis in postmenopausal women. In clinical trials, 1374 postmenopausal women were randomly divided into experimental and control groups. The subjects in the experimental group were administered 5 milligrams (mg) of Actonel, while the subjects in the control group were administered a placebo. The number of women who experienced a bone fracture over the course of one year was recorded. Of the 696 women in the experimental group, 27 experienced a fracture during the course of the year. Of the 678 women in the control group, 49 experienced a fracture during the course of the year.

- (a) What type of experimental design is this? What is the response variable? Is it qualitative or quantitative? What are the treatments?
- (b) The experiment was double-blind. What does this mean?
- (c) Does the sample evidence suggest the drug is effective in preventing bone fractures? Use the $\alpha = 0.01$ level of significance.
- (d) Construct a 95% confidence interval for the difference between the two population proportions, $p_{\text{exp}} - p_{\text{control}}$.

10. Determining Sample Size A nutritionist wants to estimate the difference between the percentage of men and women who have high cholesterol. What sample size should be obtained if she wishes the estimate to be within 2 percentage points with 90% confidence, assuming that

- (a) she uses the 1994 estimates of 18.8% male and 20.5% female from the National Center for Health Statistics?
- (b) she does not use any prior estimates?

11. Height versus Arm Span Construct and interpret a 95% confidence interval for the population mean difference between height and arm span using the data from Problem 7. What does the interval lead us to conclude regarding any differences between height and arm span?

12. McDonald's versus Wendy's Construct and interpret a 95% confidence interval about $\mu_M - \mu_W$ using the data from Problem 8. How might a marketing executive with McDonald's use this information?



Chapter Test

In Problems 1 and 2, determine whether the sampling method is independent or dependent.

1. A stock analyst wants to know if there is a difference between the mean rate of return from energy stocks and that from financial stocks. He randomly selects 13 energy stocks and computes the rate of return for the past year. He

randomly selects 13 financial stocks and computes the rate of return for the past year.

2. A prison warden wants to know if men receive longer sentences for crimes than women. He randomly samples 30 men and matches them with 30 women by type of crime committed and records their lengths of sentence.

In Problem 3, assume that the paired differences come from a population that is normally distributed.

DATA	3.	Observation	1	2	3	4	5	6	7
		X_i	18.5	21.8	19.4	22.9	18.3	20.2	23.1
		Y_i	18.3	22.3	19.2	22.3	18.9	20.7	23.9

- (a) Compute $d_i = X_i - Y_i$ for each pair of data.
- (b) Compute \bar{d} and s_d .
- (c) Test the hypothesis that $\mu_d \neq 0$ at the $\alpha = 0.01$ level of significance.
- (d) Compute a 95% confidence interval for the population mean difference μ_d .

In Problems 4 and 5, assume that the populations are normally distributed and that independent sampling occurred.

DATA	4.	Sample 1	Sample 2
		n	24
		\bar{x}	104.2
		s	12.3
			27
			110.4
			8.7

- (a) Test the hypothesis that $\mu_1 \neq \mu_2$ at the $\alpha = 0.1$ level of significance for the given sample data.
- (b) Construct a 95% confidence interval for $\mu_1 - \mu_2$.

DATA	5.	Sample 1	Sample 2
		n	13
		\bar{x}	96.6
		s	3.2
			8
			98.3
			2.5

Test the hypothesis that $\mu_1 < \mu_2$ at the $\alpha = 0.05$ level of significance for the given sample data.

6. A random sample of $n_1 = 650$ individuals results in $x_1 = 156$ successes. An independent sample of $n_2 = 550$ individuals results in $x_2 = 143$ successes. Does this represent sufficient evidence to conclude that $p_1 < p_2$ at the $\alpha = 0.05$ level of significance?

DATA	7.	A researcher wants to know whether the acidity of rain (pH) near Houston, Texas, is significantly different from that near Chicago, Illinois. He randomly selects 12 rain dates in Texas and 14 rain dates in Illinois and obtains the following data:

Texas					
4.69	5.10	5.22	4.46	4.93	4.65
5.22	4.76	4.25	5.14	4.11	4.71
Illinois					
4.40	4.69	4.22	4.64	4.54	4.35
4.40	4.75	4.63	4.45	4.49	4.36

Source: National Atmospheric Deposition Program.

- (a) Is the sampling method dependent or independent? Why?
- (b) Because the sample sizes are small, what must be true regarding the populations from which the samples were drawn?
- (c) Draw side-by-side boxplots of the data. What does the visual evidence imply about the pH of rain in the two states?

- (d) Does the evidence suggest that there is a difference in the pH of rain in Chicago and Houston? Use the $\alpha = 0.05$ level of significance.

8. In a study conducted to determine the role that sleep disorders play in academic performance, researcher Jane Gaultney conducted a survey of 1845 college students to determine if they had a sleep disorder (such as narcolepsy, insomnia, or restless leg syndrome). Of the 503 students with a sleep disorder, the mean grade point average was 2.65 with a standard deviation of 0.87. Of the 1342 students without a sleep disorder, the mean grade point average was 2.82 with a standard deviation of 0.83. Source: SLEEP 2010: Associated Professional Sleep Societies 24th Annual Meeting.

- (a) What is the response variable in this study? What is the explanatory variable?
- (b) Is there evidence to suggest sleep disorders adversely affect one's GPA at the $\alpha = 0.05$ level of significance?

9. Researchers had a car traveling 10 miles per hour collide into the rear bumper of an SUV and recorded the amount of damage, in dollars. The data are below. Do the given data suggest the repair cost of the car is higher? Use the $\alpha = 0.1$ level of significance.

Car into SUV	SUV Damage	Car Damage
Kia Forte into Hyundai Tucson	2091	1510
Dodge Caliber into Jeep Patriot	1338	2559
Honda Civic into Honda CR-V	1053	4921
Volkswagen Golf into VW Tiguan	1872	4555
Nissan Sentra into Nissan Rogue	1428	5114
Ford Focus into Ford Escape	2208	5203
Toyota Corolla into Toyota RAV4	6015	3852

Source: Insurance Institute for Highway Safety.

10. Zoloft is a drug used to treat obsessive-compulsive disorder (OCD). In randomized, double-blind clinical trials, 926 patients diagnosed with OCD were randomly divided into two groups. Subjects in group 1 (experimental group) received 200 milligrams per day (mg/day) of Zoloft, while subjects in group 2 (control group) received a placebo. Of the 553 subjects in the experimental group, 77 experienced dry mouth as a side effect. Of the 373 subjects in the control group, 34 experienced dry mouth as a side effect.

- (a) What type of experimental design is this?
- (b) What is the response variable?
- (c) Do a higher proportion of subjects experience dry mouth who are taking Zoloft versus the proportion taking the placebo? Use the $\alpha = 0.05$ level of significance.

11. Does hypnotism result in a different success rate for men and women who are trying to quit smoking? Researchers at *Science* magazine analyzed studies involving 5600 male and female smokers. Of the 2800 females, 644 quit smoking; of the 2800 males, 840 quit smoking. Construct a 90% confidence interval for the difference in proportion of males and females, $p_M - p_F$, and use the interval to judge whether there is a difference in the proportions.

12. A researcher wants to estimate the difference between the percentage of individuals without a high school diploma who smoke and the percentage of individuals with bachelor's degrees who smoke. What sample size should be obtained if she wishes the estimate to be within 4 percentage points with 95% confidence, assuming that

- (a) she uses the 1999 estimates of 32.2% of those without a high school diploma and 11.1% of those with a bachelor's degree from the National Center for Health Statistics?
- (b) she does not use any prior estimates?

13. It is commonplace to gain weight after quitting smoking. To determine the effectiveness of the drug Naltrexone in limiting weight gain after quitting smoking, 147 subjects who smoked 20 or more cigarettes daily were randomly divided into two groups. All 147 subjects received a 21-milligram (mg) transdermal nicotine patch (to curb nicotine cravings while

trying to quit smoking). However, 72 subjects received a placebo, while 75 received 25 mg of Naltrexone. After 6 weeks, the placebo subjects had a mean weight gain of 1.9 kg, with a standard deviation of 0.22 kg; the Naltrexone subjects had a mean weight gain of 0.8 kg with a standard deviation of 0.21 kg. Construct a 95% confidence interval for the mean difference in weight gain for the two groups. Based on this interval, do you believe that Naltrexone is effective in controlling weight gain following smoking cessation? Do the results have any practical significance? *Source:* Stephanie O'Malley et al. "A Controlled Trial of Naltrexone Augmentation of Nicotine Replacement Therapy for Smoking Cessation." *Archives of Internal Medicine*, Vol. 166, 667–674.

Making an Informed Decision

Which Car Should I Buy?

You have decided to purchase a car and have narrowed your choice down to two cars. However, you have two areas of concern. First, you want to purchase the car that gets the better gas mileage. Second, you want to purchase the car that holds its value better. To help make an informed decision, you decide to collect some data and run some tests.

- (a) Decide on two cars that are similar that you would consider purchasing.
- (b) Go to www.fueleconomy.gov and obtain data on the fuel economy of each car you are considering. Treat the data as a random sample of all cars.
- (c) Draw side-by-side boxplots of the fuel economy to verify there are no outliers and to verify it is reasonable to conclude the data come from a population that is normally distributed (if your sample size is small).

- (d) Conduct the appropriate test to determine if there is a significant difference in the gas mileage of the two cars.



- (e) Go to an online website that lists used cars for sale. Obtain a matched-pairs random sample of cars where each car is paired with the second car based on age of the car and mileage. For example, if you are considering a Camry or Accord, then match a two-year-old Camry with 18,000 miles with a two-year-old Accord with 18,000 miles. However, to determine how well the car holds its value, subtract the asking price of the car from the price when the car is new.
- (f) Conduct the appropriate test to determine if one car holds its value better than the other car.
- (g) Write a report detailing which car you would purchase.

12

Additional Inferential Methods

Outline

- 12.1** Goodness-of-Fit Test
- 12.2** Tests for Independence and the Homogeneity of Proportions
- 12.3** Testing the Significance of the Least-Squares Regression Model
- 12.4** Confidence and Prediction Intervals



Making an Informed Decision

Are there benefits to attending college? If so, what are they? See the Decision project on page 581.

Putting It Together

Chapters 9 through 11 introduced statistical methods that can be used to test hypotheses regarding a population parameter such as p or μ .

Often, however, rather than being interested in testing a hypothesis regarding a parameter of a probability distribution, we want to test a hypothesis regarding the entire probability distribution. For example, we might want to test whether the distribution of colors in a bag of plain M&M candies is 13% brown, 14% yellow, 13% red, 20% orange, 24% blue, and 16% green. Methods for testing such hypotheses are covered in Section 12.1.

In Section 12.2, we discuss a method for determining whether two qualitative variables are independent based on a sample. If they are not independent, the value of one variable affects the value of the other variable, so the variables are somehow related. We conclude Section 12.2 by introducing a test for homogeneity of proportions, which compare proportions from two or more populations. This test is an extension of the two-sample z -test for proportions from independent samples discussed in Section 11.1.

In Chapter 4, we learned methods for describing the relation between two quantitative variables. In Section 12.3, we test whether a linear relation exists between two quantitative variables based on the methods introduced in Chapter 10.

In Section 12.4, we construct confidence and prediction intervals about the predicted value of a response variable.

12.1 Goodness-of-Fit Test



Preparing for This Section Before getting started, review the following:

- Mutually exclusive (Section 5.2, p. 242)
- Mean of a binomial random variable (Section 6.2, p. 320)
- Expected value (Section 6.1, pp. 305–306)

- Objectives**
- ① Find critical values for the chi-square distribution
 - ② Perform a goodness-of-fit test

1 Find Critical Values for the Chi-Square Distribution

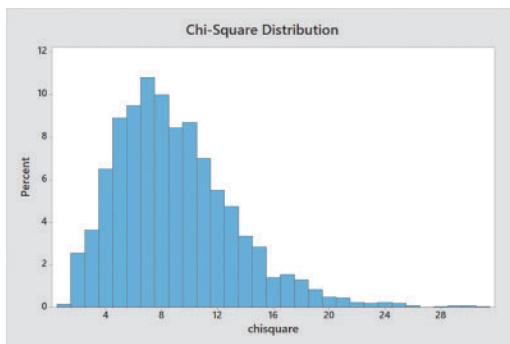
We begin by exploring the sampling distribution of s^2 through a simulation. Suppose that we obtain 2000 samples of size $n = 10$ from a population that is known to be normally distributed with mean 100 and standard deviation 15. Perform the following steps:

Step 1 Compute the sample variance of each of the 2000 samples.

Step 2 Compute $\frac{(n - 1)s^2}{\sigma^2} = \frac{9s^2}{15^2}$ for each sample.

Step 3 Draw a histogram of these values as shown in Figure 1.

Figure 1



The histogram suggests that the sampling distribution of $\frac{(n - 1)s^2}{\sigma^2}$ is skewed right. The distribution in Figure 1 follows a *chi-square distribution*.

Chi-Square Distribution

If a simple random sample of size n is obtained from a normally distributed population with mean μ and standard deviation σ , then

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

has a **chi-square distribution** with $n - 1$ degrees of freedom.

The symbol χ^2 , chi-square, is pronounced “kigh-square” (to rhyme with “sky-square”). We can find critical values of the chi-square distribution in Table VIII in Appendix A of the text. Before discussing how to read Table VIII, we introduce characteristics of the chi-square distribution.

Characteristics of the Chi-Square Distribution

1. It is not symmetric.
2. The shape of the chi-square distribution depends on the degrees of freedom, just like Student's t -distribution.
3. As the number of degrees of freedom increases, the chi-square distribution becomes more nearly symmetric. See Figure 2.
4. The values of χ^2 are nonnegative (greater than or equal to 0).

Figure 2

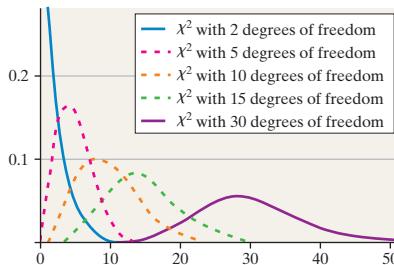


Table VIII is structured similarly to Table VII for the t -distribution. The left column represents the degrees of freedom, and the top row represents the area under the chi-square distribution to the right of the critical value. We use the notation χ_{α}^2 to denote the critical χ^2 -value such that the area under the chi-square distribution to the right of χ_{α}^2 is α .

EXAMPLE 1

Finding Critical Values for the Chi-Square Distribution

Problem Find the critical value such that the area under the chi-square distribution to the right of the critical value is 0.05, assuming 15 degrees of freedom.

Approach Perform the following steps to obtain the critical value.

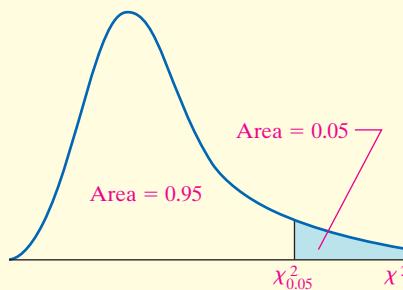
Step 1 Draw a chi-square distribution with the critical value and area labeled.

Step 2 Use Table VIII to find the critical value.

Solution

Step 1 Figure 3 shows the chi-square distribution with 15 degrees of freedom and the unknown critical value, $\chi_{0.05}^2$, labeled. The area to the right of the right critical value is 0.05.

Figure 3



Step 2 Figure 4 on the next page shows a partial representation of Table VIII. The row containing 15 degrees of freedom is boxed. The column corresponding to an area to the right of 0.05 is also boxed. The critical value is $\chi_{0.05}^2 = 24.996$.

(continued)

Figure 4

Degrees of Freedom	Area to the Right of the Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	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
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

In studying Table VIII, notice that the degrees of freedom are numbered 1 to 30, inclusive, then 40, 50, 60, . . . , 100. If the number of degrees of freedom is not in the table, choose the degrees of freedom closest to that desired. If the degrees of freedom are exactly between two values, find the mean of the values. For example, to find the critical value corresponding to 75 degrees of freedom, compute the mean of the critical values corresponding to 70 and 80 degrees of freedom.

2 Perform a Goodness-of-Fit Test



In this section, we present a procedure that can be used to test hypotheses regarding a probability distribution. For example, we might want to test whether the distribution of plain M&M candies in a bag is 13% brown, 14% yellow, 13% red, 20% orange, 24% blue, and 16% green. Or we might want to test whether the number of hits a player gets in his next four at bats follows a binomial distribution with $n = 4$ and $p = 0.298$. Both of these scenarios may be analyzed using a *goodness-of-fit test*.

Definition

A **goodness-of-fit test** is an inferential procedure used to determine whether a frequency distribution follows a specific distribution.

As an example, we might want to test whether a die is fair. This means that the probability of each outcome is $\frac{1}{6}$ when a die is cast. Because we give the die the benefit of the doubt (that is, assume that the die is fair), the null hypothesis is:

$$H_0: p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$$

As another example, according to the Census Bureau in 2000, 19.0% of the population of the United States resided in the Northeast, 22.9% in the Midwest, 35.6% in the South, and 22.5% in the West. We might want to test whether the distribution of residents is the same today as it was in 2000. Since the null hypothesis is a statement of “no change,” we have

H_0 : The distribution of residents in the United States is the same today as it was in 2000.

The idea behind testing these types of hypotheses is to compare the actual number of observations for each category of data with the number of observations we would expect if the null hypothesis were true. If a significant difference exists between the observed counts and expected counts, we have evidence against the null hypothesis.

The method for obtaining the **expected counts** is an extension of the expected value of a binomial random variable. Recall that the mean (and therefore expected value) of a binomial random variable with n independent trials and probability of success, p , is given by $E = \mu = np$.

IN OTHER WORDS

The expected count for each category is the number of trials of the experiment times the probability of success in the category.

Expected Counts

Suppose there are n independent trials of an experiment with $k \geq 3$ mutually exclusive possible outcomes. Let p_1 represent the probability of observing the first outcome and E_1 represent the expected count of the first outcome, p_2 represent the probability of observing the second outcome and E_2 represent the expected count of the second outcome, and so on. The expected counts for each possible outcome are given by

$$E_i = \mu_i = np_i \quad \text{for } i = 1, 2, \dots, k$$

EXAMPLE 2

Finding Expected Counts

Table 1 Distribution of Household Income, 2000

Income	Percent
Under \$15,000	9.9
\$15,000 to \$24,999	9.8
\$25,000 to \$34,999	9.3
\$35,000 to \$49,999	13.5
\$50,000 to \$74,999	17.9
\$75,000 to \$99,999	13.1
\$100,000 to \$149,999	14.9
\$150,000 to \$199,999	6.1
At least \$200,000	5.5

Source: U.S. Census Bureau.

Problem One growing concern regarding the U.S. economy is the inequality in the distribution of income. The data in Table 1 represent the distribution of household income for various levels of income in 2000. An economist wants to know if the distribution of income is changing, so she randomly selects 1500 households and obtains the household income. Find the expected number of households at each income level assuming that the distribution of income has not changed since 2000.

Approach Any quantitative variable (such as income) can become qualitative when categories of the variable are created. For example, in this data, we have nine categories of income. There are two steps to follow to determine the expected counts for each category.

Step 1 Determine the probabilities for each outcome.

Step 2 There are $n = 1500$ trials (the 1500 households surveyed) of the experiment. We expect $np_{\text{under } \$15,000}$ of the households surveyed to have an income under \$15,000.

Solution

Step 1 The probabilities are the relative frequencies from the 2000 distribution:

$$\begin{aligned} p_{\text{under } 15,000} &= 0.099 & p_{15,000-24,999} &= 0.098 & p_{25,000-34,999} &= 0.093 \\ p_{35,000-49,999} &= 0.135 & p_{50,000-74,999} &= 0.179 & p_{75,000-99,999} &= 0.131 \\ p_{100,000-149,999} &= 0.149 & p_{150,000-199,999} &= 0.061 & p_{\text{At least } 200,000} &= 0.055 \end{aligned}$$

Step 2 The expected counts for each income are as follows:

$$\begin{aligned} \text{Expected count of under } \$15,000: & \quad np_{\text{under } 15,000} = 1500(0.099) = 148.5 \\ \text{Expected count of } \$15,000 \text{ to } \$24,999: & \quad np_{15,000-24,999} = 1500(0.098) = 147 \\ \text{Expected count of } \$25,000 \text{ to } \$34,999: & \quad np_{25,000-34,999} = 1500(0.093) = 139.5 \\ \text{Expected count of } \$35,000 \text{ to } \$49,999: & \quad np_{35,000-49,999} = 1500(0.135) = 202.5 \\ \text{Expected count of } \$50,000 \text{ to } \$74,999: & \quad np_{50,000-74,999} = 1500(0.179) = 268.5 \\ \text{Expected count of } \$75,000 \text{ to } \$99,999: & \quad np_{75,000-99,999} = 1500(0.131) = 196.5 \\ \text{Expected count of } \$100,000 \text{ to } \$149,999: & \quad np_{100,000-149,999} = 1500(0.149) = 223.5 \\ \text{Expected count of } \$150,000 \text{ to } \$199,999: & \quad np_{150,000-199,999} = 1500(0.061) = 91.5 \\ \text{Expected count of at least } \$200,000: & \quad np_{\text{At least } 200,000} = 1500(0.055) = 82.5 \end{aligned}$$

Of the 1500 households surveyed, the economist expects to have 148.5 households that earned under \$15,000, 147 that earned \$15,000 to \$24,999, and so on, assuming the distribution of income has not changed since 2000.



To conduct a hypothesis test, we compare the observed counts to the expected counts for each category. If the observed counts are significantly different from the expected counts, we have evidence against the null hypothesis. To perform this test, we need a test statistic and sampling distribution.

Test Statistic for Goodness-of-Fit Tests

Let O_i represent the observed counts of category i , E_i represent the expected counts of category i , k represent the number of categories, and n represent the number of independent trials of an experiment. Then

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad i = 1, 2, \dots, k$$

CAUTION!

Goodness-of-fit tests are used to test hypotheses regarding the distribution of a variable based on a single population. If you wish to compare two or more populations, you must use the tests for homogeneity presented in Section 12.2.

approximately follows the chi-square distribution with $k - 1$ degrees of freedom, provided that

1. all expected frequencies are greater than or equal to 1 (all $E_i \geq 1$) and
2. no more than 20% of the expected frequencies are less than 5.

Note: $E_i = np_i$ for $i = 1, 2, \dots, k$.

In Example 2, there were $k = 9$ categories (nine levels of income).

Now that we know the distribution of goodness-of-fit tests, we can present a method for testing hypotheses regarding the distribution of a random variable.

The Goodness-of-Fit Test

To test hypotheses regarding a distribution, use the steps that follow.

Step 1 Determine the null and alternative hypotheses:

H_0 : The random variable follows a certain distribution.

H_1 : The random variable does not follow the distribution in the null hypothesis.

Step 2 Decide on a level of significance, α , depending on the seriousness of making a Type I error.

Step 3

(a) Calculate the expected counts, E_i , for each of the k categories: $E_i = np_i$ for $i = 1, 2, \dots, k$, where n is the number of trials and p_i is the probability of the i th category, assuming that the statement in the null hypothesis is true.

(b) Verify that the requirements for the goodness-of-fit test are satisfied.

1. All expected counts are greater than or equal to 1 (all $E_i \geq 1$).
2. No more than 20% of the expected counts are less than 5.

CAUTION!

If the requirements in Step 3(b) are not satisfied, one option is to combine two or more low-frequency categories into a single category.

Classical Approach

Step 3 (continued)

(c) Compute the **test statistic**

$$\chi_0^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Note: O_i is the observed count for the i th category.

Step 4 Determine the critical value using Table VIII. All goodness-of-fit tests are right-tailed tests, so the critical value is χ_{α}^2 with $k - 1$ degrees of freedom. See Figure 5 on the next page.

P-Value Approach

By Hand Step 3 (continued)

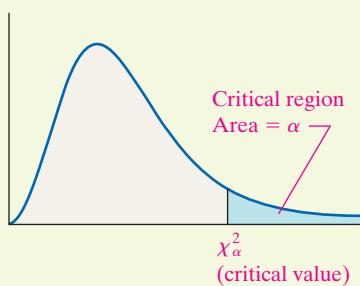
(c) Compute the **test statistic**

$$\chi_0^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Note: O_i is the observed count for the i th category.

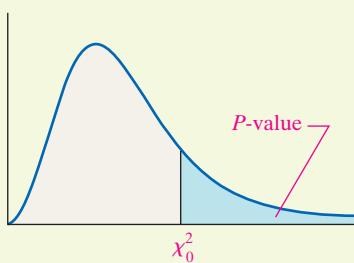
(d) Use Table VIII to approximate the *P*-value by determining the area under the chi-square distribution with $k - 1$ degrees of freedom to the right of the test statistic. See Figure 6 on the next page.

Figure 5



Compare the critical value to the statistic. If $\chi_0^2 > \chi_\alpha^2$, reject the null hypothesis.

Figure 6



Technology Step 3 (continued)

(c) Use a statistical spreadsheet or calculator with statistical capabilities to obtain the P -value. The directions for obtaining the P -value using the TI-84 Plus graphing calculators, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on pages 532–533.

Step 4 If P -value $< \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

EXAMPLE 3 Conducting a Goodness-of-Fit Test

Problem One growing concern regarding the U.S. economy is the inequality in the distribution of income. An economist wants to know if the distribution of income is changing, so she randomly selects 1500 households and obtains the household income shown in Table 2. Table 2 also contains the expected counts under the assumption the distribution has not changed since 2000 (obtained in Example 2). Does the evidence suggest that the distribution of income has changed since 2000 at the $\alpha = 0.05$ level of significance? **Note:** The data in Table 2 is based on the 2017 Current Population Survey and has been adjusted for inflation.

NOTE

The phrase “adjusted for inflation” means the purchasing power of \$1 is the same in 2017 as it was in 2000. This allows for a fair comparison of levels of income.

Table 2

Income	Observed Counts	Expected Counts
Under \$15,000	161	148.5
\$15,000 to \$24,999	144	147
\$25,000 to \$34,999	138	139.5
\$35,000 to \$49,999	184	202.5
\$50,000 to \$74,999	247	268.5
\$75,000 to \$99,999	188	196.5
\$100,000 to \$149,999	217	223.5
\$150,000 to \$199,999	105	91.5
At least \$200,000	116	82.5

Approach Follow Steps 1 through 5 just listed.

Solution

Step 1 The null hypothesis is always a statement of “no difference.” Here, that means there is no difference in the distribution of income between 2000 and today. We are looking for the sample evidence (sample data) to show that the distribution is different today.

H_0 : The distribution of household income in the United States is the same today as it was in 2000.

H_1 : The distribution of household income in the United States is different today from what it was in 2000.

(continued)

Step 2 The level of significance is $\alpha = 0.05$.

Step 3

- (a) We computed the expected counts in Example 2. The observed and expected counts are in Table 2.
- (b) Since all expected counts are greater than or equal to 5, the requirements for the goodness-of-fit test are satisfied.

Classical Approach

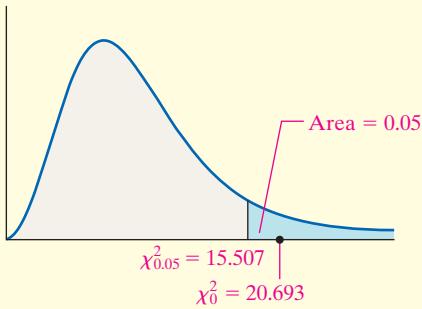
Step 3 (continued)

- (c) The test statistic is

$$\begin{aligned}\chi_0^2 &= \sum \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(161 - 148.5)^2}{148.5} + \frac{(144 - 147)^2}{147} + \frac{(138 - 139.5)^2}{139.5} \\ &\quad + \frac{(184 - 202.5)^2}{202.5} + \frac{(247 - 268.5)^2}{268.5} + \frac{(188 - 196.5)^2}{196.5} \\ &\quad + \frac{(217 - 223.5)^2}{223.5} + \frac{(105 - 91.5)^2}{91.5} + \frac{(116 - 82.5)^2}{82.5} \\ &= 20.693\end{aligned}$$

Step 4 There are $k = 9$ categories, so we find the critical value using $9 - 1 = 8$ degrees of freedom. The critical value is $\chi_{0.05}^2 = 15.507$. See Figure 7.

Figure 7



Because the test statistic, 20.693, is greater than the critical value, 15.507, we reject the null hypothesis.

P-Value Approach

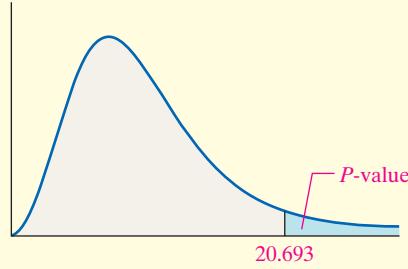
By Hand Step 3 (continued)

- (c) The test statistic is

$$\begin{aligned}\chi_0^2 &= \sum \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(161 - 148.5)^2}{148.5} + \frac{(144 - 147)^2}{147} + \frac{(138 - 139.5)^2}{139.5} \\ &\quad + \frac{(184 - 202.5)^2}{202.5} + \frac{(247 - 268.5)^2}{268.5} + \frac{(188 - 196.5)^2}{196.5} \\ &\quad + \frac{(217 - 223.5)^2}{223.5} + \frac{(105 - 91.5)^2}{91.5} + \frac{(116 - 82.5)^2}{82.5} \\ &= 20.693\end{aligned}$$

(d) There are $k = 9$ categories. The P -value is the area under the chi-square distribution with $9 - 1 = 8$ degrees of freedom to the right of $\chi_0^2 = 20.693$, as shown in Figure 8.

Figure 8



Using Table VIII, find the row that corresponds to 8 degrees of freedom. The value of 20.693 lies between 20.090, which corresponds to an area to the right of 0.01, and 21.955, which corresponds to an area to the right of 0.005. Therefore, the P -value is between 0.005 and 0.01. So, $0.005 < P\text{-value} < 0.01$.

Technology Step 3 (continued)

- (c) Figure 9 shows the result of the Goodness-of-Fit test from StatCrunch. The P -value is reported as 0.008.

Figure 9 Chi-Square goodness-of-fit results:

Observed: Observed

Expected: Expected

N	DF	Chi-Square	P-value
1500	8	20.692823	0.008

Step 4 The P -value of 0.008 means that if the null hypothesis “the distribution of income is the same today as in 2000” is true, we would expect about 8 samples in

Historical Note

The goodness-of-fit test was invented by Karl Pearson (the Pearson of correlation coefficient fame). Pearson believed that statistics should be done by determining the distribution of a random variable. Such a determination could be made only by looking at large numbers of data. This philosophy caused Pearson to "butt heads" with Ronald Fisher, because Fisher believed in analyzing small samples.

1000 to yield the results at least as extreme as those we obtained. The observed results are unusual under the assumption the null hypothesis is true. Because the P -value is less than the level of significance (Tech: $0.008 < 0.05$), we reject the null hypothesis.

Step 5 The sample evidence suggests there is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude the distribution of income in the United States is different today than it was in 2000.

NW Now Work Problem 11

If we compare the observed and expected counts, we see where the shift in income is occurring. Income below \$15,000 has observed counts above what is expected under the assumption of no change in the distribution of household income. For example, we would expect 148.5 households in the survey to have incomes under \$15,000 if there was no change in the distribution of income, but we observed 161 households in this income class. In addition, middle income levels are below what is expected. For example, for incomes from \$35,000 to \$49,999, we observed 184 households but expected 202.5 households if the distribution of incomes did not change since 2000. Lastly, at high levels of income (at least \$150,000), the observed counts are above what would be expected.

EXAMPLE 4**Conducting a Goodness-of-Fit Test****Table 3**

Day of Week	Frequency
Sunday	46
Monday	76
Tuesday	83
Wednesday	81
Thursday	81
Friday	80
Saturday	53

Problem An obstetrician wants to know whether the proportion of children born on each day of the week is the same. She randomly selects 500 birth records and obtains the data shown in Table 3 (based on data obtained from *Vital Statistics of the United States*, 2017).

Is there reason to believe that the day on which a child is born does not occur with equal frequency at the $\alpha = 0.01$ level of significance?

Approach Follow Steps 1 through 5 presented on pages 528–529.

Solution

Step 1 The null hypothesis is a statement of “no difference,” so we assume that the day on which a child is born occurs with equal frequency. If 1 represents Sunday, 2 represents Monday, and so on, we have

$$H_0: p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = \frac{1}{7}$$

H_1 : At least one of the proportions is different from the others.

Step 2 The level of significance is $\alpha = 0.01$.

Step 3

(a) The expected count for each category (day of the week), assuming the null hypothesis is true, is

$$500\left(\frac{1}{7}\right) \approx 71.4$$

(b) Since all expected counts are greater than or equal to 5, the requirements for the goodness-of-fit test are satisfied.

(c) The test statistic is

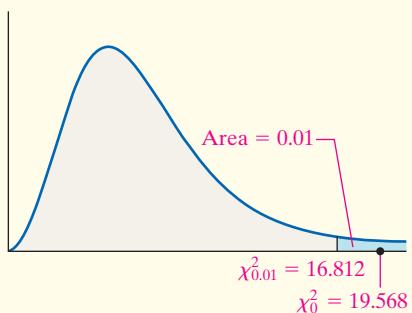
$$\begin{aligned} \chi_0^2 &= \frac{(46 - 500/7)^2}{500/7} + \frac{(76 - 500/7)^2}{500/7} + \frac{(83 - 500/7)^2}{500/7} + \frac{(81 - 500/7)^2}{500/7} \\ &\quad + \frac{(81 - 500/7)^2}{500/7} + \frac{(80 - 500/7)^2}{500/7} + \frac{(53 - 500/7)^2}{500/7} = 19.568 \end{aligned}$$

(continued)

Classical Approach

Step 4 There are $k = 7$ categories, so we find the critical value using $7 - 1 = 6$ degrees of freedom. The critical value is $\chi^2_{0.01} = 16.812$. See Figure 10.

Figure 10



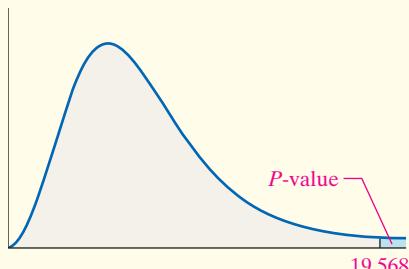
Because the test statistic, 19.568, is greater than the critical value, 16.812, we reject the null hypothesis.

P-Value Approach

Step 3 (continued)

(d) There are $k = 7$ categories. The P -value is the area under the chi-square distribution with $7 - 1 = 6$ degrees of freedom to the right of $\chi^2_0 = 19.568$, as shown in Figure 11.

Figure 11



Using Table VIII, we find the row that corresponds to 6 degrees of freedom. The value of 19.568 is greater than 18.548, which has an area under the chi-square distribution of 0.005 to the right, so the P -value is less than 0.005 (P -value < 0.005). Using technology, the exact P -value is 0.0033.

Step 4 Because the P -value is less than the level of significance, $\alpha = 0.01$, we reject the null hypothesis.

Step 5 There is sufficient evidence at the $\alpha = 0.01$ level of significance to reject the belief that the day of the week on which a child is born occurs with equal frequency. 

NW Now Work Problem 19

Whenever we are testing a hypothesis, the evidence can never prove the null hypothesis to be true. For example, in performing the goodness-of-fit test from Example 3, we tested

H_0 : the distribution of household income is the same today as in 2000

Had we failed to reject the null hypothesis, we would not be saying that the distribution today is the same as it was in 2000. We would be saying that we don't have enough evidence to conclude that the distribution has changed significantly. Unfortunately, goodness-of-fit tests cannot be used to test whether sample data follow a specific distribution. We can only say the data are consistent with a distribution stated in the null hypothesis.

Technology Step-by-Step

Goodness-of-Fit Test

TI-84 Plus*

- Enter the observed counts in L1 and enter the expected counts in L2.
- Press STAT, highlight TESTS, and select D: χ^2 GOF-Test
- Enter L1 after Observed:, and enter L2 after Expected:. Enter the appropriate degrees of freedom following df:. Highlight either Calculate or Draw and press ENTER.

*The TI-83 and older TI-84 Plus graphing calculators do not have this capability.

Minitab

- Enter the observed counts in column C1. Enter expected proportions or counts, assuming that the null hypothesis is true, in column C2, if necessary.
- Select the Stat menu, highlight Tables, then choose Chi-square Goodness-of-Fit Test (One Variable)
- Select the Observed Counts radio button and enter C1 in the cell. If you are testing equal proportions, select this radio button. If you entered expected proportions in C2, select the "Specific proportions" radio button and enter C2. If you entered expected counts in C2, select the "Proportions specified by historical counts" radio button and enter C2. Click OK.

Excel

- Load the XLSTAT Add-in.
- Enter the category names in column A. Enter the proportions used to formulate expected counts in column B. Enter the observed counts in column C.
- Select **Parametric tests**, highlight Multinomial goodness of fit.
- Place the cursor in the Frequencies cell. Highlight the data in column C. Place the cursor in the Expected proportions cell. Highlight the data in column B. Be sure the proportions radio button is selected. Check the Column labels box. Check the Chi-square test box. Enter a level of significance. Click OK.

StatCrunch

- Enter the observed counts in the first column. Enter the expected counts in the second column. Name the columns observed and expected.
- Select **Stat**, highlight **Goodness-of-fit**, then highlight **Chi-Square Test**.
- Select the column that contains the observed counts and select the column that contains the expected counts. Click Compute!.



12.1 Assess Your Understanding

Vocabulary and Skill Building

- True or False:** The shape of the chi-square distribution depends on the degrees of freedom.
- A _____ test is an inferential procedure used to determine whether a frequency distribution follows a specific distribution.
- Suppose there are n independent trials of an experiment with $k > 3$ mutually exclusive outcomes, where p_i represents the probability of observing the i th outcome. The _____ for each possible outcome are given by $E_i = \underline{\hspace{2cm}}$.
- What are the two requirements that must be satisfied to perform a goodness-of-fit test?

In Problems 5 and 6, determine the expected counts for each outcome.

NW 5.

$n = 500$				
p_i	0.2	0.1	0.45	0.25
Expected counts				

6.

$n = 700$				
p_i	0.15	0.3	0.35	0.20
Expected counts				

In Problems 7–10, determine (a) the χ^2 test statistic, (b) the degrees of freedom, (c) the critical value using $\alpha = 0.05$, and (d) test the hypothesis at the $\alpha = 0.05$ level of significance.

7. $H_0: p_A = p_B = p_C = p_D = \frac{1}{4}$

H_1 : At least one of the proportions is different from the others.

Outcome	A	B	C	D
Observed	30	20	28	22
Expected	25	25	25	25

8. $H_0: p_A = p_B = p_C = p_D = p_E = \frac{1}{5}$

H_1 : At least one of the proportions is different from the others.

Outcome	A	B	C	D	E
Observed	38	45	41	33	43
Expected	40	40	40	40	40

9. H_0 : The random variable X is binomial with $n = 4, p = 0.8$

H_1 : The random variable X is not binomial with $n = 4, p = 0.8$

X	0	1	2	3	4
Observed	1	38	132	440	389
Expected	1.6	25.6	153.6	409.6	409.6

10. H_0 : The random variable X is binomial with $n = 4, p = 0.3$

H_1 : The random variable X is not binomial with $n = 4, p = 0.3$

X	0	1	2	3	4
Observed	260	400	280	50	10
Expected	240.1	411.6	264.6	75.6	8.1

Applying the Concepts

- NW** 11. **Plain M&Ms** According to the manufacturer of M&Ms, 13% of the plain M&Ms in a bag should be brown, 14% yellow, 13% red, 24% blue, 20% orange, and 16% green. A student randomly selected a bag of plain M&Ms. He counted the number of M&Ms that were each color and obtained the results shown in the table. Test whether plain M&Ms follow the distribution stated by M&M/Mars at the $\alpha = 0.05$ level of significance.

Color	Frequency
Brown	57
Yellow	64
Red	54
Blue	75
Orange	86
Green	64

12. Peanut M&Ms According to the manufacturer of M&Ms, 12% of the peanut M&Ms in a bag should be brown, 15% yellow, 12% red, 23% blue, 23% orange, and 15% green. A student randomly selected a bag of peanut M&Ms. He counted the number of M&Ms that were each color and obtained the results shown in the table. Test whether peanut M&Ms follow the distribution stated by M&M/Mars at the $\alpha = 0.05$ level of significance.

Color	Frequency
Brown	53
Yellow	66
Red	38
Blue	96
Orange	88
Green	59

13. Benford's Law, Part I Our number system consists of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. The first significant digit in any number must be 1, 2, 3, 4, 5, 6, 7, 8, or 9 because we do not write numbers such as 12 as 012. Although we may think that each first digit appears with equal frequency so that each digit has a $\frac{1}{9}$ probability of being the first significant digit, this is not true. In 1881, Simon Newcomb discovered that first digits do not occur with equal frequency. This same result was discovered again in 1938 by physicist Frank Benford. After studying much data, he was able to assign probabilities of occurrence to the first digit in a number as shown.

Digit	1	2	3	4	5
Probability	0.301	0.176	0.125	0.097	0.079
Digit	6	7	8	9	
Probability	0.067	0.058	0.051	0.046	

Source: T. P. Hill, "The First Digit Phenomenon," *American Scientist*, July–August, 1998.

The probability distribution is now known as Benford's Law and plays a major role in identifying fraudulent data on tax returns and accounting books. For example, the following distribution represents the first digits in 200 allegedly fraudulent checks written to a bogus company by an employee attempting to embezzle funds from his employer.

First digit	1	2	3	4	5	6	7	8	9
Frequency	36	32	28	26	23	17	15	16	7

Source: *State of Arizona vs. Wayne James Nelson*.

- (a) Because these data are meant to prove that someone is guilty of fraud, what would be an appropriate level of significance when performing a goodness-of-fit test?
- (b) Using the level of significance chosen in part (a), test whether the first digits in the allegedly fraudulent checks obey Benford's Law.
- (c) Based on the results of part (b), do you think that the employee is guilty of embezzlement?

14. Benford's Law, Part II Refer to Problem 13. The distribution in the next column lists the first digit of the surface area (in square miles) of 335 rivers. Is there evidence at the $\alpha = 0.05$ level of significance to support the belief that the distribution follows Benford's Law?

First digit	1	2	3	4	5	6	7	8	9
Frequency	104	55	36	38	24	29	18	14	17

Source: Eric W. Weisstein, Benford's Law, from *MathWorld—A Wolfram Web Resource*.

15. Always Wear a Helmet The National Highway Traffic Safety Administration publishes reports about motorcycle fatalities and helmet use. The distribution shows the proportion of fatalities by location of injury for motorcycle accidents.

Location of injury	Multiple Locations	Head	Neck	Thorax	Abdomen/Lumbar/Spine
Proportion	0.57	0.31	0.03	0.06	0.03

The following data show the location of injury and number of fatalities for 2068 riders not wearing a helmet.

Location of injury	Multiple Locations	Head	Neck	Thorax	Abdomen/Lumbar/Spine
Number	1036	864	38	83	47

- (a) Does the distribution of fatal injuries for riders not wearing a helmet follow the distribution for all riders? Use the $\alpha = 0.05$ level of significance.
- (b) Compare the observed and expected counts for each category. What does this information tell you?

16. Religion in Congress Is the religious make-up of the United States Congress reflective of that in the general population? The following table shows the religious affiliation of the 535 members of the 116th Congress along with the religious affiliation of a random sample of 1200 adult Americans.

Religion	Number of Members	Sample of Residents
Protestant	293	616
Catholic	163	287
Mormon	10	20
Orthodox Christian	5	7
Jewish	34	20
Buddhist/Muslim/Hindu/Other	10	57
Unaffiliated/Don't Know/Refused	20	193

Source: Congressional Quarterly Roll Call and Pew Research.

- (a) Determine the probability distribution for the religious affiliation of the members of the 116th Congress.
- (b) Assuming the distribution of the religious affiliation of the adult American population is the same as that of the Congress, determine the number of adult Americans we would expect for each religion from a random sample of 1200 individuals.
- (c) The data in the third column represent the declared religion of a random sample of 1200 adult Americans (based on data obtained from Pew Research). Do the sample data suggest that the American population has the same distribution of religious affiliation as the 116th Congress?
- (d) Explain what the results of your analysis suggest.

17. Does It Matter Where I Sit? Does the location of your seat in a classroom play a role in attendance or grade? To answer this question, professors randomly assigned 400 students* in a general education physics course to one of four groups.

*The number of students was increased so that goodness-of-fit procedures could be used.

Source: Perkins, Katherine K. and Wieman, Carl E., "The Surprising Impact of Seat Location on Student Performance" *The Physics Teacher*, Vol. 43, Jan. 2005.

The 100 students in group 1 sat 0 to 4 meters from the front of the class, the 100 students in group 2 sat 4 to 6.5 meters from the front, the 100 students in group 3 sat 6.5 to 9 meters from the front, and the 100 students in group 4 sat 9 to 12 meters from the front.

- (a) For the first half of the semester, the attendance for the whole class averaged 83%. So, if there is no effect due to seat location, we would expect 83% of students in each group to attend. The data show the attendance history for each group. How many students in each group attended, on average? Is there a significant difference among the groups in attendance patterns? Use the $\alpha = 0.05$ level of significance.

Group	1	2	3	4
Attendance	0.84	0.84	0.84	0.81

- (b) For the second half of the semester, the groups were rotated so that group 1 students moved to the back of class and group 4 students moved to the front. The same switch took place between groups 2 and 3. The attendance for the second half of the semester averaged 80%. The data show the attendance records for the original groups (group 1 is now in back, group 2 is 6.5 to 9 meters from the front, and so on). How many students in each group attended, on average? Is there a significant difference in attendance patterns? Use the $\alpha = 0.05$ level of significance. Do you find anything curious about these data?

Group	1	2	3	4
Attendance	0.84	0.81	0.78	0.76

- (c) At the end of the semester, the proportion of students in the top 20% of the class was determined. Of the students in group 1, 25% were in the top 20%; of the students in group 2, 21% were in the top 20%; of the students in group 3, 15% were in the top 20%; of the students in group 4, 19% were in the top 20%. How many students would we expect to be in the top 20% of the class if seat location plays no role in grades? Is there a significant difference in the number of students in the top 20% of the class by group?

- (d) In earlier sections, we discussed results that were statistically significant, but did not have any practical significance. Discuss the practical significance of these results. In other words, given the choice, would you prefer sitting in the front or back?

- 18. Racial Profiling** On January 1, 2004, it became mandatory for all police departments in Illinois to record data pertaining to race from every traffic stop. Mundelein, Illinois, has been collecting data since 2000. Rather than using census data to determine the racial distribution of the village, they thought it better to use data based on who is using the roads in Mundelein. So they collected data on at-fault drivers involved in car accidents in the village (the implicit assumption here is that race is independent of fault in a crash) and obtained the following distribution of race for users of roads in Mundelein.

Race	White	African American	Hispanic	Asian
Proportion	0.719	0.028	0.207	0.046

Source: Village of Mundelein, Illinois.

The following data represent the races of all 9868 drivers who were stopped for a moving violation in the village of Mundelein in a recent year.

Race	White	African American	Hispanic	Asian
Proportion	7079	273	2025	491

Source: Village of Mundelein, Illinois.

- (a) Does the distribution of race in traffic stops reflect the distribution of drivers in Mundelein? In other words, is there any evidence of racial profiling in Mundelein? Use the $\alpha = 0.05$ level of significance.
(b) Compare observed and expected counts for each category. What does this information tell you?

- NW 19. The NHL** In his book *Outliers*, Malcolm Gladwell claims that more hockey players are born in January through March than in October through December. The following data show the number of players in the National Hockey League in the 2018–2019 season according to their birth month. Is there evidence to suggest that professional hockey players' birth dates are not uniformly distributed throughout the year at the $\alpha = 0.05$ level of significance?

Birth Month	Frequency
January–March	273
April–June	263
July–September	211
October–December	206

Source: Quanthockey.com

- 20. Bicycle Deaths** A researcher wanted to determine whether bicycle deaths were uniformly distributed over the days of the week. She randomly selected 200 deaths that involved a bicycle, recorded the day of the week on which the death occurred, and obtained the following results (the data are based on information obtained from the Insurance Institute for Highway Safety). Is there reason to believe that bicycle fatalities occur with equal frequency with respect to day of the week at the $\alpha = 0.05$ level of significance?

Day of the Week	Frequency	Day of the Week	Frequency
Sunday	16	Thursday	34
Monday	35	Friday	41
Tuesday	16	Saturday	30
Wednesday	28		

- 21. Rationalized Lies** Do people cheat or lie when the cheating or lying is not easy to identify (such as filing of taxes)? A total of 2568 college-aged subjects from various countries throughout the world rolled a single six-sided die twice. The subjects were told that the first roll counted in determining a reward and the second roll was only to determine whether the die was "working properly." Rewards were as follows: rolling a one meant earning 1 unit of the local currency (such as \$1), rolling a two meant earning 2 units, and so on—except that rolling a six meant earning nothing. The rolling was done unsupervised (although results were secretly recorded) with the subjects free to report the outcomes of their respective rolls of the die (thereby creating an opportunity to cheat or lie about the outcome).

Source: Gachter, Simon and Schulz, Jonathan, "Intrinsic Honesty and the Prevalence of Rule Violations Across Societies," *Nature* (24 March 2016) 531, 496–499.

- (a) If individuals do not lie about the outcome of the first roll of the die, what would you expect the distribution of outcomes to be?

The following distribution represents the outcomes based on the results reported in the study. The row "First Roll Frequency" represents the frequency each first roll resulted in the given outcome. The row "Reported Frequency" represents the outcome reported by the subjects.

Outcome	1	2	3	4	5	6
First Roll Frequency	420	403	432	399	442	472
Reported Frequency	192	327	508	630	823	88

- (b) Does the first roll frequency distribution conform to the model we would expect if individuals do not lie about the outcome? Use an $\alpha = 0.05$ level of significance.
 (c) Does the reported frequency distribution conform to the model we would expect if individuals do not lie about the outcome? Use an $\alpha = 0.05$ level of significance.
 (d) The following probability distribution represents the outcomes under the assumption the subjects are choosing the better outcome in the two rolls. For example, if the subject first rolls a two, then a four, the subject would simply report the four. Does the reported frequency distribution given above suggest subjects were choosing the better outcome?

Outcome	1	2	3	4	5	6
Probability	3/36	5/36	7/36	9/36	11/36	1/36

≈ 0.083 ≈ 0.139 ≈ 0.194 $= 0.25$ ≈ 0.306 ≈ 0.028

- (e) What does your data analysis suggest?



22. Pedestrian Deaths A researcher wanted to determine whether pedestrian deaths were uniformly distributed over the days of the week. She randomly selected 300 pedestrian deaths, recorded the day of the week on which the death occurred, and obtained the following results (the data are based on information obtained from the Insurance Institute for Highway Safety). Test the belief that the day of the week on which a fatality happens involving a pedestrian occurs with equal frequency at the $\alpha = 0.05$ level of significance.

Day of the Week	Frequency	Day of the Week	Frequency
Sunday	39	Thursday	41
Monday	40	Friday	49
Tuesday	30	Saturday	61
Wednesday	40		

23. Grade Distributions At Joliet Junior College, the mathematics department decided to offer a redesigned course in Intermediate Algebra, called the Math Redesign Program (MRP). Laura Egner, the coordinator of the program, wanted to determine if the grade distribution in the course differed from that of traditional courses. The following shows the grade distribution of traditional courses based on historical records and the observed grades in three pilot classes in which the MRP program was utilized.

	A	B	C	D	F	W
Traditional Distribution	0.133	0.191	0.246	0.104	0.114	0.212
Observed Counts in MRP Program	7	16	10	13	6	12

- (a) How many students were enrolled in the MRP program for the three pilot courses? Based on this result, determine the expected number of students for each grade assuming there is no difference in the distribution of MRP student grades and traditional grades.
 (b) Does the sample evidence suggest that the distribution of grades is different from the traditional classes at the $\alpha = 0.01$ level of significance?
 (c) Explain why it makes sense to use 0.01 as the level of significance.
 (d) Suppose the MRP pilot program continues in three more classes with the grades earned for all six pilot courses shown below. Notice that the sample size was simply doubled with the grade distribution remaining unchanged. Does this sample evidence suggest that the distribution of grades is different from the traditional classes at the $\alpha = 0.01$ level of significance? What does this result suggest about the role of sample size in the ability to reject a statement in the null hypothesis?

	A	B	C	D	F	W
Observed Counts in MRP Program	14	32	20	26	12	24

24. Population Shift An urban economist wonders if the distribution of U.S. residents in the United States is different today than it was in 2000. The table shows the distribution of residents in 2000 along with the observed counts of residents today based on a random sample of 1500 U.S. residents.

Region	Distribution in 2000	Observed Counts Today
Northeast	0.190	269
Midwest	0.229	327
South	0.356	554
West	0.225	350

- (a) Does the sample evidence suggest the distribution of U.S. residents has changed since 2000? Use the $\alpha = 0.05$ level of significance.
 (b) Suppose the survey was instead conducted on 5000 U.S. residents with the results shown in the following table. Compare the proportion of U.S. residents in each region based on the sample of 1500 U.S. residents versus 5000 U.S. residents. What do you notice?

Region	Observed Counts Today
Northeast	900
Midwest	1089
South	1846
West	1165

- (c) Does the sample evidence for the survey of 5000 U.S. residents suggest that the distribution of residents in the United States has changed since 2000?
 (d) Discuss the role sample size can play in determining whether the statement in the null hypothesis is rejected.

In Section 10.2, we tested hypotheses regarding a population proportion using a z -test. However, we can also use the chi-square goodness-of-fit test to test hypotheses with $k = 2$ possible outcomes. In Problems 25 and 26, we test hypotheses with the use of both methods.

25. Low Birth Weight According to the U.S. Census Bureau, 7.1% of all babies born are of low birth weight (<5 lb, 8 oz). An obstetrician wanted to know whether mothers between the ages of 35 and 39 years give birth to a higher percentage of low-birth-weight babies. She randomly selected 240 births for which the mother was 35 to 39 years old and found 22 low-birth-weight babies.

- (a) If the proportion of low-birth-weight babies for mothers in this age group is 0.071, compute the expected number of low-birth-weight births to 35- to 39-year-old mothers. What is the expected number of births to mothers 35 to 39 years old that are not low birth weight?
- (b) Answer the obstetrician's question at the $\alpha = 0.05$ level of significance using the chi-square goodness-of-fit test.
- (c) Answer the question by using the approach presented in Section 10.2.

26. Living Alone? In 2000, 25.8% of Americans 15 years of age or older lived alone, according to the Census Bureau. A sociologist, who believes that this percentage is greater today, conducts a random sample of 400 Americans 15 years of age or older and finds that 164 are living alone.

- (a) If the proportion of Americans aged 15 years or older living alone is 0.258, compute the following expected numbers:
Americans 15 years of age or older who live alone; Americans 15 years of age or older who do not live alone.
- (b) Test the sociologist's belief at the $\alpha = 0.05$ level of significance using the goodness-of-fit test.
- (c) Test the belief by using the approach presented in Section 10.2.

DATA 27. Putting It Together: The V-2 Rocket in London In Thomas Pynchon's book *Gravity Rainbow*, the characters discuss whether the Poisson probabilistic model can be used to describe the locations that Germany's feared V-2 rocket would land in. They divided London into 0.25-km² regions. They then counted the number of rockets that landed in each region, with the following results:

Number of rocket hits	0	1	2	3	4	5	6	7
Observed number of regions	229	211	93	35	7	0	0	1

Source: Lawrence Lesser. "Even More Fun Learning Statistics." *Stats: The Magazine for Students of Statistics*, Issue 49.

- (a) Estimate the mean number of rocket hits in a region by computing $\mu = \sum xP(x)$. Round your answer to four decimal places.
- (b) Explain why the requirements for conducting a goodness-of-fit test are not satisfied.
- (c) After consolidating the table, we obtain the distribution for rocket hits shown in the next column. Using the Poisson probability model, $P(x) = \frac{\mu^x}{x!} e^{-\mu}$, where μ is the mean from part (a), we can obtain the probability distribution for the number of rocket hits. Find the probability of 0 hits in a region. Then find the probability of 1 hit, 2 hits, 3 hits, and 4 or more hits.

Number of rocket hits	0	1	2	3	4 or more
Observed number of regions	229	211	93	35	8

- (d) A total of $n = 576$ rockets was fired. Determine the expected number of rocket hits, E , by computing $E = np$, where p is the probability of observing that particular number of hits in the region.
- (e) Conduct a goodness-of-fit test for the distribution using the $\alpha = 0.05$ level of significance. Do the rocket hits appear to be modeled by a Poisson random variable?

DATA 28. Putting It Together: Weldon's Dice On February 2, 1894, Frank Raphael Weldon wrote a letter to Francis Galton that included the results of 26,306 rolls of 12 dice. Weldon recorded the results such that a roll of a 5 or 6 resulted in a success, while a roll of 1, 2, 3, or 4 was a failure. The number of successes in each roll of the 12 dice are recorded in the table.

Number of Successes	Frequency	Number of Successes	Frequency
0	185	7	1331
1	1149	8	403
2	3265	9	105
3	5475	10	14
4	6114	11	4
5	5194	12	0
6	3067		

Source: *Chance Magazine*, Vol. 22, No. 4, 2009.

- (a) What is the probability of rolling a 5 or 6 when throwing a six-sided fair die?
- (b) Treating the probability determined in part (a) as the probability of success, compute the theoretical probability of 0, 1, 2, ..., 12 successes in throwing 12 dice.
- (c) Use the probabilities found in part (b) to determine the expected frequency in observing 0, 1, 2, ..., 12 successes after throwing the 12 dice 26,306 times.
- (d) Conduct a goodness-of-fit test to determine if the number of successes follows a binomial probability distribution.

Note: Combine 11 and 12 into a single bin.

Retain Your Knowledge

DATA 29. Buying a New Car How much does the typical person pay for a new 2019 Audi A4? The following data represent the selling price of a random sample of new A4s (in dollars).

41,215	41,303	41,453	41,898	40,988
40,078	41,215	39,623	42,352	41,898
40,533	42,580	40,306	41,670	39,851

Source: TrueCar.com

- (a) Is this data quantitative or qualitative?
- (b) Find the mean and median price of a new 2019 A4.
- (c) Find the standard deviation and interquartile range.
- (d) Verify it is reasonable to conclude that this data come from a population that is normally distributed.
- (e) Draw a boxplot of the data.
- (f) Estimate the typical price paid for a new 2019 Audi A4 with 90% confidence.
- (g) Would a 90% confidence interval for all new 2019 import vehicles be wider or narrower? Explain.

Explaining the Concepts

- 30.** Why is goodness of fit a good choice for the title of the procedures used in this section?
- 31.** Explain why chi-square goodness-of-fit tests are always right tailed.
- 32.** If the expected count of a category is less than 1, what can be done to the categories so that a goodness-of-fit test can still be performed?
- 33. Nine-enders** A “nine-ender” is an individual whose age ends in 9 (as in 29, 39, 49 years of age). Some studies have suggested that nine-enders tend to participate in certain activities at a higher rate than those whose age ends in the digits 0 through 8. For example, to run a marathon, participants must register with race

organizers and include their age. A study found that nine-enders are overrepresented among first-time marathoners by 48%! Across the entire life span, the age at which people were most likely to run their first marathon was twenty-nine. Another study suggested that nine-enders are less likely to register with online dating sites than non-nine-enders. Explain how a study might be structured to determine whether nine-enders enroll in certain events (such as marathons or online dating sites) at a rate different than one would expect if age is independent of participation. Include the null and alternative hypotheses along with the type of analysis you would conduct.

Source: Adam L. Alter and Hal E. Hershfield, “People Search for Meaning When They Approach a New Decade in Chronological Age,” *Proceedings of the National Academy of Sciences*, 111(48):17066–17070, 2014.

12.2 Tests for Independence and the Homogeneity of Proportions



Preparing for This Section Before getting started, review the following:

- The language of hypothesis tests (Section 10.1, pp. 435–441)
- Independent events (Section 5.3, pp. 253–254)
- Contingency tables and association (Section 4.4, pp. 206–212)
- Mean of a binomial random variable (Section 6.2, p. 320)
- Testing a hypothesis about two population proportions from independent samples (Section 11.1, pp. 479–483)

Objectives

- ① Perform a test for independence
- ② Perform a test for homogeneity of proportions

In Section 4.4, we discussed the association between two variables by examining the conditional distribution in a contingency table. The data presented in a contingency table represent the measurement of two categorical variables on each individual in the study and allow us to describe the association between the two variables.

We now introduce an inferential method to determine whether an apparent association between two categorical variables is statistically significant.

1 Perform a Test for Independence

Is there a relationship between marital status and happiness? The data in Table 4 show the marital status and happiness of individuals who participated in the General Social Survey.

Table 4

		Marital Status			
		Married	Widowed	Divorced/ Separated	Never Married
Happiness	Very Happy	600	63	112	144
	Pretty Happy	720	142	355	459
	Not Too Happy	93	51	119	127

For the data in this contingency table, we are looking at counts of a single group of individuals categorized based on the row variable “Happiness” and column variable

“Marital Status.” It seems reasonable to wonder whether one’s marital status is associated with one’s level of happiness. To determine whether there is a statistically significant association between two categorical variables, we perform a *chi-square test of independence*.

Definition

The **chi-square test for independence** is used to determine whether there is an association between a row variable and column variable in a contingency table constructed from sample data.

- The null hypothesis is that the variables are not associated, or independent.
- The alternative hypothesis is that the variables are associated, or dependent.

IN OTHER WORDS

In a chi-square independence test, the null hypothesis is always

H_0 : The variables are independent

The alternative hypothesis is always

H_1 : The variables are not independent

The idea behind testing these types of hypotheses is to compare actual counts to the counts we would expect assuming that the variables are independent (that is, assuming that the statement in the null hypothesis were true). If a significant difference between the actual counts and expected counts exists, we have evidence against the statement in the null hypothesis.

To obtain the **expected counts**, compute the number of observations expected within each cell under the assumption of independence (the statement in the null hypothesis is true). Recall that if two events E and F are independent, then $P(E \text{ and } F) = P(E) \cdot P(F)$. We can use the Multiplication Rule for Independent Events to obtain the expected proportion of observations within each cell under the assumption of independence. Then multiply each proportion by n , the sample size, to obtain the expected count within each cell.*

EXAMPLE 1

Determining the Expected Counts in a Test for Independence

Problem Is there a relationship between marital status and happiness? The data in Table 4 show the marital status and happiness of individuals who participated in the General Social Survey. Compute the expected counts within each cell, assuming that marital status and happiness are independent.

Approach

Step 1 Compute the row and column totals.

Step 2 Compute the relative marginal frequencies for the row variable and column variable.

Step 3 Use the Multiplication Rule for Independent Events to compute the expected proportion of observations within each cell, assuming independence.

Step 4 Multiply the proportions by 2985, the sample size, to obtain the expected counts within each cell.

Solution

Step 1 The row totals (blue) and column totals (red) are presented in Table 5.

Table 5

		Marital Status				
		Married	Widowed	Divorced/Separated	Never Married	Row Totals
Happiness	Very Happy	600	63	112	144	919
	Pretty Happy	720	142	355	459	1676
	Not Too Happy	93	51	119	127	390
	Column Totals	1413	256	586	730	2985

(continued)

*Recall that the expected value of a binomial random variable X for n independent trials of a binomial experiment with probability of success p is given by $E(X) = \mu = np$.

Step 2 The relative marginal frequencies for the row variable (happiness) and column variable (marital status) are presented in Table 6.

Table 6

		Marital Status				
		Married	Widowed	Divorced/Separated	Never Married	Relative Frequency
Happiness	Very Happy	600	63	112	144	$\frac{919}{2985} \approx 0.308$
	Pretty Happy	720	142	355	459	$\frac{1676}{2985} \approx 0.561$
	Not Too Happy	93	51	119	127	$\frac{390}{2985} \approx 0.131$
	Relative Frequency	$\frac{1413}{2985} \approx 0.473$	$\frac{256}{2985} \approx 0.086$	$\frac{586}{2985} \approx 0.196$	$\frac{730}{2985} \approx 0.245$	1

Step 3 Assume the variables are independent and use the Multiplication Rule for Independent Events to compute the expected proportions for each cell. For example, the proportion of individuals who are “very happy” and “married” would be

$$\begin{aligned} \left(\text{Proportion “very happy”} \right) &= (\text{proportion “very happy”}) \cdot (\text{proportion “married”}) \\ &= \left(\frac{919}{2985} \right) \left(\frac{1413}{2985} \right) \\ &= 0.145737 \end{aligned}$$

Table 7 shows the expected proportion in each cell, assuming independence.

Table 7

		Marital Status			
		Married	Widowed	Divorced/Separated	Never Married
Happiness	Very Happy	0.145737	0.026404	0.060440	0.075292
	Pretty Happy	0.265783	0.048153	0.110226	0.137312
	Not Too Happy	0.061847	0.011205	0.025649	0.031952

Step 4 Multiply the expected proportions in Table 7 by 2985, the sample size, to obtain the expected counts. See Table 8.

Table 8

		Marital Status			
		Married	Widowed	Divorced/Separated	Never Married
Happiness	Very Happy	2985(0.145737) = 435.025	2985(0.026404) = 78.816	2985(0.060440) = 180.413	2985(0.075292) = 224.747
	Pretty Happy	793.362	143.737	329.025	409.876
	Not Too Happy	184.613	33.447	76.562	95.377

If happiness and marital status are independent, we would expect a random sample of 2985 individuals to contain about 435 who are “very happy” and “married.”



The technique used in Example 1 to find the expected counts might seem rather tedious. It certainly would be more pleasant if we could determine a shortcut formula to obtain the expected counts. Consider the expected count for “very happy” and “married.” This expected count was obtained by multiplying the proportion of

individuals who are “very happy,” the proportion of individuals who are “married,” and the number of individuals in the sample. That is,

$$\begin{aligned}\text{Expected count} &= (\text{proportion “very happy”})(\text{proportion “married”})(\text{sample size}) \\ &= \frac{919}{2985} \cdot \frac{1413}{2985} \cdot 2985 \\ &= \frac{919 \cdot 1413}{2985} \quad \text{Cancel the 2985s} \\ &= \frac{(\text{row total for “very happy”})(\text{column total for “married”})}{\text{table total}}\end{aligned}$$

This leads to the following general result:

Expected Frequencies in a Chi-Square Test for Independence

To find the expected frequency in a cell when performing a chi-square independence test, multiply the cell’s row total by its column total and divide this result by the table total. That is,

$$\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{table total}} \quad (1)$$

For example, the expected frequency for “very happy and married” is

$$\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{table total}} = \frac{(919)(1413)}{2985} = 435.024$$

This result is close to that obtained in Table 8 (the difference exists because of rounding error; in fact, 435.024 is more accurate).

We need a test statistic and sampling distribution to see whether the expected and observed counts are significantly different.

Test Statistic for the Test of Independence

Let O_i represent the observed number of counts in the i th cell and E_i represent the expected number of counts in the i th cell. Then

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

approximately follows the chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom, where r is the number of rows and c is the number of columns in the contingency table, provided that (1) all expected frequencies are greater than or equal to 1 and (2) no more than 20% of the expected frequencies are less than 5.

In Example 1, there were $r = 3$ rows and $c = 4$ columns.

We now present a method for testing hypotheses regarding the association between two variables in a contingency table.

Chi-Square Test for Independence

To test hypotheses regarding the association between (or independence of) two variables in a contingency table, use the steps that follow:

Step 1 Determine the null and alternative hypotheses.

H_0 : The row variable and column variable are independent.

H_1 : The row variable and column variable are dependent.

Step 2 Choose a level of significance, α , depending on the seriousness of making a Type I error.

Step 3

(a) Calculate the expected frequencies (counts) for each cell in the contingency table using Formula (1).

(b) Verify that the requirements for the chi-square test for independence are satisfied:

1. All expected frequencies are greater than or equal to 1 (all $E_i \geq 1$).
2. No more than 20% of the expected frequencies are less than 5.

Classical Approach

Step 3 (continued)

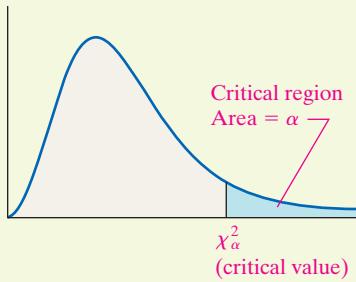
(c) Compute the test statistic

$$\chi^2_0 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Note: O_i is the observed frequency for the i th cell.

Step 4 Determine the critical value using Table VIII. All chi-square tests for independence are right-tailed tests, so the critical value is χ^2_α with $(r - 1)(c - 1)$ degrees of freedom, where r is the number of rows and c is the number of columns in the contingency table. See Figure 12.

Figure 12



Compare the critical value to the test statistic. If $\chi^2_0 > \chi^2_\alpha$, reject the null hypothesis.

Step 5 State the conclusion.

P-Value Approach

By Hand Step 3 (continued)

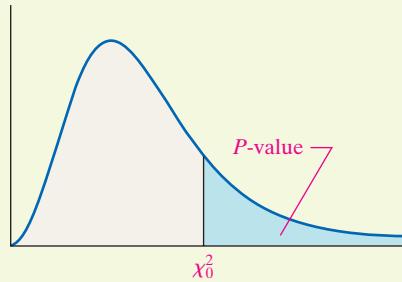
(c) Compute the test statistic

$$\chi^2_0 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Note: O_i is the observed frequency for the i th cell.

Step 4 Use Table VIII to determine an approximate P -value by determining the area under the chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom to the right of the test statistic, where r is the number of rows and c is the number of columns in the contingency table. See Figure 13.

Figure 13



Technology Step 3 (continued)

(c) Use a statistical spreadsheet or calculator with statistical capabilities to obtain the P -value. The directions for obtaining the P -value using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch are in the Technology Step-by-Step on page 549.

Step 4 If P -value $< \alpha$, reject the null hypothesis.

EXAMPLE 2 Performing a Chi-Square Test for Independence

Problem Does one's happiness depend on one's marital status? We present the data from Table 4 in Example 1 again in Table 9 to answer this question. Use the $\alpha = 0.05$ level of significance.

Table 9

		Marital Status			
		Married	Widowed	Divorced/Separated	Never Married
Happiness	Very Happy	600	63	112	144
	Pretty Happy	720	142	355	459
	Not Too Happy	93	51	119	127

Approach Follow Steps 1 through 5 just given.

Solution

Step 1 We want to determine whether happiness and marital status are dependent or independent. The null hypothesis is a statement of “no effect,” so we state the hypotheses as follows:

H_0 : Happiness and marital status are independent (not related)

H_1 : Happiness and marital status are dependent (related)

Step 2 The level of significance is $\alpha = 0.05$.

Step 3

(a) The expected frequencies were determined in Example 1. Table 10 shows the observed frequencies, with the expected frequencies in parentheses.

Table 10

		Marital Status			
		Married	Widowed	Divorced/Separated	Never Married
Happiness	Very Happy	600 (435.025)	63 (78.816)	112 (180.413)	144 (224.747)
	Pretty Happy	720 (793.362)	142 (143.737)	355 (329.025)	459 (409.876)
	Not Too Happy	93 (184.613)	51 (33.447)	119 (76.562)	127 (95.377)

(b) Since none of the expected frequencies are less than 5, the requirements for the goodness-of-fit test are satisfied.

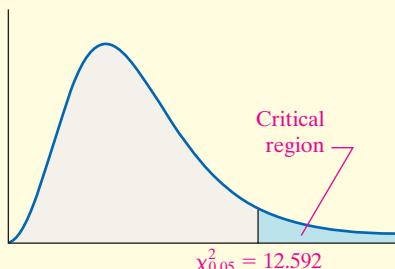
Classical Approach

Step 3 (continued)

$$\begin{aligned}
 \text{(c)} \quad \chi^2_0 &= \frac{(600 - 435.025)^2}{435.025} + \frac{(63 - 78.816)^2}{78.816} + \frac{(112 - 180.413)^2}{180.413} + \\
 &\dots + \frac{(119 - 76.562)^2}{76.562} + \frac{(127 - 95.377)^2}{95.377} \\
 &= 224.116
 \end{aligned}$$

Step 4 There are $r = 3$ rows and $c = 4$ columns. The critical value is $\chi^2_{0.05} = 12.592$ using $(r - 1)(c - 1) = (3 - 1)(4 - 1) = 6$ degrees of freedom. See Figure 14.

Figure 14



Because the test statistic, 224.116, is greater than the critical value, $\chi^2_{0.05} = 12.592$, we reject the null hypothesis.

P-Value Approach

By Hand Step 3 (continued)

$$\begin{aligned}
 \text{(c)} \quad \chi^2_0 &= \frac{(600 - 435.025)^2}{435.025} + \frac{(63 - 78.816)^2}{78.816} + \frac{(112 - 180.413)^2}{180.413} + \\
 &\dots + \frac{(119 - 76.562)^2}{76.562} + \frac{(127 - 95.377)^2}{95.377} \\
 &= 224.116
 \end{aligned}$$

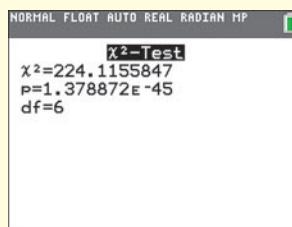
(d) There are $r = 3$ rows and $c = 4$ columns, so we find the *P*-value using $(r - 1)(c - 1) = (3 - 1)(4 - 1) = 6$ degrees of freedom. The *P*-value is the area under the chi-square distribution with 6 degrees of freedom to the right of $\chi^2_0 = 224.116$.

Using Table VIII, we find the row that corresponds to 6 degrees of freedom. The value of 224.116 is greater than 18.548. The area under the chi-square distribution to the right of 18.548 is 0.005. Because 224.116 is greater than 18.548, the *P*-value is less than 0.005, so we have *P*-value < 0.005.

Technology Step 3 (continued)

(c) Figure 15 on the next page shows the *P*-value obtained from a TI-84 Plus CE graphing calculator using the Calculate feature. The *P*-value is reported as 1.38×10^{-45} , which is very close to 0.

Figure 15



Step 4 Because the P -value is less than the level of significance, we reject the null hypothesis.

Step 5 The data suggest there is sufficient evidence, at the $\alpha = 0.05$ level of significance, to conclude that happiness and marital status are dependent. We conclude that happiness and marital status are associated with each other. 

NW Now Work Problems 7(b)–(e)

To see the association between happiness and marital status, draw a bar graph of the conditional distribution of happiness by marital status. Recall that a conditional distribution lists the relative frequency of each category of a variable, given a specific value of the other variable in a contingency table. For example, we can calculate the relative frequency of “very happy,” given that an individual is “married.” We repeat this for each remaining category of marital status.

EXAMPLE 3

Constructing a Conditional Distribution and Bar Graph

Problem Find the conditional distribution of happiness by marital status for the data in Table 9. Then draw a bar graph that represents the conditional distribution of happiness by marital status.

Approach First, compute the relative frequency for happiness, given that the individual is “married.” Then compute the relative frequency for happiness, given that the individual is “widowed,” and so on. For each marital status, draw three bars side by side. The horizontal axis represents marital status, and the vertical axis represents the relative frequency of the level of happiness.

Solution Start with the individuals who are “married.” The relative frequency with which we observe an individual who is “very happy,” given that the individual is “married,” is $\frac{600}{1413} = 0.425$. The relative frequency with which we observe an individual who is “pretty happy,” given that the individual is “married,” is $\frac{720}{1413} = 0.510$. The relative frequency with which we observe an individual who is “not too happy,” given that the individual is “married,” is $\frac{93}{1413} = 0.066$.

Now compute the relative frequency for each level of happiness, given that the individual is “widowed.” The relative frequency with which we observe an individual who is “very happy,” given that the individual is “widowed,” is $\frac{63}{256} = 0.246$. The relative frequency with which we observe an individual who is “pretty happy,” given that the individual is “widowed,” is $\frac{142}{256} = 0.555$. The relative frequency with which we observe an individual who is “not too happy,” given that the individual is “widowed,” is $\frac{51}{256} = 0.199$.

We repeat the process for “divorced/separated” and “never married” and obtain Table 11 on the next page.

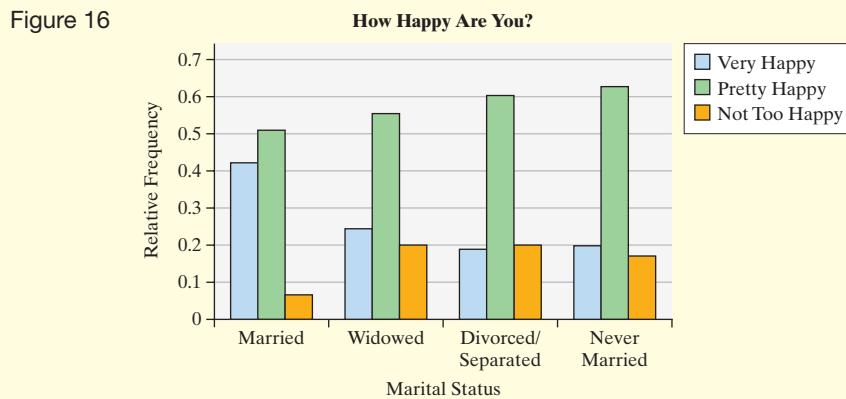
Table 11

		Marital Status			
		Married	Widowed	Divorced/Separated	Never Married
Happiness	Very Happy	$\frac{600}{1413} = 0.425$	$\frac{63}{256} = 0.246$	$\frac{112}{586} = 0.191$	$\frac{144}{730} = 0.197$
	Pretty Happy	$\frac{720}{1413} = 0.510$	$\frac{142}{256} = 0.555$	$\frac{355}{586} = 0.606$	$\frac{459}{730} = 0.629$
	Not Too Happy	$\frac{93}{1413} = 0.066$	$\frac{51}{256} = 0.199$	$\frac{119}{586} = 0.203$	$\frac{127}{730} = 0.174$

From the conditional distribution by marital status, the association between happiness and marital status should be apparent. The proportion of individuals who are “very happy” is highest in the “married” category. In addition, the proportion of individuals who are “not too happy” is lowest in the “married” category.

Figure 16 shows the graph of the conditional distribution. The blue bars represent the proportion of individuals who are “Very Happy” for each marital status, the green bars represent the proportion of individuals who are “Pretty Happy” for each marital status, and the orange bars represent the proportion of individuals who are “Not Too Happy” for each marital status. From the bar graph, it is clear that a higher proportion of married individuals are “very happy” compared with the other categories of marital status.

Figure 16


NW Now Work Problem 7(f)

② Perform a Test for Homogeneity of Proportions

A second type of chi-square test can be used to compare the population proportions from two or more independent samples. The test is an extension of the test comparing two proportions from independent samples introduced in Section 11.1.

Definition

In a **chi-square test for homogeneity of proportions**, we test whether different populations have the same proportion of individuals with some characteristic.

IN OTHER WORDS

The chi-square test for homogeneity of proportions is used to compare proportions from two or more populations.

For example, we might look at the proportion of individuals who experience headaches as a side effect for a placebo group (group 1) and for one experimental group that receives 50 milligrams (mg) per day of a medication (group 2) and another that

receives 100 mg per day (group 3). We assume that the proportion of individuals who experience a headache as a side effect is the same in each population (because the null hypothesis is a statement of “no difference”). So our hypotheses are

$$H_0: p_1 = p_2 = p_3$$

H_1 : At least one of the population proportions is different from the others.

The procedures for performing a test of homogeneity are identical to those for a test of independence.

While the procedures for the test for independence and the test of homogeneity of proportions are the same, the data differ. How? In the test for independence, we are measuring two variables (such as marital status and level of happiness) on each individual. In other words, a single population is segmented based on the value of two variables. In the test of homogeneity of proportions, we consider whether the proportion of individuals among different populations have the same value.

So, if you have a single population in which two variables are measured on each individual to assess whether one variable might be associated with another, conduct a test of independence. If you have two or more populations in which you want to determine equality of proportions among the populations, conduct a test of homogeneity of proportions.

EXAMPLE 4 A Test for Homogeneity of Proportions

Problem Zocor is a drug manufactured by Merck and Co. that is meant to reduce the level of LDL (bad) cholesterol and increase the level of HDL (good) cholesterol. In clinical trials of the drug, patients were randomly divided into three groups. Group 1 received Zocor; group 2 received a placebo; group 3 received cholestyramine, a cholesterol-lowering drug that is currently available. Table 12 contains the number of patients in each group who did and did not experience abdominal pain as a side effect. Is there evidence to indicate that the proportion of subjects in each group who experienced abdominal pain is different at the $\alpha = 0.01$ level of significance?

Table 12

	Group 1 (Zocor)	Group 2 (placebo)	Group 3 (cholestyramine)
Number of people who experienced abdominal pain	51	5	16
Number of people who did not experience abdominal pain	1532	152	163

Source: Merck and Co.

Approach Follow Steps 1 through 5 on pages 541–542.

Solution

Step 1 The null hypothesis is a statement of “no difference,” so the proportion of subjects in each group who experienced abdominal pain are equal. We state the hypotheses as follows:

$$H_0: p_1 = p_2 = p_3$$

H_1 : At least one of the proportions is different from the others.

Here, p_1 , p_2 , and p_3 are the proportions in groups 1, 2, and 3, respectively.

Step 2 The level of significance is $\alpha = 0.01$.

Step 3

(a) The expected frequency of subjects who experienced abdominal pain in group 1 is the product of the row total of individuals who experienced abdominal pain and the column total of individuals in group 1, divided by the total number of subjects in the study. So

$$E = \frac{72 \cdot 1583}{1919} = 59.393$$

Table 13 contains the row and column totals, the observed frequencies, and the expected frequencies (in parentheses).

Table 13

	Observed (and Expected) Frequencies			Row Totals
	Group 1 (Zocor)	Group 2 (placebo)	Group 3 (cholestyramine)	
Number of people who experienced abdominal pain	51 (59.393)	5 (5.891)	16 (6.716)	72
Number of people who did not experience abdominal pain	1532 (1523.607)	152 (151.109)	163 (172.284)	1847
Column totals	1583	157	179	1919

(b) Since none of the expected frequencies are less than 5, the requirements for the test of homogeneity of proportions are satisfied.

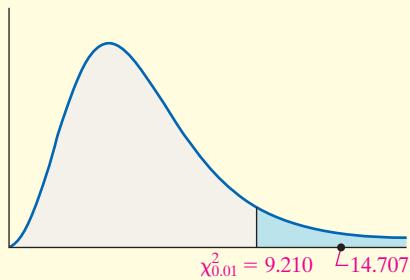
Classical Approach**Step 3 (continued)**

(c) The test statistic is

$$\begin{aligned}\chi^2_0 &= \frac{(51 - 59.393)^2}{59.393} + \frac{(5 - 5.891)^2}{5.891} + \frac{(16 - 6.716)^2}{6.716} \\ &\quad + \frac{(1532 - 1523.607)^2}{1523.607} + \frac{(152 - 151.109)^2}{151.109} + \frac{(163 - 172.284)^2}{172.284} \\ &= 14.707\end{aligned}$$

Step 4 There are $r = 2$ rows and $c = 3$ columns, so we find the critical value using $(2 - 1)(3 - 1) = 2$ degrees of freedom. The critical value is $\chi^2_{0.01} = 9.210$. See Figure 17.

Figure 17



Because the test statistic, 14.707, is greater than the critical value, 9.210, we reject the null hypothesis.

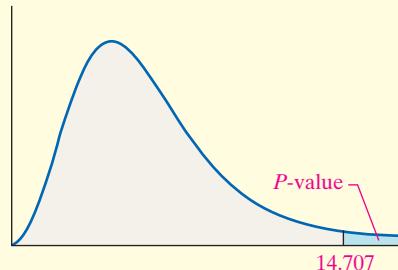
P-Value Approach**By Hand Step 3 (continued)**

(c) The test statistic is

$$\begin{aligned}\chi^2_0 &= \frac{(51 - 59.393)^2}{59.393} + \frac{(5 - 5.891)^2}{5.891} + \frac{(16 - 6.716)^2}{6.716} \\ &\quad + \frac{(1532 - 1523.607)^2}{1523.607} + \frac{(152 - 151.109)^2}{151.109} + \frac{(163 - 172.284)^2}{172.284} \\ &= 14.707\end{aligned}$$

(d) There are $r = 2$ rows and $c = 3$ columns, so we find the P -value using $(2 - 1)(3 - 1) = 2$ degrees of freedom. The P -value is the area under the chi-square distribution with 2 degrees of freedom to the right of $\chi^2_0 = 14.707$, as shown in Figure 18.

Figure 18



Using Table VIII, in the row corresponding to 2 degrees of freedom, the area under the chi-square distribution with 2 degrees of freedom, to the right of 10.597, is 0.005. Because 14.707 is greater than 10.597, the P -value is less than 0.005. So P -value < 0.005 .

(continued)

Technology Step 3

(c) Figure 19 shows the P -value obtained from StatCrunch. The P -value is reported as 0.0006.

Figure 19 Contingency table results:

Rows: var1

Columns: None

	Group 1	Group 2	Group 3	Total
Pain	51	5	16	72
No pain	1532	152	163	1847
Total	1583	157	179	1919

Chi-Square test:

Statistic	df	Value	P-value
Chi-square	2	14.706513	0.0006

Step 4 Because the P -value is less than the level of significance, $\alpha = 0.01$, we reject the null hypothesis.

CAUTION!

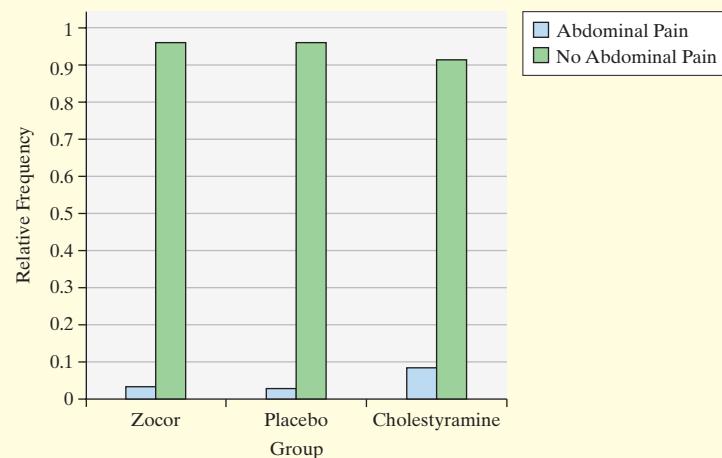
If we reject the null hypothesis in a chi-square test for homogeneity, we are saying there is sufficient evidence to believe that at least one proportion is different from the others. However, it does not tell us which proportions differ.

Step 5 The data suggest that there is sufficient evidence at the $\alpha = 0.01$ level of significance to reject the null hypothesis that the proportion of subjects in each group who experience abdominal pain are equal. We conclude that at least one of the three groups experiences abdominal pain at a rate that is different from the other two groups.

Figure 20 shows the conditional bar graph by treatment. From the graph, it is apparent that a higher proportion of patients taking cholestyramine experience abdominal pain as a side effect.

Figure 20

Patients Reporting Abdominal Pain by Treatment



NW Now Work Problem 15

**If Model Requirements Fail**

Recall that the requirements for performing a chi-square test are that all expected frequencies are greater than 1 and that at most 20% of the expected frequencies can be less than 5. If these requirements are not satisfied, the researcher has one of two options: (1) combine two or more columns (or rows) to increase the expected frequencies or (2) increase the sample size.

Technology Step-by-Step

Chi-Square Tests

TI-83/84 Plus

- Access the MATRIX menu. Highlight the EDIT menu, and select 1:[A].
- Enter the number of rows and columns of the contingency table (matrix).
- Enter the cell entries for the observed matrix, and press 2nd QUIT.
- Press STAT, highlight the TESTS menu, and select C: χ^2 – Test . . .
- With the cursor after Observed:, enter matrix [A] by accessing the MATRIX menu, highlighting NAMES, and selecting 1:[A].
- With the cursor after Expected:, enter matrix [B] by accessing the MATRIX menu, highlighting NAMES, and selecting 2:[B].
- Highlight Calculate or Draw, and press ENTER.

Minitab

- Enter the data as a contingency table into the MINITAB spreadsheet.
- Select the Stat menu, highlight Tables, and select Chi-Square Test for Association . . .
- From the pull-down menu, select “Summarized data in a two-way table.” Select the columns that contain the data, and click OK.

Excel

- Load the XLSTAT Add-in.
- Enter the data in a contingency table. Column A should be the row variable. For example, for the data in Table 4, column A is level of happiness. Each subsequent column

is the counts of each category of the column variable. For the data in Table 4, enter the counts for each marital status. Title each column (including the first column indicating the row variable).

- Select Describing data; highlight Contingency table.
- Place the cursor in the Contingency table cell. Highlight the table in the spreadsheet. Check the box Labels included. Click the Options tab.
- Check the box Chi-square test. Choose a level of significance. Click OK.

StatCrunch

- If the data are already in a contingency table, enter them into the spreadsheet. The first column is the row variable. For example, for the data in Table 4, the first column is level of happiness. Each subsequent column is the counts of each category of the column variable. For the data in Table 4, enter the counts for each marital status. Title each column (including the first column indicating the row variable). If the data are not in a contingency table, enter each variable in a column and name the column variable.
- Select Stat, highlight Tables, select Contingency, then highlight With Data or With Summary.
- Select the column variable(s). Then select the row variable. For example, the data in Table 4 has four column variables (“Married,” and so on) and the row label is “Happiness.” Decide what values you want displayed. Highlight “Chi-Square test for independence” under Hypothesis tests. Click Compute!.



12.2 Assess Your Understanding

Vocabulary and Skill Building

- 1. True or False:** The expected frequencies in a chi-square test for independence are found using the formula

$$\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{table total}}$$

- 2.** In a chi-square test for _____ of proportions, we test whether different populations have the same proportion of individuals with some characteristic.
- 3.** The following table contains observed values and expected values in parentheses for two categorical variables, X and Y , where variable X has three categories and variable Y has two categories:

	X_1	X_2	X_3
Y_1	34 (36.26)	43 (44.63)	52 (48.11)
Y_2	18 (15.74)	21 (19.37)	17 (20.89)

- (a)** Compute the value of the chi-square test statistic.
(b) Test the hypothesis that X and Y are independent at the $\alpha = 0.05$ level of significance.

- 4.** The following table contains observed values and expected values in parentheses for two categorical variables, X and Y , where variable X has three categories and variable Y has two categories.

	X_1	X_2	X_3
Y_1	87 (75.12)	74 (80.43)	34 (39.46)
Y_2	12 (23.88)	32 (25.57)	18 (12.54)

- (a)** Compute the value of the chi-square test statistic.
(b) Test the hypothesis that X and Y are independent at the $\alpha = 0.05$ level of significance.
- 5.** The following table contains the number of successes and failures for three categories of a variable.

	Category 1	Category 2	Category 3
Success	76	84	69
Failure	44	41	49

Test whether the proportions are equal for each category at the $\alpha = 0.01$ level of significance.

6. The following table contains the number of successes and failures for three categories of a variable.

Category 1	Category 2	Category 3	
Success	204	199	214
Failure	96	121	98

Test whether the proportions are equal for each category at the $\alpha = 0.01$ level of significance.

Applying the Concepts

- NW 7. Family Structure and Sexual Activity** A sociologist wants to discover whether the sexual activity of females between the ages of 15 and 19 years and family structure are associated. She randomly selects 380 females between the ages of 15 and 19 years and asks each to disclose her family structure at age 14 and whether she has had sexual intercourse. The results are shown in the table. Data are based on information obtained from the National Center for Health Statistics.

Family Structure				
Both Biological		Single Parent	Parent and Stepparent	Nonparental Guardian
Had Sexual or Adoptive Parents	Yes	64	59	44
No	86	41	36	18

- (a) Compute the expected values of each cell under the assumption of independence.
 (b) Verify that the requirements for performing a chi-square test of independence are satisfied.
 (c) Compute the chi-square test statistic.
 (d) Test whether family structure and sexual activity of 15- to 19-year-old females are independent at the $\alpha = 0.05$ level of significance.
 (e) Compare the observed frequencies with the expected frequencies. Which cell contributed most to the test statistic? Was the expected frequency greater than or less than the observed frequency? What does this information tell you?
 (f) Construct a conditional distribution by family structure and draw a bar graph. Does this evidence support your conclusion in part (d)?

- 8. Prenatal Care** An obstetrician wants to learn whether the amount of prenatal care and the wantedness of the pregnancy are associated. He randomly selects 939 women who had recently given birth and asks them to disclose whether their pregnancy was intended, unintended, or mistimed. In addition, they were to disclose when they started receiving prenatal care, if ever. The results of the survey are as follows:

Months Pregnant before Prenatal Care Began			
Wantedness of Pregnancy	Less Than 3 Months	3 to 5 Months	More Than 5 Months (or never)
Intended	593	26	33
Unintended	64	8	11
Mistimed	169	19	16

- (a) Compute the expected values of each cell under the assumption of independence.

- (b) Verify that the requirements for performing a chi-square test of independence are satisfied.
 (c) Compute the chi-square test statistic.
 (d) Test whether prenatal care and the wantedness of pregnancy are independent at the $\alpha = 0.05$ level of significance.
 (e) Compare the observed frequencies with the expected frequencies. Which cell contributed most to the test statistic? Was the expected frequency greater than or less than the observed frequency? What does this information tell you?
 (f) Construct a conditional distribution by wantedness of the pregnancy and draw a bar graph. Does this evidence support your conclusion in part (d)?

- DATA 9. Health and Happiness** Are health and happiness related? The following data represent the level of happiness and level of health for a random sample of individuals from the General Social Survey.

Health				
Happiness	Excellent	Good	Fair	Poor
	Very Happy	271	261	82
Pretty Happy	247	567	231	53
Not Too Happy	33	103	92	36

Source: General Social Survey.

- (a) Does the evidence suggest that health and happiness are related? Use the $\alpha = 0.05$ level of significance.
 (b) Construct a conditional distribution of happiness by level of health and draw a bar graph.
 (c) Write a few sentences that explain the relation, if any, between health and happiness.

- DATA 10. Health and Education** Does amount of education play a role in the healthiness of an individual? The following data represent the level of health and the highest degree earned for a random sample of individuals from the General Social Survey.

Health				
Education	Excellent	Good	Fair	Poor
Less than High School	72	202	199	62
High School	465	877	358	108
Junior College	80	138	49	11
Bachelor	229	276	64	12
Graduate	130	147	32	2

Source: General Social Survey.

- (a) Does the evidence suggest that health and education are related? Use the $\alpha = 0.05$ level of significance.
 (b) Construct a conditional distribution of health by level of education and draw a bar graph.
 (c) Write a few sentences that explain the relation, if any, between health and education. Can you think of any lurking variables that help to explain the relation between these two variables?

- 11. Social Well-being and Obesity** The Gallup Organization conducted a survey asking individuals questions pertaining to social well-being such as strength of relationship with

spouse, partner, or closest friend, making time for trips or vacations, and having someone who encourages them to be healthy. Social well-being scores were determined based on answers to these questions and used to categorize individuals as thriving, struggling, or suffering in their social well-being. In addition, body mass index (BMI) was determined based on height and weight of the individual. This allowed for classification as obese, overweight, normal weight, or underweight. The data in the following contingency table are based on the results of this survey.

	Thriving	Struggling	Suffering
Obese	202	250	102
Overweight	294	302	110
Normal Weight	300	295	103
Underweight	17	17	8

- (a) Researchers wanted to determine whether the sample data suggest there is an association between weight classification and social well-being. Explain why this data should be analyzed using a chi-square test for independence.
- (b) Do the sample data suggest that weight classification and social well-being are related? Use the $\alpha = 0.05$ level of significance.
- (c) Draw a conditional bar graph of the data by weight classification.
- (d) Write some general conclusions based on the results from parts (b) and (c).

12. Profile of Smokers The following data represent the smoking status from a random sample of 1054 U.S. residents 18 years or older by level of education.

Number of Years of Education	Smoking Status		
	Current	Former	Never
<12	178	88	208
12	137	69	143
13–15	44	25	44
16 or more	34	33	51

Source: National Health Interview Survey.

- (a) Test whether smoking status and level of education are independent at the $\alpha = 0.05$ level of significance.
- (b) Construct a conditional distribution of smoking status by number of years of education, and draw a bar graph. Does this evidence support your conclusion in part (a)?

13. Efficacy of e-Cigs Do electronic cigarettes assist in helping individuals quit smoking? Researchers found 300 current smokers to volunteer for a study in which each was randomly assigned to one of three treatment groups. Group 1 received an electronic cigarette (e-cig) in which each cartridge contained 7.2 mg of nicotine, Group 2 received an e-cig that contained 5.4 mg of nicotine, and Group 3 received an e-cig that contained no nicotine. The subjects did not know which group they were assigned. During the course of the 52-week intervention, subjects dropped out of the study. At the end of the study 65 subjects remained in Group 1, 63 in Group 2, and 55 in Group 3. After 52 weeks, it was determined via questionnaire whether the subject quit smoking entirely. Results of the study are presented in the following table.

	Group 1	Group 2	Group 3
Did Not Quit	52	54	51
Quit	13	9	4

Source: Caponnetto P, Campagna D, Cibella F, Morjaria JB, Caruso M, et al. (2013) EffiCieNcy and Safety of an eLectronic cigAreTte (ECLAT) as Tobacco Cigarettes Substitute: A Prospective 12-Month Randomized Control Design Study. *PLoS ONE* 8(6): e66317. doi:10.1371/journal.pone.0066317

- (a) What type of experimental design was used in this study?
- (b) Researchers wanted to know whether electronic cigarettes may be used to help individuals abstain from cigarette smoking. What is the response variable? Is it qualitative or quantitative?
- (c) To what population do the results of this study apply?
- (d) State the null and alternative hypotheses.
- (e) Does the evidence suggest e-cigs are effective in helping individuals abstain from cigarette smoking? Use the $\alpha = 0.05$ level of significance.
- (f) Draw a conditional bar graph by group. Explain what the graph suggests.
- (g) Write a conclusion for this hypothesis test.

14. Celebrex Celebrex, a drug manufactured by Pfizer, Inc., is used to relieve symptoms associated with osteoarthritis and rheumatoid arthritis in adults. In clinical trials of the medication, some subjects reported dizziness as a side effect. The researchers wanted to discover whether the proportion of subjects taking Celebrex who reported dizziness as a side effect differed significantly from that for other treatment groups. The following data were collected.

Side Effect	Drug				
	Celebrex	Placebo	Naproxen	Diclofenac	Ibuprofen
Dizziness	83	32	36	5	8
No dizziness	4063	1832	1330	382	337

Source: Pfizer, Inc.

- (a) Test whether the proportion of subjects within each treatment group who experienced dizziness are the same at the $\alpha = 0.01$ level of significance.
- (b) Construct a conditional distribution of side effect by treatment and draw a bar graph. Does this evidence support your conclusion in part (a)?

NW 15. What's in a Word? In a recent survey conducted by the Pew Research Center, a random sample of adults 18 years of age or older living in the continental United States was asked their reaction to the word *socialism*. In addition, the individuals were asked to disclose which political party they most associate with. Results of the survey are given in the table.

	Democrat	Independent	Republican
Positive	220	144	62
Negative	279	410	351

Source: Pew Research.

- (a) Explain why this data should be analyzed by homogeneity of proportions.
- (b) Does the evidence suggest individuals within each political affiliation react differently to the word *socialism*? Use the $\alpha = 0.05$ level of significance.
- (c) Construct a conditional distribution of reaction by political party.
- (d) Write a summary about the “partisan divide” regarding reaction to the word *socialism*.

16. What's in a Word? Part II In a recent survey conducted by the Pew Research Center, a random sample of adults 18 years of age or older living in the continental United States was asked their reaction to the word *capitalism*. In addition, the individuals were asked to disclose which political party they most associate with. Results of the survey are given in the table below.

	Democrat	Independent	Republican
Positive	235	288	256
Negative	264	266	157

Source: Pew Research.

- (a) Does the evidence suggest individuals within each political affiliation react differently to the word *capitalism*? Use the $\alpha = 0.05$ level of significance.
- (b) Construct a conditional distribution of reaction by political party.
- (c) Write a summary about the “partisan divide” regarding reaction to the word *capitalism*.
- (d) Compare the results of this problem with that of Problem 15. Write a short report detailing the findings of the two survey questions.



17. Dropping a Course A survey was conducted at a community college of 50 randomly selected students who dropped a course in the current semester to learn why students drop courses. “Personal” drop reasons include financial, transportation, family issues, health issues, and lack of child care. “Course” drop reasons include reducing one’s load, being unprepared for the course, the course was not what was expected, dissatisfaction with teaching, and not getting the desired grade. “Work” drop reasons include an increase in hours, a change in shift, and obtaining fulltime employment. Go to www.pearsonhighered.com/sullivanstats to obtain the data file 12_2_17 using the file format of your choice for the version of the text you are using.

- (a) Construct a contingency table for the two variables.
- (b) Test whether gender is independent of drop reason at the $\alpha = 0.1$ level of significance.
- (c) Construct a conditional distribution of drop reason by gender and draw a bar graph. Does this evidence support your conclusion in part (b)?



18. Political Affiliation In the Sullivan Statistics Survey, respondents were asked to disclose their political affiliation (Democrat, Independent, Republican) and also answer the question: “Would you be willing to pay higher taxes if the tax revenue went directly toward deficit reduction?” Go to www.pearsonhighered.com/sullivanstats to obtain the data file SullivanSurveyI using the file format of your choice for the version of the text you are using. Create a contingency table and determine whether the results suggest there is an association between political affiliation and willingness to pay higher taxes to directly reduce the federal debt. Use the $\alpha = 0.05$ level of significance.

19. Credit Risk Traditional underwriting to determine the risks associated with lending include credit scores, income, and employment history. The online lender ZestFinance used data analysis to find that people who fill out loan applications using all capital letters default more often than those who use all lower case letters. In addition, people who fill out the application using upper and lowercase letters accurately default at the lowest rate. Explain how to obtain and analyze data to determine whether the method used to fill out loan applications results in different default rates.

20. Insurance and Credit Scores A study by InsuranceQuotes.com found that homeowners with poor credit pay 91% more for home insurance than people with excellent credit.

- (a) A quote in the article stated, “Insurers have found a direct correlation between a consumer’s credit and the likelihood that he or she will make a home (or auto) claim.” Explain what is wrong with this quote.
- (b) Credit scores may be classified as Excellent, Good, Fair, and Poor. Explain how you might go about deciding whether credit scores might be used to determine whether an individual files a claim on his or her homeowner’s insurance policy or not. Include an explanation of the type of inferential procedure you would use.

21. The Process of Statistics—Statistics Pathway Researchers at the City University of New York (CUNY) identified 717 students who originally placed into an Elementary Algebra course. The students agreed to participate in a study related to the roll of corequisite remediation and study skills. In this study, the 717 students were randomly assigned to one of three courses: Course 1: traditional Elementary Algebra, Course 2: Elementary Algebra with workshops designed to improved study skills, or Course 3: Elementary Statistics with corequisite remediation and study skills workshops. The courses were all taught at CUNY for one semester. At the end of the semester the instructors for the course reported whether the student passed the class, or not. The data below show the outcome of the study.

Source: W., A., Watanabe-Rose, M., & Douglas, D. (2016). “Should Students Assessed as Needing Remedial Mathematics Take College-Level Quantitative Courses Instead? A Randomized Controlled Trial.” *Educational Evaluation and Policy Analysis*, 38(3):578–598. <https://doi.org/10.3102/0162373716649056>

	Course 2		
	Course 1 Elementary Algebra	Elementary Algebra with Workshop	Course 3 Elementary Statistics
Passed	95	102	138
Did Not Pass	149	125	108

- (a) What type of experimental design was this study?
- (b) What is the treatment? How many levels does it have?
- (c) What is the response variable in this study? Is it qualitative or quantitative?
- (d) The researchers wanted to know if the evidence suggested a difference in the pass rate among the three courses. Perform the appropriate test to analyze the evidence.
- (e) What variables were controlled and held fixed in this study? Are there any variables that were uncontrolled?
- (f) Suppose Teacher A was assigned to teach Course 1, Teacher B was assigned to teach Course 2, and Teacher C was assigned to teach Course 3. Explain how teacher may be a confounding variable in the study.
- (g) Historically, the pass rate of Elementary Algebra at CUNY is 0.37. Does the evidence suggest the workshop in Elementary Algebra offered in Course 2 was helpful in terms of improving pass rates?

22. Putting It Together: Women, Aspirin, and Heart Attacks

In a famous study by the Physicians Health Study Group from Harvard University from the late 1980s, 22,000 healthy male physicians were randomly divided into two groups; half the physicians took aspirin every other day, and the others were given a placebo. Of the physicians in the aspirin group, 104 heart

attacks occurred; of the physicians in the placebo group, 189 heart attacks occurred. The results were statistically significant, which led to the advice that males should take an aspirin every other day in the interest of reducing the chance of having a heart attack. Does the same advice apply to women?

In a randomized, placebo-controlled study, 39,876 healthy women 45 years of age or older were randomly divided into two groups. The women in group 1 received 100 mg of aspirin every other day; the women in group 2 received a placebo every other day. The women were monitored for 10 years to determine if they experienced a cardiovascular event (such as heart attack or stroke). Of the 19,934 in the aspirin group, 477 experienced a heart attack. Of the 19,942 women in the placebo group, 522 experienced a heart attack. *Source:* Paul M. Ridker et al. “A Randomized Trial of Low-Dose Aspirin in the Primary Prevention of Cardiovascular Disease in Women.” *New England Journal of Medicine* 352:1293–1304.

- (a) What is the population being studied? What is the sample?
- (b) What is the response variable? Is it qualitative or quantitative?
- (c) What are the treatments?
- (d) What type of experimental design is this?
- (e) How does randomization deal with the explanatory variables that were not controlled in the study?
- (f) Determine whether the proportion of cardiovascular events in each treatment group is different using a two-sample Z-test for comparing two proportions. Use the $\alpha = 0.05$ level of significance. What is the test statistic?
- (g) Determine whether the proportion of cardiovascular events in each treatment group is different using a chi-square test for homogeneity of proportions. Use the $\alpha = 0.05$ level of significance. What is the test statistic?
- (h) Square the test statistic from part (f) and compare it to the test statistic from part (g). What do you conclude?

23. Putting It Together: Corequisite College Algebra

During the fall semester of 2014, the University of North Georgia developed a corequisite College Algebra course. In this approach, students who would otherwise place in a Learning Support (LS) course in one semester and then enroll in College Algebra the subsequent semester (upon successful completion of the LS course) would instead enroll in corequisite College Algebra. In this course, students were taught prerequisite material for the College Algebra course on a just-in-time basis in an LS course while also enrolled in College Algebra. To determine if this method of instruction was effective, 163 students participated in a study from fall 2014 through spring 2015. There were 77 students in corequisite College Algebra and 86 students in traditional College Algebra. *Source:* Kim, Minsu; Hebda, Beata; Graveman, Jerry; Hebda, Piotr. “Investigating the Corequisite Model for Remedial Mathematics Courses.” *MathAMATYC Educator*, 9(3), Summer 2018.

- (a) To measure knowledge gained in each course (that is, student improvement), Hake’s gain ratio was utilized.

Gain =

$$(\text{Final Exam Score} - \text{Pretest Score}) / (100 - \text{Pretest Score}) \cdot 100$$

The table in the next column shows the mean and standard deviation gain score for both the corequisite College Algebra and traditional College Algebra. Does the evidence suggest a difference in gain between the two courses at the $\alpha = 0.05$ level of significance? If so, estimate the difference in gain between the two courses with 95% confidence.

	Corequisite	Traditional
<i>n</i>	77	86
Mean	56.3	45.9
Standard Deviation	20.3	38.3

- (b) In addition to improvement between the two courses, the researchers considered grade distribution. The table below is the grade distribution between the corequisite College Algebra and traditional College Algebra. Is grade earned in the course independent of course type? If there is an association between grade and course type, describe the association. Use a 0.05 level of significance.

	Corequisite	Traditional
A	7	19
B	29	18
C	27	22
D	6	13
F	5	7
W	3	7

- (c) Assuming that passing for this class is a grade of D or higher, what proportion of the students passed the corequisite College Algebra? What proportion of the students passed the traditional College Algebra?
- (d) Is there evidence to suggest the pass rates in corequisite College Algebra are different from those in traditional College Algebra at the $\alpha = 0.05$ level of significance? If so, estimate the difference with 95% confidence.

Retain Your Knowledge

 24. Homeruns Go to www.pearsonhighered.com/sullivanstats to obtain the data file 12_2_24 using the file format of your choice for the version of the text you are using. The variable “TrueDist” represents the distance, in feet, that the homerun traveled for all homeruns hit in the 2014 season.

- (a) Draw a relative frequency histogram of the distance a homerun traveled in 2014 using a lower class limit of the first class of 300 and a class width of 10. Describe the shape of the distribution.
- (b) Find the population mean and population standard deviation distance.
- (c) Find the quartiles of distance.
- (d) Draw a boxplot of distance. Are there any outliers?
- (e) Use a normal model to determine the proportion of homeruns that exceeded 450 feet. Compare this to the actual proportion of homeruns that exceeded 450 feet.
- (f) Use a normal model to determine the first and third quartiles. Compare this result to the quartiles found in part (c).

Explaining the Concepts

25. Explain the differences between the chi-square test for independence and the chi-square test for homogeneity. What are the similarities?
26. Why does the test for homogeneity follow the same procedures as the test for independence?

12.3 Testing the Significance of the Least-Squares Regression Model



Preparing for This Section Before getting started, review the following:

- Scatter diagrams; correlation (Section 4.1, pp. 171–172)
- Least-squares regression (Section 4.2, pp. 188–195)
- The Coefficient of Determination (Section 4.3, pp. 201–204)
- Sampling distribution of the sample mean \bar{x} (Section 8.1, pp. 371–379)
- Testing a hypothesis about μ (Section 10.3, pp. 458–463)
- Confidence intervals about a mean (Section 9.2, pp. 415–418)

Objectives

- ① State the requirements of the least-squares regression model
- ② Compute the standard error of the estimate
- ③ Verify that the residuals are normally distributed
- ④ Conduct inference on the slope of the least-squares regression model
- ⑤ Construct a confidence interval about the slope of the least-squares regression model

As a quick review of the topics discussed in Chapter 4, we present the following example:

EXAMPLE 1

Least-Squares Regression

Problem A family doctor is interested in examining the relationship between a patient's age and total cholesterol (in mg/dL). He randomly selects 14 of his female patients and obtains the data presented in Table 14. The data are based on results obtained from the National Center for Health Statistics. Draw a scatter diagram, compute the correlation coefficient, find the least-squares regression equation, and determine the coefficient of determination.

Table 14

Age, x	Total Cholesterol, y	Age, x	Total Cholesterol, y
25	180	42	183
25	195	48	204
28	186	51	221
32	180	51	243
32	210	58	208
32	197	62	228
38	239	65	269

Approach We will use a TI-84 Plus CE graphing calculator.

Solution Figure 21 on the next page displays the scatter diagram. Figure 22 displays the output. The linear correlation coefficient is 0.718. The least-squares regression equation for these data is $\hat{y} = 1.3991x + 151.3537$, where \hat{y} represents the predicted total cholesterol for a female whose age is x .

The coefficient of determination, R^2 , is 0.515, so 51.5% of the variation in total cholesterol is explained by the regression line. Figure 23 shows a graph of the least-squares regression equation on the scatter diagram.

Figure 21

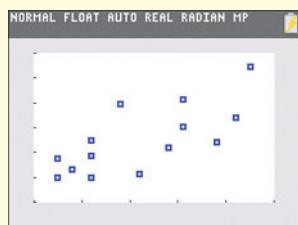


Figure 22

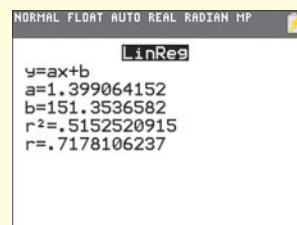
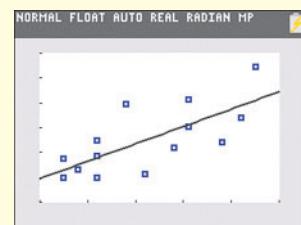


Figure 23



The information obtained in Example 1 is descriptive in nature. Notice that the descriptions are both graphical (as in the scatter diagram) and numerical (as in the correlation coefficient, the least-squares regression equation, and the coefficient of determination).

① State the Requirements of the Least-Squares Regression Model

In the least-squares regression equation $\hat{y} = b_1x + b_0$, the values for the slope, b_1 , and intercept, b_0 , are statistics, just as the sample mean, \bar{x} , and sample standard deviation, s , are statistics. The statistics b_0 and b_1 are estimates for the population intercept, β_0 , and the population slope, β_1 . The true linear relation between the explanatory variable, x , and the response variable, y , is given by $y = \beta_1x + \beta_0$.

Because b_0 and b_1 are statistics, their values vary from sample to sample, so a sampling distribution is associated with each. We use this sampling distribution to perform inference on b_0 and b_1 . For example, we might want to test whether β_1 is different from 0. If we have sufficient evidence to this effect, we conclude that there is a linear relation between the explanatory variable, x , and response variable, y .

To find the sampling distributions of b_0 and b_1 , we have some requirements of the population from which the bivariate data (x_i, y_i) were sampled. Just as in Section 8.1 when we discussed the sampling distribution of \bar{x} , we start by asking what would happen if we took many samples for a given value of the explanatory variable, x . For example, in Table 1 notice that our sample included three women aged 32 years with different corresponding values of y : 180 mg/dL, 210 mg/dL, and 197 mg/dL. This suggests that y varies for a given value of x , so there is a distribution of total cholesterol levels for $x = 32$ years. If we looked at *all* women aged 32 we could find the population mean total cholesterol for *all* 32-year-old women, denoted $\mu_{y|32}$. The notation $\mu_{y|32}$ is read “the mean value of the response variable y given that the value of the explanatory variable is 32.” We could repeat this process for any other age. In general, different ages have a different population mean total cholesterol. This brings us to our first requirement regarding inference on the least-squares regression model.

Requirement 1 for Inference on the Least-Squares Regression Model

For any particular value of the explanatory variable x (such as 32 in Example 1), the mean of the corresponding responses in the population depends linearly on x . That is,

$$\mu_{y|x} = \beta_1x + \beta_0$$

for some numbers β_0 and β_1 , where $\mu_{y|x}$ represents the population mean response when the value of the explanatory variable is x .

We also have a requirement regarding the distribution of the response variable for any particular value of the explanatory variable.

IN OTHER WORDS

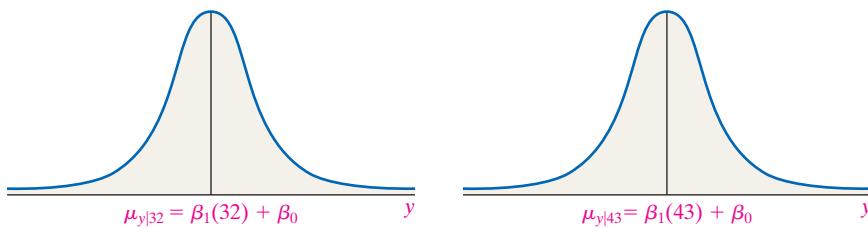
When doing inference on the least-squares regression model, we require (1) for any explanatory variable, x , the mean of the response variable, y , depends on the value of x through a linear equation, and (2) the response variable, y , is normally distributed with a constant standard deviation, σ . The mean increases/decreases at a constant rate depending on the slope, while the standard deviation remains constant.

Requirement 2 for Inference on the Least-Squares Regression Model

The response variable is normally distributed with mean $\mu_{y|x} = \beta_1 x + \beta_0$ and standard deviation σ .

The second requirement states that the mean of the response variable changes linearly, but the standard deviation remains constant, and the distribution of the response variable is normal. For example, a sample of the total cholesterol of many 32-year-old females would be normal with mean $\mu_{y|32} = \beta_1(32) + \beta_0$ and standard deviation σ . For 43-year-old females, the distribution would be normal with mean $\mu_{y|43} = \beta_1(43) + \beta_0$ and standard deviation σ . See Figure 24.

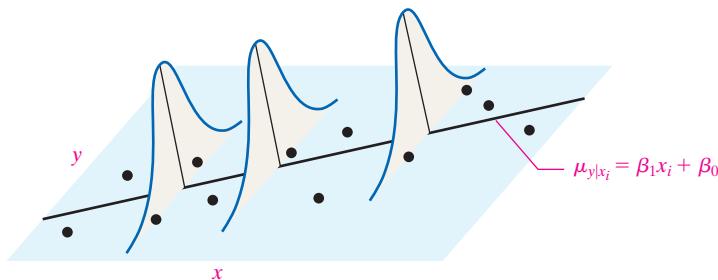
Figure 24

**IN OTHER WORDS**

The larger σ is, the more spread out the data are around the regression line.

A large value of σ , the population standard deviation, indicates that the data are widely dispersed about the regression line, and a small value of σ indicates that the data lie fairly close to the regression line. Figure 25 illustrates these ideas. The regression line represents the mean value of each normal distribution at a specified value of x . The standard deviation of each distribution is σ .

Figure 25



Of course, not all the observed values of the response variable lie on the true regression line $\mu_{y|x} = \beta_1 x + \beta_0$. The difference between the observed and predicted value of the response variable is an error term or residual, ε_i . We now present the least-squares regression model.

Definition

The **least-squares regression model** is given by

$$y_i = \beta_1 x_i + \beta_0 + \varepsilon_i \quad (1)$$

where

y_i is the value of the response variable for the i th individual

x_i is the value of the explanatory variable for the i th individual

β_0 and β_1 are the parameters to be estimated based on sample data

ε_i is a random error term with mean 0 and standard deviation $\sigma_{\varepsilon_i} = \sigma$; the error terms are independent

$i = 1, \dots, n$, where n is the sample size (number of ordered pairs in the data set)

NW Now Work Problem 13(a)

Because the expected value, or mean, of y_i is $\beta_1x_i + \beta_0$ and the expression on the left side of Equation (1) equals the expression on the right side, the expected value, or mean, of the error term, ε_i , is 0.

② Compute the Standard Error of the Estimate

In Section 4.2, we learned how to estimate β_0 and β_1 . We now present the method for obtaining the estimate of σ , the standard deviation of the response variable y for any given value of x . The unbiased estimator of σ is called the *standard error of the estimate*.

Recall the formula for the sample standard deviation from Section 3.2:

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

We compute the deviations about the mean, square them, add the squared deviations, divide by $n - 1$, and take the square root of the result. We divide by $n - 1$ because we lose 1 degree of freedom since one parameter, \bar{x} , is estimated. The same logic is used to compute the standard error of the estimate.

As we mentioned, the predicted values of y , denoted \hat{y}_i , represent the mean value of the response variable for any given value of the explanatory variable, x_i . So $y_i - \hat{y}_i$ = residual represents the difference between the observed value, y_i , and the mean value, \hat{y}_i . This calculation is used to compute the standard error of the estimate.

Definition

The **standard error of the estimate**, s_e , is found using the formula

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{\sum \text{residuals}^2}{n - 2}} \quad (2)$$

We divide by $n - 2$ because in a least-squares regression we have estimated two parameters, β_0 and β_1 . That is, we lose 2 degrees of freedom.

EXAMPLE 2 Computing the Standard Error by Hand

Problem Compute the standard error of the estimate for the data in Table 14 on page 554.

Approach Use the following steps to compute the standard error of the estimate.

Step 1 Find the least-squares regression line.

Step 2 Obtain predicted values for each observation in the data set.

Step 3 Compute the residuals for each observation in the data set.

Step 4 Compute $\sum \text{residuals}^2$.

Step 5 Compute the standard error of the estimate using Formula (2).

Solution

Step 1 In Example 1, the least-squares regression was $\hat{y} = 1.3991x + 151.3537$.

Step 2 Column 3 of Table 15 on the next page shows the predicted values for the $n = 14$ observations.

Step 3 Column 4 of Table 15 shows the residuals for the 14 observations.

(continued)

Table 15

Age, x	Total Cholesterol, y	$\hat{y} = 1.3991x + 151.3537$	Residuals, $y - \hat{y}$	Residuals 2 , $(y - \hat{y})^2$
25	180	186.33	-6.33	40.0689
25	195	186.33	8.67	75.1689
28	186	190.53	-4.53	20.5209
32	180	196.12	-16.12	259.8544
32	210	196.12	13.88	192.6544
32	197	196.12	0.88	0.7744
38	239	204.52	34.48	1188.8704
42	183	210.12	-27.12	735.4944
48	204	218.51	-14.51	210.5401
51	221	222.71	-1.71	2.9241
51	243	222.71	20.29	411.6841
58	208	232.50	-24.50	600.2500
62	228	238.10	-10.10	102.0100
65	269	242.30	26.70	712.8900
Σ residuals $^2 = 4553.705$				

Step 4 Sum the squared residuals in column 5 to find the sum of squared errors:

$$\Sigma \text{ residuals}^2 = 4553.705$$

Step 5 Use Formula (2) to compute the standard error of the estimate.

$$s_e = \sqrt{\frac{\Sigma \text{ residuals}^2}{n - 2}} = \sqrt{\frac{4553.705}{14 - 2}} = 19.48$$

CAUTION!

Be sure to divide by $n - 2$ when computing the standard error of the estimate.

EXAMPLE 3**Obtaining the Standard Error of the Estimate Using Technology**

Figure 26

Regression Statistics	
Multiple R	0.7178106
R square	0.5152521
Adjusted R square	0.4748564
Standard error	19.480535
Observations	14

Problem Obtain the standard error of the estimate for the data in Table 14 using statistical software.

Approach Use Excel to obtain the standard error. The steps for obtaining the standard error of the estimate using TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on page 564.

Solution Figure 26 shows the partial output from Excel. The results agree with the by-hand computation.

NW Now Work Problem 13(b)

③ Verify That the Residuals Are Normally Distributed

CAUTION!

The residuals must be normally distributed to perform inference on the least-squares regression line.

The least-squares regression model $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$ requires the response variable, y_i , to be normally distributed. Because $\beta_1 x_i + \beta_0$ is constant for any x_i , if y_i is normal, then the residuals, ε_i , must be normal. To perform statistical inference on the regression line, we verify that the residuals are normally distributed by examining a normal probability plot.

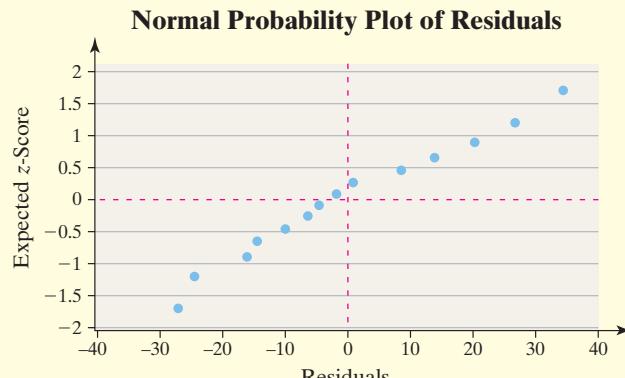
EXAMPLE 4**Verifying That the Residuals Are Normally Distributed**

Problem Verify that the residuals obtained in Table 15 from Example 2 are normally distributed.

Approach Construct a normal probability plot to assess normality. If the correlation between the residuals and expected z -scores is greater than the critical value in Table VI, the residuals are said to be normal.

Solution Figure 27 contains the normal probability plot. The correlation between the residuals and expected z -scores is $0.988 > 0.935$ (Table VI), it is reasonable to conclude the residuals are normally distributed. Therefore, we can perform inference on the least-squares regression equation.

Figure 27



NW Now Work Problem 13(c)



④ Conduct Inference on the Slope of the Least-Squares Regression Model

At this point, we know how to estimate the intercept and slope of the least-squares regression model. We can also compute the standard error of the estimate, s_e , which is an estimate of σ , the standard deviation of the response variable about the true least-squares regression model, and we know how to assess the normality of the residuals. We will now use this information to test whether a linear relation exists between the explanatory and the response variables.

We want to answer the following question: Do the sample data provide sufficient evidence to conclude that a linear relation exists between the two variables? If there is no linear relation between the response and explanatory variables, the slope of the true regression line will be zero. Do you know why? A slope of zero means that information about the explanatory variable, x , does not change our estimate of the value of the response variable, y .

Using the notation of hypothesis testing, we can perform one of three tests:

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$	$H_1: \beta_1 < 0$	$H_1: \beta_1 > 0$

The null hypothesis is $\beta_1 = 0$, the statement of “no effect.” We want to find evidence of a relation in the alternative hypothesis. The two-tailed test determines whether a linear relation exists between two variables without regard to the sign of the slope. The left-tailed test determines whether the slope of the true regression line is negative. The right-tailed test determines whether the slope of the true regression line is positive.

To test any one of these hypotheses, we need to know the sampling distribution of b_1 . It turns out that when certain conditions are met,

$$t = \frac{b_1 - \beta_1}{\frac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}}} = \frac{b_1 - \beta_1}{s_{b_1}}$$

NOTE

Because $s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$,

the standard deviation of b_1 can be calculated using

$$s_{b_1} = \frac{s_e}{\sqrt{n - 1} s_x}$$

follows Student's t -distribution with $n - 2$ degrees of freedom, where n is the number of observations, b_1 is the estimate of the slope of the regression line β_1 , and s_{b_1} is the sample standard deviation of b_1 .

Note that

$$s_{b_1} = \frac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}}$$

Hypothesis Test Regarding the Slope Coefficient, β_1

To test whether two quantitative variables are linearly related, use the following steps provided that

- The sample is obtained using random sampling or from a randomized experiment.
- The residuals are normally distributed with constant error variance.

Step 1 Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways:

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$	$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$	$H_1: \beta_1 < 0$	$H_1: \beta_1 > 0$

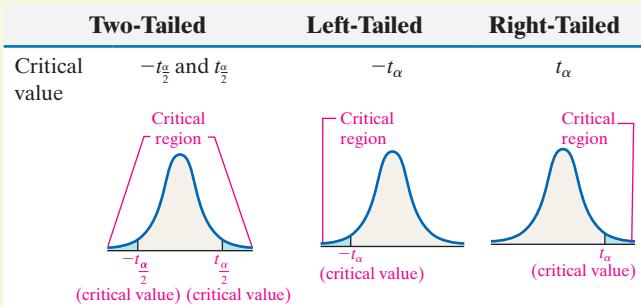
Step 2 Select a level of significance, α , depending on the seriousness of making a Type I error.

Classical Approach

Step 3 Compute the **test statistic**

$$t_0 = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1}{s_{b_1}}$$

which follows Student's t -distribution with $n - 2$ degrees of freedom. Use Table VII to determine the critical value.



Step 4 Compare the critical value to the test statistic.

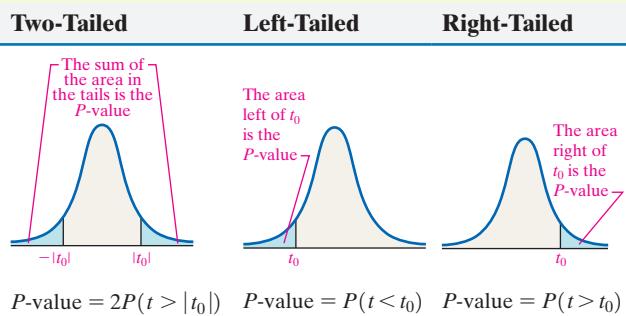
Two-Tailed	Left-Tailed	Right-Tailed
If $t_0 < -t_{\frac{\alpha}{2}}$ or $t_0 > t_{\frac{\alpha}{2}}$, reject the null hypothesis.	If $t_0 < -t_\alpha$, reject the null hypothesis.	If $t_0 > t_\alpha$, reject the null hypothesis.

P-Value Approach

By-Hand Step 3 Compute the **test statistic**

$$t_0 = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{b_1}{s_{b_1}}$$

which follows Student's t -distribution with $n - 2$ degrees of freedom. Use Table VII to approximate the P -value.



Technology Step 3 Use a statistical spreadsheet or calculator with statistical capabilities to obtain the P -value. The directions for obtaining the P -value using the TI-83/84 Plus graphing calculators, Minitab, Excel, and StatCrunch are in the Technology Step-by-Step on page 564.

Step 4 If $P\text{-value} < \alpha$, reject the null hypothesis.

Step 5 State the conclusion.

Because these procedures are **robust**, minor departures from normality will not adversely affect the results of the test. In fact, for large samples ($n \geq 30$), inferential procedures regarding b_1 can be used even with significant departures from normality.

EXAMPLE 5 Testing for a Linear Relation

Problem Test whether a linear relation exists between age and total cholesterol at the $\alpha = 0.05$ level of significance using the data in Table 14 from Example 1.

Approach Verify that the requirements to perform the inference are satisfied. Then follow Steps 1 through 5.

Solution In Example 1, we were told that the individuals were randomly selected. In Example 4, we confirmed that the residuals are normally distributed. Assume the requirement of constant error variance is satisfied.

Now follow Steps 1 through 5.

Step 1 We want to know if there is a linear relation between age and total cholesterol without regard to the sign of the slope. This is a two-tailed test and we have

$$H_0: \beta_1 = 0 \quad \text{versus} \quad H_1: \beta_1 \neq 0$$

Step 2 The level of significance is $\alpha = 0.05$.

CAUTION!

In Step 3, use unrounded values of the sample mean in the computation of $\sum(x_i - \bar{x})^2$ to avoid round-off error.

Classical Approach

Step 3 We obtained an estimate of β_1 in Example 1 to be $b_1 = 1.3991$, and we computed the standard error, $s_e = 19.48$, in Example 2. To determine the standard deviation of b_1 , compute $\sum(x_i - \bar{x})^2$, where the x_i are the values of the explanatory variable, age, and \bar{x} is the sample mean. We compute this value in Table 16.

Table 16

Age, x	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
25	42.07143	-17.07143	291.4337
25	42.07143	-17.07143	291.4337
28	42.07143	-14.07143	198.0051
32	42.07143	-10.07143	101.4337
32	42.07143	-10.07143	101.4337
32	42.07143	-10.07143	101.4337
38	42.07143	-4.07143	16.5765
42	42.07143	-0.07143	0.0051
48	42.07143	5.92857	35.1479
51	42.07143	8.92857	79.7194
51	42.07143	8.92857	79.7194
58	42.07143	15.92857	253.7193
62	42.07143	19.92857	397.1479
65	42.07143	22.92857	525.7193
$\Sigma(x_i - \bar{x})^2 = 2472.9284$			

We have

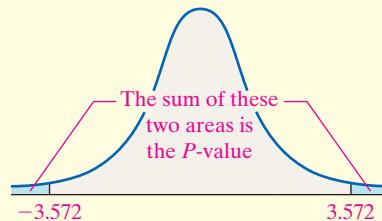
$$s_{b_1} = \frac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}} = \frac{19.48}{\sqrt{2472.9284}} = 0.3917$$

P-Value Approach

By-Hand Step 3 From Step 3 of the Classical Approach, we have that the test statistic is $t_0 = 3.572$.

Because this is a two-tailed test, the *P*-value is the sum of the area under the *t*-distribution with $14 - 2 = 12$ degrees of freedom to the left of $-t_0 = -3.572$, and to the right of $t_0 = 3.572$, as shown in Figure 29. That is, $P\text{-value} = P(t < -3.572) + P(t > 3.572) = 2P(t > 3.572)$, with 12 degrees of freedom.

Figure 29



Using Table VII, we find the row that corresponds to 12 degrees of freedom. The value 3.572 lies between 3.428 and 3.930. The area under the *t*-distribution with 12 degrees of freedom to the right of 3.428 is 0.0025. The area under the *t*-distribution with 12 degrees of freedom to the right of 3.930 is 0.001.

Because 3.572 is between 3.428 and 3.930, the *P*-value is between 2(0.001) and 2(0.0025). So,

$$0.002 < P\text{-value} < 0.005$$

Using Technology Step 3 Using Minitab, we find the *P*-value is 0.004. See Figure 30 on the next page.

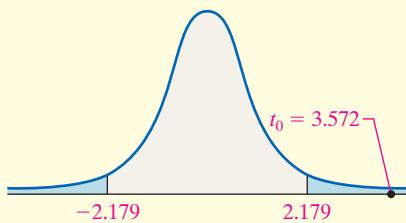
(continued)

The test statistic is

$$t_0 = \frac{b_1}{s_{b_1}} = \frac{1.3991}{0.3917} = 3.572$$

Because this is a two-tailed test, we determine the critical t -values at the $\alpha = 0.05$ level of significance with $n - 2 = 14 - 2 = 12$ degrees of freedom to be $-t_{0.05/2} = -t_{0.025} = -2.179$ and $t_{0.05/2} = t_{0.025} = 2.179$. The critical regions are displayed in Figure 28.

Figure 28



Step 4 The test statistic is $t_0 = 3.572$. We label this point in Figure 28. Because the test statistic is greater than the critical value $t_{0.025} = 2.179$, we reject the null hypothesis.

CAUTION!

If we do not reject H_0 , then we use the sample mean of y to predict the value of the response for any value of the explanatory variable.

Step 4 The P -value of 0.004 suggests that about 4 samples in 1000 would yield a slope estimate that is as extreme as or more extreme than the one obtained if the null hypothesis of no linear relation were true. Because the P -value is less than the level of significance, $\alpha = 0.05$, we reject the null hypothesis.

Step 5 There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that a linear relation exists between age and total cholesterol.



NW Now Work Problems 13(d) and 13(e)

5 Construct a Confidence Interval about the Slope of the Least-Squares Regression Model

We can also obtain confidence intervals for the slope of the least-squares regression line. The procedure is identical to that for obtaining confidence intervals for a mean. As was the case with confidence intervals for a population mean, the confidence interval for the slope of the least-squares regression line is of the form

Point estimate \pm margin of error

Confidence Intervals for the Slope of the Regression Line

A $(1 - \alpha) \cdot 100\%$ confidence interval for the slope of the true regression line, β_1 , is given by the following formulas:

$$\begin{aligned} \text{Lower bound: } b_1 &- t_{\alpha/2} \cdot \frac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}} \\ \text{Upper bound: } b_1 &+ t_{\alpha/2} \cdot \frac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}} \end{aligned} \tag{3}$$

Here, $t_{\alpha/2}$ is computed with $n - 2$ degrees of freedom.

Note: This interval can be computed only if the data are randomly obtained, the residuals are normally distributed, and there is constant error variance.

Figure 30

Regression Analysis: Cholesterol versus Age

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	4840.5	4840.5	12.76	0.004
Age	1	4840.5	4840.5	12.76	0.004
Error	12	4553.9	379.5		
Lack-of-Fit	8	3746.7	468.3	2.32	0.217
Pure Error	4	807.2	201.8		
Total	13	9394.4			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
19.4805	51.53%	47.49%	33.55%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	151.4	17.3	8.76	0.000	
Age	1.399	0.392	3.57	0.004	1.00

Regression Equation

$$\text{Cholesterol} = 151.4 + 1.399 \text{ Age}$$

EXAMPLE 6**Constructing a Confidence Interval for the Slope of the True Regression Line**

Problem Determine a 95% confidence interval for the slope of the true regression line for the data presented in Table 14 in Example 1.

By-Hand Approach

Step 1 Determine the least-squares regression line.

Step 2 Verify that the requirements for inference on the regression line are satisfied.

Step 3 Compute s_e .

Step 4 Determine the critical value $t_{\alpha/2}$ with $n - 2$ degrees of freedom.

Step 5 Compute the bounds on the $(1 - \alpha) \cdot 100\%$ confidence interval for β_1 using Formula (3).

Step 6 Interpret the result by stating, "We are 95% confident that β_1 is between *lower bound* and *upper bound*."

By-Hand Solution

Step 1 The least-squares regression line was determined in Example 1 and is $\hat{y} = 1.3991x + 151.3537$.

Step 2 The requirements were verified in Examples 1–5.

Step 3 We computed s_e in Example 2, obtaining $s_e = 19.48$.

Step 4 Because we wish to determine a 95% confidence interval, we have $\alpha = 0.05$. Therefore, we need to find $t_{0.05/2} = t_{0.025}$ with $14 - 2 = 12$ degrees of freedom. Referring to Table VII, we find that $t_{0.025} = 2.179$.

Step 5 We use Formula (3) to find the lower and upper bounds.

$$\text{Lower bound: } b_1 - t_{\alpha/2} \cdot \frac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}} = 1.3991 - 2.179 \cdot \frac{19.48}{\sqrt{2472.9284}} \\ = 1.3991 - 0.8536 = 0.5455$$

$$\text{Upper bound: } b_1 + t_{\alpha/2} \cdot \frac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}} = 1.3991 + 2.179 \cdot \frac{19.48}{\sqrt{2472.9284}} \\ = 1.3991 + 0.8536 = 2.2526$$

Step 6 We are 95% confident that the mean increase in cholesterol for each additional year of life is somewhere between 0.5455 mg/dL and 2.2527 mg/dL. Because the 95% confidence interval does not include 0, we reject $H_0: \beta_1 = 0$.

Technology Approach

Step 1 Use a statistical spreadsheet or graphing calculator with advanced statistical features to obtain the confidence interval. We will use StatCrunch. The steps for constructing confidence intervals using StatCrunch, the TI-83/84 Plus graphing calculators, Minitab, and Excel are given in the Technology Step-by-Step on page 564.

Step 2 Interpret the result.

Technology Solution

Step 1 Figure 31 shows the results obtained from StatCrunch.

Figure 31

Simple linear regression results:

Dependent Variable: Cholesterol
Independent Variable: Age
 $\text{Cholesterol} = 151.35365 + 1.3990642 \text{Age}$
Sample size: 14
 R (correlation coefficient) = 0.7178
 $R\text{-sq} = 0.5152521$
Estimate of error standard deviation: 19.480536

Parameter estimates:

Parameter	Estimate	Std. Err.	DF	95% L. Limit	95% U. Limit
Intercept	151.35365	17.28376	12	113.69558	189.01173
Slope	1.3990642	0.39173746	12	0.5455416	2.2525868

The lower bound (95% L. Limit) is 0.5455 and the upper bound (95% U. Limit) is 2.2526.

Step 2 We are 95% confident that the mean increase in cholesterol for each additional year of life is between 0.5455 mg/dL and 2.2526 mg/dL. Because the confidence interval does not contain 0, we reject $H_0: \beta_1 = 0$.

NW Now Work Problem 13(f)**CAUTION!**

It is best that the values of the explanatory variable be spread out when doing regression analysis.

In looking carefully at the formula for the standard deviation of b_1 , notice that the larger the value of $\sum(x_i - \bar{x})^2$, the smaller the value of s_{b_1} . This result implies that whenever we are finding a least-squares regression line, attempt to make the values of the explanatory variable, x , as evenly spread out as possible so that b_1 , our estimate of β_1 , is as precise as possible.

Inference on the Linear Correlation Coefficient

Perhaps you are wondering why we have not presented hypothesis tests regarding the linear correlation coefficient in this section. Recall in Chapter 4 that we introduced a quick method for testing the significance of the correlation coefficient, even though we did not have a full appreciation of statistical inference. At this point, we intentionally avoid discussion of inference on the correlation coefficient for two reasons: (1) the hypothesis test on the slope and a hypothesis test on the linear correlation coefficient will yield the same conclusion, and (2) inferential methods on the linear correlation coefficient, ρ , require that the y 's at any given x be normally distributed and that the x 's at any given y be normally distributed. That is, testing a hypothesis such as $H_0: \rho = 0$ versus $H_1: \rho \neq 0$ requires that the two variables follow a **bivariate normal distribution** or be **jointly normally distributed**. Verifying this requirement is a difficult task. Although a normal probability plot of the x_i 's and a separate normal probability plot of the y_i 's generally mean that the joint distribution is normal, it is not guaranteed. For these two reasons, we will be content in verifying the linearity of the data by performing inference on the slope coefficient only.

Technology Step-by-Step

Testing the Least-Squares Regression Model

TI-83/84 Plus

Hypothesis Test on the Slope

- Enter the explanatory variable in L1 and the response variable in L2.
- Press STAT, highlight TESTS, and select F:LinRegTTest....
- Be sure that Xlist is L1 and Ylist is L2. Make sure that Freq: is set to 1. Select the direction of the alternative hypothesis. Place the cursor on Calculate and press ENTER.

Confidence Interval for the Slope

- Enter the explanatory variable in L1 and the response variable in L2.
- Press STAT, highlight TESTS, and select G: LinRegTInt
- Be sure that Xlist is L1 and Ylist is L2. Make sure the Freq: is set to 1. Select the confidence level. Highlight Calculate. Press ENTER.

Minitab

- With the explanatory variable in C1 and the response variable in C2, select the Stat menu and highlight Regression. Highlight Regression, then select Fit Regression Model....
- Place the cursor in the "Responses:" box. Highlight the column containing the response variable. Click Select. Place the cursor in the "Continuous predictors:" box. Highlight the column containing the explanatory variable. Click Select. Click OK.

Excel

- Make sure the Data Analysis Tool Pack is activated by selecting File. Click Options, then click Add-Ins. From the drop-down menu, select "Excel Add-ins." Click Go. Check the boxes "Analysis ToolPak" and "Analysis ToolPak-VBA." Click OK.
- Enter the explanatory variable in column A and the response variable in column B.
- Select the Data menu and then select Data Analysis
- Select the Regression option.
- With the cursor in the Y-range cell, highlight the range of cells that contains the response variable. With the cursor in the X-range cell, highlight the range of cells that contains the explanatory variable. Click OK.

StatCrunch

Hypothesis Test on the Slope

- Enter the explanatory variable in column var1 and the response variable in column var2.
- Select Stat, highlight Regression, choose Simple Linear. Choose var1 for the X-variable, choose var2 for the Y-variable. Select the Hypothesis tests radio button. Choose the appropriate values in the null hypothesis for both the intercept and slope. Choose the direction of the alternative hypothesis. Click Compute!.

Confidence Interval for the Slope

- Enter the explanatory variable in column var1 and the response variable in column var2.
- Select Stat, highlight Regression, choose Simple Linear. Choose var1 for the X-variable, choose var2 for the Y-variable. Select the Confidence intervals radio button. Choose the confidence level. Click Compute!.



12.3 Assess Your Understanding

Vocabulary and Skill Building

- Suppose a least-squares regression line is given by $\hat{y} = 4.302x - 3.293$. What is the mean value of the response variable if $x = 20$?
- True or False:** In a least-squares regression, the response variable is normally distributed with mean $\mu_{y|x}$ and standard deviation σ .
- In the least-squares regression model, $y_i = \beta_1 x_i + \beta_0 + \varepsilon_i$, ε_i is a random error term with mean _____ and standard deviation $\sigma_{\varepsilon_i} = _____$.
- If $H_0: \beta_1 = 0$ is not rejected, what is the best estimate for the value of the response variable for any value of the explanatory variable?

In Problems 5–10, use the results of Problems 7–12, respectively, from Section 4.2 to answer the following questions:

- What are the estimates of β_0 and β_1 ?
- Compute the standard error, the point estimate for σ .
- Determine s_{b_1} .
- Assuming the residuals are normally distributed, test $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$ at the $\alpha = 0.05$ level of significance.

5.

x	3	4	5	7	8
y	4	6	7	12	14

7.

x	-2	-1	0	1	2
y	-4	0	1	4	5

9.

x	20	30	40	50	60
y	100	95	91	83	70

6.

x	3	5	7	9	11
y	0	2	3	6	9

8.

x	-2	-1	0	1	2
y	7	6	3	2	0

10.

x	5	10	15	20	25
y	2	4	7	11	18

Applying the Concepts



11. **An Unhealthy Commute** The following data represent commute times (in minutes) and a score on a well-being survey.

Commute Time (minutes), <i>x</i>	Gallup-Healthways Well-Being Index Composite Score, <i>y</i>
5	69.2
15	68.3
25	67.5
35	67.1
50	66.4
72	66.1
105	63.9

Source: The Gallup Organization.

Use the results from Problem 17 in Section 4.2 to answer the following questions:

- Treating commute time as the explanatory variable, x , determine the estimates of β_0 and β_1 .
- Compute the standard error of the estimate, s_e .
- Determine s_{b_1} .
- A normal probability plot of the residuals indicates it is reasonable to conclude the residuals are normally distributed. Test whether a linear relation exists between commute time and well-being index composite score at the $\alpha = 0.05$ level of significance.
- Construct a 95% confidence interval about the slope of the true least-squares regression line.



12. **Credit Scores** An economist wants to determine the relation between one's FICO score, x , and the interest rate of a 36-month auto loan, y . The data represent the interest rate (in percent) a bank might offer on a 36-month auto loan for various FICO scores.

Credit Score, <i>x</i>	Interest Rate (percent), <i>y</i>
545	18.982
595	17.967
640	12.218
675	8.612
705	6.680
750	5.150

Source: www.myfico.com

Use the results from Problem 18 in Section 4.2 to answer the following questions:

- Treating credit score as the explanatory variable, x , determine the estimates of β_0 and β_1 .
- Compute the standard error of the estimate, s_e .
- Determine s_{b_1} .
- A normal probability plot of the residuals indicates it is reasonable to conclude the residuals are normally distributed. Test whether a linear relation exists between credit score and interest rate at the $\alpha = 0.05$ level of significance.
- Construct a 95% confidence interval about the slope of the true least-squares regression line.



13. **Height versus Head Circumference** A pediatrician wants to determine the relation that may exist between a child's height and head circumference. She randomly selects 11 children from her practice, measures their heights and head circumferences, and obtains the following data:

Height (inches), <i>x</i>	Head Circumference (inches), <i>y</i>	Height (inches), <i>x</i>	Head Circumference (inches), <i>y</i>
27.75	17.5	26.5	17.3
24.5	17.1	27	17.5
25.5	17.1	26.75	17.3
26	17.3	26.75	17.5
25	16.9	27.5	17.5
27.75	17.6		

Source: Denise Slucki, student at Joliet Junior College.

Use the results from Problem 19 in Section 4.2 to answer the following questions:

- Treating height as the explanatory variable, x , determine the estimates of β_0 and β_1 .
- Compute the standard error of the estimate, s_e .
- Determine whether the residuals are normally distributed.
- Determine s_{b_1} .
- If the residuals are normally distributed, test whether a linear relation exists between height and head circumference at the $\alpha = 0.01$ level of significance.
- If the residuals are normally distributed, construct a 95% confidence interval about the slope of the true least-squares regression line.
- A child comes in for a physical, and the nurse determines his height to be 26.5 inches. However, the child is being rather uncooperative, so the nurse is unable to measure the head circumference of the child. What would be a good estimate of this child's head circumference? Why is this a good estimate?

DATA **14. Hurricanes** The following data represent the maximum wind speed (in knots) and atmospheric pressure (in millibars) for a random sample of hurricanes that originated in the Atlantic Ocean. Does atmospheric pressure play a role in the wind speed of a hurricane?

Atmospheric Pressure (mb), x	Wind Speed (knots), y	Atmospheric Pressure (mb), x	Wind Speed (knots), y
993	50	1006	40
995	60	942	120
994	60	1002	40
997	45	986	50
1003	45	983	70
1004	40	994	65
1000	55	940	120
994	55	976	80
942	105	966	100
1006	30	982	55

Source: National Hurricane Center.

Use the results from Problem 22 in Section 4.2 to answer the following questions:

- Treating atmospheric pressure as the explanatory variable, x , determine the estimates of β_0 and β_1 .
- Compute the standard error of the estimate.
- Determine whether the residuals are normally distributed.
- Determine s_{b_1} .
- If the residuals are normally distributed, test whether a linear relation exists between the atmospheric pressure and wind speed at the $\alpha = 0.01$ level of significance.
- If the residuals are normally distributed, construct a 99% confidence interval for the slope of the true least-squares regression line.
- What is the mean wind speed of a hurricane whose atmospheric pressure is 995 mb?

DATA **15. Concrete** As concrete cures, it gains strength. The following data represent the 7-day and 28-day strength (in pounds per square inch) of a certain type of concrete:

7-Day Strength, x	28-Day Strength, y	7-Day Strength, x	28-Day Strength, y
2300	4070	2480	4120
3390	5220	3380	5020
2430	4640	2660	4890
2890	4620	2620	4190
3330	4850	3340	4630

- Treating the 7-day strength as the explanatory variable, x , determine the estimates of β_0 and β_1 .
- Compute the standard error of the estimate.
- Determine s_{b_1} .
- Assuming the residuals are normally distributed, test whether a linear relation exists between 7-day strength and 28-day strength at the $\alpha = 0.05$ level of significance.
- Assuming the residuals are normally distributed, construct a 95% confidence interval for the slope of the true least-squares regression line.
- What is the estimated mean 28-day strength of this concrete if the 7-day strength is 3000 psi?

DATA **16. Tar and Nicotine** Every year the Federal Trade Commission (FTC) must report tar and nicotine levels in cigarettes to Congress. Tar and nicotine levels of over 1200 brands of cigarettes are given to Congress and a random sample of those appear in the following table:

Brand	Tar (mg), x	Nicotine (mg), y
Barclay 100	5	0.4
Benson and Hedges King	16	1.1
Camel Regular	24	1.7
Chesterfield King	24	1.4
Doral	8	0.5
Kent Golden Lights	9	0.8
Kool Menthol	9	0.8
Lucky Strike	24	1.5
Marlboro Gold	15	1.2
Newport Menthol	18	1.3
Salem Menthol	17	1.3
Virginia Slims Ultra Light	5	0.5
Winston Light	10	0.8

Source: Federal Trade Commission.

- Treating the amount of tar as the explanatory variable, x , determine the estimates of β_0 and β_1 .
- Compute the standard error of the estimate.
- Determine s_{b_1} .
- Assuming the residuals are normally distributed, test whether a linear relation exists between the amount of tar, x , and the amount of nicotine, y , at the $\alpha = 0.1$ level of significance.
- Assuming the residuals are normally distributed, construct a 90% confidence interval for the slope of the true least-squares regression line.
- What is the mean amount of nicotine in a cigarette that has 12 milligrams of tar?

- DATA 17. Invest in Education** Go to www.pearsonhighered.com/sullivanstats to obtain the data file 12_3_17. The variable “Cost” represents the four-year cost including tuition, supplies, room and board, the variable “Annual ROI” represents the return on investment for graduates of the school—essentially how much you would earn on the investment of attending the school. The variable “Grad Rate” represents the graduation rate of the school.

- (a) In Problem 49 from Section 4.1, a scatter diagram between “Cost” and “Grad Rate” treating “Cost” as the explanatory variable suggested a positive association between the two variables. Treating “Cost” as the explanatory variable, x , test whether a negative association exists between the cost and annual ROI for graduates of four-year schools at the $\alpha = 0.01$ level of significance. Normal probability plots suggest the residuals are normally distributed.
- (b) Construct a 90% confidence interval for the slope of the true least-squares regression line.
- (c) What is the mean annual ROI for a four-year school whose cost is \$180,000?

- DATA 18. American Black Bears** In 1969, Dr. Michael R. Pelton of the University of Tennessee initiated a long-term study of the American black bear (*Ursus americanus*) population in Great Smoky Mountains National Park. One aspect of the study was to develop a model that could be used to predict a bear’s weight (since it is not practical to weigh bears in the field). One variable that is thought to be related to weight is the length of the bear. The following data represent the lengths and weights of 12 American black bears.

Total Length (cm), x	Weight (kg), y
139.0	110
138.0	60
139.0	90
120.5	60
149.0	85
141.0	100
141.0	95
150.0	85
166.0	155
151.5	140
129.5	105
150.0	110

Source: fieldtripearth.org

Use the results from Problem 20 in Section 4.2 to answer the following questions:

- (a) Treating total length as the explanatory variable, x , determine the estimates of β_0 and β_1 .
- (b) Assuming the residuals are normally distributed, test whether a linear relation exists between total length and weight at the $\alpha = 0.05$ level of significance.
- (c) Assuming the residuals are normally distributed, construct a 95% confidence interval for the slope of the true least-squares regression line.
- (d) What is the mean weight of American black bears of length 146.0 cm?

- DATA 19. CEO Performance** (Refer to Problem 33 in Section 4.1) The following data represent the total compensation for 12 randomly selected chief executive officers (CEOs) and the company’s stock performance in 2017.

Company	Compensation (millions of dollars)	Stock Return (%)
MACERICH CO	12.8	-2.9
REGENCY CENTERS CORP	5.6	3.6
ROCKWELL COLLINS INC	8.1	57
KOHLS CORP	11.3	71
INTERPUBLIC GROUP OF COMPANIES, INC.	16.9	-11
MICROSOFT CORP	20	38
HENRY SCHEIN INC	7.2	-7.9
INTERNATIONAL FLAVORS & FRAGRANCES INC	7.7	32
HANESBRANDS INC.	9.6	-0.3
WALGREENS BOOTS ALLIANCE, INC.	14.7	2.8
CERNER CORP	2.6	42
M&T BANK CORP	4.2	11

Source: *The Wall Street Journal*.

- (a) Treating compensation as the explanatory variable, x , determine the estimates of β_0 and β_1 .
- (b) Assuming the residuals are normally distributed, test whether a linear relation exists between compensation and stock return at the $\alpha = 0.05$ level of significance.
- (c) Assuming the residuals are normally distributed, construct a 95% confidence interval for the slope of the true least-squares regression line.
- (d) Based on your results to parts (b) and (c), would you recommend using the least-squares regression line to predict the stock return of a company based on the CEO’s compensation? Why?

- DATA 20. Bear Markets** (Refer to Problem 34, Section 4.1) A bear market is a market condition in which the price of the security falls. A bear market in the stock market is defined as a condition in which the market declines by 20% or more over the course of at least two months. The following data represent the number of months and percentage change in the S&P500 (a group of 500 stocks).

Months	Percent Change	Months	Percent Change
1.9	-44.57	12.1	-20.57
8.3	-44.29	14.9	-21.63
3.3	-32.86	6.5	-27.97
3.4	-42.54	8.0	-22.18
6.8	-61.81	18.1	-36.06
5.8	-40.60	21.0	-48.20
3.1	-29.43	20.7	-27.11
13.4	-31.81	3.4	-33.51
12.9	-54.47	18.2	-36.77
5.1	-24.44	9.3	-33.75
7.6	-28.69	13.6	-51.93
17.9	-34.42	2.1	-27.62
11.8	-28.47		

Source: Gold-Eagle.

- (a) Treating months as the explanatory variable, x , determine the estimates for β_0 and β_1 .
- (b) Assuming the residuals are normally distributed, test whether a linear relation exists between the number of months of a bear market and percent change at the $\alpha = 0.05$ level of significance.
- (c) Assuming the residuals are normally distributed, construct a 95% confidence interval for the slope of the true least-squares regression line.
- (d) Based on your results to parts (b) and (c), would you recommend using the least-squares regression line to predict the percent change in the S&P500 during a bear market? Why?

DATA **21. Age versus HDL Cholesterol** A doctor wanted to determine whether there is a relation between a male's age and his HDL (so-called good) cholesterol. He randomly selected 17 of his patients and determined their HDL levels. He obtained the following data.

Age, x	HDL Cholesterol, y	Age, x	HDL Cholesterol, y
38	57	38	44
42	54	66	62
46	34	30	53
32	56	51	36
55	35	27	45
52	40	52	38
61	42	49	55
61	38	39	28
26	47		

Source: Data based on information obtained from the National Center for Health Statistics.

- (a) Draw a scatter diagram of the data, treating age as the explanatory variable. What type of relation, if any, appears to exist between age and HDL cholesterol?
- (b) Determine the least-squares regression equation from the sample data.
- (c) Assuming the residuals are normally distributed, test whether a linear relation exists between age and HDL cholesterol levels at the $\alpha = 0.01$ level of significance.
- (d) Assuming the residuals are normally distributed, construct a 95% confidence interval for the slope of the true least-squares regression line.
- (e) For a 42-year-old male patient who visits the doctor's office, would you recommend using the least-squares regression line obtained in part (b) to predict the HDL cholesterol of this patient? Why? What would be a good estimate for the HDL cholesterol of this patient?

22. The output shown was obtained from Minitab.

The regression equation is $y = 12.4 + 1.40 x$	Predictor	Coef	StDev	T	P
	Constant	12.396	1.381	8.97	0.000
	x	1.3962	0.1245	11.21	0.000
S = 2.167	R-Sq = 91.3%	R-Sq(adj) = 90.6%			

- (a) The least-squares regression equation is $\hat{y} = 1.3962x + 12.396$. What is the predicted value of y at $x = 10$?
- (b) What is the mean of y at $x = 10$?
- (c) The standard error, s_e , is 2.167. What is an estimate of the standard deviation of y at $x = 10$?
- (d) If the requirements for inference on the least-squares regression model are satisfied, what is the distribution of y at $x = 10$?

DATA **23. Influential Observations** Zillow.com is a site that can be used to assess the value of homes in your neighborhood. The organization provides a list of homes for sale as well as a Zestimate, which is the price Zillow believes the home will sell for. The following data represent the Zestimate and sale price (in thousands of dollars) of a random sample of recently sold homes in Charleston, South Carolina.

Zestimate	Sale Price
362	370
309	315
365.5	371.9
215	218
184	186.5
252.5	260
247.5	250.8
244	251

Source: zillow.com

- (a) Draw a scatter diagram of the data, treating the Zestimate as the explanatory variable and sale price as the response variable.
- (b) Determine the least-squares regression line. Test whether there is a relation between the Zestimate and sale price at the $\alpha = 0.05$ level of significance.
- (c) An observation is said to be influential if the inclusion of the observation significantly changes the value of the slope and/or intercept of the least-squares regression line. A home with a Zestimate of \$370,000 recently sold for \$150,000. Determine the least-squares regression line with this home included. Test whether there is a relation between the Zestimate and sale price at the $\alpha = 0.05$ level of significance. Do you think this observation is influential?

Explaining the Concepts

- 24.** Why is it important to perform graphical as well as analytical analyses when analyzing relations between two quantitative variables?
- 25.** What do the y -coordinates on the least-squares regression line represent?
- 26.** Why is it desirable to have the explanatory variables spread out to test a hypothesis regarding β_1 or to construct confidence intervals about β_1 ?
- 27.** Why don't we conduct inference on the linear correlation coefficient?

12.4 Confidence and Prediction Intervals



Preparing for This Section Before getting started, review the following:

- Confidence intervals about μ (Section 9.2, pp. 415–418)

- Objectives**
- ① Construct confidence intervals for a mean response
 - ② Construct prediction intervals for an individual response

We know how to obtain the least-squares regression equation of best fit from data. We also know how to use the least-squares regression equation to obtain a predicted value. For example, the least-squares regression equation for the cholesterol data introduced in Example 1 from Section 12.3 is

$$\hat{y} = 1.3991x + 151.3537$$

where \hat{y} represents the predicted total cholesterol for a female whose age is x . The predicted value of total cholesterol for a given age x actually has two interpretations:

1. It represents the mean total cholesterol for all females whose age is x .
2. It represents the predicted total cholesterol for a randomly selected female whose age is x .

So if we let $x = 42$ in the least-squares regression equation $\hat{y} = 1.3991x + 151.3537$, we obtain $\hat{y} = 1.3991(42) + 151.3537 = 210.1$ mg/dL. We can interpret this result in one of two ways:

1. The mean total cholesterol for all 42-year-old females is 210.1 mg/dL.
2. The predicted total cholesterol for a randomly selected 42-year-old female is 210.1 mg/dL.

Of course, there is a margin of error in making predictions, so we construct intervals about any predicted value to describe its accuracy. The type of interval constructed will depend on whether we are predicting a mean total cholesterol for all 42-year-old females or the total cholesterol for an individual 42-year-old female. In other words, the margin of error is going to be different for predicting the mean total cholesterol for all females who are 42 years old versus the total cholesterol for one individual. Which prediction (the mean or the individual) do you think will have a wider confidence interval? It seems logical that the distribution of means should have less variability (and therefore a lower margin of error) than the distribution of individuals. After all, in the distribution of means, high total cholesterols can be offset by low total cholesterols.

Definitions

Confidence intervals for a mean response are intervals constructed about the predicted value of y , at a given level of x , that are used to measure the accuracy of the mean response of all the individuals in the population where the value of the explanatory variable is x .

Prediction intervals for an individual response are intervals constructed about the predicted value of y that are used to measure the accuracy of a single individual's predicted value.

IN OTHER WORDS

Confidence intervals are intervals for the mean of the population. Prediction intervals are intervals for an individual from the population.

If we use the least-squares regression equation to predict the mean total cholesterol for all 42-year-old females, we construct a confidence interval for a mean response. If we use the least-squares regression equation to predict the total cholesterol for a particular 42-year-old female, we construct a prediction interval for an individual response.

1 Construct Confidence Intervals for a Mean Response

The structure of a confidence interval is the same as it was in Section 9.1. The interval is of the form

$$\text{Point estimate} \pm \text{margin of error}$$

Confidence Interval for the Mean Response of y , \hat{y}

A $(1 - \alpha) \cdot 100\%$ confidence interval for \hat{y} , the mean response of y for a specified value of x , is given by

$$\text{Lower bound: } \hat{y} - t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \quad (1)$$

$$\text{Upper bound: } \hat{y} + t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

where x^* is the given value of the explanatory variable, n is the number of observations, and $t_{\alpha/2}$ is the critical value with $n - 2$ degrees of freedom.

NOTE

The interval may be constructed provided the residuals are normally distributed or the sample size is large.

EXAMPLE 1

Constructing a Confidence Interval for a Mean Response by Hand

Problem Construct a 95% confidence interval for the predicted mean total cholesterol of all 42-year-old females using the data in Table 14 on page 554.

Approach Determine a confidence interval for the predicted mean total cholesterol at $x^* = 42$ using Formula (1), since our estimate is for the mean cholesterol of all 42-year-old females.

Solution The least-squares regression equation is $\hat{y} = 1.3991x + 151.3537$. To find the predicted mean total cholesterol of all 42-year-olds, let $x^* = 42$ in the regression equation and obtain $\hat{y} = 1.3991(42) + 151.3537 = 210.1$ mg/dL. From Example 2 in Section 12.3, we found that $s_e = 19.48$, and from Example 5 in Section 12.3, we found that $\sum(x_i - \bar{x})^2 = 2472.9284$ and $\bar{x} = 42.07143$. The critical t -value, $t_{\alpha/2} = t_{0.025}$, with $n - 2 = 14 - 2 = 12$ degrees of freedom is 2.179. The 95% confidence interval for the predicted mean total cholesterol for all 42-year-old females is therefore

$$\begin{aligned} \text{Lower bound: } \hat{y} - t_{\alpha/2} \cdot s_e \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} &= 210.1 - 2.179 \cdot 19.48 \cdot \sqrt{\frac{1}{14} + \frac{(42 - 42.07143)^2}{2472.9284}} \\ &= 198.8 \end{aligned}$$

$$\begin{aligned} \text{Upper bound: } \hat{y} + t_{\alpha/2} \cdot s_e \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} &= 210.1 + 2.179 \cdot 19.48 \cdot \sqrt{\frac{1}{14} + \frac{(42 - 42.07143)^2}{2472.9284}} \\ &= 221.4 \end{aligned}$$

NW Now Work Problems 3(a) and (b)

We are 95% confident that the mean total cholesterol of all 42-year-old females is between 198.8 mg/dL and 221.4 mg/dL.

2 Construct Prediction Intervals for an Individual Response

The procedure for obtaining a prediction interval for an individual response is similar to that for finding a confidence interval for a mean response. The only difference is the standard error. More variability is associated with individuals than with means. Therefore, the computation of the interval must account for this increased variability. Again, the form of the interval is

$$\text{Point estimate} \pm \text{margin of error}$$

IN OTHER WORDS

Prediction intervals are wider than confidence intervals because it is tougher to guess the value for an individual than the mean of a population.

Prediction Interval for an Individual Response about \hat{y}

A $(1 - \alpha) \cdot 100\%$ prediction interval for \hat{y} , the individual response of y , is given by

$$\text{Lower bound: } \hat{y} - t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} \quad (2)$$

$$\text{Upper bound: } \hat{y} + t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

where x^* is the given value of the explanatory variable, n is the number of observations, and $t_{\alpha/2}$ is the critical value with $n - 2$ degrees of freedom.

NOTE

The interval may be constructed provided the residuals are normally distributed or the sample size is large.

Notice that the only difference between Formulas (1) and (2) is the “ $1 +$ ” under the radical in Formula (2).

EXAMPLE 2 Constructing a Prediction Interval for an Individual Response by Hand

Problem Construct a 95% prediction interval for the predicted total cholesterol of a 42-year-old female.

Approach Determine the predicted total cholesterol at $x^* = 42$ and use Formula (2), since our estimate is for a particular 42-year-old female.

Solution The least-squares regression equation is $\hat{y} = 1.3991x + 151.3537$. To find the predicted total cholesterol of a 42-year-old, let $x^* = 42$ in the regression equation and obtain $\hat{y} = 1.3991(42) + 151.3537 = 210.1$ mg/dL. From Example 2 in Section 12.3, we found that $s_e = 19.48$; from Example 5 in Section 12.3, we found that $\sum(x_i - \bar{x})^2 = 2472.9284$ and $\bar{x} = 42.07143$. We find $t_{\alpha/2} = t_{0.025}$ with $n - 2 = 14 - 2 = 12$ degrees of freedom to be 2.179.

The 95% prediction interval for the predicted total cholesterol for a 42-year-old female is

$$\text{Lower bound: } \hat{y} - t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = 210.1 - 2.179 \cdot 19.48 \cdot \sqrt{1 + \frac{1}{14} + \frac{(42 - 42.07143)^2}{2472.9284}} = 166.2$$

$$\text{Upper bound: } \hat{y} + t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = 210.1 + 2.179 \cdot 19.48 \cdot \sqrt{1 + \frac{1}{14} + \frac{(42 - 42.07143)^2}{2472.9284}} = 254.0$$

NW Now Work Problems

3(c) and (d)

We are 95% confident that the total cholesterol of a randomly selected 42-year-old female is between 166.2 mg/dL and 254.0 mg/dL.

Notice that the interval about the individual (prediction interval for an individual response) is wider than the interval about the mean (confidence interval for a mean response). The reason for this should be clear: More variability is associated with individuals than with groups of individuals. That is, it is more difficult to predict a particular 42-year-old female's total cholesterol than it is to predict the mean total cholesterol for all 42-year-old females.

EXAMPLE 3 Confidence and Prediction Intervals Using Technology

Using Technology

The bounds for confidence and prediction intervals obtained using statistical software may differ from bounds computed by hand due to rounding error.

Problem Construct a 95% confidence interval for the predicted mean total cholesterol of all 42-year-old females using statistical software. Construct a 95% prediction interval for the predicted total cholesterol for a 42-year-old female using statistical software.

Approach We will use Minitab to obtain the intervals. The steps for obtaining confidence and prediction intervals using Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on page 572.

(continued)

Solution Figure 32 shows the results obtained from Minitab.

Figure 32

Prediction

Fit	SE Fit	95% CI	95% PI
210.114	5.20647	(198.770, 221.458)	(166.180, 254.049)

Technology Step-by-Step

Confidence and Prediction Intervals

TI-83/84 Plus

The TI-83/84 Plus graphing calculators do not compute confidence or prediction intervals.

Minitab

- With the model determined as in Section 12.3, select the **Stat** menu, highlight **Regression**, highlight **Regression** again, and select **Predict ...**.
- Select the response variable. From the pull-down menu select “Enter individual values.” Enter the value of the explanatory variable for which you want to make a prediction in the table.
- Click the Options ... button. Enter the desired level of confidence. Click OK twice.

Excel

- Enter the values of the explanatory variable in Column A and the corresponding values of the response variable in Column B. Enter the value of the explanatory variable for which you want to make a prediction in a cell.

- Select the XLSTAT menu. Choose the Modeling Data menu and select Linear Regression.
- With the cursor in the Y/Dependent variables quantitative cell, highlight the data in Column B. With the cursor in the X/Explanatory variables quantitative cell, highlight the data in Column A.
- Select the Prediction menu. Check the Prediction box. With the cursor in the X/Explanatory variable quantitative cell, highlight the cell that contains the value of the explanatory variable. Click OK.

StatCrunch

Follow the steps given in Section 12.3 for testing the significance of the least-squares regression model. Enter the value of the explanatory variable for which you want to make a prediction in the box following “X value(s)” under “Prediction of Y:” Enter a level of confidence. Click Compute!.



12.4 Assess Your Understanding

Vocabulary and Skill Building

- Intervals constructed about the predicted value of y , at a given level of x , that are used to measure the accuracy of the mean response of all individuals in the population are called _____ intervals for a(n) _____ response.
- Intervals constructed about the predicted value of y , at a given level of x , that are used to measure the accuracy of a single individual's prediction are called _____ intervals for a(n) _____ response.

In Problems 3–6, use the results of Problems 5–8 in Section 12.3.

- NW** 3. Using the sample data from Problem 5 in Section 12.3,
- Predict the mean value of y if $x = 7$.
 - Construct a 95% confidence interval for the mean value of y if $x = 7$.
 - Predict the value of y if $x = 7$.
 - Construct a 95% prediction interval for the value of y if $x = 7$.
 - Explain the difference between the prediction in parts (a) and (c).

- Using the sample data from Problem 6 in Section 12.3,
 - Predict the mean value of y if $x = 8$.
 - Construct a 95% confidence interval for the mean value of y if $x = 8$.
 - Predict the value of y if $x = 8$.
 - Construct a 95% prediction interval for the value of y if $x = 8$.
 - Explain the difference between the prediction in parts (a) and (c).
- Using the sample data from Problem 7 in Section 12.3,
 - Predict the mean value of y if $x = 1.4$.
 - Construct a 95% confidence interval for the mean value of y if $x = 1.4$.
 - Predict the value of y if $x = 1.4$.
 - Construct a 95% prediction interval for the value of y if $x = 1.4$.
- Using the sample data from Problem 8 in Section 12.3,
 - Predict the mean value of y if $x = 1.8$.
 - Construct a 95% confidence interval for the mean value of y if $x = 1.8$.
 - Predict the value of y if $x = 1.8$.
 - Construct a 95% prediction interval for the value of y if $x = 1.8$.

Applying the Concepts

7. An Unhealthy Commute Use the results of Problem 11 from Section 12.3 to answer the following questions:

- (a) Predict the mean well-being index composite score of all individuals whose commute time is 20 minutes.
- (b) Construct a 90% confidence interval for the mean well-being index composite score of all individuals whose commute time is 20 minutes.
- (c) Predict the well-being index composite score of Jane, whose commute time is 20 minutes.
- (d) Construct a 90% prediction interval for the well-being index composite score of Jane, whose commute time is 20 minutes.
- (e) Explain the difference between the predictions made in parts (a) and (c).

8. Credit Scores Use the results of Problem 12 from Section 12.3 to answer the following questions:

- (a) Predict the mean interest rate of all individuals whose credit score is 730.
- (b) Construct a 90% confidence interval for the mean interest rate of all individuals whose credit score is 730.
- (c) Predict the interest rate of Kaleigh, whose credit score is 730.
- (d) Construct a 90% prediction interval for the interest rate of Kaleigh, whose credit score is 730.
- (e) Explain the difference between the prediction made in parts (a) and (c).

9. Height versus Head Circumference Use the results of Problem 13 from Section 12.3 to answer the following questions:

- (a) Predict the mean head circumference of children who are 25.75 inches tall.
- (b) Construct a 95% confidence interval for the mean head circumference of children who are 25.75 inches tall.
- (c) Predict the head circumference of a randomly selected child who is 25.75 inches tall.
- (d) Construct a 95% prediction interval for the head circumference of a child who is 25.75 inches tall.
- (e) Explain the difference between the predictions in parts (a) and (c).

10. Hurricanes Use the results of Problem 14 in Section 12.3 to answer the following questions:

- (a) Predict the mean wind speed of all hurricanes whose atmospheric pressure is 950 mb.
- (b) Construct a 95% confidence interval for the mean wind speed found in part (a).
- (c) Predict the wind speed of a randomly selected hurricane whose atmospheric pressure is 950 mb.
- (d) Construct a 95% prediction interval for the wind speed found in part (c).
- (e) Explain why the predicted lengths in parts (a) and (c) are the same, yet the intervals constructed in parts (b) and (d) are different.

11. Concrete Use the results of Problem 15 from Section 12.3 to answer the following questions:

- (a) Predict the mean 28-day strength of concrete whose 7-day strength is 2550 psi.
- (b) Construct a 95% confidence interval for the mean 28-day strength of concrete whose 7-day strength is 2550 psi.
- (c) Predict the 28-day strength of concrete whose 7-day strength is 2550 psi.

- (d) Construct a 95% prediction interval for the 28-day strength of concrete whose 7-day strength is 2550 psi.

- (e) Explain the difference between the predictions in parts (a) and (c).

12. Tar and Nicotine Use the results of Problem 16 in Section 12.3 to answer the following questions:

- (a) Predict the mean nicotine content of all cigarettes whose tar content is 12 mg.
- (b) Construct a 95% confidence interval for the tar content found in part (a).
- (c) Predict the nicotine content of a randomly selected cigarette whose tar content is 12 mg.
- (d) Construct a 95% prediction interval for the nicotine content found in part (c).
- (e) Explain why the predicted nicotine contents found in parts (a) and (c) are the same, yet the intervals constructed in parts (b) and (d) are different.

13. Invest in Education Use the results of Problem 17 in Section 12.3 to answer the following questions:

- (a) Predict the mean annual ROI of all four-year schools whose cost is \$180,000.
- (b) Construct a 95% confidence interval for the mean annual ROI found in part (a).
- (c) Predict the annual ROI of a randomly selected four-year school whose cost is \$180,000.
- (d) Construct a 95% prediction interval for the annual ROI found in part (c).
- (e) Explain why the predicted lengths in parts (a) and (c) are the same, yet the intervals constructed in parts (b) and (d) are different.

14. American Black Bears Use the results of Problem 18 from Section 12.3 to answer the following questions:

- (a) Predict the mean weight of American black bears with a total length of 154.5 cm.
- (b) Construct a 95% confidence interval for the mean weight of American black bears with a total length of 154.5 cm.
- (c) Predict the weight of a randomly selected American black bear that is 154.5 cm long.
- (d) Construct a 95% prediction interval for the weight of an American black bear that is 154.5 cm long.
- (e) Explain why the predicted weights in parts (a) and (c) are the same, yet the intervals constructed in parts (b) and (d) are different.

15. CEO Performance Using the results of Problem 19 from Section 12.3, explain why it does not make sense to construct confidence or prediction intervals based on the least-squares regression equation.

16. Bear Markets Using the results of Problem 20 from Section 12.3, explain why it does not make sense to construct confidence or prediction intervals based on the least-squares regression equation.

 **17. Threaded Problem: Tornado** Is the width of a tornado related to the distance for which the tornado is on the ground? Go to www.pearsonhighered.com/sullivanstats to obtain the data file 12_4_17. These data represent the width (in yards) and length (in miles) for tornadoes in Louisiana in 2017. These data are from the “Tornadoes_2017” data that we have been analyzing throughout the course.

- (a) Treating width of the tornado as the explanatory variable and distance on the ground as the response variable, determine estimates for β_0 and β_1 .

- (b) Draw a boxplot of the residuals. Are there any outliers?
- (c) While the residuals are not approximately normally distributed, the sample size is large. Therefore, it is appropriate to use Student's t -distribution to test whether a positive linear relation exists between width and distance. Conduct this test.
- (d) Estimate the mean length of all tornadoes whose width is 500 yards with 95% confidence.
- (e) Estimate the length of a tornado whose width is 500 yards with 95% confidence.
- (f) What do you think is the cause for the wide interval estimate in part (e)?

DATA **18. Home Runs** (Refer to Problem 31, Section 4.2) The following data represent the speed at which a ball was hit (in miles per hour) and the distance it traveled (in feet) for a random sample of home runs in a Major League baseball game in 2018.

Speed (mph)	Distance (feet)
107.9	441
110.4	427
103.5	422
105.4	418
105.5	414
101.7	411
103.3	408
101.0	405
103.6	402
101.4	399
100.7	396
101.3	393

Source: baseballsavant.mlb.com

- (a) Treating speed at which the ball was hit as the explanatory variable and distance the ball traveled as the response variable, determine estimates for β_0 and β_1 .
- (b) The residuals are approximately normally distributed. Test whether a positive linear relation exists between speed at which a ball is hit and distance the ball travels.
- (c) Estimate the mean distance a ball will travel among all home runs hit with a speed of 103 mph with 95% confidence.
- (d) Estimate the distance a home run ball will travel if it is hit with a speed of 103 mph with 95% confidence.

DATA **19. Putting It Together: Predicting Intelligence** Can a photograph of an individual be used to predict their intelligence? Researchers at Charles University in Prague, Czech Republic, had 160 raters analyze the photos of 80 students and asked each rater to rate the intelligence and attractiveness of the individual in the photo on a scale from one to seven. To eliminate individual bias in ratings, each rater's scores were converted to z -scores using each individual's mean rating. The perceived intelligence and attractiveness of each photo was calculated as the mean z -score. Go to www.pearsonhighered.com/sullivanstats to obtain the data file 12_4_19 using the file format of your choice. The following explains the variables in the data:

- sex: Gender of the individual in the photo
- age: Age of the individual in the photo
- perceived intelligence (ALL): Mean z -score of the perceived intelligence of all 160 raters
- perceived intelligence (WOMEN): Mean z -score of the perceived intelligence of the female raters
- perceived intelligence (MEN): Mean z -score of the perceived intelligence of the male raters
- attractiveness (ALL): Mean z -score of the attractiveness rating of all 160 raters
- attractiveness (MEN): Mean z -score of the attractiveness rating of the male raters
- attractiveness (WOMEN): Mean z -score of the attractiveness rating of the female raters
- IQ: Intelligence quotient based on the Czech version of Intelligence Structure Test

Source: Kleisner K, Chvátalová V, Flegr J (2014) Perceived Intelligence Is Associated with Measured Intelligence in Men but Not Women. PLoS One 9(3): e81237. doi:10.1371/journal.pone.0081237

- (a) Are attractive people perceived as more intelligent? Draw a scatter diagram between attractiveness (ALL) and perceived intelligence (ALL) for all 160 raters treating perceived intelligence as the response variable.
- (b) What is the linear correlation coefficient between attractiveness and perceived intelligence for all 160 raters? Based on the linear correlation coefficient, does a linear relation exist between attractiveness and perceived intelligence?
- (c) Treating perceived intelligence (ALL) as the response variable and attractiveness (ALL) as the explanatory variable, find the least-squares regression equation between these two variables.
- (d) Provide an interpretation of the intercept.
- (e) A normal probability plot confirms the residuals are normally distributed. Test whether a positive linear relation exists between perceived intelligence and attractiveness.
- (f) Are higher IQs associated with higher perceived intelligence? Draw a scatter diagram between IQ and perceived intelligence for all 160 raters treating IQ as the response variable. What is the linear correlation coefficient between IQ and perceived intelligence (ALL)? Is this linear correlation coefficient suggestive of a linear relation between the two variables? Explain.
- (g) Treating IQ as the response variable, find the least-squares regression between IQ and perceived intelligence (ALL) for females only ($sex = F$). Test whether a positive linear relation exists between perceived intelligence for females only and IQ. Use an $\alpha = 0.1$ level of significance.
- (h) Treating IQ as the response variable, find the least-squares regression between IQ and perceived intelligence (ALL) for males only ($sex = M$). Test whether a positive linear relation exists between perceived intelligence for males only and IQ. Use an $\alpha = 0.1$ level of significance.
- (i) Construct a 95% confidence interval for the mean IQ of males who have perceived intelligence of 1.28.
- (j) Construct a 95% prediction interval for the IQ of a particular male whose perceived intelligence is 1.28.



Chapter 12 Review

Summary

In this chapter, we continued our discussion of inferential statistics. Here, we introduced chi-square methods and inference on the least-squares regression line.

The first chi-square method involved tests for goodness-of-fit. The null hypothesis is always a statement of no change/no effect/no difference. So the null hypothesis in a goodness-of-fit test is that the random variable follows some specified distribution. That is, there is no difference in the distribution of the sample data and that stated in the null hypothesis. The alternative hypothesis (or statement we are gathering evidence to demonstrate) is that the random variable does not follow the specified distribution. We used the chi-square distribution to test these hypotheses. This is done by comparing the expected values based on the distribution of the random variable to the observed values. If the observed values differ significantly from those expected, we have evidence against the statement in the null hypothesis.

Next, we introduced the chi-square test for independence. In such a test, the researcher obtains random data for two variables and tests whether the variables are associated. The null hypothesis in these tests is always that the variables are not associated (independent). The test statistic compares the expected values if the variables were independent to those observed. If the expected and observed values differ significantly, we reject the null hypothesis and conclude that there is evidence to support the belief that the variables are not independent (they are associated). We draw bar graphs of the conditional distributions to help us see the association, if any.

The chi-square test for homogeneity of proportions is an extension of the test for comparing two independent proportions. Here, we are testing the null hypothesis that

three or more proportions are equal versus the alternative that at least one proportion differs from the others. This test uses the exact same approach as that for independence. While the procedures for the test for independence and the test of homogeneity of proportions are the same, the data differ.

In the test for independence, we are measuring two variables (such as marital status and level of happiness) on each individual. In other words, a *single* population is segmented based on the value of two variables. In the test of homogeneity of proportions, we consider whether the proportion of individuals among *different* populations have the same value. So, if you have a single population in which two variables are measured on each individual to assess whether one variable might be associated with another, conduct a test of independence. If you have two or more populations in which you want to determine equality of proportions among the choices, conduct a test of homogeneity of proportions.

The last two sections of this chapter dealt with inferential techniques that can be used on the least-squares regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

In Section 12.3, we used sample data to obtain estimates of an intercept and slope. The residuals are required to be normally distributed, with mean 0 and constant standard deviation σ . Provided that these requirements are satisfied, we can test hypotheses regarding the slope to determine whether the relation between the explanatory and response variables is linear.

In Section 12.4, we learned how to construct confidence and prediction intervals for a predicted value. We constructed confidence intervals for a mean response and prediction intervals for an individual response.

Vocabulary

Chi-square distribution (p. 524)
Goodness-of-fit test (p. 526)
Expected counts (pp. 527, 541)
Test statistic (pp. 528, 542)
Chi-square test for independence (p. 539)

Chi-square test for homogeneity of proportions (p. 545)
Least-squares regression model (p. 556)
Standard error of the estimate (p. 557)
Robust (p. 560)

Bivariate normal distribution (p. 564)
Jointly normally distributed (p. 564)
Confidence interval for a mean response (p. 569)
Prediction interval for an individual response (p. 569)

Formulas

Expected Counts in a Goodness-of-Fit Test

$$E_i = \mu_i = np_i \quad \text{for } i = 1, 2, \dots, k$$

Chi-Square Test Statistic

$$\chi^2_0 = \sum \frac{(O_i - E_i)^2}{E_i} \quad i = 1, 2, \dots, k$$

Expected Frequencies in a Test for Independence or Homogeneity of Proportions

$$\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{table total}}$$

Standard Error of the Estimate

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum \text{residuals}^2}{n-2}}$$

Standard Error of b_1

$$s_{b_1} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Confidence Intervals for the Slope of the Regression Line

A $(1 - \alpha) \cdot 100\%$ confidence interval for the slope of the true regression line, β_1 , is given by the following formulas:

$$\text{Lower bound: } b_1 - t_{\alpha/2} \cdot \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}} = b_1 - t_{\alpha/2} \cdot s_{b_1} \quad \text{Upper bound: } b_1 + t_{\alpha/2} \cdot \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}} = b_1 + t_{\alpha/2} \cdot s_{b_1}$$

Here, $t_{\alpha/2}$ is computed with $n - 2$ degrees of freedom.

Confidence Interval about the Mean Response of \hat{y}

A $(1 - \alpha) \cdot 100\%$ confidence interval about the mean response of y , \hat{y} , is given by

$$\text{Lower bound: } \hat{y} - t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \quad \text{Upper bound: } \hat{y} + t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

where x^* is the given value of the explanatory variable and $t_{\alpha/2}$ is the critical value with $n - 2$ degrees of freedom.

Prediction Interval about an Individual Response, \hat{y}

A $(1 - \alpha) \cdot 100\%$ prediction interval for the individual response of y , \hat{y} , is given by

$$\text{Lower bound: } \hat{y} - t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \quad \text{Upper bound: } \hat{y} + t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

where x^* is the given value of the explanatory variable and $t_{\alpha/2}$ is the critical value with $n - 2$ degrees of freedom.

Objectives

Section	You should be able to . . .	Example(s)	Review Exercises
12.1	1 Find critical values for the chi-square distribution (p. 524) 2 Perform a goodness-of-fit test (p. 526)	1 2–4	1 and 2
12.2	1 Perform a test for independence (p. 538) 2 Perform a test for homogeneity of proportions (p. 545)	1–3 4	3 and 4 5
12.3	1 State the requirements of the least-squares regression model (p. 555) 2 Compute the standard error of the estimate (p. 557) 3 Verify that the residuals are normally distributed (p. 558) 4 Conduct inference on the slope of the least-squares regression model (p. 559) 5 Construct a confidence interval about the slope of the least-squares regression model (p. 562)	p. 555 2 and 3 4 5 6	6 7(b), 8(b), 9(c) 7(c), 8(c), 9(d) 7(e), 8(e), 9(f), 7(f), 8(f), 9(g)
12.4	1 Construct confidence intervals for a mean response (p. 570) 2 Construct prediction intervals for an individual response (p. 570)	1 and 3 2 and 3	7(g), 8(g) 7(i), 8(i)

Review Exercises

- 1. Roulette Wheel** A pit boss suspects that a roulette wheel is out of balance. A roulette wheel has 18 black slots, 18 red slots, and 2 green slots. The pit boss spins the wheel 500 times and records the following frequencies:

Outcome	Frequency
Black	233
Red	237
Green	30

Is the wheel out of balance? Use the $\alpha = 0.05$ level of significance.

- 2. World Series** Are the teams that play in the World Series evenly matched? To win a World Series, a team must win four games. If the teams are evenly matched, we would expect the

number of games played in the World Series to follow the distribution shown in the first two columns of the following table. The third column represents the actual number of games played in each World Series from 1930 to 2019. Do the data support the distribution that would exist if the teams are evenly matched and the outcome of each game is independent? Use the $\alpha = 0.05$ level of significance.

Number of Games	Probability	Observed Frequency
4	0.125	16
5	0.25	18
6	0.3125	19
7	0.3125	36

Source: Major League Baseball.

3. Titanic With 20% of men, 74% of women, and 52% of children surviving the infamous *Titanic* disaster, it is clear that the saying “women and children first” was followed. But what, if any, role did the class of service play in the survival of passengers? The data shown represent the survival status of passengers by class of service.

		Class		
		First	Second	Third
Survival Status	Survived	203	118	178
	Did Not Survive	122	167	528

- (a) Is class of service independent of survival rate? Use the $\alpha = 0.05$ level of significance.
 (b) Construct a conditional distribution of survival status by class of service and draw a bar graph. What does this summary tell you?

4. Premature Birth and Education Does the length of term of pregnancy play a role in the level of education of the baby? Researchers in Norway followed over 1 million births between 1967 and 1988 and looked at the educational attainment of the children. The following data are based on the results of their research. Note that a full-term pregnancy is 38 weeks. Is gestational period independent of completing a high school diploma? Use the $\alpha = 0.05$ level of significance.

		Gestational Period (weeks)				
		22–27	28–32	33–36	37–42	43+
Less Than High School Degree	Yes	14	34	140	1010	81
	No	26	65	343	3032	208

Source: *Chicago Tribune*, March 26, 2008.

5. Roosevelt versus Landon One of the most famous presidential elections (from a statistician’s point of view) is the 1936 contest between incumbent Franklin D. Roosevelt (FDR) and Republican challenger Alf Landon. The notoriety of the election comes from the fact that polling done by *Literary Digest* suggested that Landon would win by a 3 to 2 margin. However, Roosevelt actually crushed Landon in the general election.

After the election, author and editor David Lawrence studied the vote. Lawrence asked the following question: “To what extent was the campaign a reflection of a new trend in American politics, a trend in which the federal government’s paternalistic interest in the citizen brought an amazing reward to the party in power?”

As part of FDR’s New Deal policies to help get the United States out of the Great Depression, the federal government created the Agricultural Adjustment Administration (AAA). The AAA tried to force higher prices for commodities by paying farmers not to farm and not to bring cattle to market.

To determine whether farm subsidies may have played a role in the election results, Lawrence looked at election results in 2052 counties. He segmented the counties based on AAA funding between 1934 and 1935 as follows:

- High-funded AAA counties received \$500,000 or more in AAA money
- Medium-funded AAA counties received between \$100,000 and \$500,000 in AAA money
- Low-funded AAA counties received some AAA money, but less than \$100,000
- Counties with no AAA-fund did not receive any funds.

Treat the following data as a random sample of voters within each county type. The results are based on the election results in the various counties.

	High-Funded	Medium-Funded	Low-Funded	No Funding
Roosevelt	745	641	513	411
Landon	498	484	437	465

Source: Folsom, Burton, “New Deal or Raw Deal?” Simon & Schuster, 2008.

- (a) Do the data suggest that the level of funding received by counties through the AAA is associated with the candidate? Use the $\alpha = 0.05$ level of significance.
 (b) Construct a conditional distribution of candidate by level of AAA funding and draw a bar graph.

6. Obligations to Vote and Serve In the General Social Survey, individuals were asked whether civic duty included voting and whether it included serving on a jury. The results of the survey are shown in the table. Is there a difference in the proportion of individuals who feel jury duty is a civic duty and the proportion of individuals who feel voting is a civic duty? Use the $\alpha = 0.05$ level of significance.

		Jury	
		Duty (success)	Not Duty (failure)
Voting	Duty (success)	1322	65
	Not duty (failure)	45	17

7. What is the simple least-squares regression model? What are the requirements to perform inference on a simple least-squares regression line? How do we verify that these requirements are met?

8. Seat Choice and GPA A biology professor wants to investigate the relation between the seat location chosen by a student on the first day of class and their cumulative grade point average (GPA). He randomly selected an introductory biology class and obtained the following information for the 38 students in the class.

Row Chosen, x	GPA, y	Row Chosen, x	GPA, y
1	4.00	6	2.63
2	3.35	6	3.15
2	3.50	6	3.69
2	3.67	6	3.71
2	3.75	7	2.88
3	3.37	7	2.93
3	3.62	7	3.00
4	2.35	7	3.21
4	2.71	7	3.53
4	3.75	7	3.74
5	3.10	7	3.75
5	3.22	7	3.90
5	3.36	8	2.30
5	3.58	8	2.54
5	3.67	8	2.61
5	3.69	9	2.71
5	3.72	9	3.74
5	3.84	9	3.75
6	2.35	11	1.71

Source: S. Kalinowski and Taper M. “The Effect of Seat Location on Exam Grades and Student Perceptions in an Introductory Biology Class.” *Journal of College Science Teaching*, 36(4):54–57, 2007.

- (a) Treating row as the explanatory variable, determine the estimates of β_0 and β_1 . What is the mean GPA of students who choose a seat in the fifth row?
- (b) Compute the standard error of the estimate, s_e .
- (c) Determine whether the residuals are normally distributed.
- (d) Determine s_{b_1} .
- (e) If the residuals are normally distributed, test whether a linear relation exists between the explanatory variable, row choice, and response variable, GPA, at the $\alpha = 0.05$ level of significance.
- (f) If the residuals are normally distributed, construct a 95% confidence interval for the slope of the true least-squares regression line.
- (g) Construct a 95% confidence interval for the mean GPA of students who choose a seat in the fifth row.
- (h) Predict the GPA of a randomly selected student who chooses a seat in the fifth row.
- (i) Construct a 95% prediction interval for the GPA found in part (h).
- (j) Explain why the predicted GPAs found in parts (a) and (h) are the same, yet the intervals are different.

DATA **9. Apartments** The following data represent the square footage and rents for apartments in Queens, New York and Nassau County, New York. For this problem, only consider the Queens data.

Queens (New York City)		Nassau County (Long Island)	
Square Footage, x	Rent per Month, y	Square Footage, x	Rent per Month, y
500	650	1100	1875
588	1215	588	1075
1000	2000	1250	1775
688	1655	556	1050
825	1250	825	1300
1259	2700	743	1475
650	1200	660	1315
560	1250	975	1400
1073	2350	1429	1900
1452	3300	800	1650
1305	3100	1077	1395

Source: apartments.com

- (a) What are the estimates of β_0 and β_1 ? What is the mean rent of a 900-square-foot apartment in Queens?
- (b) Compute the standard error of the estimate, s_e .
- (c) Determine whether the residuals are normally distributed.
- (d) Determine s_{b_1} .

- (e) If the residuals are normally distributed, test whether a linear relation exists between the explanatory variable, x , and response variable, y , at the $\alpha = 0.05$ level of significance.
- (f) If the residuals are normally distributed, construct a 95% confidence interval for the slope of the true least-squares regression line.
- (g) Construct a 90% confidence interval for the mean rent of all 900-square-foot apartments in Queens.
- (h) Predict the rent of a particular 900-square-foot apartment in Queens.
- (i) Construct a 90% prediction interval for the rent of a particular 900-square-foot apartment in Queens.
- (j) Explain why the predicted rents found in parts (a) and (h) are the same, yet the intervals are different.

10. Calories versus Sugar The following data represent the number of calories per serving and the number of grams of sugar per serving for a random sample of high-protein and moderate-protein energy bars.

Calories, x	Sugar, y	Calories, x	Sugar, y
180	10	270	20
200	18	320	2
210	14	110	10
220	20	180	12
220	0	200	22
230	28	220	24
240	2	230	24

Source: Consumer Reports.

- (a) Draw a scatter diagram of the data, treating calories as the explanatory variable. What type of relation, if any, appears to exist between calories and sugar?
- (b) Determine the least-squares regression equation from the sample data.
- (c) Compute the standard error of the estimate.
- (d) Determine whether the residuals are normally distributed.
- (e) Determine s_{b_1} .
- (f) If the residuals are normally distributed, test whether a linear relation exists between calories and sugar content at the $\alpha = 0.01$ level of significance.
- (g) If the residuals are normally distributed, construct a 95% confidence interval about the slope of the true least-squares regression line.
- (h) For a randomly selected energy bar, would you recommend using the least-squares regression line obtained in part (b) to predict the sugar content of the energy bar? Why? What would be a good estimate for the sugar content of the energy bar?



Chapter Test

1. A pit boss is concerned that a pair of dice being used in a craps game is not fair. The distribution of the expected sum of two fair dice is as follows:

Sum of Two Dice	Probability	Sum of Two Dice	Probability
2	$\frac{1}{36}$	8	$\frac{5}{36}$
3	$\frac{2}{36}$	9	$\frac{4}{36}$
4	$\frac{3}{36}$	10	$\frac{3}{36}$
5	$\frac{4}{36}$	11	$\frac{2}{36}$
6	$\frac{5}{36}$	12	$\frac{1}{36}$
7	$\frac{6}{36}$		

The pit boss rolls the dice 400 times and records the sum of the dice. The table shows the results. Do you think the dice are fair? Use the $\alpha = 0.01$ level of significance.

Sum of Two Dice	Frequency	Sum of Two Dice	Frequency
2	16	8	59
3	23	9	45
4	31	10	34
5	41	11	19
6	62	12	11
7	59		

2. A researcher wanted to determine if the distribution of educational attainment of Americans today is different from the distribution in 2000. The distribution of educational attainment in 2000 was as follows:

Education	Relative Frequency
Not a high school graduate	0.158
High school graduate	0.331
Some college	0.176
Associate's degree	0.078
Bachelor's degree	0.170
Advanced degree	0.087

Source: *Statistical Abstract of the United States*.

5. Many municipalities are passing legislation that forbids smoking in restaurants and bars. Bar owners claim that these laws hurt their business. Are their concerns legitimate? The following data represent the smoking status and frequency of visits to bars from the General Social Survey. Do smokers tend to spend more time in bars? Use the $\alpha = 0.05$ level of significance.

	Almost Daily	Several Times a Week	Several Times a Month	Once a Month	Several Times a Year	Once a Year	Never
Smoker	80	409	294	362	433	336	1265
Nonsmoker	57	350	379	471	573	568	3297

The researcher randomly selects 500 Americans, learns their levels of education, and obtains the data shown in the table. Do the data suggest that the distribution of educational attainment has changed since 2000? Use the $\alpha = 0.1$ level of significance.

Education	Frequency
Not a high school graduate	72
High school graduate	159
Some college	85
Associate's degree	44
Bachelor's degree	92
Advanced degree	48

3. The Harris Poll asked a random sample of adult Americans, "How important are moral values when deciding how to vote?" The results of the survey by disclosed political affiliation are shown in the table.

		Political Affiliation		
		Republican	Independent	Democrat
Morality	Important	644	662	670
	Not important	56	155	147

- (a) Do the sample data suggest that the proportion of adults who feel morality is important differ based on political affiliation? Use the $\alpha = 0.05$ level of significance.
(b) Construct a conditional distribution of morality by party affiliation and draw a bar graph. What does this summary tell you?
4. The General Social Survey regularly asks individuals to disclose their religious affiliation. The following data represent the religious affiliation of young adults, aged 18 to 29, in the 1970s, 1980s, 1990s, and 2000s. Do the data suggest different proportions of 18- to 29-year-olds have been affiliated with religion in the past four decades? Use the $\alpha = 0.05$ level of significance.

	1970s	1980s	1990s	2000s
Affiliated	2395	3022	2121	624
Unaffiliated (no religion)	327	412	404	2087

Source: General Social Survey.

6. State the requirements to perform inference on a simple least-squares regression line.

- DATA** 7. Crickets make a chirping noise by sliding their wings rapidly over each other. Perhaps you have noticed that the number of chirps seems to increase with the temperature. The following table lists the temperature (in degrees Fahrenheit, °F) and the number of chirps per second for the striped ground cricket.

Temperature, x	Chirps per Second, y	Temperature, x	Chirps per Second, y
88.6	20.0	71.6	16.0
93.3	19.8	84.3	18.4
80.6	17.1	75.2	15.5
69.7	14.7	82.0	17.1
69.4	15.4	83.3	16.2
79.6	15.0	82.6	17.2
80.6	16.0	83.5	17.0
76.3	14.4		

Source: George W. Pierce. *The Songs of Insects*, Cambridge, MA: Harvard University Press, 1949, pp. 12–21.

- (a) What are the estimates of β_0 and β_1 ? What is the mean number of chirps when the temperature is 80.2°F?
 (b) Compute the standard error of the estimate, s_e .
 (c) Determine whether the residuals are normally distributed.
 (d) Determine s_{β_1} .
 (e) If the residuals are normally distributed, test whether a linear relation exists between the explanatory variable, x , and response variable, y , at the $\alpha = 0.05$ level of significance.
 (f) If the residuals are normally distributed, construct a 95% confidence interval for the slope of the true least-squares regression line.
 (g) Construct a 90% confidence interval for the mean number of chirps found in part (a).
 (h) Predict the number of chirps on a day when the temperature is 80.2°F.
 (i) Construct a 90% prediction interval for the number of chirps found in part (h).
 (j) Explain why the predicted number of chirps found in parts (a) and (h) are the same, yet the intervals are different.

- DATA** 8. The following data represent the height (inches) of boys between the ages of 2 and 10 years.

Boy Age, x	Height, y	Boy Age, x	Height, y	Boy Age, x	Height, y
2	36.1	5	45.6	8	48.3
2	34.2	5	44.8	8	50.9
2	31.1	5	44.6	9	52.2
3	36.3	6	49.8	9	51.3
3	39.5	7	43.2	10	55.6
4	41.5	7	47.9	10	59.5
4	38.6	8	51.4		

Source: National Center for Health Statistics.

- (a) Treating age as the explanatory variable, determine the estimates of β_0 and β_1 . What is the mean height of a 7-year-old boy?
 (b) Compute the standard error of the estimate, s_e .
 (c) Assuming the residuals are normally distributed, test whether a linear relation exists between the explanatory variable, age, and response variable, height, at the $\alpha = 0.05$ level of significance.
 (d) Assuming the residuals are normally distributed, construct a 95% confidence interval for the slope of the true least-squares regression line.
 (e) Construct a 90% confidence interval for the mean height found in part (a).
 (f) Predict the height of a 7-year-old boy.
 (g) Construct a 90% prediction interval for the height found in part (f).
 (h) Explain why the predicted heights found in parts (a) and (f) are the same, yet the intervals are different.

- DATA** 9. A researcher believes that as age increases, the grip strength (pounds per square inch, psi) of an individual's dominant hand decreases. From a random sample of 17 females, he obtains the following data.

Age, x	Grip Strength, y	Age, x	Grip Strength, y
15	65	34	45
16	60	37	58
28	58	41	70
61	60	43	73
53	46	49	45
43	66	53	60
16	56	61	56
25	75	68	30
28	46		

Source: Kevin McCarthy, student at Joliet Junior College.

- (a) Treating age as the explanatory variable, determine the estimates of β_0 and β_1 .
 (b) Assuming the residuals are normally distributed, test whether a linear relation exists between the explanatory variable, age, and response variable, grip strength, at the $\alpha = 0.05$ level of significance.
 (c) Based on your answer to (b), what would be a good estimate of the grip strength of a randomly selected 42-year-old female?

Making an Informed Decision

Benefits of College

Are there benefits to attending college? If so, what are they? In this project, we will identify some of the perks that a college education provides. Obtain a random sample of at least 50 people aged 21 years or older and administer the following survey.

Please answer the following questions:

1. What is the highest level of education you have attained?

Have not completed high school

High school graduate

College graduate

2. What is your employment status?

Employed

Unemployed, but actively seeking work

Unemployed, but not actively seeking work

3. If you are employed, what is your annual income?

Less than \$20,000

\$20,000–\$39,999

\$40,000–\$60,000

More than \$60,000

4. If you are employed, which statement best describes the level of satisfaction you have with your career? Answer this question only if you are employed.

Satisfied—I enjoy my job and am happy with my career.

Somewhat satisfied—Work is work, but I am not unhappy with my career.

Somewhat dissatisfied—I do not enjoy my work, but I also have no intention of leaving.

Dissatisfied—Going to work is painful. I would quit tomorrow if I could.

(a) Use the results of the survey to create a contingency table for each of the following categories:

- Level of education/employment status
- Level of education/annual income
- Level of education/job satisfaction
- Annual income/job satisfaction

(b) Perform a chi-square test for independence on each contingency table from part (a).

(c) Draw bar graphs for each contingency table from part (a).

(d) Write a report that details your findings.



This page intentionally left blank

Credits

Cover Image Credit: Vasabii/Shutterstock

Chapter 1: Image Credits

SHUTTERSTOCK: Focal point/Shutterstock 2; **SHUTTERSTOCK:** Ministr-84/Shutterstock 8; **Texas Instruments:** Texas Instruments 27; **Texas Instruments:** Texas Instruments 27; **Minitab:** Minitab 28; **Microsoft Corporation:** © Microsoft Corporation 28; **StatCrunch:** StatCrunch 28; **Minitab:** Minitab 31; **Shutterstock:** Zimmytws/Shutterstock 39; **Pearson Education:** Pearson Education 45; **123RF GB LIMITED:** Cathy Yeulet/Stock Photo/123RF Ltd 49; **Shutterstock:** Samuel Borges Photography/Shutterstock 52; **SHUTTERSTOCK:** Matka_Wariatka/Shutterstock 54; **SHUTTERSTOCK:** Focal point/Shutterstock 62;

Chapter 1: Text Credits

Houghton Mifflin Harcourt: American Heritage Dictionary 3; **Duke University:** Duke University, “Breast-feeding Boosts IQ in Infants with ‘Helpful’ Genetic Variant,” Science Daily 6 November 2007 04; **CIA World Factbook:** CIA World Factbook 9; **National Institute on Drug Abuse:** National Institute on Drug Abuse 11; **Bureau of Justice Statistics:** Bureau of Justice Statistics 11; **National Institute on Drug Abuse:** National Institute on Drug Abuse 11; **Baseball Almanac, Inc:** baseball-almanac.com 11; **GUARDIAN NEWS AND MEDIA LIMITED:** www.theguardian.com 11; **American Society of Microbiology and the Soap and Detergent Association:** American Society for Microbiology and the Soap and Detergent Association 11; **Newsweek:** Newsweek Magazine 11; **American Medical Association:** John P. Forman, MD; Eric B. Rimm, ScD; Meir J. Stampfer, MD; Gary C. Curhan, MD, ScD, “Folate Intake and the Risk of Incident Hypertension among US Women,” Journal of the American Medical Association 293:320–329, 2005. 12; **Governors Highway Safety Association:** Governors Highway Safety Association. 12; **Motor Trend Group, LLC:** Motor Trend Group, LLC. 12; U.S. **National Library of Medicine:** Weiser, M., Zarka, S., Werbeloff, N., Kravitz, E. and Lubin, G. (2010). “Cognitive Test Scores in Male Adolescent Cigarette Smokers Compared to Non-smokers: A Population-Based Study.” Addiction. 105:358–363. doi: 10.1111/j.1360-0443.2009.02740.x.) 12; **American Medical Association:** Dean R. Focht III, Carole Spicer, Mary P. Fairchok. “The Efficacy of Duct Tape vs. Cryotherapy in the Treatment of Verruca Vulgaris (The Common Wart),” Archives of Pediatrics and Adolescent Medicine, 156(10), 2002 12; **Telegraph Media Group Limited:** Season of Birth Affects Your Mood Later In Life by Nicola Fifield from The Telegraph. Copyright © 2014 by Telegraph Media Group Limited. 13 14; **Oxford University Press:** Benson, V. S. et al. “Mobile Phone Use and Risk of Brain Neoplasms and Other Cancers: Prospective Study,” International Journal of Epidemiology 2013 Jun; 42(3): 792–802 14; **National Institute of Environmental Health Sciences:** Report of Partial findings from the National Toxicology Program Carcinogenesis Studies of Cell Phone Radiofrequency Radiation in Hsd: Sprague Dawley SD Rats (Whole Body Exposure) Michael Wyde, Mark Cesta, Chad Blystone, Susan Elmore, Paul Foster, Michelle Hooth, Grace Kissling, David Malarkey, Robert Sills, Matthew Stout, Nigel Walker, Kristine Witt, Mary Wolfe, John Bucher, June 23, 2016 15; **Massachusetts Medical Society:** Kristin L. Nichol, MD, MPH, MBA, James D. Nordin, MD, MPH, David B. Nelson, PhD, John P. Mullooly, PhD, Eelko Hak, PhD. “Effectiveness of Influenza Vaccine in the Community-Dwelling Elderly,” New England Journal of Medicine 357:1373–1381, 2007 16; **Massachusetts Medical Society:** Kjell Benson, BA, and Arthur J. Hartz, MD, PhD. “A Comparison of Observational Studies and Randomized, Controlled Trials,” New England Journal of Medicine 342:1878–1886, 2000 17; **Gannett Satellite Information Network, LLC:** USA Today 21; **Oxford University Press:** European Heart Journal 31 (9):1065–1070, February 2010 21; **Lippincott Williams & Wilkins:** European Journal of Cancer Prevention, 16(5): 446–452, October 2007 21; **John Wiley & Sons, Inc:** Christelle Delmas, Carine Platat, Brigitte Schweitzer, Aline Wagner, Mohamed Ouaja, and Chantal Simon. “Association Between Television in Bedroom and Adiposity Throughout Adolescence,” Obesity, 15:2495–2503, 2007. 21; **BioMed Central Ltd:** Sally K Tracy, Alec Welsh, Donna Hartz, Anne Lainchbury, Andrew Bisits, Jan White, and Mark Tracy “Caseload midwifery compared to standard or private obstetric care for first time mothers in a public teaching hospital in Australia: a cross sectional study of cost and birth outcomes” BMC Pregnancy and Childbirth 2014, 14:46 21; **e: Joseph Moor:** Copyright © by Joseph Moore. 22; **Oxford University Press:** Mireille B. Toledano, Rachel B. Smith, Irene Chang, Margaret Douglass, Paul Elliott; Cohort Profile: UK Cosmos – a UK cohort for study of environment and health, International Journal of Epidemiology, Volume 46, Issue 3, 1 June 2017, Pages 775-787 22; **Boy Scouts of America:** Boy Scouts of America 29; **The Herald and Weekly Times:** Herald Sun, September 9, 2007 42; **Pew Research Center:** “Comparing Survey Sampling Strategies: Random-Digit Dial vs Voter Files” by Kennedy, Courtney et. al. Pew Research, Oct. 9, 2018 43; **Elsevier:** Yong Zhu and James H. Hollis. “Increasing the Number of Chews before Swallowing Reduces Meal Size in Normal-Weight, Overweight, and Obese Adults,” Journal of the Academy of Nutrition and Dietetics, 11 November 2013 52; **American Medical Association:** Bankole A. Johnson, Norman Rosenthal, et al. “Topiramate for Treating Alcohol Dependence: A Randomized Controlled Trial,” Journal of the American Medical Association, 298(14):1641–1651, 2007 52; **American Psychological Association:** Barth, John A., Chen, Mark, Burrows, Lara. Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action. Journal of Personality and Social Psychology. 1996. Vol. 71. No. 2, 230-244 53; **American Medical Association:** Jack D. Edinger, PhD; William K. Wohlgemuth, PhD; Rodney A. Radtke, MD; Gail R. Marsh, PhD; Ruth E. Quillian, PhD. “Cognitive Behavioral Therapy for Treatment of Chronic Primary Insomnia,” Journal of the American Medical Association 285:1856–1864, 2001. 53; **American Medical Association:** Paul R. Solomon et al. “Ginkgo for Memory Enhancement,” Journal of the American Medical Association 288:835–840, 2002. 54; **Nutrition Press:** A. Geliebter et al. “Reduced Stomach Capacity in Obese Subjects after Dieting,” American Journal of Clinical Nutrition 63(2):170–173, 1996 54; **John Wiley & Sons, Inc:** Tawfik AA, Osman MAR. The effect of autologous activated platelet-rich plasma injection on female pattern hair loss: A randomized placebo-controlled study. J Cosmet Dermatol. 2018;17:47-53. 54; **Nature Research:** E. K. Papies and associates, “Using Health Primes to Reduce Unhealthy Snack Purchases Among Overweight

Consumers in a Grocery Store," International Journal of Obesity (2013), 1–6. 55; **American Cancer Society:** American Cancer Society. 59; **BioMed Central Ltd:** Michael Moore et al. "In Vitro Tooth Whitening Effect of Two Medicated Chewing Gums Compared to a Whitening Gum and Saliva," BioMed Central Oral Health 8:23, 2008. 59; **BioMed Central Ltd:** Judith Hegeman et al. "Even Low Alcohol Concentrations Affect Obstacle Avoidance Reactions in Healthy Senior Individuals," BMC Research Notes 3:243, 2010 60; **Public Library of Science:** A. S. Attwood, N. E. Scott-Samuel, G. Stothart, M. R. Munafò (2012) "Glass Shape Influences Consumption Rate for Alcoholic Beverages." PLoS ONE 7(8) 61; **Pharmaceuticals Corp:** Pharmaceuticals Corp. 62; **American Society for Nutrition:** "Colas, but not other carbonated beverages, are associated with low bone mineral density in older women: The Framingham Osteoporosis Study," American Journal of Clinical Nutrition 84: 936–942, 2006. 62;

Chapter 2: Image Credits

123RF: Pakete/123RF Ltd 67 , 119; **Sandler Training Survey of 1,053 adults:** Sandler Training Survey of 1,053 adults 74; **Kiwanis International and Novartis Vaccines:** Kiwanis International and Novartis Vaccines 75; **Alamy Images:** Kumar Sriskandan/Alamy Stock Photo 80; **Pearson Education:** Pearson Education 84.

Chapter 2: Text Credits

Krystal Catton: Krystal Catton, student at Joliet Junior College 68; **U.S. Census Bureau:** U.S. Census Bureau 71; **U.S. Census Bureau:** U.S. Census Bureau 75; **Gallup, Inc.:** Gallup 75; **Baseball Almanac, Inc.:** <http://www.baseball-almanac.com/> 75; **Federal Trade Commission:** Federal Trade Commission 76; **Harris Interactive:** Harris Interactive 76; **Harris Interactive:** Harris Interactive 76; **Amy Loves It:** www.amylovesit.com 77; **Pew Research Center:** Pew Internet 77; **Marist Poll:** Marist Poll 77; **Crunchbase Inc.:** www.whatsgoodly.com 77; **Sandbox Networks, Inc.:** Based on results from www.infoplease.com 78; **Instacart:** www.instacart.com 78; **Verizon Media:** Yahoo! Finance 78; **Wakefield Research:** Based on results from Wakefield Research 78; **Sandbox Networks, Inc.:** Based on results from www.infoplease.com 79; **Association for the Advancement of Computing in Education:** Journal of Computers in Mathematics and Science Teaching 26(1):55–73, 2007 79; **U.S. Department of Health & Human Services:** National Vital Statistics Reports, Vol. 67, No. 1.82; **U.S. Census Bureau.:** U.S. Census Bureau 82; **NYC Open Data:** NYC Open Data 83; **Federal Reserve Bank of Philadelphia:** Federal Reserve Bank of Philadelphia 99; **PayScale, Inc.:** [payscale.com](http://www.payscale.com) 101; **NPA Services, Inc.:** www.miseryindex.us 101; **MLB Advanced Media, LP:** Statcast 90; **U.S. Geological Survey:** U.S. Geological Survey, Earthquake Hazards Program 90; **U.S. Department of Energy:** Based on data from the U.S. Department of Energy 90; **Federal Reserve Bank of St. Louis:** Federal Reserve Bank of St. Louis 91; **CIA World Factbook:** CIA World Factbook 91; **MLB Advanced Media, LP:** Statcast 91; **Tax Foundation:** Tax Foundation 92; **U.S. Department of Education:** U.S. Department of Education 92; **TD Ameritrade IP Company, Inc.:** TD Ameritrade 92; **Nielsen:** Based on information provided by Nielsen/NetRatings 92; **Treasury Direct:** treasurydirect.gov 103; **Stormfax:** Stormfax Weather Almanac 104; **Florida Museum of Natural History:** Florida Museum of Natural History 104; **U.S. Census Bureau:** U.S. Census Bureau 110; **W. W. Norton & Company, Inc.:** How to Lie with Statistics (W. W. Norton & Company, Inc., 1982) by Darrell Huff 110; **Graphics Press:** The Visual Display of Quantitative Information (Graphics Press,2001) by Edward Tufte 110; **U.S. Bureau of Labor Statistics:** U.S. Bureau of Labor Statistics 111; **U.S. Census Bureau:** U.S. Census Bureau 111; **U.S. Census Bureau:** U.S. Census Bureau 111; **U.S. Statistical Abstract:** U.S. Statistical Abstract 111; **U.S. Centers for Medicare & Medicaid Services:** U.S. Centers for Medicare & Medicaid Services 112; **Centers for Disease Control and Prevention:** Centers for Disease Control and Prevention 113; **Federal Bureau of Investigation:** FBI, Uniform Crime Reports 114; **U.S. Federal Statistical System:** National Center for Health Statistics 115; **U.S. Census Bureau:** U.S. Census Bureau 115; **Trina S. McNamara:** Trina S. McNamara, a student at Joliet Junior College 115; **Grade Inflation:** gradeinflation.com 116; **Tax Foundation:** The Tax Foundation 116; **HNTB Corporation:** HNTB Corporation 117; **Paul Oswiecimski:** Paul Oswiecimski 118; **U.S. Census Bureau:** Centers for Disease Control and U.S. Census Bureau 118.

Chapter 3: Image Credits

Texas Instruments: Texas Instruments 123; **Texas Instruments:** Texas Instruments 127; **Texas Instruments:** Texas Instruments 127; **Minitab:** Minitab 141; **Minitab:** Minitab 141; **Pearson Education:** Pearson Education 146; **Texas Instruments:** Texas Instruments 157; **Pearson Education:** Pearson Education 169; **Boston Gazette LLC:** Texas Instruments 174; **SHUTTERSTOCK:** Cleanfotos/Shutterstock 121 , 183.

Chapter 3: Text Credits

Yolanda Sullivan: Yolanda Sullivan's cell phone records 126; **Krystal Catton:** Krystal Catton, student at Joliet Junior College 129; **U.S. Department of Energy:** fueleconomy.gov 130; **Carol Wesolowski:** Carol Wesolowski, student at Joliet Junior College 132; **Emily McCarney:** Emily McCarney, student at Joliet Junior College 131; **Ladonna Hansen:** Ladonna Hansen, Park Curator 132; **TouringPlans.com:** TouringPlans.com 132; **Instacart.:** Instacart 133; **NORC at the University of Chicago:** Based on data from the General Social Survey 133; **Tamela Ohm:** Tamela Ohm, student at Joliet Junior College 133; **U.S. Department of Energy:** fueleconomy.gov 147; **Emily McCarney:** Emily McCarney, student at Joliet Junior College 148; **Andrew Dieter:** Andrew Dieter and Brad Schmidgall, students at Joliet Junior College 148; **Ladonna Hansen:** Ladonna Hansen, Park Curator 149; **U.S. Federal Statistical System:** National Climatic Data Center 149; **autotrader.com:** autotrader.com 151; **U.S. Federal Statistical System:** National Center for Health Statistics 151; **T. Rowe Price:** T. Rowe Price 151; **LearnVest, Inc.:** Based on a poll by LearnVest 157; **U.S. Census Bureau:** Based on data from the U.S. Census Bureau 157; **U.S. Department of Health & Human Services:** National Vital Statistics Reports 158; **U.S. Department of Energy:** Based on data from the U.S. Department of Energy 158; **U.S. Census Bureau:** U.S. Census Bureau 158; **U.S. Department of Health & Human Services:** National Vital Statistics Reports 159; **U.S. Census Bureau:** U.S. Census Bureau 159; **U.S. Department of Health & Human Services:** National Vital Statistics Reports 159; **U.S. Department of Health & Human Services:** CDC Vital and Health Statistics, Advance Data, Number 361, July 5, 2005. 166; **U.S. Department of Health & Human Services:** CDC Vital and Health Statistics, Advance Data, Number 361, July 5, 2005. 166; **U.S. Department of Energy:** fueleconomy.gov 168; **Joliet Junior College Veterinarian Technology Program:** Joliet Junior College Veterinarian Technology Program 168; **TouringPlans.com:**

TouringPlans.com 168; **Crunchbase Inc.**: www.whatsgoodly.com 168; **Laura Gillogly**: Laura Gillogly, student at Joliet Junior College 170; **NASA**: NASA Life Sciences Data Archive 173; **Sandbox Networks, Inc.**: factmonster.com 175; **U.S. Census Bureau**: American Community Survey by the U.S. Census Bureau 175; **NORC at the University of Chicago**: General Social Survey 175; **Kelly Roe**: Kelly Roe, student at Joliet Junior College 175; **Ladonna Hansen**: Ladonna Hansen, Park Curator 176; **Amanda A. Sindewald**: Amanda A. Sindewald, student at Joliet Junior College 176; **Trina McNamara**: Trina McNamara, student at Joliet Junior College 176; **American Public Health Association**: “The Effect of Paternal Smoking on the Birthweight of Newborns Whose Mothers Did Not Smoke,” Fernando D. Martinez, Anne L. Wright, Lynn M. Taussig, American Journal of Public Health Vol. 84, No. 9. 176; **PLOS**: de Ridder, D., Kroese, F., Adriaanse, M., & Evers, C., “Always Gamble on an Empty Stomach: Hunger Is Associated with Advantageous Decision Making,” PLOS One 9(10). doi: 10.1371/journal.pone.0111081. 177; **American Statistical Association**: Christenson, Ronald, and Blackwood, Larry, “Tests for Precision and Accuracy of Multiple Measuring Devices,” Technometrics, 35(4): 411–421, 1993. 180; **Cars.com**: Cars.com 180; **Infoplease**: Information Please Almanac 180; **U.S. Census Bureau**: Based on data from the American Community Survey 180; **Infoplease**: Infoplease.com 181.

Chapter 4: Image Credits

123RF: Vadim Sadovski/123RF 185; **Microsoft Corporation**: © Microsoft Corporation 192; **Minitab**: Minitab 196; **StatCrunch**: StatCrunch 196; **Texas Instruments**: Texas Instruments 206; **StatCrunch**: StatCrunch 207; **StatCrunch**: StatCrunch 207; **Microsoft Corporation**: © Microsoft Corporation 207; **Pearson Education**: Pearson Education 207; **Microsoft Corporation**: © Microsoft Corporation 238; **Microsoft Corporation**: © Microsoft Corporation 238 ; **123RF**:Vadim Sadovski/123RF 244; **Alamy Stock Photo**: History and Art Collection/Alamy Stock Photo 191; **Alamy Stock Photo**: Classic Image/Alamy Stock Photo 205.

Chapter 4: Text Credits

Paul Stephenson: Paul Stephenson, student at Joliet Junior College. 186; **Nutrition Press**: Based on data obtained from Katherine L. Tucker et al., “Colas, but not other carbonated beverages, are associated with low bone mineral density in older women: The Framingham Osteoporosis Study.” American Journal of Clinical Nutrition 2006, 84:936–942. 194; **Gallup, Inc.**: The Gallup Organization.197; **Fair Isaac Corporation**: www.myfico.com 197; **Denise Slucki**: Denise Slucki, student at Joliet Junior College. 197; **North Carolina Zoological Society**: www.fieldtripearth.org 197; **MLB Advanced Media, LP**: baseballsavant.mlb.com 198; **Dow Jones & Company**:The Wall Street Journal. 198; **Gold-Eagle Inc**: Gold-Eagle. 198; **U.S. Department of Transportation**:National Highway and Traffic Safety Institute. 199; **Elsevier**: L. Willerman, R. Schultz, J. N. Rutledge, and E. Bigler (1991). “In Vivo Brain Size and Intelligence,” Intelligence, 15, 223–228.199; **The American Statistician**: Frank Anscombe. “Graphs in Statistical Analysis,” American Statistician 27: 17–21, 1993. 200; **PAREonline.net**: Theodore Coladarci and Irv Kornfield. “RateMyProfessors.com versus Formal In-class Student Evaluations of Teaching,” Practical Assessment, Research, & Evaluation, 12:6, May, 2007.201; **Paul Stephenson**: Paul Stephenson, student at Joliet Junior College. 202; **F. Didot**: Legende in his text Nouvelles méthodes pour la détermination des orbites des comètes, published in 1806 204; **State University of New York** : Xu Zhang and Solomon W. Polachek, State University of New York at Binghamton “The Husband-Wife Age Gap at First Marriage: A Cross-Country Analysis.” 211; **BP p.l.c.**: www.bp.com 211; **Denise Slucki**: Denise Slucki, student at Joliet Junior College. 212; **Gallup, Inc.**: The Gallup Organization. 212; **Fair Isaac Corporation**: www.myfico.com 212; **North Carolina Zoological Society** : www.fieldtripearth.org212; **MLB Advanced Media, LP**: baseballsavant.mlb.com 213; **National Hurricane Center**: National Hurricane Center. 213; **U.S. Department of Transportation**: National Highway and Traffic Safety Institute. 214; **Zillow Group, Inc.**: zillow.com 214; **The American College of Obstetricians and Gynecologists**: Ira M. Bernstein et. al. “Maternal Smoking and Its Association with Birthweight.” Obstetrics & Gynecology 106 (Part 1) 5, 2005. 214; **B. Tauchnitz**: Mark Twain, in his book Life on the Mississippi (1884) 215; **Ladonna Hansen**: Ladonna Hansen, Park Curator. 222; **Ladonna Hansen**: Ladonna Hansen, Park Curator. 226; **Harris Poll**: The Harris Poll. 235; **Harris Poll**: The Harris Poll.235; **Pew Research Center**: Pew Research Center for the People and the Press. 235; **The American Statistician**:David R. Appleton et al. “Ignoring a Covariate: An Example of Simpson’s Paradox.” American Statistician 50(4), 1996. 236; **John Wiley & Sons, Inc.**: John Blume, Theodore Eisenberg, and Martin T. Wells. “Explaining Death Row’s Population and Racial Composition,” Journal of Empirical Legal Studies, 1(1), 165–207, March, 2004. 236; **UC Regents**: Justin Wolfers. “Point Shaving: Corruption in NCAA Basketball.” 238; **CoStar Group, Inc**: apartments.com 239; **Florida Museum of Natural History**: Florida Museum of Natural History. 240; **Bureau of Labor Statistics**: Bureau of Labor Statistics. 240; **Bureau of Labor Statistics**: Bureau of Labor Statistics.240; **Harvard University Press**: George W. Pierce. The Songs of Insects. Cambridge, MA: Harvard University Press, 1949, pp. 12–21. 242.

Chapter 5: Image Credits

Shutterstock: Ryzhkov Photography/Shutterstock 250; **Microsoft Corporation**: © Microsoft Corporation 253; **Pearson Education**: Pearson Education 257; **Pearson Education**: Pearson Education 259; **Pearson Education**: Pearson Education 260; **Shutterstock**: Zolnieriek/Shutterstock 262; **Shutterstock**: Zolnieriek/Shutterstock 279; **Microsoft Corporation**: © Microsoft Corporation 309; **Microsoft Corporation**: © Microsoft Corporation 297; **Microsoft Corporation**: © Microsoft Corporation 307; **Microsoft Corporation**: © Microsoft Corporation307; **Texas Instrument**: Texas Instrument 297; **Microsoft Corporation**: © Microsoft Corporation 299; **Texas Instrument**: Texas Instrument 299; **Microsoft Corporation**: © Microsoft Corporation 309; **Microsoft Corporation**: © Microsoft Corporation 309; **Microsoft Corporation**: © Microsoft Corporation 309; **Shutterstock** : Ryzhkov Photography/Shutterstock 322; **Shutterstock**: Ryzhkov Photography/Shutterstock 323; **Alamy Stock Photo**: SPUTNIK/Alamy Stock Photo 289; **Shutterstock**: ID1974/Shutterstock298; **123RF**: Cylonphoto/123RF Ltd. 302.

Chapter 5: Text Credits

Mars, Incorporated and its Affiliates: M&Ms 254; **ABC News Internet Ventures**: Fivethirtyeight.com 261; **Dow Jones & Company**: Wall Street Journal, September 24, 2014.261; **Chicago Tribune**: Chicago Tribune, June 14, 2015. 261; **Federal Bureau of Investigation**: U.S. Federal Bureau of Investigation. 262; **Centers for Disease**

Control and Prevention: National Vital Statistics Report 263; **Verizon Media:** Yahoo! Finance. 263; **Wakefield Research:** Based on results from Wakefield Research. 263; **Sigma Xi, The Scientific Research Honor Society:** The First Digit Phenomenon, T. P. Hill, American Scientist, July–August, 1998.267; **U.S. Census Bureau:** U.S. Census Bureau 270; **U.S. Census Bureau:** U.S. Census Bureau. 272; **SportMedBC:** <https://sportsmedbc.com/article/baseball-injuries273>; **Pew Research Center:** Pew Research. 274; **U.S. Census Bureau:** U.S. Census Bureau. 274; **Center for Medicare and Medicaid Services:** Center for Medicare and Medicaid Services. 274; **Oxford University Press:** “Shapiro, Jacobs, and Thun. ‘Cigar Smoking in Men and Risk of Death from Tobacco-Related Cancers,’ Journal of the National Cancer Institute, February 16, 2000.” 275; **The Harris Poll Hard Data:** Harris Interactive. 276; **National Highway Traffic Safety Administration:** Fatality Analysis Reporting System. 276; **SpringPublisher Team:** Krig, Moran, Regan. “An Analysis of a Red-Light Camera Program in the City of Milwaukee,” Spring 2006, prepared for the city of Milwaukee Budget and Management Division 276; **Dow Jones & Company:** Wall Street Journal, December 8, 2016. 282; **U.S. Department of Health and Human Services:** “U.S. Department of Health and Human Services, reported in USA Today.” 282; **U.S. Census Bureau:** U.S. Census Bureau, American Community Survey285; **American Community Survey:** American Community Survey, 2013 290; **The Harris Poll Hard Data:** The Harris Poll. 290; **The Harris Poll Hard Data:** Harris Interactive. 291; **National Highway Traffic Safety Administration:** Fatality Analysis Reporting System. 291; **Dow Jones & Company:** “The Sequence’ Is the Secret to Success.” Wendy Wang, Wall Street Journal, March 28, 2018. 293; **Centers for Disease Control and Prevention:** National Vital Statistics Report.319; **Sporting News Publishing Co:** Miklasz, B., et al. Celebrating 70: Mark McGwire’s Historic Season, Sporting News Publishing Co., 1998, p. 179. 320.

Chapter 6: Image Credits

123RF GB LIMITED: Zimmytw/123RF Ltd. 324; **Texas Instruments:** Texas Instruments 332; **Pearson Education:** Pearson Education 338; **SHUTTERSTOCK:** Shutterstock 339; **Texas Instruments:** Texas Instruments 345; **Texas Instruments:** Texas Instruments 345; **Microsoft Corporation:** © Microsoft Corporation 348.

Chapter 6: Text Credits

United States Census Bureau: Based on data from the U.S. Census Bureau 334; **Chicago Tribune:** Chicago Tribune 334; **MLB Advanced Media, LP:** Major League Baseball 334; **Gallup, Inc.:** Based on data from a Gallup poll 335; **U.S. Department of Health & Human Services:** National Vital Statistics Report. 335; **U.S. Department of Commerce.:** U.S. Department of Commerce. 335; **Taylor & Francis, Ltd.:** “Examining a Gambler’s Claims: Probabilistic Fact-Checking and Don Johnson’s Extraordinary Winning Streak” by W.J. Hurley, Jack Brimberg, and Richard Kohar. Chance Vol. 27.1, 2014. 335; **Wizard of Odds Consulting, Inc.:** The Wizard of Odds. 336; **Multi-State Lottery Association:** www.powerball.com 336; **Association for Computing Machinery, Inc.:** Passenger-Based Predictive Modeling of Airline No-show Rates by Richard D. Lawrence, Se June Hong, and Jacques Cherrier 351; **Northeastern University:** Racial Profiling Data Collection Resource Center., Published by Northeastern University. 351; **National Federation of State High School Associations.:** National Federation of State High School Associations. 351; **Prentice Hall:** Adapted from An Introduction to Mathematical Statistics by Larsen et al., Prentice Hall, Upper Saddle River, NJ, 2001 352; **InfoPlease:** Information Please Almanac 355; **Wizard of Odds Consulting, Inc.:** <http://wizardofodds.com/threecardpoker> 356; **Penguin Random House:** SuperFreakonomics. 356; **American Academy of Pediatrics:** Janis Whitlock, John Eckenrode, and Daniel Silverman. “Self-injurious Behaviors in a College Population,” Pediatrics 117: 1939–1948. 356; **IBM Corp:** www.wimbledon.org 357; **Picton Press:** Wilford W. Whitaker and Gary T. Horlacher, Broad Bay Pioneers (Rockport, Maine: Picton Press, 1998), 63–68. Distributions created from the partially reconstructed 1752 passenger list of the St. Andrew presented by Whitaker and Horlacher. 002.

Chapter 7: Image Credits

Shutterstock: Jennifer Griner/Shutterstock 365; **Alamy Stock Photo:** History and Art Collection/Alamy Stock Photo 369; **Pearson Education:** Pearson Education 370; **Pearson Education:** Pearson Education 372; **Texas Instrument:** Texas Instrument 379; **StatCrunch:** StatCrunch 381; **Microsoft Corporation:** © Microsoft Corporation 382; **Microsoft Corporation:** © Microsoft Corporation 382; **Texas Instrument:** Texas Instrument 383; **Texas Instrument:** Texas Instrument 395; **Pearson Education:** Pearson Education 393; **Shutterstock:** Jennifer Griner/Shutterstock 401;

Chapter 7: Text Credits

BioMed Central Ltd: Storem, Christian, et al. “Mild Therapeutic Hypothermia Shortens Intensive Care Unit Stay of Survivors After Out-of-Hospital Cardiac Arrest Compared to Historical Controls.” Critical Care 2008, 12:R78 BioMed Central. 375; **University of Illinois:** University of Illinois Extension 385; **Taylor & Francis, Ltd.:** Brad Warner and Jim Rutledge, Chance 12(1): 10–14, 1999. 385; **John Wiley & Sons, Inc.:** Tim Cole, “People Smugglers, Statistics, and Bone Age,” Significance Magazine, June 2012: 9(3). 385; **Taylor & Francis, Ltd.:** W. J. Hurley and Andrey Pavlov, “There Will Be Blood: On the Risk-Return Characteristics of a Blackjack Counting System,” Chance, 24:(2), Spring 2011.386; **UC Regents:** Justin Wolfers, “Point Shaving: Corruption in NCAA Basketball.”386; **UC Regents:** Justin Wolfers, “Point Shaving: Corruption in NCAA Basketball.”386; University of Illinois: University of Illinois Extension.386; **Taylor & Francis, Ltd.:** Brad Warner and Jim Rutledge, Chance 12(1): 10–14, 1999. 386; **Centers for Disease Control and Prevention:** National Vital Statistics Report 387; **Iowa Greyhound Park:** Greyhound Park, Dubuque, IA 389; **National Oceanic and Atmospheric Administration:** National Oceanic and Atmospheric Administration 392; **Taylor & Francis, Ltd.:** Chance 12(1): 10–14, 1999. 392; **Baseball Almanac, Inc.:** www.baseball-almanac.com399; **Institute of Mathematical Statistics:** S. M. Stigler. “Do Robust Estimators Work with Real Data?” Annals of Statistics 5(1977), 1055–1078 399; **National Center for Health Statistics:** M. A. McDowell, C. D. Fryar, R. Hirsch, and C. L. Ogden. Anthropometric Reference Data for Children and Adults: U. S. Population, 1999–2002. Advance data from vital and health statistics: No. 361. **Hyattsville, MD:** National Center for Health Statistics, 2005. 400

Chapter 8: Image Credits

SHUTTERSTOCK: Rawpixel.com/Shutterstock 404; **PEARSON:** Pearson Education 411; **SHUTTERSTOCK:** Zolnierenk/Shutterstock 417; **Microsoft Corporation:** © Microsoft Corporation 420; **SHUTTERSTOCK:** Rawpixel.com/Shutterstock 428;

Chapter 8: Text Credits

NORC at the University of Chicago: General Social Survey 409; **U.S. Department of Health and Human Service:** “Anthropometric Reference Data for Children and Adults: U.S. Population, 1999–2002”; Volume 361, July 7, 2005. 414; **Academy of Motion Picture Arts and Science:** awardsdatabase.oscars.org 416; **Internet Movie Database:** The Internet Movie Database. 425; Advanced Football Analytics 425; **American Dietetic Association:** American Dietetic Association. 426;

Chapter 9: Image Credits

123RF: Feverpitched/123RF Ltd. 430; Feverpitched/123RF Ltd. 471; **Texas Instrument :** Texas Instrument 438; **Minitab:** Minitab 452; **StatCrunch:** StatCrunch 467.

Chapter 9: Text Credits

AbbVie Inc. : <http://www.androgelpro.com/clinical-studies/default.aspx> 443; **Canadian Journal of Forest Research:** L. Winter. “Live Tree and Tree-Ring Records to Reconstruct the Structural Development of an Old-Growth Douglas Fir/Western Hemlock Stand in the Western Washington Cascades.” Corvallis, OR: Forest Science Data Bank, 2005. 469; **BrooksBaseball:** Brooksbaseball.net 463; **Cars.com:** Cars.com 465; **U.S. Department of Energy:** www.fueleconomy.gov 443; www.fueleconomy.gov 451; www.fueleconomy.gov 471; **John Wiley & Sons, Inc.:** Heid, Cory. “Tootsie Pops: How Many Licks to the Chocolate?” Significance, October, 2013 Volume 10 Issue 5 456; Seo, WH et al. “The Association between Periodontitis and Obstructive Sleep Apnea: A Preliminary Study”, J Periodontal Res 2013 Aug; 48 (4) 463; **National Atmospheric Deposition Program:** National Atmospheric Deposition Program 457; **TouringPlans.com:** touringplans.com 457; **Insurance Institute for Highway Safety:** Insurance Institute for Highway Safety 457; Insurance Institute for Highway Safety 465; **TD Ameritrade:** TD Ameritrade 457; **Essentialbaby:** www.essentialbaby.com 457; www.essentialbaby.com 465; **Society for the Study of Addiction:** “Happy Ending: A Randomized Controlled Trial of a Digital MultiMedia Smoking Cessation Intervention.” Havar Brendryen and Pal Kraft. Addiction 103(3):478–484, 2008 459; **PLoS:** Dixit VV, Chand S, Nair DJ (2016) “Autonomous Vehicles: Disengagements, Accidents, and Reaction Times,” PLoS ONE, 11(12):e0168054. <https://doi.org/10.1371/journal.pone.0168054> 463; **Open Data Houston:** Open Data Houston 464; **Massachusetts Medical Society:** Dariush Mozaffarian, M.D. et al. “Changes in Diet and Lifestyle and Long-Term Weight Gain in Women and Men,” The New England Journal of Medicine 364;25 464; **Human Kinetics Publishers Inc.:** Kirk J. Cureton, Gordon L. Warren, et al., “Caffeinated Sports Drink: Ergogenic Effects and Possible Mechanisms.” International Journal of Sport Nutrition and Exercise Metabolism 17(1):35–55, 2007 468; **U.S. Department of Health & Human Services:** The Centers for Disease Control 469; **U.S. Department of Education:** Based on data from The Toolbox Revisited by Clifford Adelman, U.S. Department of Education 469; **Virginia Piekarski:** Virginia Piekarski, Joliet Junior College 469; **Environmental Protection Agency;** Environmental Protection Agency 470; **Wimbledon:** www.wimbledon.org 470; **National Atmospheric Deposition Program:** National Atmospheric Deposition Program 465; National Atmospheric Deposition Program 475; **Zillow:** Zillow.com 475; **Sports City:** SportsCity 475.

Chapter 10: Image Credits

SHUTTERSTOCK: JohnKwan/Shutterstock. 484; **GETTY IMAGES INCORPORATED:** Ingram Publishing/Getty Images. 489; **StatCrunch:** StatCrunch. 501; **Texas Instrument:** Texas Instrument. 503; **Texas Instrument:** Texas Instrument. 511; **SHUTTERSTOCK:** JohnKwan/Shutterstock. 527;

Chapter 10: Text Credits

American Statistical Association: Letter by Ronald A Fisher in Nature. Copyright © by Nature Publishing Group, 493; **Public Agenda:** Based on “Reality Check: Are Parents and Students Ready for More Math and Science?” Public Agenda, 2006. 505; **Nature Research:** Pegna, Alan J. et al., “Discriminating Emotional Faces without Primary Visual Cortices Involves the Right Amygdala.” Nature Neuroscience, 8(1), 2005. 505; **Pearson Education:** Kirk, Honey and Lerma, Diane, “Reading Your Way to Success in Mathematics: A Paired Course of Developmental Mathematics and Reading.” MathAMATYC Educator, Vol. 1. No. 2, 2010. 505; **Gallup, Inc:** Gallup Organization, April 14, 2014. 506; **VegasInsider.com Inc:** <http://www.vegasinsider.com> . 506; **Sage Publications:** L.N. Moorthy, M.G.E. Peterson, K.B. Onel, and T.J.A. Lehman. “Do Children with Lupus Have Fewer Male Siblings?” Lupus 2008 17:128–131, 2008. 507; **Nature Research:** J. Kiley Hamlin et al., “Social Evaluation by Preverbal Infants.” Nature, Nov. 2007. 508; **Michael Carlisle:** Michael Carlisle, student at Joliet Junior College. 511; **BioMed Central Ltd:** Vendula Kyselova et al., “Effects of p-nonylphenol and resveratrol on body and organ weight and in vivo fertility of outbred CD-1 mice,” Reproductive Biology and Endocrinology, 2003. 516; **Stanford University:** Norman H. Nie and D. Sunshine Hillygus. “Where Does Internet Time Come From? A Reconnaissance.” IT & Society, 1(2). 516; **MLB Advanced Media, LP:** baseballsavant.mlb.com. 517; **Statcast:** Statcast. 517; **National Atmospheric Deposition Program:** National Atmospheric Deposition Program. 517; **United States National Academy of Sciences:** Anne-Marie Chang, et. al. “Evening Use of Light-Emitting eReaders Negatively Affects Sleep, Circadian Timing, and Next-Morning Alertness” PNAS 2015 112(4) 1232–1277; doi:10.1073/pnas.1418490112. 519; **U.S. Department of Energy:** www.fueleconomy.gov . 521 ; **American Student Assistance:** American Student Assistance. 522; **ESPN:** espn.com. 524; **Michael McCraith:** Michael McCraith, Joliet Junior College. 525; **Pearson Education:** MyMathLab. 525; **Dennis Johnson:** Dennis Johnson, student at Joliet Junior College. 527; **American Medical Association:** Christopher D. Gardner, Alexandre Kiazand, and Sofiya Alhassan, et al. “Comparison of the Atkins, Zone, Ornish, and LEARN Diets for Change in Weight and Related Risk Factors among Overweight Premenopausal Women. The A to Z Weight Loss Study: A Randomized Trial.” Journal of the American Medical Association, March 2007.527; **English Heritage:** This fictional account is based on information obtained from Archaeometry and Stonehenge (www.eng.h.gov.uk/stoneh). The means and standard deviations used throughout this case study were constructed by calculating the statistics from the midpoint of the calibrated date range supplied for each artefact. 538.

Chapter 11: Image Credits

Getty Images: Don Mason/Getty Images 585; **Getty Images:** Don Mason/Getty Images 597; **Texas Instrument:** Texas Instrument 546; **StatCrunch:** StatCrunch 556; **Texas Instrument:** Texas Instrument 557; **StatCrunch:** StatCrunch 567;

Chapter 11: Text Credits

Merck Sharp & Dohme Corp.: www.clarinex.com 550; **John Wiley & Sons, Inc.**: Fausey C.M., and Matlock, T. Can Grammar Win Elections? Political Psychology, no. s!Qi: 10.1111/j.1467-9221.2010.00802.x 551; **Professor Andy Neill**: Professor Andy Neill, Joliet Junior College 554; **Medknow Publications**: Gkini MA, Kouskoukis AE, Tripsianis G, Rigopoulos D, Kouskoukis K., "Study of Platelet-Rich Plasma Injections in the Treatment of Androgenetic Alopecia through a One-Year Period." J Cutan Aesthet Surg, 2014; 7:213–219. 559; **Informa UK Limited**: Ronald Christenson and Larry Blackwood, "Tests for Precision and Accuracy of Multiple Measuring Devices." Technometrics, 35(4):411–421, 1993.559; **Nature Research**: J. Kiley Hamlin et al., "Social Evaluation by Preverbal Infants," Nature, Nov. 2007. 559; **The University of Mississippi**: PsychExperiments at the University of Mississippi 559; **Anna Behounek**: Anna Behounek, student at Joliet Junior College. 560; **Insurance Institute for Highway Safety**: Insurance Institute for Highway Safety 560; Virginia Piekarski: Virginia Piekarski, Joliet Junior College 560; **Yahoo!Travel**: Yahoo!Travel 561; **City of Chicago**: City of Chicago Data Portal 561; **American Statistical Association**: Catherine Elizabeth Cavagnaro. "Glide Testing: A Paired Samples Experiment." Stats 46, Fall 2006. 561; **American Statistical Association**: Catherine Elizabeth Cavagnaro. "Glide Testing: A Paired Samples Experiment." Stats 46, Fall 2006. 562; **NASA**: NASA Life Sciences Data Archive 565; **Informa UK Limited**: Moser and Stevens, "Homogeneity of Variance in the Two-Sample Means Test." American Statistician 46(1). 569; **Elsevier Ltd.**: Stine A. Hogmboe et. al., "Influence of Marital Status on Testosterone Levels—A Ten Year Follow-Up of 1113 Men" Psychoneuroendocrinology 80(2017):155–161.571; **United States Department of Education**: Clifford Adelman, The Toolbox Revisited. United States Department of Education, 2006.571; **National Academy of Sciences**: Seth B. Young, "Evaluation of Pedestrian Walking Speeds in Airport Terminals." Transportation Research Record. Paper 99-0824.571; **Sage Publications**: Seth B. Young, "Evaluation of Pedestrian Walking Speeds in Airport Terminals." Transportation Research Record. Paper 99-0824.572; **The University of Mississippi**: PsychExperiments at the University of Mississippi 572; **American Psychological Association**: Based on the research of Dijksterhuis, Ap, and Ad van Knippenberg. "The relation between perception and behavior, or how to win a game of Trivial Pursuit." Journal of Personality and Social Psychology 74.4 (1998): 865+. Academic OneFile. Web. 5 July 2010. 572; **Elsevier Ltd.**: Wansink, B. Junyoung, K. "Bad Popcorn in Big Buckets: Portion Size Can Influence Intake as Much as Taste." Journal of Nutrition Education & Behavior, September 2005; 35(5):242–245. 572; **NEHA**: William G. Walter and Angie Stober. "Microbial Air Sampling in a Carpeted Hospital." Journal of Environmental Health, 30 (1968), p. 405. 572; **Informa UK Limited**: Mark Gellevij et al. "Multimodal versus Unimodal Instruction in a Complex Learning Context." Journal of Experimental Education 70(3):215–239, 2002. 573; **U.S. Bureau of Labor Statistics**: American Time Use Survey. 573; **Brian Ortiz**: Brian Ortiz, student at Joliet Junior College 577; **Elsevier Ltd.**: Data from de Lucas, R.D., Balildan, P., Neiva, C.M., Greco, C.C., & Denadai, B.S. (2000). The effects of wet suits on physiological and biomechanical indices during swimming. Journal of Science and Medicine in Sport 3 (1): 1–8. 577; **American Association for the Advancement of Science**: Todorov, Mandisodza, Goren, Hall, "Inferences of Competence from Faces Predict Election Outcomes." Science Vol. 308. 578; **Academic and Business Research Institute**: Ellis, Y., Daniels, W. and Jauregui, A. (2010) "The Effect of Multitasking on the Grade Performance of Business Students." Research in Higher Education Journal, 8. 578; **University of Wisconsin Press**: Christopher Cornwell, David B. Mustard, and Jessica Van Parys, "Noncognitive Skills and Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School," J. Human Resources, Winter 2013, 48(1):236–264. doi: 10.3386/jhr.48.1.236. 579; **BioMed Central Ltd**: Lee Park, Danielle Andrade, Andrew Mastey, James Sun, LeRoi Hicks "Institution Specific Risk Factors for 30 Day Readmission at a Community Hospital: A Retrospective Observational Study" BioMedCentral 2014 14:40. 579; **Yahoo!Finance**: Yahoo!Finance 579; **John Climent**: John Climent, Cecil Community College 582; **Catherine M. Simmons**: Catherine M. Simmons, student at Joliet Junior College 582; **Catherine M. Simmons**: Caponnetto P, Campagna D, Cibella F, Morjaria JB, Caruso M, et al. (2013) EffiCiency and Safety of an eLectronic cigAreTte (ECLAT) as Tobacco Cigarettes Substitute: A Prospective 12-Month Randomized Control Design Study. PLoS ONE 8(6): e66317. doi:10.1371/journal.pone.0066317. 583; **Insurance Institute for Highway Safety**: Insurance Institute for Highway Safety; National Atmospheric Deposition Program: National Atmospheric Deposition Program 584; **American Medical Association**: Stephanie O'Malley et al. "A Controlled Trial of Naltrexone Augmentation of Nicotine Replacement Therapy for Smoking Cessation." Archives of Internal Medicine, Vol. 166, 667–674. 585; **Yahoo!Finance**: Yahoo!Finance 580; **The University of Mississippi**: PsychExperiments at the University of Mississippi 584; **American Society of Nephrology**: Jacques Blacher et al., "Association between Plasma Homocysteine Concentrations and Cardiac Hypertrophy in End-Stage Renal Disease." Journal of Nephrology 12(4): 248–255, 1999. Article available at www.sin_italia.org/jnonline/vol12n4/blacher/blacher.htm.595;

Chapter 12: Image Credits

SHUTTERSTOCK: Spiroview Inc/Shutterstock 596; **GETTY IMAGES INCORPORATED**: Don Mason/ Getty Images 597; **StatCrunch**: StatCrunch 601; **Alamy Images**: History and Art Collection/Alamy Stock Photo 602; **Texas Instruments**: Texas Instruments 615; **StatCrunch**: StatCrunch 619; **SHUTTERSTOCK**: Spiroview Inc/Shutterstock 658; **Texas Instruments**: Texas Instruments 692; **Texas Instruments**: Texas Instruments 692; **Texas Instruments**: Texas Instruments 692; **StatCrunch**: StatCrunch 700.

Chapter 12: Text Credits

United States Census Bureau: U.S. Census Bureau 602; **Sigma Xi, The Scientific Research Honor Society**: T. P. Hill, "The First Digit Phenomenon," American Scientist, July–August, 1998. 609; **Wolfram Research, Inc.**: Eric W. Weisstein, Benford's Law, from MathWorld—A Wolfram Web Resource. 609; **Pew Research Center**: Congressional Quarterly Roll Call and Pew Research 609; **Village of Mundelein**: Village of Mundelein, Illinois 610; **QuanThockey.com**: QuanThockey.com 610; **Nature Research**: Gachter, Simon and Schulz, Jonathan, "Intrinsic Honesty and the Prevalence of Rule Violations Across Societies," Nature (24 March 2016) 531, 496–499. 610; **Informa UK Limited**: Lawrence Lesser. "Even More Fun Learning Statistics." Stats: The Magazine for Students of Statistics, Issue 49. 612; **Informa UK Limited**: Chance Magazine, Vol. 22, No. 4, 2009

612; **TrueCar, Inc.**: TrueCar.com 612; **United States National Academy of Sciences**: Adam L. Alter and Hal E. Hershfield, "People Search for Meaning When They Approach a New Decade in Chronological Age," Proceedings of the National Academy of Sciences, 111(48):17066–17070, 2014. 613; **Merck & Co.**: Merck and Co. 621; **NORC at the University of Chicago**: General Social Survey 625; **NORC at the University of Chicago**: General Social Survey 625; **Centers for Disease Control and Prevention**: National Health Interview Survey 626; **PLOS**: Caponnetto P, Campagna D, Cibella F, Morjaria JB, Caruso M, et al. (2013) EffiCency and Safety of an eLectronic cigAreTte (ECLAT) as Tobacco Cigarettes Substitute: A Prospective 12-Month Randomized Control Design Study. PLoS ONE 8(6): e66317 doi:10.1371/journal.pone.0066317 626; **Pfizer Inc.**: Pfizer, Inc. 626; **Pew Research Center**: Pew Research 626; **Pew Research Center**: Pew Research 627; **Sage Publications**: W., A., Watanabe-Rose, M., & Douglas, D. (2016). "Should Students Assessed as Needing Remedial Mathematics Take College- Level Quantitative Courses Instead? A Randomized Controlled Trial." Educational Evaluation and Policy Analysis, 38(3):578–598. <https://doi.org/10.3102/0162373716649056> 627; **Massachusetts Medical Society**: Paul M. Ridker et al. "A Randomized Trial of Low-Dose Aspirin in the Primary Prevention of Cardiovascular Disease in Women." New England Journal of Medicine 352:1293–1304. 627; **Pearson Education**: Kim, Minsu; Hebda, Beata; Graveman, Jerry; Hebda, Piotr. "Investigating the Corequisite Model for Remedial Mathematics Courses." MathAMATYC Educator, 9(3), Summer 2018. 628; **Gallup, Inc.**: The Gallup Organization. 639; **Fair Isaac Corporation**: www.myfico.com 639; **National Hurricane Center**: National Hurricane Center. 640; **Denise Slucki**: Denise Slucki, student at Joliet Junior College. 640; **Federal Trade Commission**: Federal Trade Commission. 641; **North Carolina Zoological Society**: fieldtripearth.org 641; **Dow Jones & Company**: The Wall Street Journal. 641; **Gold-Eagle Inc**: Gold-Eagle. 642; **National Center for Health Statistics**: Data based on information obtained from the National Center for Health Statistics. 642; **Zillow**: Zillow. com 643; **MLB Advanced Media, LP**: baseballsavant.mlb.com 649; **PLoS**: Kleisner K, Chvátalová V, Flegr J (2014) Perceived Intelligence Is Associated with Measured Intelligence in Men but Not Women. PLoS One 9(3): e81237 doi:10.1371/journal.pone.0081237 649; **MLB Advanced Media, LP**: Major League Baseball 652; **Newspapers.com**: Chicago Tribune, March 26, 2008 652; **Simon & Schuster**: Folsom, Burton, "New Deal or Raw Deal?" Simon & Schuster, 2008. 652; **National Science Teachers Association**: S. Kalinowski and Taper M. "The Effect of Seat Location on Exam Grades and Student Perceptions in an Introductory Biology Class." Journal of College Science Teaching, 36(4):54–57, 2007. 653; **Apartments.com**: apartments.com 653; **NORC at the University of Chicag**: General Social Survey 655; **NORC at the University of Chicag**: General Social Survey 655; **Harvard University Press**: George W. Pierce. The Songs of Insects, Cambridge, MA: Harvard University Press, 1949, pp. 12–21. 656; **National Center for Health Statistics**: National Center for Health Statistics. 656; **Kevin McCarthy**: Kevin McCarthy, student at Joliet Junior College. 657;

This page intentionally left blank

Appendix



Tables

Table I

Row Number	Random Numbers									
	Column Number									
01–05	06–10	11–15	16–20	21–25	26–30	31–35	36–40	41–45	46–50	
01	89392	23212	74483	36590	25956	36544	68518	40805	09980	00467
02	61458	17639	96252	95649	73727	33912	72896	66218	52341	97141
03	11452	74197	81962	48443	90360	26480	73231	37740	26628	44690
04	27575	04429	31308	02241	01698	19191	18948	78871	36030	23980
05	36829	59109	88976	46845	28329	47460	88944	08264	00843	84592
06	81902	93458	42161	26099	09419	89073	82849	09160	61845	40906
07	59761	55212	33360	68751	86737	79743	85262	31887	37879	17525
08	46827	25906	64708	20307	78423	15910	86548	08763	47050	18513
09	24040	66449	32353	83668	13874	86741	81312	54185	78824	00718
10	98144	96372	50277	15571	82261	66628	31457	00377	63423	55141
11	14228	17930	30118	00438	49666	65189	62869	31304	17117	71489
12	55366	51057	90065	14791	62426	02957	85518	28822	30588	32798
13	96101	30646	35526	90389	73634	79304	96635	06626	94683	16696
14	38152	55474	30153	26525	83647	31988	82182	98377	33802	80471
15	85007	18416	24661	95581	45868	15662	28906	36392	07617	50248
16	85544	15890	80011	18160	33468	84106	40603	01315	74664	20553
17	10446	20699	98370	17684	16932	80449	92654	02084	19985	59321
18	67237	45509	17638	65115	29757	80705	82686	48565	72612	61760
19	23026	89817	05403	82209	30573	47501	00135	33955	50250	72592
20	67411	58542	18678	46491	13219	84084	27783	34508	55158	78742

Table II
Critical Values for Correlation Coefficient

<i>n</i>	
3	0.997
4	0.950
5	0.878
6	0.811
7	0.754
8	0.707
9	0.666
10	0.632
11	0.602
12	0.576
13	0.553
14	0.532
15	0.514
16	0.497
17	0.482
18	0.468
19	0.456
20	0.444
21	0.433
22	0.423
23	0.413
24	0.404
25	0.396
26	0.388
27	0.381
28	0.374
29	0.367
30	0.361

Table III**Binomial Probability Distribution***p*

<i>n</i>	<i>x</i>	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
2	0	0.9801	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500	0.2025	0.1600	0.1225	0.0900	0.0625	0.0400	0.0225	0.0100	0.0025
	1	0.0198	0.0950	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000	0.4950	0.4800	0.4550	0.4200	0.3750	0.3200	0.2550	0.1800	0.0950
	2	0.0001	0.0025	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500	0.3025	0.3600	0.4225	0.4900	0.5625	0.6400	0.7225	0.8100	0.9025
3	0	0.9703	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250	0.0911	0.0640	0.0429	0.0270	0.0156	0.0080	0.0034	0.0010	0.0001
	1	0.0294	0.1354	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750	0.3341	0.2880	0.2389	0.1890	0.1406	0.0960	0.0574	0.0270	0.0071
	2	0.0003	0.0071	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750	0.4084	0.4320	0.4436	0.4410	0.4219	0.3840	0.3251	0.2430	0.1354
4	0	0.9606	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625	0.0410	0.0256	0.0150	0.0081	0.0039	0.0016	0.0005	0.0001	0.0000+
	1	0.0388	0.1715	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500	0.2005	0.1536	0.1115	0.0756	0.0469	0.0256	0.0115	0.0036	0.0005
	2	0.0006	0.0135	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750	0.3675	0.3456	0.3105	0.2646	0.2109	0.1536	0.0975	0.0486	0.0135
5	0	0.9510	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313	0.0185	0.0102	0.0053	0.0024	0.0010	0.0003	0.0001	0.0000+	0.0000+
	1	0.0480	0.2036	0.3281	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1563	0.1128	0.0768	0.0488	0.0284	0.0146	0.0064	0.0022	0.0005	0.0000+
	2	0.0010	0.0214	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125	0.2757	0.2304	0.1811	0.1323	0.0879	0.0512	0.0244	0.0081	0.0011
6	0	0.9415	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156	0.0083	0.0041	0.0018	0.0007	0.0002	0.0001	0.0000+	0.0000+	0.0000+
	1	0.0571	0.2321	0.3543	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938	0.0609	0.0369	0.0205	0.0102	0.0044	0.0015	0.0004	0.0001	0.0000+
	2	0.0014	0.0305	0.0984	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344	0.1861	0.1382	0.0951	0.0595	0.0330	0.0154	0.0055	0.0012	0.0001
7	0	0.9321	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078	0.0037	0.0016	0.0006	0.0002	0.0001	0.0000+	0.0000+	0.0000+	0.0000+
	1	0.0659	0.2573	0.3720	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547	0.0320	0.0172	0.0084	0.0036	0.0013	0.0004	0.0001	0.0000+	0.0000+
	2	0.0020	0.0406	0.1240	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641	0.1172	0.0774	0.0466	0.0250	0.0115	0.0043	0.0012	0.0002	0.0000+
8	0	0.9229	0.6583	0.4383	0.2806	0.1697	0.1035	0.0584	0.0350	0.0198	0.0092	0.0047	0.0023	0.0011	0.0005	0.0002	0.0001	0.0000+	0.0000+	0.0000+	0.0000+
	1	0.0757	0.2727	0.4083	0.4460	0.4115	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938	0.0609	0.0369	0.0205	0.0102	0.0044	0.0015	0.0004	0.0001	0.0000+
	2	0.0024	0.0416	0.1227	0.2083	0.2740	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641	0.1172	0.0774	0.0466	0.0250	0.0115	0.0043	0.0012	0.0002	0.0000+
9	0	0.9135	0.6183	0.3983	0.2406	0.1297	0.0735	0.0384	0.0198	0.0092	0.0047	0.0023	0.0011	0.0005	0.0002	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
	1	0.0845	0.2927	0.4383	0.4760	0.4415	0.3860	0.3325	0.2737	0.2166	0.1698	0.1250	0.0812	0.0472	0.0227	0.0125	0.0054	0.0016	0.0004	0.0001	0.0000+
	2	0.0025	0.0416	0.1227	0.2083	0.2740	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641	0.1172	0.0774	0.0466	0.0250	0.0115	0.0043	0.0012	0.0002	0.0000+
10	0	0.9043	0.5783	0.3583	0.1906	0.0797	0.0335	0.0184	0.0092	0.0047	0.0023	0.0011	0.0005	0.0002	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
	1	0.0957	0.3127	0.4383	0.4660	0.4315	0.3790	0.3255	0.2737	0.2166	0.1698	0.1250	0.0812	0.0472	0.0227	0.0125	0.0054	0.0016	0.0004	0.0001	0.0000+
	2	0.0025	0.0416	0.1227	0.2083	0.2740	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641	0.1172	0.0774	0.0466	0.0250	0.0115	0.0043	0.0012	0.0002	0.0000+
11	0	0.8951	0.5383	0.3183	0.1506	0.0597	0.0235	0.0114	0.0057	0.0028	0.0013	0.0006	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1057	0.3327	0.4383	0.4660	0.4315	0.3790	0.3255	0.2737	0.2166	0.1698	0.1250	0.0812	0.0472	0.0227	0.0125	0.0054	0.0016	0.0004	0.0001	0.0000+
	2	0.0025	0.0416	0.1227	0.2083	0.2740	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641	0.1172	0.0774	0.0466	0.0250	0.0115	0.0043	0.0012	0.0002	0.0000+
12	0	0.8859	0.4983	0.3783	0.1906	0.0797	0.0335	0.0184	0.0092	0.0047	0.0023	0.0011	0.0005	0.0002	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1157	0.3727	0.4383	0.4660	0.4315	0.3790	0.3255	0.2737	0.2166	0.1698	0.1250	0.0812	0.0472	0.0227	0.0125	0.0054	0.0016	0.0004	0.0001	0.0000+
	2	0.0025	0.0416	0.1227	0.2083	0.2740	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641	0.1172	0.0774	0.0466	0.0250	0.0115	0.0043	0.0012	0.0002	0.0000+
13	0	0.8767	0.4583	0.3383	0.1506	0.0597	0.0235	0.0114	0.0057	0.0028	0.0013	0.0006	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1263	0.3327	0.4383	0.4660	0.4315	0.3790	0.3255	0.2737	0.2166	0.1698	0.1250	0.0812	0.0472	0.0227	0.0125	0.0054	0.0016	0.0004	0.0001	0.0000+
	2	0.0025	0.0416	0.1227	0.2083	0.2740	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641	0.1172	0.0774	0.0466	0.0250	0.0115	0.0043	0.0012	0.0002	0.0000+
14	0	0.8675	0.4183	0.2983	0.1506	0.0597	0.0235	0.0114	0.0057	0.0028	0.0013	0.0006	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1361	0.3127	0.4383	0.4660	0.4315	0.3790	0.3255	0.2737	0.2166	0.1698	0.1250	0.0812	0.0472	0.0227	0.0125	0.0054	0.0016	0.0004	0.0001	0.0000+
	2	0.0025	0.0416	0.1227	0.2083	0.2740	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641	0.1172	0.0774	0.0466	0.0250	0.0115	0.0043	0.0012	0.0002	0.0000+
15	0	0.8583	0.3783	0.2583	0.1506	0.0597	0.0235	0.0114	0.0057	0.0028	0.0013	0.0006	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1457	0.2827	0.3720	0.4383	0.4660	0.4315	0.3790	0.3255	0.2737	0.2166	0.1698	0.1250	0.0812	0.0472	0.0227	0.0125	0.0054	0.0016	0.0004	0.0001
	2	0.0025	0.0416	0.1227	0.2083	0.2740	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641	0.1172	0.0774	0.0466	0.0250	0.0115	0.0043	0.0012	0.0002	0.0000+
16	0	0.8491	0.3383	0.1983	0.1506	0.0597	0.0235	0.0114	0.0057	0.0028	0.0013	0.0006	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	1	0.1543	0.2327	0.3127	0.3720	0.4383	0.4660	0.4315	0.3790	0.3255	0.2737	0.21									

Table III (continued)

		<i>p</i>																			
<i>n</i>	<i>x</i>	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
8	0	0.9227	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039	0.0017	0.0007	0.0002	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
	1	0.0746	0.2793	0.3826	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0313	0.0164	0.0079	0.0033	0.0012	0.0004	0.0001	0.0000+	0.0000+	0.0000+
	2	0.0026	0.0515	0.1488	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094	0.0703	0.0413	0.0217	0.0100	0.0038	0.0011	0.0002	0.0000+	0.0000+
	3	0.0001	0.0054	0.0331	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2188	0.1719	0.1239	0.0808	0.0467	0.0231	0.0092	0.0026	0.0004	0.0000+
	4	0.0000+	0.0004	0.0046	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734	0.2627	0.2322	0.1875	0.1361	0.0865	0.0459	0.0185	0.0046	0.0004
	5	0.0000+	0.0000+	0.0004	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2188	0.2568	0.2787	0.2786	0.2541	0.2076	0.1468	0.0839	0.0331	0.0054
	6	0.0000+	0.0000+	0.0000+	0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094	0.1569	0.2090	0.2587	0.2965	0.3115	0.2936	0.2376	0.1488	0.0515
	7	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0313	0.0548	0.0896	0.1373	0.1977	0.2670	0.3355	0.3847	0.3826	0.2793
9	8	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0002	0.0007	0.0017	0.0039	0.0084	0.0168	0.0319	0.0576	0.1001	0.1678	0.2725	0.4305	0.6634
	0	0.9135	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020	0.0008	0.0003	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
	1	0.0830	0.2985	0.3874	0.3679	0.3020	0.2253	0.1556	0.1004	0.0605	0.0339	0.0176	0.0083	0.0035	0.0013	0.0004	0.0001	0.0000+	0.0000+	0.0000+	0.0000+
	2	0.0034	0.0629	0.1722	0.2597	0.3020	0.3003	0.2668	0.2162	0.1612	0.1110	0.0703	0.0407	0.0212	0.0098	0.0039	0.0012	0.0003	0.0000+	0.0000+	0.0000+
	3	0.0001	0.0077	0.0446	0.1069	0.1762	0.2336	0.2668	0.2716	0.2508	0.2119	0.1641	0.1160	0.0743	0.0424	0.0210	0.0087	0.0028	0.0006	0.0001	0.0000+
	4	0.0000+	0.0006	0.0074	0.0283	0.0661	0.1168	0.1715	0.2194	0.2508	0.2600	0.2461	0.2128	0.1672	0.1181	0.0735	0.0389	0.0165	0.0050	0.0008	0.0000+
	5	0.0000+	0.0000+	0.0008	0.0050	0.0165	0.0389	0.0735	0.1181	0.1672	0.2128	0.2461	0.2600	0.2508	0.2194	0.1715	0.1168	0.0661	0.0283	0.0074	0.0006
	6	0.0000+	0.0000+	0.0001	0.0006	0.0028	0.0087	0.0210	0.0424	0.0743	0.1160	0.1641	0.2119	0.2508	0.2716	0.2668	0.2336	0.1762	0.1069	0.0446	0.0077
10	7	0.0000+	0.0000+	0.0000+	0.0000+	0.0003	0.0012	0.0039	0.0098	0.0212	0.0407	0.0703	0.1110	0.1612	0.2162	0.2668	0.3003	0.3020	0.2597	0.1722	0.0629
	8	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0004	0.0013	0.0035	0.0083	0.0176	0.0339	0.0605	0.1004	0.1556	0.2253	0.3020	0.3679	0.3874	0.2985
	9	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0003	0.0008	0.0020	0.0046	0.0101	0.0207	0.0404	0.0751	0.1342	0.2316	0.3874	0.6302	
	0	0.9044	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010	0.0003	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	1	0.0914	0.3151	0.3874	0.3474	0.2684	0.1877	0.1211	0.0725	0.0403	0.0207	0.0098	0.0042	0.0016	0.0005	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
	2	0.0042	0.0746	0.1937	0.2759	0.3020	0.2816	0.2335	0.1757	0.1209	0.0763	0.0439	0.0229	0.0106	0.0043	0.0014	0.0004	0.0001	0.0000+	0.0000+	0.0000+
	3	0.0001	0.0105	0.0574	0.1298	0.2013	0.2503	0.2668	0.2522	0.2150	0.1665	0.1172	0.0746	0.0425	0.0212	0.0090	0.0031	0.0008	0.0001	0.0000+	0.0000+
	4	0.0000+	0.0010	0.0112	0.0401	0.0881	0.1460	0.2001	0.2377	0.2508	0.2384	0.2051	0.1596	0.1115	0.0689	0.0368	0.0162	0.0055	0.0012	0.0001	0.0000+
10	5	0.0000+	0.0001	0.0015	0.0085	0.0264	0.0584	0.1029	0.1536	0.2007	0.2340	0.2461	0.2340	0.2007	0.1536	0.1029	0.0584	0.0264	0.0085	0.0015	0.0001
	6	0.0000+	0.0000+	0.0001	0.0012	0.0055	0.0162	0.0368	0.0689	0.1115	0.1596	0.2051	0.2384	0.2508	0.2377	0.2001	0.1460	0.0881	0.0401	0.0112	0.0010
	7	0.0000+	0.0000+	0.0000+	0.0001	0.0008	0.0031	0.0090	0.0212	0.0425	0.0746	0.1172	0.1665	0.2150	0.2522	0.2668	0.2503	0.2013	0.1298	0.0574	0.0105
	8	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0004	0.0014	0.0043	0.0106	0.0229	0.0439	0.0763	0.1209	0.1757	0.2335	0.2818	0.3020	0.2759	0.1937	0.0746
	9	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0005	0.0016	0.0042	0.0098	0.0207	0.0403	0.0725	0.1211	0.1877	0.2684	0.3474	0.3874	0.3151	
	10	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0003	0.0010	0.0025	0.0060	0.0135	0.0282	0.0563	0.1074	0.1969	0.3487	0.6302	

Note: 0.0000+ means the probability is 0.0000 rounded to four decimal places. However, the probability is *not* zero.

Table III (continued)

n	x	p																			
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
11	0	0.8953	0.5688	0.3138	0.1673	0.0859	0.0422	0.0198	0.0088	0.0036	0.0014	0.0005	0.0002	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	1	0.0995	0.3293	0.3835	0.3248	0.2362	0.1549	0.0932	0.0518	0.0266	0.0125	0.0054	0.0021	0.0007	0.0002	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	2	0.0050	0.0867	0.2131	0.2866	0.2953	0.2581	0.1998	0.1395	0.0887	0.0513	0.0269	0.0126	0.0052	0.0018	0.0005	0.0001	0.0000+	0.0000+	0.0000+	
	3	0.0002	0.0137	0.0710	0.1517	0.2215	0.2581	0.2568	0.2254	0.1774	0.1259	0.0806	0.0462	0.0234	0.0102	0.0037	0.0011	0.0002	0.0000+	0.0000+	
	4	0.0000+	0.0014	0.0158	0.0536	0.1107	0.1721	0.2201	0.2428	0.2365	0.2060	0.1611	0.1128	0.0701	0.0379	0.0173	0.0064	0.0017	0.0003	0.0000+	0.0000+
	5	0.0000+	0.0001	0.0025	0.0132	0.0388	0.0803	0.1321	0.1830	0.2207	0.2360	0.2256	0.1931	0.1471	0.0985	0.0566	0.0268	0.0097	0.0023	0.0003	0.0000+
	6	0.0000+	0.0000+	0.0003	0.0023	0.0097	0.0268	0.0566	0.0985	0.1471	0.1931	0.2256	0.2360	0.2207	0.1830	0.1321	0.0803	0.0388	0.0132	0.0025	0.0001
	7	0.0000+	0.0000+	0.0000+	0.0003	0.0017	0.0064	0.0173	0.0379	0.0701	0.1128	0.1611	0.2060	0.2365	0.2428	0.2201	0.1721	0.1107	0.0536	0.0158	0.0014
	8	0.0000+	0.0000+	0.0000+	0.0000+	0.0002	0.0011	0.0037	0.0102	0.0234	0.0462	0.0806	0.1259	0.1774	0.2254	0.2568	0.2581	0.2215	0.1517	0.0710	0.0137
	9	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0005	0.0018	0.0052	0.0126	0.0269	0.0513	0.0887	0.1395	0.1998	0.2581	0.2953	0.2866	0.2131	0.0867
	10	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0002	0.0007	0.0021	0.0054	0.0125	0.0266	0.0518	0.0932	0.1549	0.2362	0.3248	0.3835	0.3293
	11	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0002	0.0005	0.0014	0.0036	0.0088	0.0198	0.0422	0.0859	0.1673	0.3138	0.5688	
12	0	0.8864	0.5404	0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	1	0.1074	0.3413	0.3766	0.3012	0.2062	0.1267	0.0712	0.0368	0.0174	0.0075	0.0029	0.0010	0.0003	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	2	0.0060	0.0988	0.2301	0.2924	0.2835	0.2323	0.1678	0.1088	0.0639	0.0339	0.0161	0.0068	0.0025	0.0008	0.0002	0.0000+	0.0000+	0.0000+	0.0000+	
	3	0.0002	0.0173	0.0852	0.1720	0.2362	0.2581	0.2397	0.1954	0.1419	0.0923	0.0537	0.0277	0.0125	0.0048	0.0015	0.0004	0.0001	0.0000+	0.0000+	
	4	0.0000+	0.0021	0.0213	0.0683	0.1329	0.1936	0.2311	0.2367	0.2128	0.1700	0.1208	0.0762	0.0420	0.0199	0.0078	0.0024	0.0005	0.0001	0.0000+	
	5	0.0000+	0.0002	0.0038	0.0193	0.0532	0.1032	0.1585	0.2039	0.2270	0.2225	0.1934	0.1489	0.1009	0.0591	0.0291	0.0115	0.0033	0.0006	0.0000+	
	6	0.0000+	0.0000+	0.0005	0.0040	0.0155	0.0401	0.0792	0.1281	0.1766	0.2124	0.2256	0.2124	0.1766	0.1281	0.0792	0.0401	0.0155	0.0040	0.0005	
	7	0.0000+	0.0000+	0.0000+	0.0006	0.0033	0.0115	0.0291	0.0591	0.1009	0.1489	0.1934	0.2225	0.2270	0.2039	0.1585	0.1032	0.0532	0.0193	0.0038	0.0002
	8	0.0000+	0.0000+	0.0000+	0.0001	0.0005	0.0024	0.0078	0.0199	0.0420	0.0762	0.1208	0.1700	0.2128	0.2367	0.2311	0.1936	0.1329	0.0683	0.0213	0.0021
	9	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0004	0.0015	0.0048	0.0125	0.0277	0.0537	0.0923	0.1419	0.1954	0.2397	0.2581	0.2362	0.1720	0.0852	0.0173
	10	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0002	0.0008	0.0025	0.0068	0.0161	0.0339	0.0639	0.1088	0.1678	0.2323	0.2835	0.2924	0.2301	0.0988
	11	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0003	0.0010	0.0029	0.0075	0.0174	0.0368	0.0712	0.1267	0.2062	0.3012	0.3766	0.3413
	12	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0002	0.0008	0.0022	0.0057	0.0138	0.0317	0.0687	0.1422	0.2824	0.5404	

Note: 0.0000+ means the probability is 0.0000 rounded to four decimal places. However, the probability is *not* zero.

n	x	p																		
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	
15	0	0.8601	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	1	0.1303	0.3658	0.3432	0.2312	0.1319	0.0668	0.0305	0.0126	0.0047	0.0016	0.0005	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	2	0.0092	0.1348	0.2669	0.2856	0.2309	0.1559	0.0916	0.0476	0.0219	0.0090	0.0032	0.0010	0.0003	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	
	3	0.0004	0.0307	0.1285	0.2184	0.2501	0.2252	0.1700	0.1110	0.0634	0.0318	0.0139	0.0052	0.0016	0.0004	0.0001	0.0000+	0.0000+	0.0000+	
	4	0.0000+	0.0049	0.0428	0.1156	0.1876	0.2252	0.2186	0.1792	0.1268	0.0780	0.0417	0.0191	0.0074	0.0024	0.0006	0.0001	0.0000+	0.0000+	
	5	0.0000+	0.0006	0.0105	0.0449	0.1032	0.1651	0.2061	0.2123	0.1859	0.1404	0.0916	0.0515	0.0245	0.0096	0.0030	0.0007	0.0001	0.0000+	0.0000+
	6	0.0000+	0.0000+	0.0019	0.0132	0.0430	0.0917	0.1472	0.1906	0.2066	0.1914	0.1527	0.1048	0.0612	0.0298	0.0116	0.0034	0.0007	0.0001	0.0000+
	7	0.0000+	0.0000+	0.0003	0.0030	0.0138	0.0393	0.0811	0.1319	0.1771	0.2013	0.1964	0.1647	0.1181	0.0710	0.0348	0.0131	0.0035	0.0005	0.0000+
	8	0.0000+	0.0000+	0.0000	0.0005	0.0035	0.0131	0.0348	0.0710	0.1181	0.1647	0.1964	0.2013	0.1771	0.1319	0.0811	0.0393	0.0138	0.0030	0.0003
	9	0.0000+	0.0000+	0.0000	0.0001	0.0007	0.0034	0.0116	0.0298	0.0612	0.1048	0.1527	0.1914	0.2066	0.1906	0.1472	0.0917	0.0430	0.0132	0.0019
	10	0.0000+	0.0000+	0.0000	0.0000+	0.0001	0.0007	0.0030	0.0096	0.0245	0.0515	0.0916	0.1404	0.1859	0.2123	0.2061	0.1651	0.1032	0.0449	0.0105
	11	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0001	0.0006	0.0024	0.0074	0.0191	0.0417	0.0780	0.1268	0.1792	0.2186	0.2252	0.1876	0.1156	0.0428
	12	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0000+	0.0001	0.0004	0.0016	0.0052	0.0139	0.0318	0.0634	0.1110	0.1700	0.2252	0.2501	0.2184	0.1285
	13	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0003	0.0010	0.0032	0.0090	0.0219	0.0476	0.0916	0.1559	0.2309	0.2856	0.2669
	14	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0005	0.0016	0.0047	0.0126	0.0305	0.0668	0.1319	0.2312	0.3432
	15	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0005	0.0016	0.0047	0.0134	0.0352	0.0874	0.2059	0.4633
20	0	0.8179	0.3585	0.1216	0.0388	0.0115	0.0032	0.0008	0.0002	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	1	0.1652	0.3774	0.2702	0.1368	0.0576	0.0211	0.0068	0.0020	0.0005	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	2	0.0159	0.1887	0.2852	0.2293	0.1369	0.0669	0.0278	0.0100	0.0031	0.0008	0.0002	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	3	0.0010	0.0596	0.1901	0.2428	0.2054	0.1339	0.0716	0.0323	0.0123	0.0040	0.0011	0.0002	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	4	0.0000+	0.0133	0.0898	0.1821	0.2182	0.1897	0.1304	0.0738	0.0350	0.0139	0.0046	0.0013	0.0003	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	5	0.0000+	0.0022	0.0319	0.1028	0.1746	0.2023	0.1789	0.1272	0.0746	0.0365	0.0148	0.0049	0.0013	0.0003	0.0000+	0.0000+	0.0000+	0.0000+	
	6	0.0000+	0.0003	0.0089	0.0454	0.1091	0.1686	0.1916	0.1712	0.1244	0.0746	0.0370	0.0150	0.0049	0.0012	0.0002	0.0000+	0.0000+	0.0000+	
	7	0.0000+	0.0000+	0.0020	0.0160	0.0545	0.1124	0.1643	0.1844	0.1659	0.1221	0.0739	0.0366	0.0146	0.0045	0.0010	0.0002	0.0000+	0.0000+	
	8	0.0000+	0.0000+	0.0004	0.0046	0.0222	0.0609	0.1144	0.1614	0.1797	0.1623	0.1201	0.0727	0.0355	0.0136	0.0039	0.0008	0.0001	0.0000+	
	9	0.0000+	0.0000+	0.0001	0.0011	0.0074	0.0271	0.0654	0.1158	0.1597	0.1771	0.1602	0.1185	0.0710	0.0336	0.0120	0.0030	0.0005	0.0000+	
	10	0.0000+	0.0000+	0.0000	0.0002	0.0020	0.0099	0.0308	0.0686	0.1171	0.1593	0.1762	0.1593	0.1171	0.0686	0.0308	0.0099	0.0020	0.0002	
	11	0.0000+	0.0000+	0.0000	0.0000+	0.0005	0.0030	0.0120	0.0336	0.0710	0.1185	0.1602	0.1771	0.1597	0.1158	0.0654	0.0271	0.0074	0.0011	
	12	0.0000+	0.0000+	0.0000	0.0000+	0.0001	0.0008	0.0039	0.0136	0.0355	0.0727	0.1201	0.1623	0.1797	0.1614	0.1144	0.0609	0.0222	0.0046	
	13	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0002	0.0010	0.0045	0.0146	0.0366	0.0739	0.1221	0.1659	0.1844	0.1643	0.1124	0.0545	0.0160	
	14	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0000+	0.0002	0.0012	0.0049	0.0150	0.0370	0.0746	0.1244	0.1712	0.1916	0.1686	0.1091	0.0454	
	15	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0000+	0.0000+	0.0003	0.0013	0.0049	0.0148	0.0365	0.0746	0.1272	0.1789	0.2023	0.1746	0.1028	
	16	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0003	0.0013	0.0046	0.0139	0.0350	0.0738	0.1304	0.1897	0.2182	0.1821	
	17	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0002	0.0011	0.0040	0.0123	0.0323	0.0716	0.1339	0.2054	0.2428	0.1901	
	18	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0002	0.0008	0.0031	0.0100	0.0278	0.0669	0.1369	0.2293	0.2852	
	19	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0001	0.0005	0.0020	0.0068	0.0211	0.0576	0.1368	0.2702	0.3774	
	20	0.0000+	0.0000+	0.0000	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0002	0.0008	0.0032	0.0115	0.0388	0.1216	0.3585

Note: 0.0000+ means the probability is 0.0000 rounded to four decimal places. However, the probability is *not* zero.

Table IV**Cumulative Binomial Probability Distribution**

n	x	p																			
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
2	0	0.9801	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500	0.2025	0.1600	0.1225	0.0900	0.0625	0.0400	0.0225	0.0100	0.0025
	1	0.9999	0.9975	0.9900	0.9775	0.9600	0.9375	0.9100	0.8775	0.8400	0.7975	0.7500	0.6975	0.6400	0.5775	0.5100	0.4375	0.3600	0.2775	0.1900	0.0975
3	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	0	0.9703	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250	0.0911	0.0640	0.0429	0.0270	0.0156	0.0080	0.0034	0.0010	0.0001
3	1	0.9997	0.9928	0.9720	0.9393	0.8960	0.8438	0.7840	0.7183	0.6480	0.5748	0.5000	0.4253	0.3520	0.2818	0.2160	0.1563	0.1040	0.0608	0.0280	0.0073
	2	1.0000-	0.9999	0.9990	0.9966	0.9920	0.9844	0.9730	0.9571	0.9360	0.9089	0.8750	0.8336	0.7840	0.7254	0.6570	0.5781	0.4880	0.3859	0.2710	0.1426
3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	0	0.9606	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625	0.0410	0.0256	0.0150	0.0081	0.0039	0.0016	0.0005	0.0001	0.0000+
4	1	0.9994	0.9860	0.9477	0.8905	0.8192	0.7383	0.6517	0.5630	0.4752	0.3910	0.3125	0.2415	0.1792	0.1265	0.0837	0.0508	0.0272	0.0120	0.0037	0.0005
	2	1.0000-	0.9995	0.9963	0.9880	0.9728	0.9492	0.9163	0.8735	0.8208	0.7585	0.6875	0.6090	0.5248	0.4370	0.3483	0.2617	0.1808	0.1095	0.0523	0.0140
3	1	1.0000-	1.0000-	0.9999	0.9995	0.9984	0.9961	0.9919	0.9850	0.9744	0.9590	0.9375	0.9085	0.8704	0.8215	0.7599	0.6836	0.5904	0.4780	0.3439	0.1855
	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	0	0.9510	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313	0.0185	0.0102	0.0053	0.0024	0.0010	0.0003	0.0001	0.0000+	0.0000+
	1	0.9990	0.9774	0.9185	0.8352	0.7373	0.6328	0.5282	0.4284	0.3370	0.2562	0.1875	0.1312	0.0870	0.0540	0.0308	0.0156	0.0067	0.0022	0.0005	0.0000+
2	1	1.0000-	0.9988	0.9914	0.9734	0.9421	0.8965	0.8369	0.7648	0.6826	0.5931	0.5000	0.4069	0.3174	0.2352	0.1631	0.1035	0.0579	0.0266	0.0086	0.0012
	3	1.0000-	1.0000-	0.9995	0.9978	0.9933	0.9844	0.9692	0.9460	0.9130	0.8688	0.8125	0.7438	0.6630	0.5716	0.4718	0.3672	0.2627	0.1648	0.0815	0.0226
4	1	1.0000-	1.0000-	1.0000-	0.9999	0.9997	0.9990	0.9976	0.9947	0.9898	0.9815	0.9688	0.9497	0.9222	0.8840	0.8319	0.7627	0.6723	0.5563	0.4095	0.2262
	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	0	0.9415	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156	0.0083	0.0041	0.0018	0.0007	0.0002	0.0001	0.0000+	0.0000+	0.0000+
	1	0.9985	0.9672	0.8857	0.7765	0.6554	0.5339	0.4202	0.3191	0.2333	0.1636	0.1094	0.0692	0.0410	0.0223	0.0109	0.0046	0.0016	0.0004	0.0001	0.0000+
2	1	1.0000-	0.9978	0.9842	0.9527	0.9011	0.8306	0.7443	0.6471	0.5443	0.4415	0.3438	0.2553	0.1792	0.1174	0.0705	0.0376	0.0170	0.0059	0.0013	0.0001
	3	1.0000-	0.9999	0.9987	0.9941	0.9830	0.9624	0.9295	0.8826	0.8208	0.7447	0.6563	0.5585	0.4557	0.3529	0.2557	0.1694	0.0989	0.0473	0.0159	0.0022
4	1	1.0000-	1.0000-	0.9999	0.9996	0.9984	0.9954	0.9891	0.9777	0.9590	0.9308	0.8906	0.8364	0.7667	0.6809	0.5798	0.4661	0.3446	0.2235	0.1143	0.0328
	5	1.0000-	1.0000-	1.0000-	0.9999	0.9999	0.9998	0.9993	0.9982	0.9959	0.9917	0.9844	0.9723	0.9533	0.9246	0.8824	0.8220	0.7379	0.6229	0.4686	0.2649
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	7	0	0.9321	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078	0.0037	0.0016	0.0006	0.0002	0.0001	0.0000+	0.0000+	0.0000+
7	1	0.9980	0.9556	0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625	0.0357	0.0188	0.0090	0.0038	0.0013	0.0004	0.0001	0.0000+	0.0000+
	2	1.0000-	0.9962	0.9743	0.9262	0.8520	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266	0.1529	0.0963	0.0556	0.0288	0.0129	0.0047	0.0012	0.0002	0.0000+
3	1	1.0000-	0.9998	0.9973	0.9879	0.9667	0.9294	0.8740	0.8002	0.7102	0.6083	0.5000	0.3917	0.2898	0.1998	0.1260	0.0706	0.0333	0.0121	0.0027	0.0002
	4	1.0000-	1.0000-	0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734	0.6836	0.5801	0.4677	0.3529	0.2436	0.1480	0.0738	0.0257	0.0038
5	1	1.0000-	1.0000-	1.0000-	0.9999	0.9996	0.9987	0.9962	0.9910	0.9812	0.9643	0.9375	0.8976	0.8414	0.7662	0.6706	0.5551	0.4233	0.2834	0.1497	0.0444
	6	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9998	0.9994	0.9984	0.9963	0.9922	0.9848	0.9720	0.9510	0.9176	0.8665	0.7903	0.6794	0.5217	0.3017	
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note: 0.0000+ means the probability is 0.0000 rounded to four decimal places. However, the probability is *not* zero.

1.0000- means the probability is 1.0000 rounded to four decimal places. However, the probability is *not* one.

This table computes the cumulative probability of obtaining x successes in n trials of a binomial experiment with probability of success p .

n	x	p																			
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
8	0	0.9227	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039	0.0017	0.0007	0.0002	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
	1	0.9973	0.9428	0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.0632	0.0352	0.0181	0.0085	0.0036	0.0013	0.0004	0.0001	0.0000+	0.0000+	0.0000+
	2	0.9999	0.9942	0.9619	0.8948	0.7969	0.6785	0.5518	0.4278	0.3154	0.2201	0.1445	0.0885	0.0498	0.0253	0.0113	0.0042	0.0012	0.0002	0.0000+	0.0000+
	3	1.0000-	0.9996	0.9950	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.4770	0.3633	0.2604	0.1737	0.1061	0.0580	0.0273	0.0104	0.0029	0.0004	0.0000+
	4	1.0000-	1.0000-	0.9996	0.9971	0.9896	0.9727	0.9420	0.8389	0.8263	0.7396	0.6367	0.5230	0.4059	0.2936	0.1941	0.1138	0.0563	0.0214	0.0050	0.0004
	5	1.0000-	1.0000-	1.0000-	0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555	0.7799	0.6846	0.5722	0.4482	0.3215	0.2031	0.1052	0.0381	0.0058
	6	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648	0.9368	0.8936	0.8309	0.7447	0.6329	0.4967	0.3428	0.1869	0.0572
	7	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9998	0.9993	0.9983	0.9961	0.9916	0.9832	0.9681	0.9424	0.8999	0.8322	0.7275	0.5695
	8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	9	0	0.9135	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020	0.0008	0.0003	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
9	1	0.9966	0.9288	0.7748	0.5995	0.4362	0.3003	0.1960	0.1211	0.0705	0.0385	0.0195	0.0091	0.0038	0.0014	0.0004	0.0001	0.0000+	0.0000+	0.0000+	0.0000+
	2	0.9999	0.9916	0.9470	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.0898	0.0498	0.0250	0.0112	0.0043	0.0013	0.0003	0.0000+	0.0000+	0.0000+
	3	1.0000-	0.9994	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539	0.1658	0.0994	0.0536	0.0253	0.0100	0.0031	0.0006	0.0001	0.0000+
	4	1.0000-	1.0000-	0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5000	0.3786	0.2666	0.1717	0.0988	0.0489	0.0196	0.0056	0.0009	0.0000+
	5	1.0000-	1.0000-	0.9999	0.9994	0.9969	0.9900	0.9747	0.9464	0.9006	0.8342	0.7461	0.6386	0.5174	0.3911	0.2703	0.1657	0.0856	0.0339	0.0083	0.0006
	6	1.0000-	1.0000-	1.0000-	1.0000-	0.9997	0.9987	0.9957	0.9888	0.9750	0.9502	0.9102	0.8505	0.7682	0.6627	0.5372	0.3993	0.2618	0.1409	0.0530	0.0084
	7	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805	0.9615	0.9295	0.8789	0.8040	0.6997	0.5638	0.4005
	8	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9997	0.9992	0.9980	0.9954	0.9899	0.9793	0.9596	0.9249	0.8658	0.7684	0.6126	0.3698
	9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	10	0	0.9044	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010	0.0003	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
10	1	0.9957	0.9139	0.7361	0.5443	0.3758	0.2440	0.1493	0.0860	0.0464	0.0233	0.0107	0.0045	0.0017	0.0005	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
	2	0.9999	0.9885	0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996	0.0547	0.0274	0.0123	0.0048	0.0016	0.0004	0.0001	0.0000+	0.0000+	0.0000+
	3	1.0000-	0.9990	0.9872	0.9500	0.8791	0.7759	0.6496	0.5138	0.3823	0.2660	0.1719	0.1020	0.0548	0.0260	0.0106	0.0035	0.0009	0.0001	0.0000+	0.0000+
	4	1.0000-	0.9999	0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.3770	0.2616	0.1662	0.0949	0.0473	0.0197	0.0064	0.0014	0.0001	0.0000+
	5	1.0000-	1.0000-	0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.6230	0.4956	0.3669	0.2485	0.1503	0.0781	0.0328	0.0099	0.0016	0.0001
	6	1.0000-	1.0000-	1.0000-	0.9999	0.9991	0.9965	0.9894	0.9740	0.9452	0.8980	0.8281	0.7340	0.6177	0.4862	0.3504	0.2241	0.1209	0.0500	0.0128	0.0010
	7	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453	0.9004	0.8327	0.7384	0.6172	0.4744	0.3222	0.1798	0.0702	0.0115
	8	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9995	0.9983	0.9955	0.9893	0.9767	0.9536	0.9140	0.8507	0.7560	0.6242	0.4557	0.2639
	9	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9997	0.9990	0.9975	0.9940	0.9865	0.9718	0.9437	0.8926	0.8031	0.6513	0.4013
	10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note: 0.0000+ means the probability is 0.0000 rounded to four decimal places. However, the probability is *not* zero.

1.0000- means the probability is 1.0000 rounded to four decimal places. However, the probability is *not* one.

Table IV (continued)

n	x	p																	
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
11	0	0.8953	0.5688	0.3138	0.1673	0.0859	0.0422	0.0198	0.0088	0.0036	0.0014	0.0005	0.0002	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
	1	0.9948	0.8981	0.6974	0.4922	0.3221	0.1971	0.1130	0.0606	0.0302	0.0139	0.0059	0.0022	0.0007	0.0002	0.0000+	0.0000+	0.0000+	0.0000+
	2	0.9998	0.9848	0.9104	0.7788	0.6174	0.4552	0.3127	0.2001	0.1189	0.0652	0.0327	0.0148	0.0059	0.0020	0.0006	0.0001	0.0000+	0.0000+
	3	1.0000-	0.9984	0.9815	0.9306	0.8389	0.7133	0.5696	0.4256	0.2963	0.1911	0.1133	0.0610	0.0293	0.0122	0.0043	0.0012	0.0002	0.0000+
	4	1.0000-	0.9999	0.9972	0.9841	0.9496	0.8854	0.7897	0.6683	0.5326	0.3971	0.2744	0.1738	0.0994	0.0501	0.0216	0.0076	0.0020	0.0003
	5	1.0000-	1.0000-	0.9997	0.9973	0.9883	0.9657	0.9218	0.8513	0.7535	0.6331	0.5000	0.3669	0.2465	0.1487	0.0782	0.0343	0.0117	0.0027
	6	1.0000-	1.0000-	1.0000-	0.9997	0.9980	0.9924	0.9784	0.9499	0.9006	0.8262	0.7256	0.6029	0.4672	0.3317	0.2103	0.1146	0.0504	0.0159
	7	1.0000-	1.0000-	1.0000-	1.0000-	0.9998	0.9988	0.9957	0.9878	0.9707	0.9390	0.8867	0.8089	0.7037	0.5744	0.4304	0.2867	0.1611	0.0694
	8	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9994	0.9980	0.9941	0.9852	0.9673	0.9348	0.8811	0.7999	0.6873	0.5448	0.3826	0.2212
	9	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9998	0.9993	0.9976	0.9941	0.9861	0.9698	0.9394	0.8870	0.8029	0.6779	0.5078
10	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9996	0.9995	0.9986	0.9964	0.9912	0.9802	0.9578	0.9141	0.8327
	11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	12	0	0.8864	0.5404	0.2824	0.1422	0.0687	0.0317	0.0138	0.0057	0.0022	0.0008	0.0002	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
	1	0.9938	0.8816	0.6590	0.4435	0.2749	0.1584	0.0850	0.0424	0.0196	0.0083	0.0032	0.0011	0.0003	0.0001	0.0000+	0.0000+	0.0000+	0.0000+
	2	0.9998	0.9804	0.8891	0.7358	0.5583	0.3907	0.2528	0.1513	0.0834	0.0421	0.0193	0.0079	0.0028	0.0008	0.0002	0.0000+	0.0000+	0.0000+
	3	1.0000-	0.9978	0.9744	0.9078	0.7946	0.6488	0.4925	0.3467	0.2253	0.1345	0.0730	0.0356	0.0153	0.0056	0.0017	0.0004	0.0001	0.0000+
	4	1.0000-	0.9998	0.9957	0.9761	0.9274	0.8424	0.7237	0.5833	0.4382	0.3044	0.1938	0.1117	0.0573	0.0255	0.0095	0.0028	0.0006	0.0001
	5	1.0000-	1.0000-	0.9995	0.9954	0.9806	0.9456	0.8822	0.7873	0.6652	0.5269	0.3872	0.2607	0.1582	0.0846	0.0386	0.0143	0.0039	0.0007
	6	1.0000-	1.0000-	0.9999	0.9993	0.9961	0.9857	0.9614	0.9154	0.8418	0.7393	0.6128	0.4731	0.3348	0.2127	0.1178	0.0544	0.0194	0.0046
	7	1.0000-	1.0000-	1.0000-	0.9999	0.9994	0.9972	0.9905	0.9745	0.9427	0.8883	0.8062	0.6956	0.5618	0.4167	0.2763	0.1576	0.0726	0.0239
	8	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9996	0.9983	0.9944	0.9847	0.9644	0.9270	0.8655	0.7747	0.6533	0.5075	0.3512	0.2054	0.0922
	9	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9998	0.9992	0.9972	0.9921	0.9807	0.9579	0.9166	0.8487	0.7472	0.6093	0.4417
10	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9997	0.9989	0.9968	0.9917	0.9804	0.9576	0.9150	0.8416	0.7251	0.5565
	11	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9998	0.9992	0.9978	0.9943	0.9862	0.9683	0.9313	0.8578
12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note: 0.0000+ means the probability is 0.0000 rounded to four decimal places. However, the probability is *not* zero.

1.0000- means the probability is 1.0000 rounded to four decimal places. However, the probability is *not* one.

Table IV (continued)

n	x	p																						
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95			
15	0	0.8601	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0016	0.0005	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+			
	1	0.9904	0.8290	0.5490	0.3186	0.1671	0.0802	0.0353	0.0142	0.0052	0.0017	0.0005	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+			
	2	0.9996	0.9638	0.8159	0.6042	0.3980	0.2361	0.1268	0.0617	0.0271	0.0107	0.0037	0.0011	0.0003	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+			
	3	1.0000-	0.9945	0.9444	0.8227	0.6482	0.4613	0.2969	0.1727	0.0905	0.0424	0.0176	0.0063	0.0019	0.0005	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+		
	4	1.0000-	0.9994	0.9873	0.9383	0.8358	0.6865	0.5155	0.3519	0.2173	0.1204	0.0592	0.0255	0.0093	0.0028	0.0007	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	5	1.0000-	0.9999	0.9978	0.9832	0.9389	0.8518	0.7216	0.5643	0.4032	0.2608	0.1509	0.0769	0.0338	0.0124	0.0037	0.0008	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+
	6	1.0000-	1.0000-	0.9997	0.9964	0.9819	0.9434	0.8689	0.7548	0.6098	0.4522	0.3036	0.1818	0.0950	0.0422	0.0152	0.0042	0.0008	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	
	7	1.0000-	1.0000-	1.0000-	0.9994	0.9958	0.9827	0.9500	0.8868	0.7869	0.6535	0.5000	0.3465	0.2131	0.1132	0.0500	0.0173	0.0042	0.0006	0.0000+	0.0000+	0.0000+	0.0000+	
	8	1.0000-	1.0000-	1.0000-	0.9999	0.9992	0.9958	0.9848	0.9578	0.9050	0.8182	0.6964	0.5478	0.3902	0.2452	0.1311	0.0566	0.0181	0.0036	0.0003	0.0000+	0.0000+	0.0000+	
	9	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9992	0.9963	0.9876	0.9662	0.9231	0.8491	0.7392	0.5968	0.4357	0.2784	0.1484	0.0611	0.0168	0.0022	0.0001	0.0000+	0.0000+	
	10	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9993	0.9972	0.9907	0.9745	0.9408	0.8796	0.7827	0.6481	0.4845	0.3135	0.1642	0.0617	0.0127	0.0006	0.0000+	0.0000+	
	11	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9995	0.9981	0.9937	0.9824	0.9576	0.9095	0.8273	0.7031	0.5387	0.3518	0.1773	0.0556	0.0055	0.0000+	0.0000+	
	12	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9997	0.9989	0.9963	0.9893	0.9729	0.9383	0.8732	0.7639	0.6020	0.3958	0.1841	0.0362	0.0000+	0.0000+	
	13	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9995	0.9983	0.9948	0.9858	0.9647	0.9198	0.8329	0.6814	0.4510	0.1710	0.0000+	0.0000+	
	14	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9995	0.9984	0.9953	0.9866	0.9648	0.9126	0.7941	0.5367	0.0000+	0.0000+	
	15	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.0000+	0.0000+	
20	0	0.8179	0.3585	0.1216	0.0388	0.0115	0.0032	0.0008	0.0002	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	1	0.9831	0.7358	0.3917	0.1756	0.0692	0.0243	0.0076	0.0021	0.0005	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	2	0.9990	0.9245	0.6769	0.4049	0.2061	0.0913	0.0355	0.0121	0.0036	0.0009	0.0002	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	3	1.0000-	0.9841	0.8670	0.6477	0.4114	0.2252	0.1071	0.0444	0.0160	0.0049	0.0013	0.0003	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	4	1.0000-	0.9974	0.9568	0.8298	0.6296	0.4148	0.2375	0.1182	0.0510	0.0189	0.0059	0.0015	0.0003	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	5	1.0000-	0.9997	0.9887	0.9327	0.8042	0.6172	0.4164	0.2454	0.1256	0.0553	0.0207	0.0064	0.0016	0.0003	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	6	1.0000-	1.0000-	0.9976	0.9781	0.9133	0.7858	0.6080	0.4166	0.2500	0.1299	0.0577	0.0214	0.0065	0.0015	0.0003	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	7	1.0000-	1.0000-	0.9996	0.9941	0.9679	0.8982	0.7723	0.6010	0.4159	0.2520	0.1316	0.0580	0.0210	0.0060	0.0013	0.0002	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	8	1.0000-	1.0000-	0.9999	0.9987	0.9900	0.9591	0.8867	0.7624	0.5956	0.4143	0.2517	0.1308	0.0565	0.0196	0.0051	0.0009	0.0001	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	9	1.0000-	1.0000-	1.0000-	0.9998	0.9974	0.9861	0.9520	0.8782	0.7553	0.5914	0.4119	0.2493	0.1275	0.0532	0.0171	0.0039	0.0006	0.0000+	0.0000+	0.0000+	0.0000+	0.0000+	
	10	1.0000-	1.0000-	1.0000-	1.0000-	0.9994	0.9961	0.9829	0.9468	0.8725	0.7507	0.5881	0.4086	0.2447	0.1218	0.0480	0.0139	0.0026	0.0002	0.0000+	0.0000+	0.0000+	0.0000+	
	11	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9991	0.9949	0.9804	0.9435	0.8692	0.7483	0.5857	0.4044	0.2376	0.1133	0.0409	0.0100	0.0013	0.0001	0.0000+	0.0000+	0.0000+	
	12	1.0000-	1.0000-	1.0000-	1.0000-	0.9998	0.9987	0.9940	0.9790	0.9420	0.8684	0.7480	0.5841	0.3990	0.2277	0.1018	0.0321	0.0059	0.0004	0.0000+	0.0000+	0.0000+	0.0000+	
	13	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9997	0.9985	0.9935	0.9786	0.9423	0.8701	0.7500	0.5834	0.3920	0.2142	0.0867	0.0219	0.0024	0.0000+	0.0000+	0.0000+	
	14	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9997	0.9984	0.9936	0.9793	0.9447	0.8744	0.7546	0.5836	0.3828	0.1958	0.0673	0.0113	0.0003	0.0000+	0.0000+	
	15	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9997	0.9985	0.9941	0.9811	0.9490	0.8818	0.7625	0.5852	0.3704	0.1702	0.0432	0.0026	0.0000+	0.0000+	
	16	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9997	0.9987	0.9951	0.9840	0.9556	0.8929	0.7748	0.5886	0.3523	0.1330	0.0159	0.0000+	0.0000+	0.0000+	
	17	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9998	0.9991	0.9964	0.9879	0.9645	0.9087	0.7939	0.5951	0.3231	0.0755	0.0000+	0.0000+	0.0000+	0.0000+	
	18	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	0.9999	0.9995	0.9979	0.9924	0.9757	0.9308	0.8244	0.6083	0.2642	0.0000+	0.0000+	0.0000+
	19	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	1.0000-	
	20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Note: 0.0000+ means the probability is 0.0000 rounded to four decimal places. However, the probability is *not* zero.

1.0000- means the probability is 1.0000 rounded to four decimal places. However, the probability is *not* one.

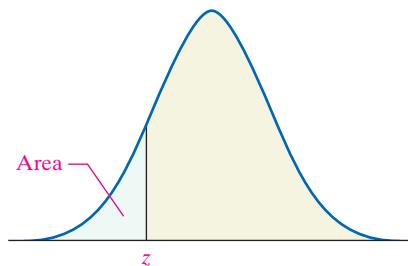


Table V

z	Standard Normal Distribution									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-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

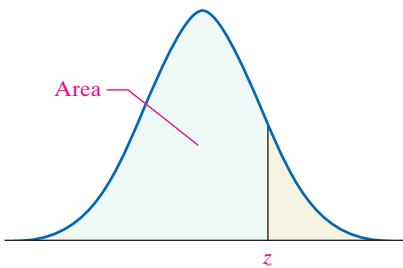


Table V (continued)

z	Standard Normal Distribution									
	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

(b) Confidence Interval Critical Values, $z_{\alpha/2}$

Level of Confidence	Critical Value, $z_{\alpha/2}$
0.90 or 90%	1.645
0.95 or 95%	1.96
0.98 or 98%	2.33
0.99 or 99%	2.575

(c) Hypothesis Testing Critical Values

Level of Significance, α	Left-Tailed	Right-Tailed	Two-Tailed
0.10	-1.28	1.28	± 1.645
0.05	-1.645	1.645	± 1.96
0.01	-2.33	2.33	± 2.575

Table VI

Sample Size, <i>n</i>	Critical Value
5	0.880
6	0.888
7	0.898
8	0.906
9	0.912
10	0.918
11	0.923
12	0.928
13	0.932
14	0.935
15	0.939
16	0.941
17	0.944
18	0.946
19	0.949
20	0.951
21	0.952
22	0.954
23	0.956
24	0.957
25	0.959
30	0.960

Source: S. W. Looney and T. R. Gullledge, Jr.
“Use of the Correlation Coefficient with
Normal Probability Plots,” *American
Statistician* 39(Feb. 1985): 75–79.

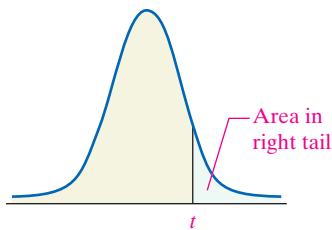
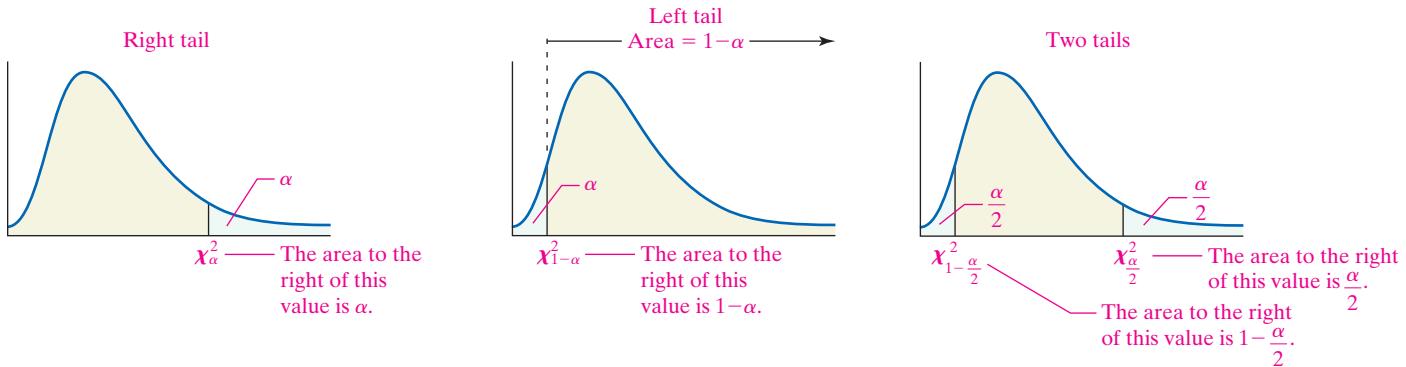


Table VII

Degrees of Freedom	t-Distribution Area in Right Tail											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.894	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
31	0.682	0.853	1.054	1.309	1.696	2.040	2.144	2.453	2.744	3.022	3.375	3.633
32	0.682	0.853	1.054	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.365	3.622
33	0.682	0.853	1.053	1.308	1.692	2.035	2.138	2.445	2.733	3.008	3.356	3.611
34	0.682	0.852	1.052	1.307	1.691	2.032	2.136	2.441	2.728	3.002	3.348	3.601
35	0.682	0.852	1.052	1.306	1.690	2.030	2.133	2.438	2.724	2.996	3.340	3.591
36	0.681	0.852	1.052	1.306	1.688	2.028	2.131	2.434	2.719	2.990	3.333	3.582
37	0.681	0.851	1.051	1.305	1.687	2.026	2.129	2.431	2.715	2.985	3.326	3.574
38	0.681	0.851	1.051	1.304	1.686	2.024	2.127	2.429	2.712	2.980	3.319	3.566
39	0.681	0.851	1.050	1.304	1.685	2.023	2.125	2.426	2.708	2.976	3.313	3.558
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
70	0.678	0.847	1.044	1.294	1.667	1.994	2.093	2.381	2.648	2.899	3.211	3.435
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
90	0.677	0.846	1.042	1.291	1.662	1.987	2.084	2.368	2.632	2.878	3.183	3.402
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
<i>z</i>	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.291

Table VIII

Degrees of Freedom	Chi-Square (χ^2) Distribution Area to the Right of Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	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
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169



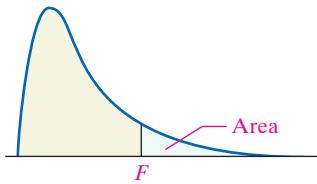


Table IX

		F-Distribution Critical Values							
		Degrees of Freedom in the Numerator							
Area in Right Tail		1	2	3	4	5	6	7	8
1	0.100	39.86	49.59	53.59	55.83	57.24	58.20	58.91	59.44
	0.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88
	0.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66
	0.010	4052.20	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07
	0.001	405284.07	499999.50	540379.20	562499.58	576404.56	585937.11	592873.29	598144.16
2	0.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37
	0.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37
	0.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37
	0.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37
3	0.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25
	0.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54
	0.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49
	0.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62
4	0.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95
	0.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98
	0.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80
	0.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00
5	0.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34
	0.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76
	0.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29
	0.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65
6	0.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98
	0.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60
	0.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10
	0.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03
7	0.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75
	0.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73
	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90
	0.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84
	0.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63
8	0.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59
	0.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43
	0.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03
	0.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05

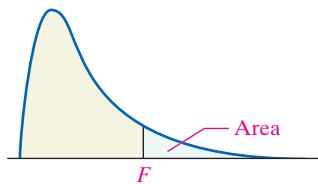


Table IX (continued)

		F-Distribution Critical Values							
		Degrees of Freedom in the Numerator							
Area in Right Tail		9	10	15	20	30	60	120	1000
1	0.100	59.86	60.19	61.22	61.74	62.26	62.79	63.06	63.30
	0.050	240.54	241.88	245.95	248.01	250.10	252.20	253.25	254.19
	0.025	963.28	968.63	984.87	993.10	1001.41	1009.80	1014.02	1017.75
	0.010	6022.47	6055.85	6157.28	6208.73	6260.65	6313.03	6339.39	6362.68
	0.001	602283.99	605620.97	615763.66	620907.67	626098.96	631336.56	633972.40	636301.21
2	0.100	9.38	9.39	9.42	9.44	9.46	9.47	9.48	9.49
	0.050	19.38	19.40	19.43	19.45	19.46	19.48	19.49	19.49
	0.025	39.39	39.40	39.43	39.45	39.46	39.48	39.49	39.50
	0.010	99.39	99.40	99.43	99.45	99.47	99.48	99.49	99.50
	0.001	999.39	999.40	999.43	999.45	999.47	999.48	999.49	999.50
3	0.100	5.24	5.23	5.20	5.18	5.17	5.15	5.14	5.13
	0.050	8.81	8.79	8.70	8.66	8.62	8.57	8.55	8.53
	0.025	14.47	14.42	14.25	14.17	14.08	13.99	13.95	13.91
	0.010	27.35	27.23	26.87	26.69	26.50	26.32	26.22	26.14
	0.001	129.86	129.25	127.37	126.42	125.45	124.47	123.97	123.53
4	0.100	3.94	3.92	3.87	3.84	3.82	3.79	3.78	3.76
	0.050	6.00	5.96	5.86	5.80	5.75	5.69	5.66	5.63
	0.025	8.90	8.84	8.66	8.56	8.46	8.36	8.31	8.26
	0.010	14.66	14.55	14.20	14.02	13.84	13.65	13.56	13.47
	0.001	48.47	48.05	46.76	46.10	45.43	44.75	44.40	44.09
5	0.100	3.32	3.30	3.24	3.21	3.17	3.14	3.12	3.11
	0.050	4.77	4.74	4.62	4.56	4.50	4.43	4.40	4.37
	0.025	6.68	6.62	6.43	6.33	6.23	6.12	6.07	6.02
	0.010	10.16	10.05	9.72	9.55	9.38	9.20	9.11	9.03
	0.001	27.24	26.92	25.91	25.39	24.87	24.33	24.06	23.82
6	0.100	2.96	2.94	2.87	2.84	2.80	2.76	2.74	2.72
	0.050	4.10	4.06	3.94	3.87	3.81	3.74	3.70	3.67
	0.025	5.52	5.46	5.27	5.17	5.07	4.96	4.90	4.86
	0.010	7.98	7.87	7.56	7.40	7.23	7.06	6.97	6.89
	0.001	18.69	18.41	17.56	17.12	16.67	16.21	15.98	15.77
7	0.100	2.72	2.70	2.63	2.59	2.56	2.51	2.49	2.47
	0.050	3.68	3.64	3.51	3.44	3.38	3.30	3.27	3.23
	0.025	4.82	4.76	4.57	4.47	4.36	4.25	4.20	4.15
	0.010	6.72	6.62	6.31	6.16	5.99	5.82	5.74	5.66
	0.001	14.33	14.08	13.32	12.93	12.53	12.12	11.91	11.72
8	0.100	2.56	2.54	2.46	2.42	2.38	2.34	2.32	2.30
	0.050	3.39	3.35	3.22	3.15	3.08	3.01	2.97	2.93
	0.025	4.36	4.30	4.10	4.00	3.89	3.78	3.73	3.68
	0.010	5.91	5.81	5.52	5.36	5.20	5.03	4.95	4.87
	0.001	11.77	11.54	10.84	10.48	10.11	9.73	9.53	9.36

Table IX (continued)

F-Distribution Critical Values											
Degrees of Freedom in the Denominator	Area in Right Tail	Degrees of Freedom in the Numerator									
		1	2	3	4	5	6	7	8	9	10
9	0.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
	0.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
	0.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
	0.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89
10	0.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
	0.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
	0.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
	0.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75
12	0.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
	0.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	0.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
	0.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
	0.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29
15	0.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
	0.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	0.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
	0.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
	0.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08
20	0.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
	0.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
	0.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
	0.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08
25	0.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
	0.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	0.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
	0.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
	0.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56
50	0.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80	1.76	1.73
	0.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
	0.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46	2.38	2.32
	0.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70
	0.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82	3.67
100	0.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73	1.69	1.66
	0.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93
	0.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32	2.24	2.18
	0.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50
	0.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44	3.30
200	0.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63
	0.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88
	0.025	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26	2.18	2.11
	0.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41
	0.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43	3.26	3.12
1000	0.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68	1.64	1.61
	0.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84
	0.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20	2.13	2.06
	0.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43	2.34
	0.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30	3.13	2.99

Table IX (continued)

F-Distribution Critical Values											
Degrees of Freedom in the Denominator	Area in Right Tail	Degrees of Freedom in the Numerator									
		12	15	20	25	30	40	50	60	120	1000
9	0.100	2.38	2.34	2.30	2.27	2.25	2.23	2.22	2.21	2.18	2.16
	0.050	3.07	3.01	2.94	2.89	2.86	2.83	2.80	2.79	2.75	2.71
	0.025	3.87	3.77	3.67	3.60	3.56	3.51	3.47	3.45	3.39	3.34
	0.010	5.11	4.96	4.81	4.71	4.65	4.57	4.52	4.48	4.40	4.32
	0.001	9.57	9.24	8.90	8.69	8.55	8.37	8.26	8.19	8.00	7.84
10	0.100	2.28	2.24	2.20	2.17	2.16	2.13	2.12	2.11	2.08	2.06
	0.050	2.91	2.85	2.77	2.73	2.70	2.66	2.64	2.62	2.58	2.54
	0.025	3.62	3.52	3.42	3.35	3.31	3.26	3.22	3.20	3.14	3.09
	0.010	4.71	4.56	4.41	4.31	4.25	4.17	4.12	4.08	4.00	3.92
	0.001	8.45	8.13	7.80	7.60	7.47	7.30	7.19	7.12	6.94	6.78
12	0.100	2.15	2.10	2.06	2.03	2.01	1.99	1.97	1.96	1.93	1.91
	0.050	2.69	2.62	2.54	2.50	2.47	2.43	2.40	2.38	2.34	2.30
	0.025	3.28	3.18	3.07	3.01	2.96	2.91	2.87	2.85	2.79	2.73
	0.010	4.16	4.01	3.86	3.76	3.70	3.62	3.57	3.54	3.45	3.37
	0.001	7.00	6.71	6.40	6.22	6.09	5.93	5.83	5.76	5.59	5.44
15	0.100	2.02	1.97	1.92	1.89	1.87	1.85	1.83	1.82	1.79	1.76
	0.050	2.48	2.40	2.33	2.28	2.25	2.20	2.18	2.16	2.11	2.07
	0.025	2.96	2.86	2.76	2.69	2.64	2.59	2.55	2.52	2.46	2.40
	0.010	3.67	3.52	3.37	3.28	3.21	3.13	3.08	3.05	2.96	2.88
	0.001	5.81	5.54	5.25	5.07	4.95	4.80	4.70	4.64	4.47	4.33
20	0.100	1.89	1.84	1.79	1.76	1.74	1.71	1.69	1.68	1.64	1.61
	0.050	2.28	2.20	2.12	2.07	2.04	1.99	1.97	1.95	1.90	1.85
	0.025	2.68	2.57	2.46	2.40	2.35	2.29	2.25	2.22	2.16	2.09
	0.010	3.23	3.09	2.94	2.84	2.78	2.69	2.64	2.61	2.52	2.43
	0.001	4.82	4.56	4.29	4.12	4.00	3.86	3.77	3.70	3.54	3.40
25	0.100	1.82	1.77	1.72	1.68	1.66	1.63	1.61	1.59	1.56	1.52
	0.050	2.16	2.09	2.01	1.96	1.92	1.87	1.84	1.82	1.77	1.72
	0.025	2.51	2.41	2.30	2.23	2.18	2.12	2.08	2.05	1.98	1.91
	0.010	2.99	2.85	2.70	2.60	2.54	2.45	2.40	2.36	2.27	2.18
	0.001	4.31	4.06	3.79	3.63	3.52	3.37	3.28	3.22	3.06	2.91
50	0.100	1.68	1.63	1.57	1.53	1.50	1.46	1.44	1.42	1.38	1.33
	0.050	1.95	1.87	1.78	1.73	1.69	1.63	1.60	1.58	1.51	1.45
	0.025	2.22	2.11	1.99	1.92	1.87	1.80	1.75	1.72	1.64	1.56
	0.010	2.56	2.42	2.27	2.17	2.10	2.01	1.95	1.91	1.80	1.70
	0.001	3.44	3.20	2.95	2.79	2.68	2.53	2.44	2.38	2.21	2.05
100	0.100	1.61	1.56	1.49	1.45	1.42	1.38	1.35	1.34	1.28	1.22
	0.050	1.85	1.77	1.68	1.62	1.57	1.52	1.48	1.45	1.38	1.30
	0.025	2.08	1.97	1.85	1.77	1.71	1.64	1.59	1.56	1.46	1.36
	0.010	2.37	2.22	2.07	1.97	1.89	1.80	1.74	1.69	1.57	1.45
	0.001	3.07	2.84	2.59	2.43	2.32	2.17	2.08	2.01	1.83	1.64
200	0.100	1.58	1.52	1.46	1.41	1.38	1.34	1.31	1.29	1.23	1.16
	0.050	1.80	1.72	1.62	1.56	1.52	1.46	1.41	1.39	1.30	1.21
	0.025	2.01	1.90	1.78	1.70	1.64	1.56	1.51	1.47	1.37	1.25
	0.010	2.27	2.13	1.97	1.87	1.79	1.69	1.63	1.58	1.45	1.30
	0.001	2.90	2.67	2.42	2.26	2.15	2.00	1.90	1.83	1.64	1.43
1000	0.100	1.55	1.49	1.43	1.38	1.35	1.30	1.27	1.25	1.18	1.08
	0.050	1.76	1.68	1.58	1.52	1.47	1.41	1.36	1.33	1.24	1.11
	0.025	1.96	1.85	1.72	1.64	1.58	1.50	1.45	1.41	1.29	1.13
	0.010	2.20	2.06	1.90	1.79	1.72	1.61	1.54	1.50	1.35	1.16
	0.001	2.77	2.54	2.30	2.14	2.02	1.87	1.77	1.69	1.49	1.22

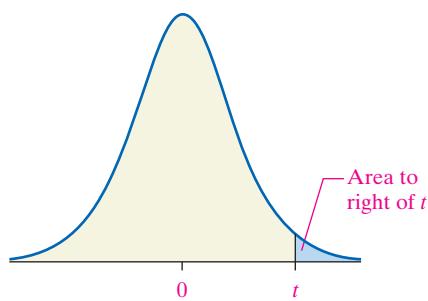


Table X

Student's *t*-Distribution

<i>t</i> \ df	1	2	3	4	5	6	7	8	9	10	11	12
0.0	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
0.1	0.468	0.465	0.463	0.463	0.462	0.462	0.462	0.461	0.461	0.461	0.461	0.461
0.2	0.437	0.430	0.427	0.426	0.425	0.424	0.424	0.423	0.423	0.423	0.423	0.422
0.3	0.407	0.396	0.392	0.390	0.388	0.387	0.386	0.386	0.385	0.385	0.385	0.385
0.4	0.379	0.364	0.358	0.355	0.353	0.352	0.351	0.350	0.349	0.349	0.348	0.348
0.5	0.352	0.333	0.326	0.322	0.319	0.317	0.316	0.315	0.315	0.314	0.313	0.313
0.6	0.328	0.305	0.295	0.290	0.287	0.285	0.284	0.283	0.282	0.281	0.280	0.280
0.7	0.306	0.278	0.267	0.261	0.258	0.255	0.253	0.252	0.251	0.250	0.249	0.249
0.8	0.285	0.254	0.241	0.234	0.230	0.227	0.225	0.223	0.222	0.221	0.220	0.220
0.9	0.267	0.232	0.217	0.210	0.205	0.201	0.199	0.197	0.196	0.195	0.194	0.193
1.0	0.250	0.211	0.196	0.187	0.182	0.178	0.175	0.173	0.172	0.170	0.169	0.169
1.1	0.235	0.193	0.176	0.167	0.161	0.157	0.154	0.152	0.150	0.149	0.147	0.146
1.2	0.221	0.177	0.158	0.148	0.142	0.138	0.135	0.132	0.130	0.129	0.128	0.127
1.3	0.209	0.162	0.142	0.132	0.125	0.121	0.117	0.115	0.113	0.111	0.110	0.109
1.4	0.197	0.148	0.128	0.117	0.110	0.106	0.102	0.100	0.098	0.096	0.095	0.093
1.5	0.187	0.136	0.115	0.104	0.097	0.092	0.089	0.086	0.084	0.082	0.081	0.080
1.6	0.178	0.125	0.104	0.092	0.085	0.080	0.077	0.074	0.072	0.070	0.069	0.068
1.7	0.169	0.116	0.094	0.082	0.075	0.070	0.066	0.064	0.062	0.060	0.059	0.057
1.8	0.161	0.107	0.085	0.073	0.066	0.061	0.057	0.055	0.053	0.051	0.050	0.049
1.9	0.154	0.099	0.077	0.065	0.058	0.053	0.050	0.047	0.045	0.043	0.042	0.041
2.0	0.148	0.092	0.070	0.058	0.051	0.046	0.043	0.040	0.038	0.037	0.035	0.034
2.1	0.141	0.085	0.063	0.052	0.045	0.040	0.037	0.034	0.033	0.031	0.030	0.029
2.2	0.136	0.079	0.058	0.046	0.040	0.035	0.032	0.029	0.028	0.026	0.025	0.024
2.3	0.131	0.074	0.052	0.041	0.035	0.031	0.027	0.025	0.023	0.022	0.021	0.020
2.4	0.126	0.069	0.048	0.037	0.031	0.027	0.024	0.022	0.020	0.019	0.018	0.017
2.5	0.121	0.065	0.044	0.033	0.027	0.023	0.020	0.018	0.017	0.016	0.015	0.014
2.6	0.117	0.061	0.040	0.030	0.024	0.020	0.018	0.016	0.014	0.013	0.012	0.012
2.7	0.113	0.057	0.037	0.027	0.021	0.018	0.015	0.014	0.012	0.011	0.010	0.010
2.8	0.109	0.054	0.034	0.024	0.019	0.016	0.013	0.012	0.010	0.009	0.009	0.008
2.9	0.106	0.051	0.031	0.022	0.017	0.014	0.011	0.010	0.009	0.008	0.007	0.007
3.0	0.102	0.048	0.029	0.020	0.015	0.012	0.010	0.009	0.007	0.007	0.006	0.006
3.1	0.099	0.045	0.027	0.018	0.013	0.011	0.009	0.007	0.006	0.006	0.005	0.005
3.2	0.096	0.043	0.025	0.016	0.012	0.009	0.008	0.006	0.005	0.005	0.004	0.004
3.3	0.094	0.040	0.023	0.015	0.011	0.008	0.007	0.005	0.005	0.004	0.004	0.003
3.4	0.091	0.038	0.021	0.014	0.010	0.007	0.006	0.005	0.004	0.003	0.003	0.003
3.5	0.089	0.036	0.020	0.012	0.009	0.006	0.005	0.004	0.003	0.003	0.002	0.002
3.6	0.086	0.035	0.018	0.011	0.008	0.006	0.004	0.003	0.003	0.002	0.002	0.002
3.7	0.084	0.033	0.017	0.010	0.007	0.005	0.004	0.003	0.002	0.002	0.002	0.002
3.8	0.082	0.031	0.016	0.010	0.006	0.004	0.003	0.003	0.002	0.002	0.001	0.001
3.9	0.080	0.030	0.015	0.009	0.006	0.004	0.003	0.002	0.002	0.001	0.001	0.001
4.0	0.078	0.029	0.014	0.008	0.005	0.004	0.003	0.002	0.002	0.001	0.001	0.001

Table X (continued)**Student's *t*-Distribution**

<i>t</i>	<i>df</i>	13	14	15	16	17	18	19	20	21	22	23	24
0.0		0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
0.1		0.461	0.461	0.461	0.461	0.461	0.461	0.461	0.461	0.461	0.461	0.461	0.461
0.2		0.422	0.422	0.422	0.422	0.422	0.422	0.422	0.422	0.422	0.422	0.422	0.422
0.3		0.384	0.384	0.384	0.384	0.384	0.384	0.384	0.384	0.384	0.383	0.383	0.383
0.4		0.348	0.348	0.347	0.347	0.347	0.347	0.347	0.347	0.347	0.347	0.346	0.346
0.5		0.313	0.312	0.312	0.312	0.312	0.312	0.311	0.311	0.311	0.311	0.311	0.311
0.6		0.279	0.279	0.279	0.278	0.278	0.278	0.278	0.278	0.277	0.277	0.277	0.277
0.7		0.248	0.248	0.247	0.247	0.247	0.246	0.246	0.246	0.246	0.246	0.245	0.245
0.8		0.219	0.219	0.218	0.218	0.217	0.217	0.217	0.217	0.216	0.216	0.216	0.216
0.9		0.192	0.192	0.191	0.191	0.190	0.190	0.190	0.189	0.189	0.189	0.189	0.189
1.0		0.168	0.167	0.167	0.166	0.166	0.165	0.165	0.165	0.164	0.164	0.164	0.164
1.1		0.146	0.145	0.144	0.144	0.143	0.143	0.143	0.142	0.142	0.142	0.141	0.141
1.2		0.126	0.125	0.124	0.124	0.123	0.123	0.122	0.122	0.122	0.121	0.121	0.121
1.3		0.108	0.107	0.107	0.106	0.105	0.105	0.105	0.104	0.104	0.104	0.103	0.103
1.4		0.092	0.092	0.091	0.090	0.090	0.089	0.089	0.088	0.088	0.088	0.087	0.087
1.5		0.079	0.078	0.077	0.077	0.076	0.075	0.075	0.075	0.074	0.074	0.074	0.073
1.6		0.067	0.066	0.065	0.065	0.064	0.064	0.063	0.063	0.062	0.062	0.062	0.061
1.7		0.056	0.056	0.055	0.054	0.054	0.053	0.053	0.052	0.052	0.052	0.051	0.051
1.8		0.048	0.047	0.046	0.045	0.045	0.044	0.044	0.043	0.043	0.043	0.042	0.042
1.9		0.040	0.039	0.038	0.038	0.037	0.037	0.036	0.036	0.036	0.035	0.035	0.035
2.0		0.033	0.033	0.032	0.031	0.031	0.030	0.030	0.030	0.029	0.029	0.029	0.028
2.1		0.028	0.027	0.027	0.026	0.025	0.025	0.025	0.024	0.024	0.024	0.023	0.023
2.2		0.023	0.023	0.022	0.021	0.021	0.021	0.020	0.020	0.020	0.019	0.019	0.019
2.3		0.019	0.019	0.018	0.018	0.017	0.017	0.016	0.016	0.016	0.016	0.015	0.015
2.4		0.016	0.015	0.015	0.014	0.014	0.014	0.013	0.013	0.013	0.013	0.012	0.012
2.5		0.013	0.013	0.012	0.012	0.011	0.011	0.011	0.011	0.010	0.010	0.010	0.010
2.6		0.011	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.008	0.008	0.008	0.008
2.7		0.009	0.009	0.008	0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006	0.006
2.8		0.008	0.007	0.007	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.005	0.005
2.9		0.006	0.006	0.005	0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.004	0.004
3.0		0.005	0.005	0.004	0.004	0.004	0.004	0.004	0.004	0.003	0.003	0.003	0.003
3.1		0.004	0.004	0.004	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.002
3.2		0.003	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002
3.3		0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
3.4		0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001
3.5		0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.6		0.002	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.7		0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.8		0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000
3.9		0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
4.0		0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table X (continued)Student's *t*-Distribution

<i>t</i>	<i>df</i>	25	26	27	28	29	30	35	40	60	120	$\infty (=z)$
0.0		0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
0.1		0.461	0.461	0.461	0.461	0.461	0.460	0.460	0.460	0.460	0.460	0.460
0.2		0.422	0.422	0.421	0.421	0.421	0.421	0.421	0.421	0.421	0.421	0.421
0.3		0.383	0.383	0.383	0.383	0.383	0.383	0.383	0.383	0.383	0.382	0.382
0.4		0.346	0.346	0.346	0.346	0.346	0.346	0.346	0.346	0.345	0.345	0.345
0.5		0.311	0.311	0.311	0.310	0.310	0.310	0.310	0.310	0.309	0.309	0.309
0.6		0.277	0.277	0.277	0.277	0.277	0.276	0.276	0.275	0.275	0.274	0.274
0.7		0.245	0.245	0.245	0.245	0.245	0.244	0.244	0.244	0.243	0.243	0.242
0.8		0.216	0.215	0.215	0.215	0.215	0.215	0.214	0.214	0.213	0.213	0.212
0.9		0.188	0.188	0.188	0.188	0.188	0.187	0.187	0.186	0.185	0.185	0.184
1.0		0.163	0.163	0.163	0.163	0.163	0.162	0.162	0.161	0.160	0.159	0.159
1.1		0.141	0.141	0.141	0.140	0.140	0.140	0.139	0.139	0.138	0.137	0.136
1.2		0.121	0.120	0.120	0.120	0.120	0.120	0.119	0.119	0.117	0.116	0.115
1.3		0.103	0.103	0.102	0.102	0.102	0.102	0.101	0.101	0.099	0.098	0.097
1.4		0.087	0.087	0.086	0.086	0.086	0.086	0.085	0.085	0.083	0.082	0.081
1.5		0.073	0.073	0.073	0.072	0.072	0.072	0.071	0.071	0.069	0.068	0.067
1.6		0.061	0.061	0.061	0.060	0.060	0.060	0.059	0.059	0.057	0.056	0.055
1.7		0.051	0.051	0.050	0.050	0.050	0.050	0.049	0.048	0.047	0.046	0.045
1.8		0.042	0.042	0.042	0.041	0.041	0.041	0.040	0.040	0.038	0.037	0.036
1.9		0.035	0.034	0.034	0.034	0.034	0.034	0.033	0.032	0.031	0.030	0.029
2.0		0.028	0.028	0.028	0.028	0.027	0.027	0.027	0.026	0.025	0.024	0.023
2.1		0.023	0.023	0.023	0.022	0.022	0.022	0.022	0.021	0.020	0.019	0.018
2.2		0.019	0.018	0.018	0.018	0.018	0.018	0.017	0.017	0.016	0.015	0.014
2.3		0.015	0.015	0.015	0.015	0.014	0.014	0.014	0.013	0.012	0.012	0.011
2.4		0.012	0.012	0.012	0.012	0.012	0.011	0.011	0.011	0.010	0.009	0.008
2.5		0.010	0.010	0.009	0.009	0.009	0.009	0.009	0.008	0.008	0.007	0.006
2.6		0.008	0.008	0.007	0.007	0.007	0.007	0.007	0.006	0.006	0.005	0.005
2.7		0.006	0.006	0.006	0.006	0.006	0.006	0.005	0.005	0.004	0.004	0.003
2.8		0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.004	0.003	0.003	0.003
2.9		0.004	0.004	0.004	0.004	0.004	0.003	0.003	0.003	0.003	0.002	0.002
3.0		0.003	0.003	0.003	0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.001
3.1		0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001
3.2		0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001
3.3		0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000
3.4		0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000
3.5		0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000
3.6		0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000
3.7		0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3.8		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3.9		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4.0		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Answers

CHAPTER 1 Data Collection

1.1 Assess Your Understanding (page 11)

1. (a) III (b) VIII (c) IV (d) VII (e) I
(f) VI (g) II (h) V
3. Parameter 5. Statistic
7. Parameter 9. Statistic
11. Qualitative 13. Quantitative
15. Quantitative 17. Qualitative
19. Discrete 21. Continuous
23. Continuous 25. Discrete
27. Nominal 29. Ratio
31. Ordinal 33. Ratio

35. Population: teenagers 13 to 17 years of age who live in the United States. Sample: 1028 teenagers 13 to 17 years of age who live in the United States

37. Population: entire soybean crop. Sample: 100 plants selected

39. Population: women 27 to 44 years of age with hypertension. Sample: 7373 women 27 to 44 years of age with hypertension

41. Individuals: Alabama, Colorado, Indiana, North Carolina, Wisconsin

Variables: minimum age for driver's license (unrestricted), mandatory belt use seating positions, maximum allowable speed limit on rural interstates

Data for minimum age for driver's license: 17, 17, 18, 16, 18

Data for mandatory belt use seating positions: front, front, all, all, all
Data for maximum allowable speed limit on rural interstates: 70, 75, 70, 70, 65 (mph)

The variable *minimum age for driver's license* is continuous; the variable *mandatory belt use seating positions* is qualitative; the variable *maximum allowable speed limit on rural interstates* is continuous.

43. (a) To determine if adolescents 18–21 who smoke have a lower IQ than nonsmokers
(b) All adolescents 18–21; the 20,211 18-year-old Israeli male military recruits.
(c) The average IQ of smokers was 94; the average IQ of nonsmokers was 101.
(d) Lower IQ individuals are more likely to choose to smoke.
45. (a) To determine the proportion of adult Americans who believe the federal government wastes 51 cents or more of every dollar
(b) American adults aged 18 years or older
(c) 1017 American adults aged 18 years or older
(d) Of the 1017 individuals surveyed, 35% indicated that 51 cents or more is wasted.
(e) Gallup is 95% certain that the percentage of all adult Americans who believe the federal government wastes 51 cents or more of every dollar received is between 31% and 39%.

47. (a) Qualitative (b) Qualitative
(c) Quantitative; Discrete (d) Quantitative; Continuous

49. Nominal: ordinal, the level of measurement changes because the goal of the research has changed. Rather than the number being used to identify the player, it is now also used to explain the caliber of the player, with a higher number implying a lower caliber of player.

51. (a) Does season of birth affect mood?
(b) 400 people
(c) Qualitative
(d) Those born in summer are prone to mood swings; those born in winter are less likely to be irritable.
(e) Season of birth plays a role in one's temperament.

53. Qualitative variables describe characteristics of individuals. Quantitative variables provide numerical counts or measures of individuals. A discrete variable is a quantitative variable that has a finite

or countable number of possible values. A discrete variable cannot take on every possible value between any two possible values. Continuous variables are also quantitative variables, but there are an infinite number of possible values that are not countable. A continuous variable may take on every possible value between any two values.

55. This means that the values of the variable change from individual to individual. In addition, certain variables can change over time for certain individuals. Because data vary, two different statistical analyses of the same variable can lead to different results.

57. No. We measure age to as much accuracy as we wish.

1.2 Assess Your Understanding (page 20)

1. The explanatory variable is the variable that affects the value of the response variable. In research, we want to see how changes in the value of the explanatory variable affect the value of the response variable.

3. (a) II (b) III (c) I

5. It depends. Sometimes there are ethical reasons why an experiment cannot be conducted. Other times the researcher may want to conduct an observational study first to validate a belief before investing time and money in an experiment. Certainly, a designed experiment is the superior analytical tool if ethics, time, and money are not an issue.

7. There is a perceived benefit to obtaining a flu shot, so there are ethical issues in intentionally denying certain seniors access.

9. Observational study 11. Experiment

13. Observational study 15. Experiment

17. (a) Cohort since people were followed for 10 years.

(b) Whether the individual has heart disease or not; whether the individual is happy or not

(c) Confounding due to lurking variables

19. (a) The researchers administered a questionnaire to obtain their results, so there is no control of the explanatory variable. This is a cross-sectional study.

(b) Body mass index; whether a TV is in the bedroom or not

(c) Answers will vary. Some lurking variables might be the amount of exercise per week and eating habits.

(d) The researchers made an effort to avoid confounding by accounting for potential lurking variables.

(e) No. This is an observational study, so all we can say is that a television in the bedroom is associated with a higher body mass index.

21. (a) The data was collected at a specific point in time.

(b) Delivery scenario

(c) Method of delivery (qualitative); Cost (quantitative)

23. Answers will vary. This is a prospective, cohort observational study. Some possible lurking variables include smoking habits, eating habits, exercise, and family history of cancer. The study concluded that there might be an increased risk of certain blood cancers in people with prolonged exposure to electromagnetic fields. The author of the article reminds us that this is an observational study, so there is no control of variables that may affect the likelihood of getting certain cancers. In addition, the author states that we should do things in our lives that promote health.

25. Because individuals in the early 1900s were pressured to become right-handed, we would see a lower proportion of left-handers who are older in the study. This would make it seem as though left-handers die younger because the older individuals in the study are primarily right-handed.

27. Web scraping can be used to extract data from tables on web pages and then upload the data to a file. Or, web scraping can be used to create a data set of words from an online article (that is, fetching unstructured information and transforming it into a structured format through something called parsing and reformatting processes). Web scraping can also be used to dynamically call information from websites with links.

ANS-2 ANSWERS 1.5 Assess Your Understanding

1.3 Assess Your Understanding (page 28)

1. The frame is a list of the individuals in the population we are studying.
3. Once an individual is selected, he or she cannot be selected again.
5. Answers will vary. One option would be to write each name on a sheet of paper and choose the names out of a hat. A second option would be to number the books from 1 to 9 and randomly select three distinct numbers.
7. (a) 616, 630; 616, 631; 616, 632; 616, 645; 616, 649; 616, 650; 630, 631; 630, 632; 630, 645; 630, 649; 630, 650; 631, 632; 631, 645; 631, 649; 631, 650; 632, 645; 632, 649; 632, 650; 645, 649; 645, 650; 649, 650
(b) There is a 1 in 21 chance the pair of courses will be EPR 630 and EPR 645.
9. (a) 83, 67, 84, 38, 22, 24, 36, 58, 34
(b) Answers may vary depending on the type of technology used.
11. (a) Answers will vary. (b) Answers will vary.
13. (a) The list provided by the administration serves as the frame. Number the students on the list from 1 through 19,935. Using a random-number generator, set a seed (or select a starting point if using Table I in Appendix A). Generate 25 different numbers randomly. The students corresponding to these numbers will be the 25 students in the sample.
(b) Answers will vary.
15. Answers will vary. However, the members should be numbered from 1 to 32. Then, using the table of random numbers or a random-number generator, four different numbers should be selected. The names corresponding to these numbers will represent the simple random sample.
17. Answers will vary.

1.4 Assess Your Understanding (page 36)

1. If the population can be divided into homogeneous, nonoverlapping groups
3. Convenience samples are not random samples. Individuals who participate are often self-selected, so the results are not likely to be representative of the population.
5. stratified 7. False 9. True
11. Systematic 13. Cluster 15. Simple random
17. Cluster 19. Convenience 21. Stratified
23. 16, 41, 66, 91, 116, 141, 166, 191, 216, 241, 266, 291, 316, 341, 366, 391, 416, 441, 466, 491
25. Answers will vary. To obtain the sample, number the Democrats 1 to 16 and obtain a simple random sample of size 2. Then number the Republicans 1 to 16 and obtain a simple random sample of size 2. Be sure to use a different starting point in Table I or a different seed for each stratum.
27. (a) 90
(b) Randomly select a number between 1 and 90. Suppose that we randomly select 15; then the individuals in the survey will be 15, 105, 195, 285, . . . , 4425.
29. SRS: number from 1 to 1280. Randomly select 128 students to survey.
Stratified: Randomly select four students from each section to survey.
Cluster: Randomly select four sections; survey all students in these four sections.
Answers will vary.
31. Answers will vary. A good choice might be stratified sampling with the strata being commuters and noncommuters.
33. Answers will vary. One option would be cluster sampling. The clusters could be the city blocks. Randomly select clusters and then survey all the households on the selected city blocks.
35. Answers will vary. Simple random sampling will work fine here, especially because a list of 6600 individuals who meet the needs of our study already exists (the frame).
37. (a) Registered voters who have voted in the past few elections.

- (b) Because each individual has an equally likely chance of being selected, there is a chance that one group may be over- or underrepresented.
- (c) By using a stratified sample, the strategist can obtain a simple random sample within each strata so that the number of individuals in the sample is proportionate to the number of individuals in the population.

1.5 Assess Your Understanding (page 42)

1. A closed question has fixed choices for answers, whereas an open question is a free-response question. Closed questions are easier to analyze, but limit the responses. Open questions allow respondents to state exactly how they feel, but are harder to analyze due to the variety of answers and possible misinterpretation of answers.
3. (a) III (b) I (c) IV (d) II
5. (a) Sampling bias due to undercoverage since the first 60 customers may not be representative of the customer population.
(b) Since a complete frame is not possible, systematic random sampling could be used to make the sample more representative of the customer population.
7. (a) Response bias due to a poorly worded question.
(b) The survey should begin by stating the current penalty for selling a gun illegally. The question might be rewritten as “Do you approve or disapprove of harsher penalties for individuals who sell guns illegally?” The words *approve* and *disapprove* should be rotated from individual to individual.
9. (a) Nonresponse bias; if the survey is only written in English, non-English speaking homes will not complete the survey, resulting in undercoverage.
(b) The survey can be improved through face-to-face or telephone interviews.
11. (a) Sampling bias due to undercoverage, since the readers of the magazine may not be representative of all Australian women, and interviewer error, since advertisements and images in the magazine could affect the women’s view of themselves.
(b) A well-designed sampling plan (not in a magazine), such as a cluster sample, could make the sample more representative of the intended population.
13. (a) Response bias due to a poorly worded question
(b) The question should be reworded so that it doesn’t imply the opinion of the editors. One possibility might be “Do you believe that a marriage can be maintained after an extramarital relation?”
15. (a) Response bias because the students are not likely to be truthful when the teacher is administering the survey
(b) The survey should be administered by an impartial party so that the students will be more likely to respond truthfully.
17. No; the survey still suffers from undercoverage (sampling bias), nonresponse bias, and potentially response bias.
19. The ordering of the questions is likely to affect the survey results. Perhaps question B should be asked first. Another possibility is to rotate the questions randomly.
21. The company is using a reward in the form of the \$5.00 payment and an incentive by telling the readers that his or her input will make a difference.
23. (a) Sampling bias due to undercoverage. Individuals who choose not to register to vote may have some characteristics that differ from those who do register.
(b) Undercoverage also would exist for RDD polls. It is likely the case that the poll is not capturing individuals in a lower socioeconomic class. RBS likely has more of this type of bias because access to cell phones is fairly prevalent today.
(c) RDD had the lower response rate at 6% versus 8% in the RBS survey. This is likely due to the fact that the RBS survey had actual phone numbers to choose from, rather than random digits.
(d) The RDD survey oversampled Republicans. This could be due to socioeconomic considerations.

- 25.** Definitely, especially if households that are on the do-not-call registry have a trait that is not part of those households that are not on the do-not-call registry.
- 27.** Answers will vary. However, the research should include the fact that exit polls tended to undersample non-college-educated whites and oversampled college-educated whites. In this election, non-college-educated voters broke for Trump, while college-educated voters were carried by Clinton.

29. The words used in a survey question can have a significant impact on the response. Whenever you read survey results, be mindful of the way the survey question was written. In this particular situation, the words “along with allies in Europe” had a large impact on survey results.

31. Answers will vary.

33. Sampling bias: using an incorrect frame led to undercoverage. Nonresponse bias: the low response rate.

- 35.** **(a)** Answers may vary. However, you should stratify by political affiliation (Democrat, Republican, Independent, for example). **(b)** Answers may vary, but historically Democratic turnout is higher in presidential election cycles. **(c)** Answers may vary. A higher percentage of Democrats in polls versus turnout will lead to overstating the predicted percentage of Democratic votes.

37. Callbacks, incentives

39. The researcher can learn common answers.

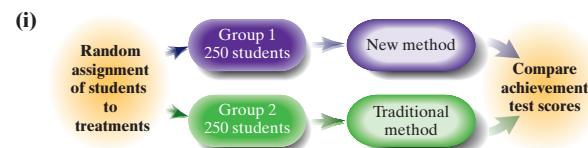
41. Better survey results. A low response rate may mean that some segments of the population are underrepresented or that only individuals with strong opinions have participated.

43. The question is ambiguous because it could be interpreted as hours per day, or hours per week, or hours for a particular class. The question could be improved by being more specific, such as, “On average, how many hours do you study each day for your statistics course?”

1.6 Assess Your Understanding (page 50)

- 1. (a)** A person, object, or some other well-defined item upon which a treatment is applied
(b) Any combination of the values of the factors (explanatory variables)
(c) A quantitative or qualitative variable that represents the variable of interest
(d) The variable whose effect on the response variable is to be assessed by the experimenter
(e) A treatment that looks just like the “real” treatments in a study. Could be an innocuous medication or a procedure that follows the same steps as the experimental procedure, but leaves out a key intervention.
(f) The effect of two factors (explanatory variables) on the response variable cannot be distinguished.
(g) Nondisclosure of the treatment an experimental unit is receiving
- 3.** In a single-blind experiment, the subject does not know which treatment is received. In a double-blind experiment, neither the subject nor the researcher in contact with the subject knows which treatment is received.
- 5.** False
- 7. (a)** To determine the association between number of times one chews food and food consumption.
(b) The response variable is food consumption; quantitative
(c) The explanatory variable is chew level (100%, 150%, 200%); qualitative
(d) The 45 individuals 18 to 45 years of age.
(e) Determine a baseline number of chews before swallowing; same type of food used in baseline as in experiment; same time of day (lunch)
(f) Randomization reduces the effect of the order in which number of chews required plays. For example, perhaps the first time through subjects are more diligent about their chewing than the last time through the study.

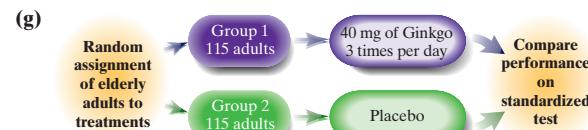
- 9. (a)** Score on achievement test
(b) Method of teaching, grade level, intelligence, school district, teacher. Fixed: grade level, school district, teacher. Set at a predetermined level: method of teaching
(c) New teaching method and traditional method; 2
(d) Random assignment
(e) Group 2
(f) Completely randomized design
(g) 500 first-grade students in District 203
(h) If students tend to perform worse in classes later in the day (due to being tired or anxious to get out of school), then time of day may be a confounding variable. Suppose group 1 is taught in the morning and group 2 is taught in the afternoon; if group 1 scores better on the achievement test, we won’t know whether this is due to the new method of teaching, or due to time of day the course is offered. One solution would be to have a rotating schedule so classes are not always taught at the same time of day.



- 11. (a)** Matched pair
(b) Whiteness level
(c) Crest Whitestrips Premium with brushing and flossing versus brushing and flossing alone
(d) Answers will vary. One possible variable would be diet. Certain foods and tobacco products are more likely to stain teeth. This could affect the whiteness level.
(e) Answers will vary. One possible answer is that using twins helps control for genetic factors (like weak teeth) that may affect the results of the study.

- 13. (a)** Completely randomized design
(b) The response variable is the time it takes to travel 9.75 meters, which is a quantitative variable.
(c) Priming is the treatment and it is set at two levels—scrambled sentence task using words associated with old age or words not associated with old age.
(d) The 30 male and female undergraduate students
(e) The undergraduates did not know which group they were assigned to, and the individual assigning the students did not know which group the student was assigned to.
(f) The elderly priming condition subjects had a travel time significantly higher than that of the neutral priming condition.

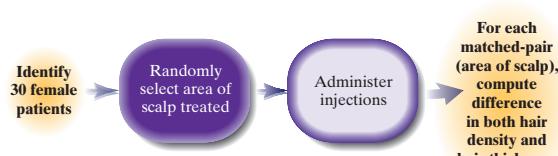
- 15. (a)** Completely randomized design
(b) Adults older than 60 years in good health
(c) Score on a standardized test of learning and memory
(d) Drug: 40 mg of ginkgo 3 times per day or a matching placebo
(e) 98 men and 132 women older than 60 in good health
(f) The placebo group



- 17. (a)** Matched-pairs design because it is the same patient receiving both treatments
(b) Females with hair loss 20 to 45 years of age
(c) Hair density (number of hairs per square centimeter) and hair diameter
(d) The treatment is the injection and it is set at two levels: platelet-rich plasma (PRP) injection or saline injection.
(e) The thirty female patients

(f) Randomly choose the area of the scalp that receives the treatment.

(g)



19. Answers will vary.

21. (a) This is an observational study because there is no intent to manipulate an explanatory variable. The explanatory variable is whether the individual is a green tea drinker or not, which is a qualitative variable.
 (b) Some lurking variables include diet, exercise, genetics, gender, age
 (c) Completely randomized design
 (d) To double-blind this experiment, we would need the placebo to look, taste, and smell like green tea. Subjects would not know which treatment is being delivered. Plus, the individual responsible for measuring changes in LDL cholesterol would not know the treatment either (odds are this blinding is not necessary due to the way the experiment is set up since the subjects have completed the experiment by the time the LDL is measured).
 (e) The treatment is the tea, which is set at three levels.
 (f) Exercise, diet, age

(g) Number the subjects 1 to 120. Randomly select 40 subjects and assign them to the placebo group. Then, randomly select 40 from the remaining 80 and assign them to the one cup of green tea group. The remaining subjects will be in the two cups of green tea group. By randomly assigning subjects the expectation is that those variables not controlled (such as genetic history) are neutralized (even out).
 (h) Exercise is a confounding variable because any difference in the change in LDL cholesterol cannot be attributed to the tea. It may be the exercise that caused the change in LDL cholesterol.

23. Answers will vary. Completely randomized design is likely best.

25. (a) Blood pressure
 (b) Daily consumption of salt, daily consumption of fruits and vegetables, body's ability to process salt
 (c) Salt: controlled; fruits/vegetables: controlled, body: cannot be controlled. To deal with variability in body types, randomly assign experimental units to each treatment group.
 (d) Answers will vary. Three might be a good choice; one level below RDA, one equal to RDA, one above RDA.

27. Answers will vary.

29. Control groups are needed in a designed experiment to serve as a baseline against which other treatments can be compared.

31. Answers will vary.

33. Randomization is meant to even out those variables that are not controlled for in a designed experiment. Answers to the randomization question may vary; however, each experimental unit must be assigned randomly. For example, a researcher might randomly select 30 experimental units from the 90 using the methods of simple random sampling and assign them to treatment 1. Then the researcher could randomly select another 30 experimental units from the remaining 60 and assign them to treatment 2. The remaining experimental units would go to treatment 3.

Chapter 1 Review Exercises (page 56)

1. (a) The variable of interest in the study
 (b) A characteristic of an individual
 (c) Allows for classification of individuals based on an attribute or characteristic
 (d) Provides numerical measures of individuals. The values of quantitative variables can be added or subtracted and provide meaningful results.

(e) Measures the value of the response variable without attempting to influence the value of the response or explanatory variables.

(f) If a researcher randomly assigns the individuals in a study to groups, intentionally manipulates the value of an explanatory variable and controls other explanatory variables at fixed values, and then records the value of the response variable for each individual, the study is a designed experiment.

(g) Confounding occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory variable and response variable may be due to some other variable or variables not accounted for in the study.

(h) An explanatory variable that was not considered in the study, but that affects the value of the response variable in the study

2. (1) *Cross-sectional studies*: collect information about individuals at a specific point in time or over a very short period of time. (2) *Case-control studies*: look back in time and match individuals possessing a certain characteristic with those that do not. (3) *Cohort studies*: collect information about a group of individuals over a period of time.

3. (1) Identify the research objective, (2) collect the data needed to answer the research questions, (3) describe the data, (4) perform inference

4. (1) *Sampling bias* occurs when the techniques used to select individuals to be in the sample favor one part of the population over another. (2) *Nonresponse bias* occurs when the individuals selected to be in the sample that do not respond to the survey have different opinions from those that do respond. (3) *Response bias* exists when the answers on a survey do not reflect the true feelings of the respondent.

5. *Nonsampling errors* are errors that result from undercoverage, nonresponse bias, response bias, or data-entry errors. *Sampling errors* are errors that result from using a sample to estimate information about a population. They result because samples contain incomplete information regarding a population.

6. (1) Identify the problem to be solved. (2) Determine the factors that affect the response variable. (3) Determine the number of experimental units. (4) Determine the level of each factor. (5) Conduct the experiment. (6) Test the claim.

- | | |
|--|-----------------------------|
| 7. Quantitative; discrete | 8. Quantitative; continuous |
| 9. Qualitative | 10. Statistic |
| 11. Parameter | 12. Interval |
| 13. Nominal | 14. Ordinal |
| 15. Ratio | 16. Observational study |
| 17. Experiment | 18. Cohort study |
| 19. Convenience sample | 20. Cluster sample |
| 21. Stratified sample | 22. Systematic sample |
| 23. (a) Sampling bias; undercoverage or nonrepresentative sample due to poor sampling frame
(b) Response bias; interviewer error
(c) Data-entry error | |
| 24. Answers will vary. | |
| 25. Answers will vary. | |
| 26. Answers will vary. Label each goggle with pairs of digits from 00 to 99. Using row 12, column 1 of Table I in Appendix A and reading down, the selected labels would be 55, 96, 38, 85, 10, 67, 23, 39, 45, 57, 82, 90, and 76. The goggles with these labels would be inspected for defects. | |
| 27. (a) To determine the ability of chewing gum to remove stains from teeth
(b) Designed experiment, because treatments were intentionally imposed on the experimental units (teeth) to determine their effect on a response variable (percentage of stain removed)
(c) Completely randomized design
(d) Percentage of stain removed
(e) Type of stain remover (gum or saliva); qualitative
(f) The 64 stained teeth (bovine incisors)
(g) The amount of time chewing (fixed at 120 minutes); method of chewing (fixed by simulator) | |

- (h)** Gums A and B remove significantly more stain than gum C or saliva. Gum C removes more stain than saliva.
- 28.** **(a)** Matched-pairs design
(b) Reaction time; quantitative
(c) Alcohol (two drinks: placebo and 40% vodka mixed with orange juice)
(d) Food consumption; caffeine intake
(e) Weight; gender
(f) To act as a placebo to control for psychosomatic effects of alcohol
(g) Alcohol delays reaction time significantly in seniors for low levels of alcohol consumption. The study applies to healthy seniors who are not regular drinkers.
- 29.** Answers will vary.
- 30.** Answers will vary.
- 31.** In a completely randomized design, the experimental units are randomly assigned to one of the treatments. The value of the response variable is compared for each treatment. In a matched-pairs design, experimental units are matched up on the basis of some common characteristic (such as husband–wife or twins). The difference in the matched-up experimental units is analyzed.
- 32.** Answers will vary.
- 33.** Answers will vary.
- 34.** Randomization is meant to even out the effect of those variables that are not controlled for in a designed experiment. Answers to the randomization question may vary; however, each experimental unit must be assigned randomly. For example, a researcher might randomly select 25 experimental units from the 100 using the methods of simple random sampling and assign them to treatment 1. Then the researcher could randomly select another 25 experimental units from the remaining 75 and assign them to treatment 2 and so on.
- Chapter 1 Test (page 59)**
- Collect information, organize and summarize the information, analyze the information to draw conclusions, provide a measure of confidence in the conclusions drawn from the information.
 - (1) Identify the research objective, (2) collect the data needed to answer the research questions, (3) describe the data, (4) perform inference
 - Quantitative, continuous, ratio
 - Qualitative, ordinal
 - Quantitative, discrete, ratio
 - Experiment, battery life
 - Observational study; whether the gap between the rich and poor will grow or shrink
 - A *cross-sectional study* collects data at a specific point in time or over a short period of time; a *cohort study* collects data over a period of time, sometimes over a long period of time; a *case-controlled study* is retrospective, looking back in time to collect data.
 - An experiment involves the researcher actively imposing treatments on experimental units in order to observe any difference between the treatments in terms of effect on the response variable. In an observational study, the researcher observes the individuals in the study without attempting to influence the response variable in any way. Only an experiment will allow a researcher to establish causality.
 - A control group is necessary for a baseline comparison. Comparing other treatments to the control group allows the researcher to identify which, if any, of the other treatments are superior to the current treatment (or no treatment at all). Blinding is important to eliminate bias due to the individual or experimenter knowing which treatment is being applied.
 - (1) Identify the problem to be solved, (2) determine the factors that affect the response variable, (3) determine the number of experimental units, (4) determine the level of each factor, (5) conduct the experiment, (6) test the claim.
 - Number the franchise locations 1 to 15. Use Table I in Appendix A or a random-number generator to determine four unique numbers. The

franchise locations corresponding to these numbers are the franchises in the sample. Results will vary.

- 13.** Obtain a simple random sample for each stratum. Be sure to use a different starting point in Table I in Appendix A or a different seed for each stratum. Results will vary.
- 14.** Number the blocks from 1 to 2500 and obtain a simple random sample of size 10. The blocks corresponding to these numbers represent the blocks analyzed. Analyze all trees on each of the selected blocks. Results will vary.
- 15.** Ideally k would be $600/14 = 42$ (rounded down). Randomly select a number between 1 and 42. This represents the first slot machine inspected. Then inspect every 42nd machine thereafter. Results will vary.
- 16.** In a completely randomized design, the experimental units are randomly assigned to one of the treatments. The value of the response variable is compared for each treatment.
- 17.** **(a)** Sampling bias (voluntary response)
(b) Nonresponse bias
(c) Response bias (poorly worded question)
(d) Sampling bias (undercoverage)
- 18.** **(a)** Matched-pairs design
(b) 159 social drinkers
(c) Type of beer glass
(d) Time to complete the drink; quantitative
(e) The type of glass used in the first week is determined randomly. This is to neutralize the effect of drinking out of a specific glass first.
- (f)**

```

graph LR
    A[Randomly select a straight glass or curved glass.] --> B[Measure time to complete the drink.]
    B --> C[One week later, measure time to complete the drink using other glass.]
    C --> D[For each matched-pair, compute the difference in drink time.]
  
```
- 19.** **(a)** Completely randomized design
(b) Topical cream; 0.5%, 1.0%, 0%
(c) Neither the subjects nor the person applying the treatments were aware of which treatment was being given.
(d) Placebo group (0% cream)
(e) 225 patients with skin irritation
(f)

```

graph TD
    A[Randomly assign patients to creams] --> B[0.5% cream (75 patients)]
    A --> C[1.0% cream (75 patients)]
    A --> D[0% cream (placebo) (75 patients)]
    B --> E[Compare improvement in skin irritation]
    C --> E
    D --> E
  
```
- 20.** **(a)** Subjects were observed over a long period of time and certain characteristics were recorded. The response variable was recorded at the end of the study.
(b) Bone mineral density; weekly cola consumption
(c) Quantitative
(d) The researchers observed values of variables that could potentially impact bone mineral density (besides cola consumption) so their effect could be isolated from the variable of interest.
(e) Smoking status; alcohol consumption, physical activity, calcium intake.
(f) Women who consumed at least one cola per day (on average) had a bone mineral density that was significantly lower at the femoral neck than those who consumed less than one cola per day. No, they are associated.
- 21.** Lurking variables tend to occur in observational studies. In addition, lurking variables are related to both an explanatory variable and the response variable. Confounding variables tend to occur in experiments. The effect of a confounding variable on the response variable cannot be distinguished from a second explanatory variable.

CHAPTER 2 Organizing and Summarizing Data

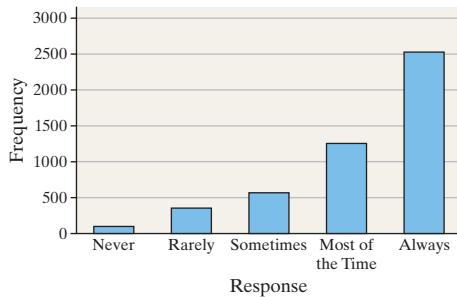
2.1 Assess Your Understanding (page 70)

- Raw data are data that are not organized.
- One (although rounding may cause the result to vary slightly)
- (a) Boss
(b) 158
(c) Values do not sum to one. That is, there is no “whole.”
- (a) OF (b) 15 (c) 5
(d) Each of these positions should be reported as MVPs, rather than treating the three positions as a single position.
- (a) 69% (b) 55.2 million (c) Inferential
- (a) 0.42; 0.61 (b) 55+
(c) 18–34 (d) As age increases, so does likelihood to buy American

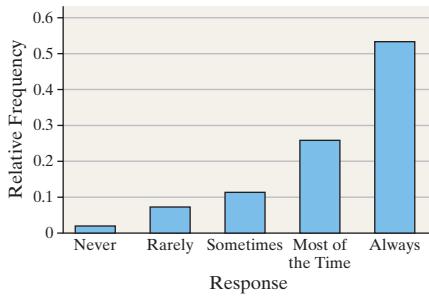
Response	Relative Frequency
Never	0.0262
Rarely	0.0678
Sometimes	0.1156
Most of the time	0.2632
Always	0.5272

(b) 52.7% (c) 9.4%

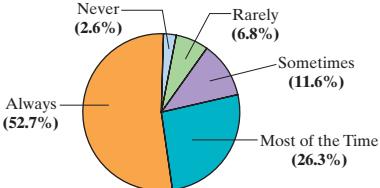
(d) **How Often Do You Wear Your Seat Belt?**



(e) **How Often Do You Wear Your Seat Belt?**



(f) **How Often Do You Wear Your Seat Belt?**

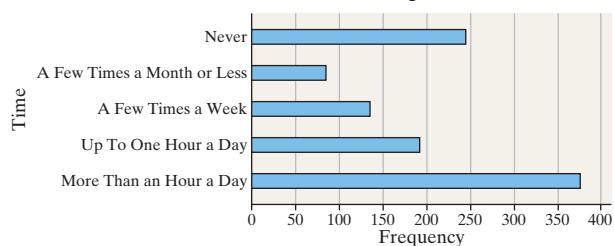


(g) This is a descriptive statement because it is reporting a result of the sample.

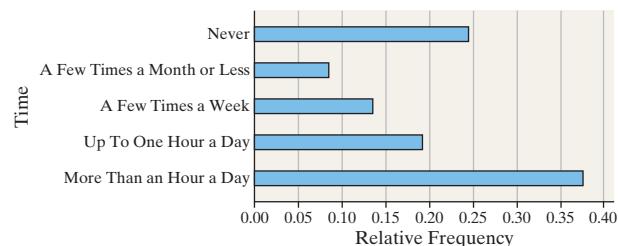
Response	Relative Frequency
More than 1 hour a day	0.3678
Up to 1 hour a day	0.1873
A few times a week	0.1288
A few times a month or less	0.0790
Never	0.2371

(b) 0.2371 (about 24%)

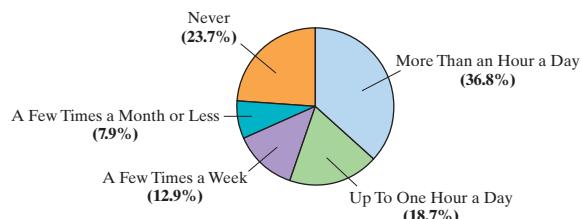
(c) **Time Spent Online**



(d) **Time Spent Online**



(e) **Time Spent Online**

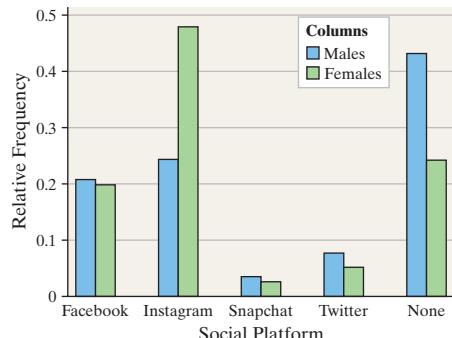


(f) No level of confidence is provided along with the estimate.

17. (a), (b)

Platform	Males	Females
Facebook	0.2087	0.1993
Instagram	0.2433	0.4803
Snapchat	0.0359	0.0260
Twitter	0.0781	0.0512
None	0.4341	0.2433

(c) **What Social Platform Influenced Your Online Shopping?**

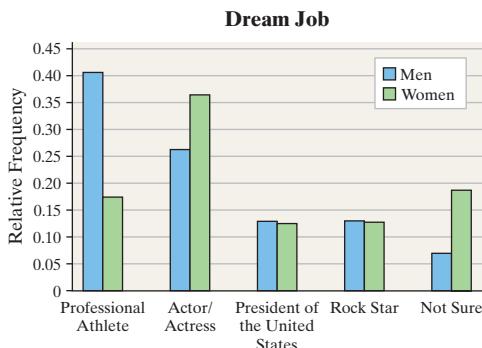


- (d) Females are much more likely for Instagram to influence their online shopping, while males are more likely to have none of these platforms influence their online shopping.

19. (a)

Dream Job	Males	Females
Professional Athlete	0.4040	0.18
Actor/Actress	0.2626	0.37
President of the United States	0.1313	0.13
Rock Star	0.1313	0.13
Not Sure	0.0707	0.19

(b)



(c) Answers will vary.

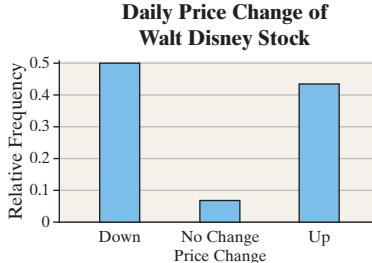
21. (a), (b)

Price Change	Frequency	Relative Frequency
Down	15	0.5
No Change	2	0.067
Up	13	0.433

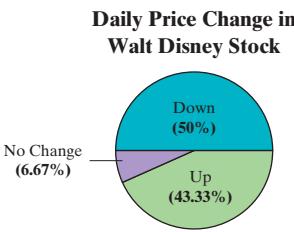
(c)



(d)



(e)

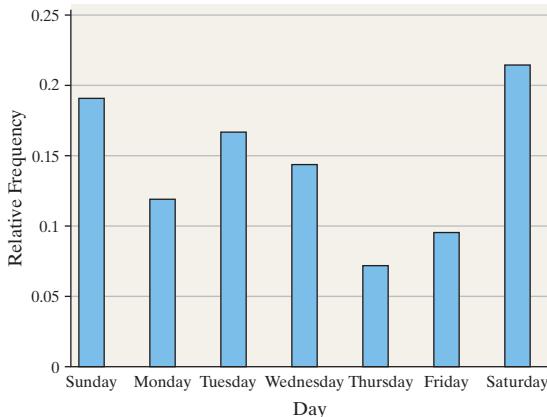


23. (a)

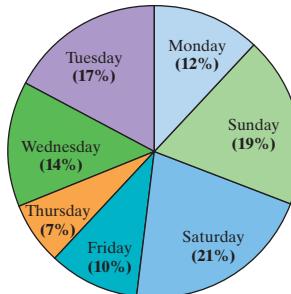
Day	Frequency	Relative Frequency
Sunday	8	0.1905
Monday	5	0.1190
Tuesday	7	0.1667
Wednesday	6	0.1429
Thursday	3	0.0714
Friday	4	0.0952
Saturday	9	0.2143

(b) Saturday

(c) **Online Grocery Deliveries**



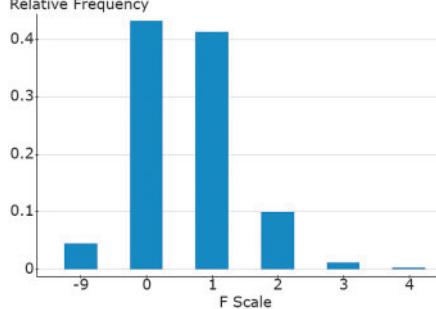
(d) **Online Grocery Deliveries**



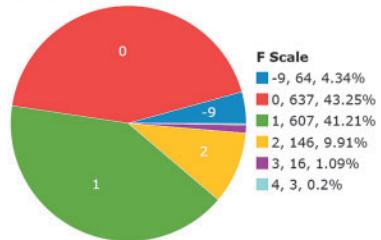
25. (a)

F Scale	Frequency	Relative Frequency
0	64	0.0434
1	637	0.4325
2	607	0.4121
3	146	0.0991
-9	16	0.0109
-4	3	0.0020

(b) **Tornadoes in the United States, 2017**

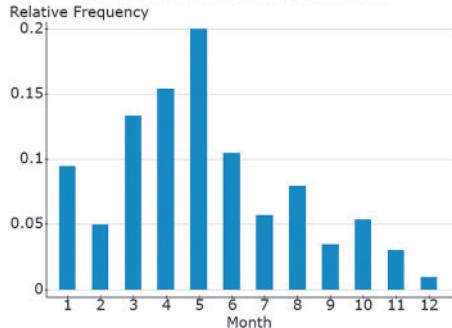


(c) **Tornadoes in the United States, 2017**



(d) Answers may vary, but a bar chart is easier to read with twelve observations.

Tornadoes in the United States, 2017



(e) Texas (168)

27. Answers will vary.

29. (a) To determine if online homework improves student learning over traditional pencil-and-paper homework.

(b) Experiment

(c) Answers will vary. Some examples are same teacher, same semester, same course.

(d) There could be differences between the classes. The instructor may give more instruction to one class than the other. The instructor is not blinded, so he or she may treat one group differently from the other.

(e) *Number of students*: quantitative, discrete; *average age*: quantitative, continuous; *average exam score*: quantitative, continuous; *type of homework*: qualitative; *college experience*: qualitative

(f) Qualitative; ordinal. Answers will vary. Likely yes because it gives more “weight” to the higher grade and the researcher is attempting to convey that a higher percent of students passed using online homework.

(g) Side-by-side relative frequency bar graph

(h) Yes; the “whole” is the set of students who received a grade for the course for each class type.

(i) The table shows the two groups with no prior college experience had roughly the same average exam grade. From the bar graph, we see that the students using online homework had a lower percent for As, but had a higher percent who passed with a C or better.

31. Answers may vary. Decreasing order of importance seems to be a good choice if the goal is to emphasize the importance. A pie chart does not allow for order.

33. No. The percentages do not add to 100%.

2.2 Assess Your Understanding (page 86)

1. classes 3. class width 5. True

7. False. The distribution shape shown is skewed right.

- | | | |
|---|-------------|-----------------|
| 9. (a) 8 | (b) 2 | (c) 15 |
| (d) 4 | (e) 15% | (f) Bell shaped |
| 11. (a) 200 | (b) 10 | |
| (c) 60–69, 2; 70–79, 3; 80–89, 13; 90–99, 42; 100–109, 58; 110–119, 40; 120–129, 31; 130–139, 8; 140–149, 2; 150–159, 1 | | |
| (d) 100–109 | (e) 150–159 | |
| (f) 5.5% | (g) No | |

13. (a) Likely skewed right. Most household incomes will be to the left (perhaps in the \$50,000 to \$150,000 range), with fewer higher incomes to the right (in the millions).

(b) Likely bell-shaped. Most scores will occur near the middle range, with scores tapering off equally in both directions.

(c) Likely skewed right. Most households will have, say, 1 to 4 occupants, with fewer households having a higher number of occupants.

(d) Likely skewed left. Most Alzheimer’s patients will fall in older-aged categories, with fewer patients being younger.

15. (a) 3 (b) 1 (c) 14 (d) 20 (e) 8.75%; 12.5%

17. (a) 9% (b) 2010 (c) 2008 (d) 2009 (e) 2011
(f) Declining due to the decreases in unemployment

Number of Children under 5	Relative Frequency
0	0.32
1	0.36
2	0.24
3	0.06
4	0.02

(b) 24% (c) 60%

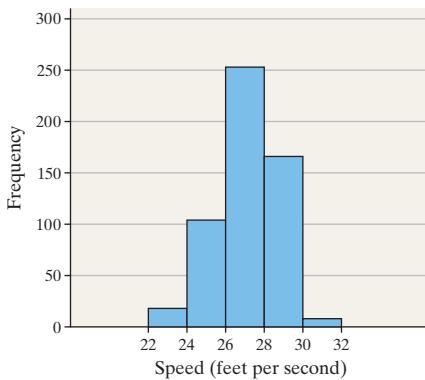
21. (a) Five classes

(b) Lower class limits: 22, 24, 26, 28, 30; upper class limits: 23.9, 25.9, 27.9, 29.9, 31.9

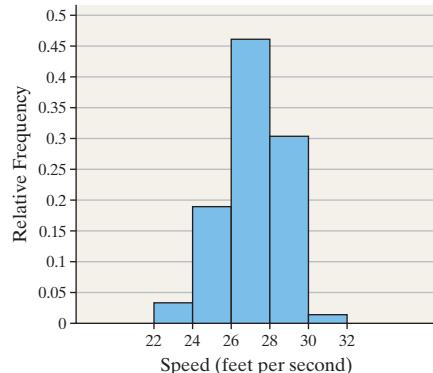
(c) Class width: 2

Speed (ft/sec)	Relative Frequency
22–23.9	0.0328
24–25.9	0.1894
26–27.9	0.4608
28–29.9	0.3024
30–31.9	0.0146

(b) Sprint Speed of Major League Baseball Players



(c) Sprint Speed of Major League Baseball Players



The percentage of players with a sprint speed between 24 and 25.9 feet per second is 18.94%. The percentage of players with a sprint speed less than 23.9 feet per second is 3.28%.

25. (a) Discrete. The possible values for the number of televisions are countable.

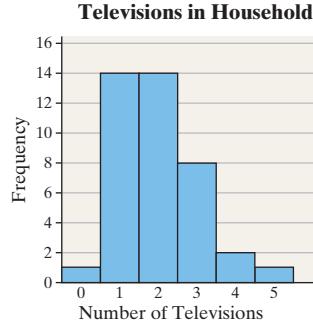
(b), (c)

Number of Televisions	Frequency	Relative Frequency
0	1	0.025
1	14	0.35
2	14	0.35
3	8	0.2
4	2	0.05
5	1	0.025

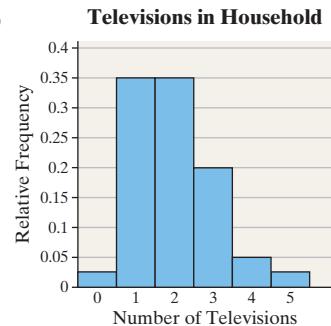
(d) 20%

(e) 7.5%

(f)



(g)

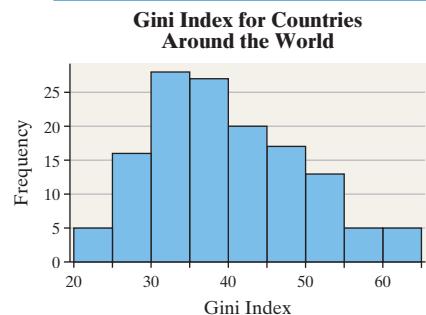


(h) Skewed right

27. (a), (b)

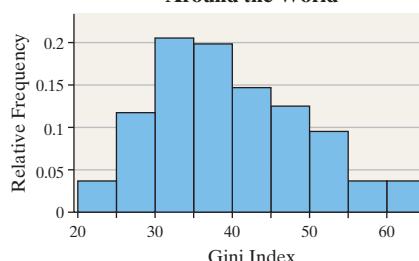
Gini Index	Frequency	Relative Frequency
20–24.9	5	0.037
25–29.9	16	0.118
30–34.9	28	0.206
35–39.9	27	0.199
40–44.9	20	0.147
45–49.9	17	0.125
50–54.9	13	0.096
55–59.9	5	0.037
60–64.9	5	0.037

(c)



(d)

Gini Index for Countries Around the World

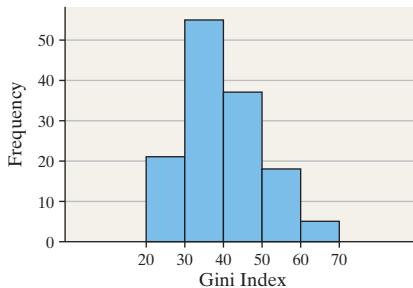


(e) Skewed right

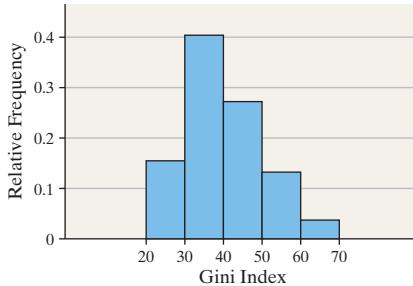
(f)

Gini Index	Frequency	Relative Frequency
20–29.9	21	0.154
30–39.9	55	0.404
40–49.9	37	0.272
50–59.9	18	0.132
60–69.9	5	0.037

Gini Index for Countries Around the World



Gini Index for Countries Around the World



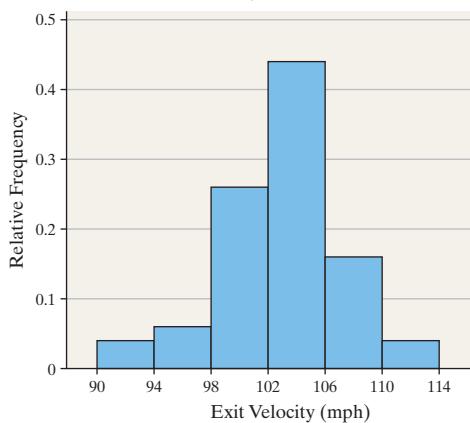
Skewed right

(g) Answers may vary. However, the summary with a class width of 5 seems superior.

29. (a)

Exit Velocity (mph)	Relative Frequency
90–93.9	0.04
94–97.9	0.06
98–101.9	0.26
102–105.9	0.44
106–109.9	0.16
110–113.9	0.04

(b) **Exit Velocity of Home Runs**

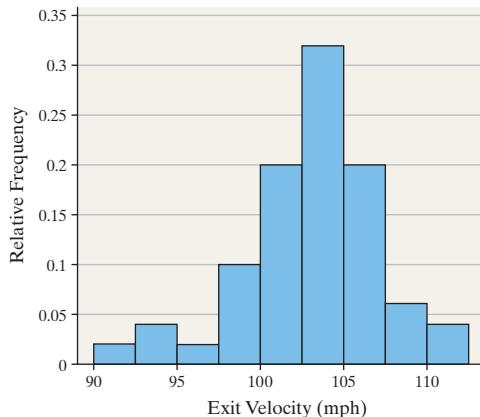


(c) Symmetric and bell-shaped.

(d)

Exit Velocity (mph)	Relative Frequency
90–92.4	0.02
92.5–94.9	0.04
95–97.4	0.02
97.5–99.9	0.10
100–102.4	0.20
102.5–104.9	0.32
105–107.4	0.20
107.5–109.9	0.06
110–112.4	0.04

Exit Velocity of Home Runs



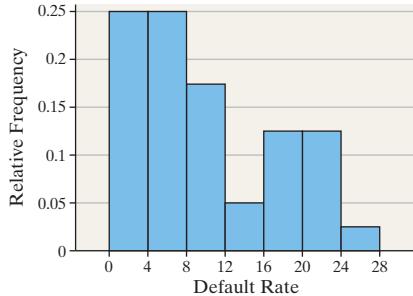
With a class width of 2.5, the distribution looks slightly skewed to the left.

(e) Answers may vary. However, the class width of 4 appears to provide a better summary.

- 31. (a)** Use 0 as the lower class limit of the first class and a class width of 4.

(b)

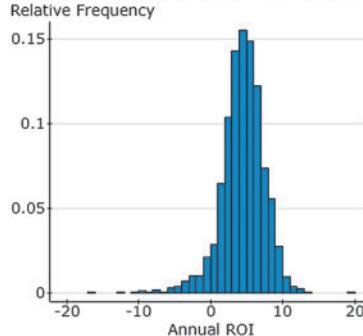
Student Loan Default Rates



(c) Skewed right

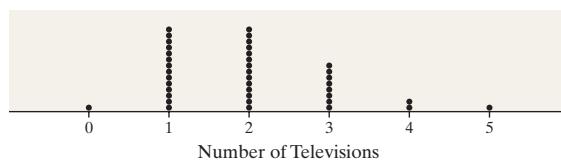
- 33.** It is disconcerting to note that some schools have a negative ROI.

Return on Investment for Higher Education



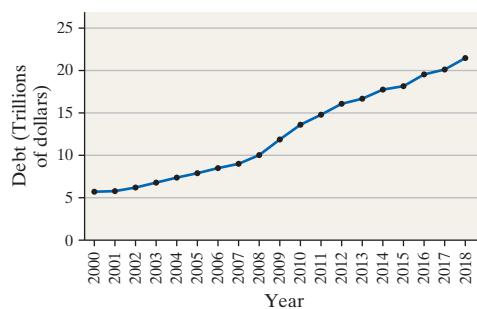
35.

Televisions in Household



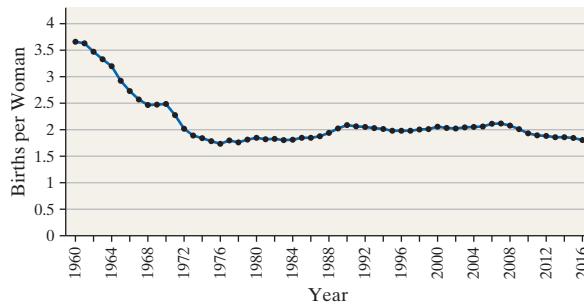
37. (a)

Total Federal Debt



(b) 19.0%; No

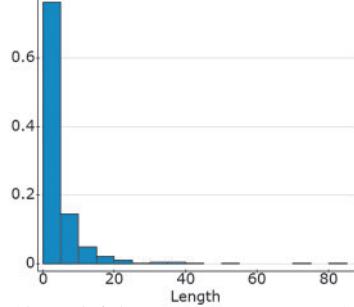
39. Births per Woman in the United States



Births per woman was lowest in 1976.

41. (a) Tornadoes in the United States, 2017

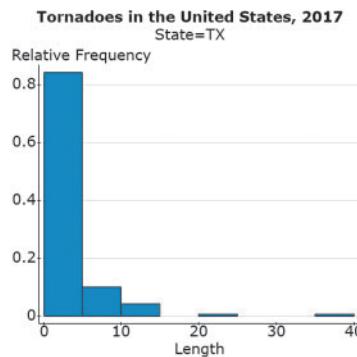
Relative Frequency



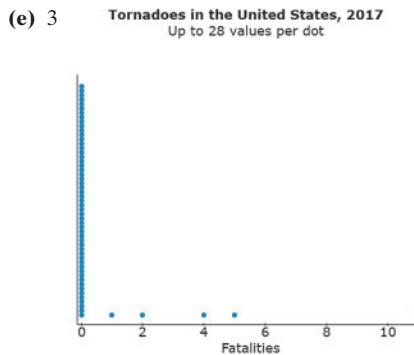
(b) Skewed right

(c) 0.145

(d) 0.006



(e) 3

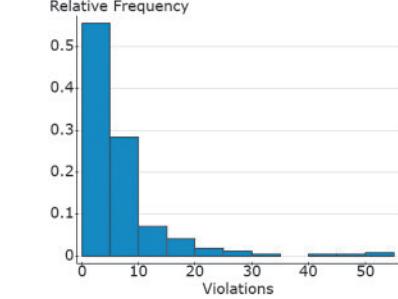
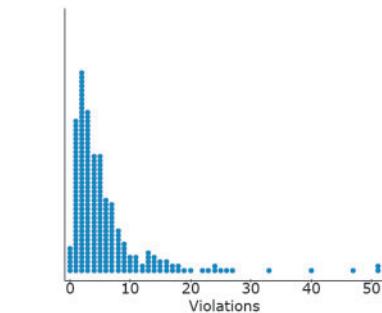


(f) 66

43. (a) Quantitative; discrete

(b) This is population data because it is all recorded violations for the day.

(c) 271

(d) **Red Light Camera Violations in Chicago**(e) **Red Light Camera Violations in Chicago**

Answers may vary.

(f) Yes; 6

45. Answers will vary.

47. There is no such thing as a correct class width, however some choices are better than others.

49. Yes. This exercise illustrates the fact that there is no such thing as the “correct” histogram. However, some histograms are better than others and class width can affect the shape of the graph.

51. Time-series plots are drawn with quantitative variables. They are drawn to see trends in the data.

2.3 Assess Your Understanding (page 98)

1. The number of shark attacks is going to be higher in the summer months due to there being more swimmers. This graphic should be changed to account for the number of swimmers in the water.

3. (a) The vertical axis starts at \$34,000 instead of \$0. Doing this suggests that income for females increased much more than it actually did.
 (b) This graph indicates that the median income for females has increased, but not as significantly as suggested by the graph in part (a).

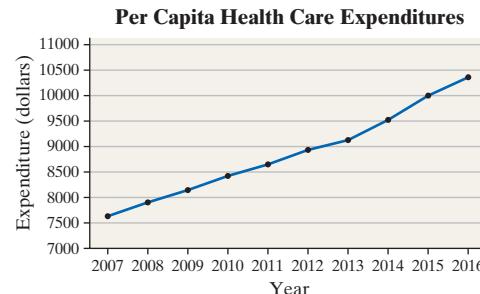


5. The bar for 12p–6p covers twice as many hours as the other bars. By combining two 3-hour periods, this bar looks larger compared to the others, making afternoon hours look more dangerous. When the bar is split into two periods, the graph may give a different impression.

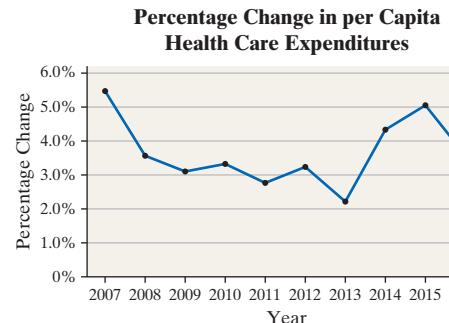
7. This graph is misleading because it does not take into account the size of the population of each state. Certainly, Vermont is going to pay less in total taxes than California simply because its population is so much lower. The moral of the story here is that many variables should be considered on per capita (per person) basis. For example, this graph should be drawn to represent taxes paid per capita (per person).

9. (a) The bar for housing should be a little more than twice the length of the bar for transportation, but it is not.
 (b) Adjust the graph so that the lengths of the bars are proportional.

11. (a) The politician's view:

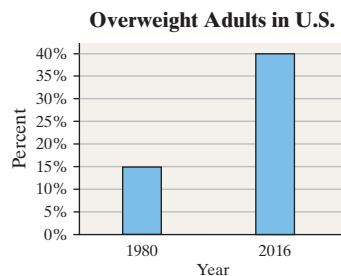


(b) The health care industry's view:

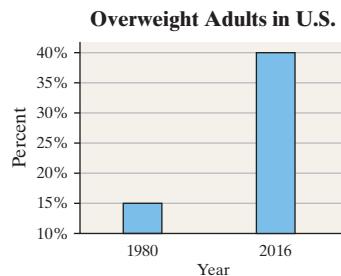


(c) Answers may vary. Not only does the scale affect the message, so does the variable used to measure the argument.

13. (a) Graphic that is not misleading:



- (b) Graphic that is misleading (graphics may vary):

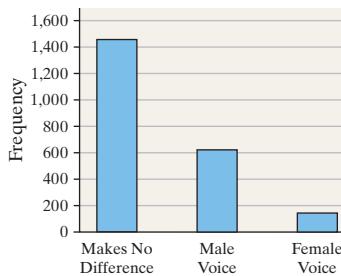


15. Three-dimensional graphs are deceptive. The area for P (pitcher) looks substantially larger than the area for 3B (third base) even though both are the same percentage. Graphs should not be drawn using three dimensions. Instead, use a two-dimensional graph.

Chapter 2 Review Exercises (page 101)

1. (a) 2216 participants (b) 0.653

- (c) **Convincing Voice in Purchasing a Car**

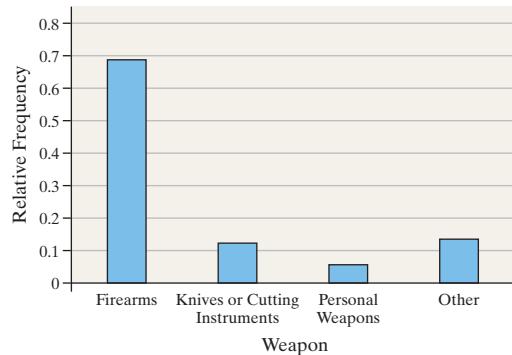


- (d) Answers may vary.

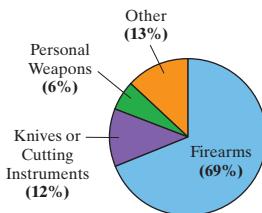
2. (a)

Type of Weapon	Relative Frequency
Firearms	0.6895
Knives or cutting instruments	0.1217
Personal weapons	0.0564
Other weapon	0.1324

- (b) **Weapons Used in Homicides**



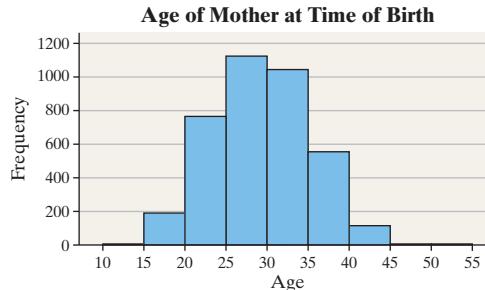
- (c) **Weapons Used in Homicides**



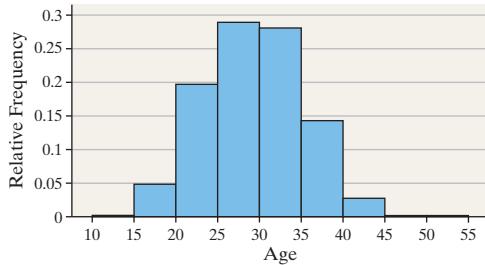
3. (a) **Age of Mother** **Relative Frequency**

10–14	0.0005
15–19	0.0503
20–24	0.1984
25–29	0.2915
30–34	0.2832
35–39	0.1439
40–44	0.0298
45–49	0.0021
50–54	0.0003

- (b) Fairly bell-shaped



- (c) **Age of Mother at Time of Birth**

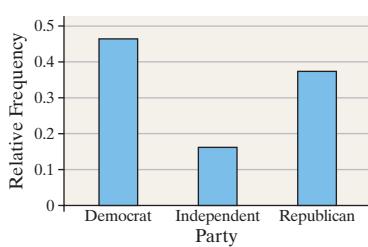


- (d) 19.84% (e) 45.93%

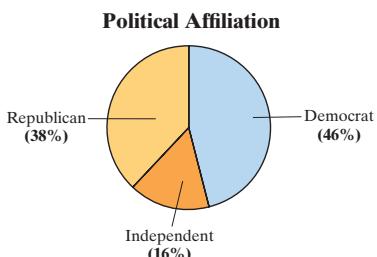
4. (a), (b)

Affiliation	Frequency	Relative Frequency
Democrat	46	0.46
Independent	16	0.16
Republican	38	0.38

- (c) **Political Affiliation**



(d)

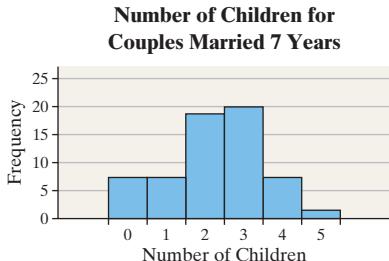


(e) Democrat appears to be the most common affiliation.

5. (a), (b), (c), (d)

Number of Children	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0	7	0.1167	7	0.1167
1	7	0.1167	14	0.2333
2	18	0.3000	32	0.5333
3	20	0.3333	52	0.8667
4	7	0.1167	59	0.9833
5	1	0.0167	60	1

(e)



The distribution is symmetric.

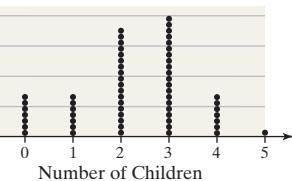
(f)



(g) 30% of the couples has two children.

(h) 76.7% of the couples has at least two children.

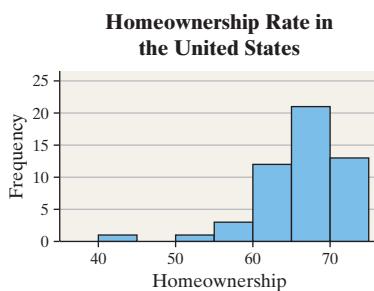
(i)



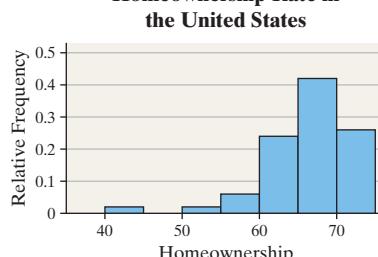
6. (a), (b)

Homeownership Rate	Frequency	Relative Frequency
40–44.9	1	0.0196
45–49.9	0	0
50–54.9	1	0.0196
55–59.9	3	0.0588
60–64.9	12	0.2353
65–69.9	21	0.4118
70–74.9	13	0.2549

(c)



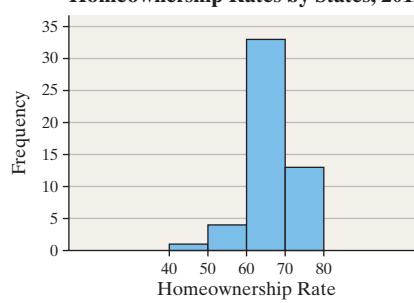
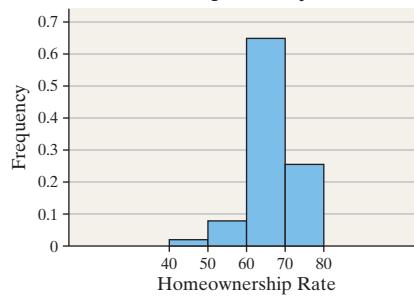
(d)



(e) The distribution is skewed to the left.

(f)

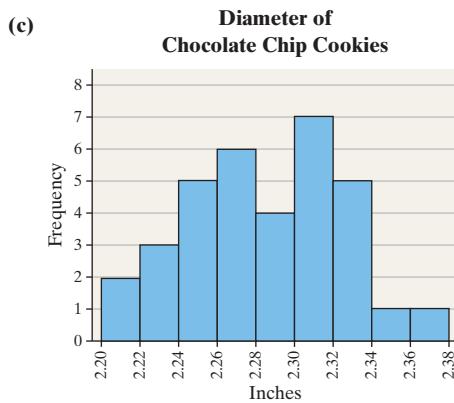
Homeownership Rate	Frequency	Relative Frequency
40–44.9	1	0.0196
50–54.9	4	0.0784
60–64.9	33	0.6471
70–74.9	13	0.2549

Homeownership Rates by States, 2017**Homeownership Rates by States, 2017**

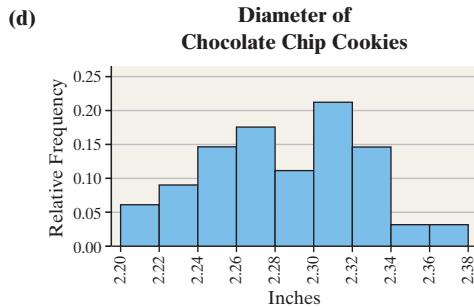
(g) Answers may vary. However, the class width of 5 appears to be the better summary.

7. (a), (b) Answers will vary. Using 2.2000 as the lower class limit of the first class and 0.0200 as the class width, we obtain the following:

Diameter of a Cookie		
Class	Frequency	Relative Frequency
2.2000–2.2199	2	0.0588
2.2200–2.2399	3	0.0882
2.2400–2.2599	5	0.1471
2.2600–2.2799	6	0.1765
2.2800–2.2999	4	0.1176
2.3000–2.3199	7	0.2059
2.3200–2.3399	5	0.1471
2.3400–2.3599	1	0.0294
2.3600–2.3799	1	0.0294



The distribution is roughly symmetric.



8. Hours Spent Online

13	467
14	05578
15	1236
16	456
17	113449
18	066889
19	2
20	168
21	119
22	29
23	48
24	4
25	7

Legend: 13 | 4 = average 13.4 hours per week.

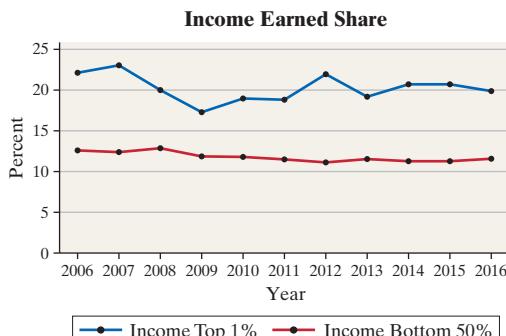
The distribution is skewed right.

9. (a) Yes. Grade point averages have increased every time period for all schools.

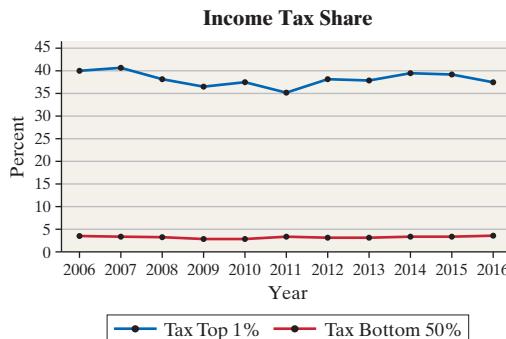
(b) GPAs increased 5.1% for public schools. GPAs increased 6.1% for private schools. Private schools have higher inflation both because the GPAs are higher, and they are increasing faster.

(c) The graph may be misleading because it starts at 2.6 on the vertical axis.

10. (a)



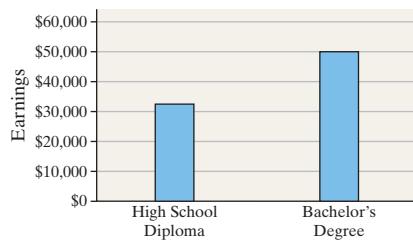
- (b)



11. (a) Answers will vary.

(b) An example of a graph that does not mislead:

2017 Average Earnings



12. (a) Flats are preferred most; extra-high heels are preferred least.

(b) The bar heights and areas are not proportional.

Chapter 2 Test (page 104)

1. (a) 5 Stars (b) 2747 (c) 951 (d) 61%

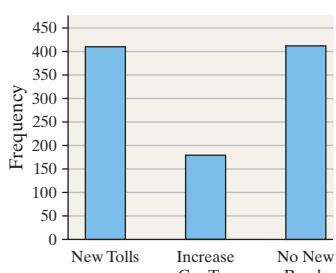
(e) No. This is a bar graph. We only talk about distribution shape for graphs of quantitative data, such as histograms.

2. (a)

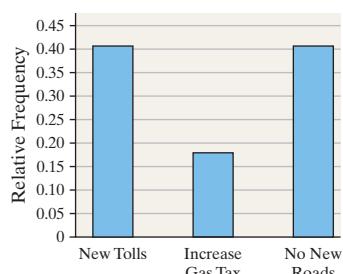
Response	Frequency	Relative Frequency
New tolls	412	0.4100
Increase gas tax	181	0.1801
No new roads	412	0.4100

(b) About 18% of respondents indicated they would like to see an increase in gas taxes.

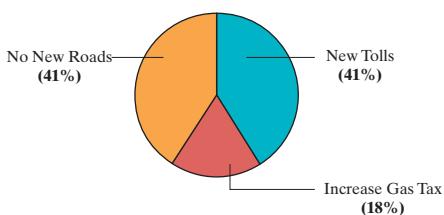
(c) How Would You Prefer to Pay for New Road Construction?



(d) How Would You Prefer to Pay for New Road Construction?



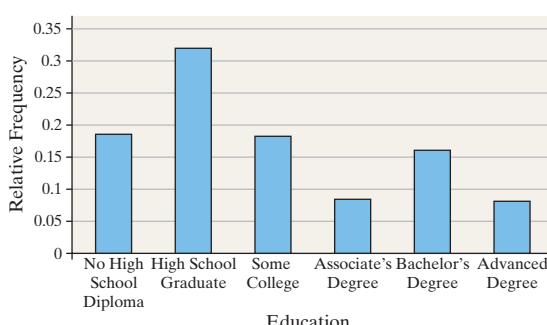
(e) How Would You Prefer to Pay for New Road Construction?



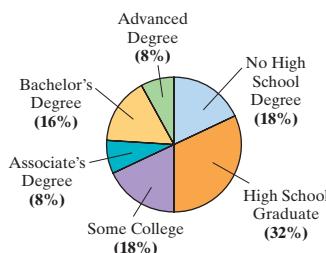
3. (a), (b)

Educational Attainment	Frequency	Relative Frequency
No high school diploma	9	0.18
High school graduate	16	0.32
Some college	9	0.18
Associate's degree	4	0.08
Bachelor's degree	8	0.16
Advanced degree	4	0.08

(c) Educational Attainment of Commuters



(d) Educational Attainment of Commuters

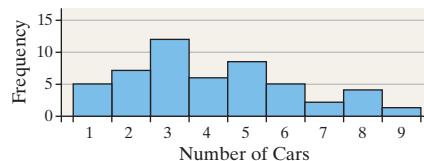


(e) High school graduate is the most common educational level.

4. (a), (b)

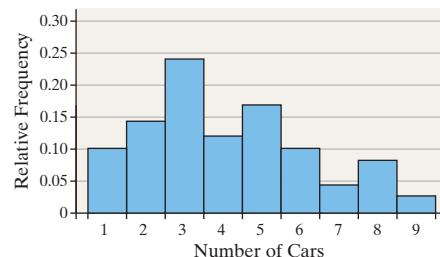
Number of Customers	Frequency	Relative Frequency
1	5	0.10
2	7	0.14
3	12	0.24
4	6	0.12
5	8	0.16
6	5	0.10
7	2	0.04
8	4	0.08
9	1	0.02

(c) Number of Cars Arriving at McDonald's



The distribution is skewed right.

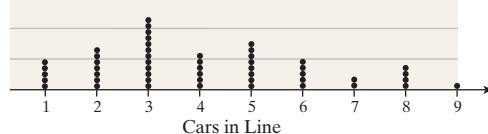
(d) Number of Cars Arriving at McDonald's



(e) 24% of weeks, three cars arrived between 11:50 A.M. and noon.

(f) 76% of weeks, at least three cars arrived between 11:50 A.M. and noon.

(g)

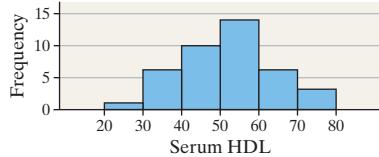


5. Answers may vary. One possibility follows:

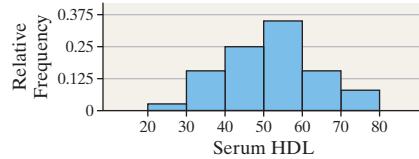
(a), (b) Using a lower class limit of the first class of 20 and a class width of 10:

Serum HDL	Frequency	Relative Frequency
20–29	1	0.025
30–39	6	0.15
40–49	10	0.25
50–59	14	0.35
60–69	6	0.15
70–79	3	0.075

(c) Serum HDL of 20–29 Year Olds



(d) Serum HDL of 20–29 Year Olds



(e) Bell shaped

6. (a) Time Spent on Homework

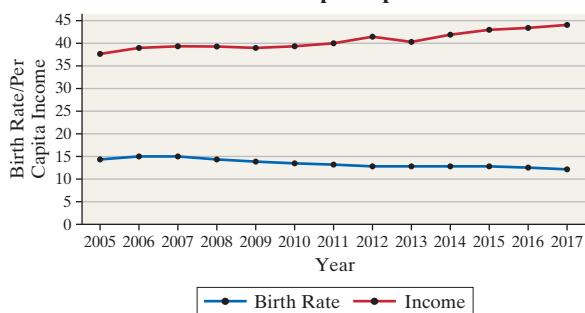
4 | 0567
5 | 26
6 | 13
7 | 01338
8 | 59
9 | 1369
10 | 3899
11 | 0018
12 | 556

Legend: 4 | 0 represents 40 minutes.

The distribution is symmetric (uniform).

7.

Birth Rate and per Capita Income



As per capita income increases, birth rate appears to decrease.

8. Answers will vary.

CHAPTER 3 Numerically Summarizing Data

3.1 Assess Your Understanding (page 116)

1. A statistic is resistant if it is not sensitive to extreme values. The median is resistant because it is a positional measure of central tendency, and increasing the largest value or decreasing the smallest value does not affect the position of the center. The mean is not resistant because it is a function of the sum of the values of data. Changing the magnitude of one value changes the sum of the values.

3. HUD uses the median because the data are skewed. Explanations will vary.

5. The median is between the 5000th and the 5001st ordered values.

7. $\bar{x} = 11$

9. $\mu = 9$

11. Mean price per ticket was \$1725.

13. Mean: 30.52 mpg; Median: 31.3 mpg; there is no mode.

15. The mean, median, and mode strengths are 3670, 3830, and 4090 pounds per square inch, respectively.

17. (a) mean > median

(c) mean < median

Justification will vary.

19. (a) Traditional: $\bar{x} = 71.82$; $M = 70.8$;

Flipped: $\bar{x} = 77.48$; $M = 76.8$

(b) $\bar{x} = 113.22$; $M = 71.5$; resistance

21. (a) The mean pulse rate is 72.2 beats per minute.

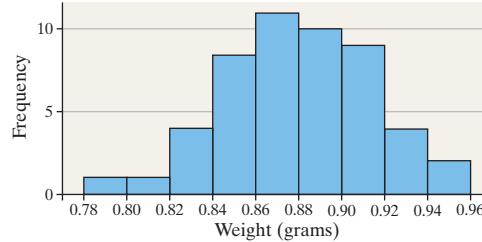
(b) Samples and sample means will vary.

(c) Answers will vary.

23. The distribution is symmetric. The mean is the better measure of central tendency.

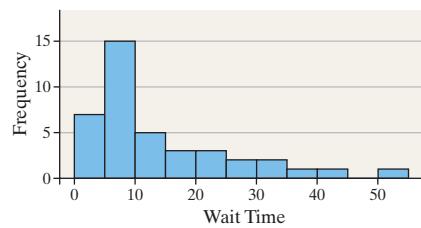
25. $\bar{x} = 0.875$ gram; $M = 0.875$ gram. The distribution is symmetric, so the mean is the better measure of central tendency.

Weight of Plain M&Ms



27. The histogram shows that wait time is skewed right. $\bar{x} = 13.5$ minutes; $M = 8.5$ minutes; the median is the better measure of central tendency because the data are skewed right.

Wait Time (in Minutes) for Dinosaur Ride



29. (a) Moderate

(b) Yes, to avoid response bias

31. Sample of size 5: All data recorded correctly: $\bar{x} = 99.8$; $M = 100$; 106 recorded as 160: $\bar{x} = 110.6$; $M = 100$

Sample of size 12: All data recorded correctly:

$\bar{x} = 100.4$; $M = 101$; 106 recorded as 160: $\bar{x} = 104.9$; $M = 101$

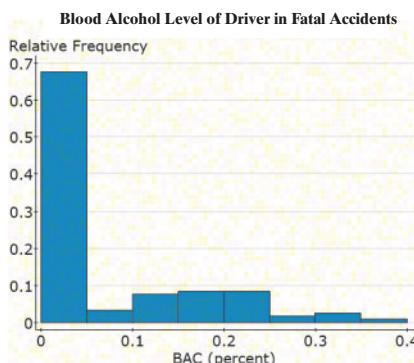
Sample of size 30: All data recorded correctly:

$\bar{x} = 100.6$; $M = 99$; 106 recorded as 160: $\bar{x} = 102.4$; $M = 99$

For each sample size, the mean becomes larger, but the median remains constant. As the sample size increases, the effect of the misrecorded data on the mean decreases.

33. The unreadable score is 44.

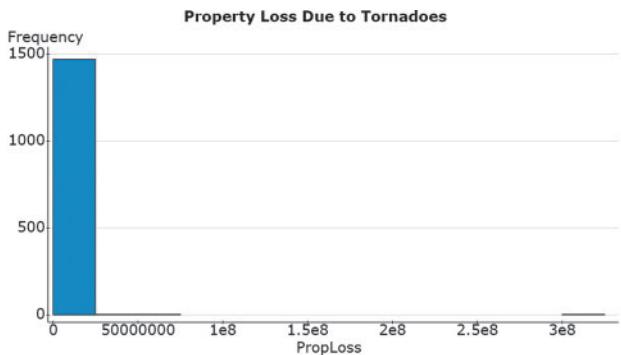
35. The histogram is skewed right. $\bar{x} = 0.061$; $M = 0$. Because the distribution is skewed right, we would expect to report the median as the measure of central tendency. However, a median of 0 does not tell the whole story of the fatal accidents, so it could be argued that both the mean and median be reported. Because the median is 0, at least half of all fatal accidents involved drivers with no alcohol in their system.



- 37.** **(a)** Mean = \$50,000; median = \$50,000; mode = \$50,000
(b) New data set: 32.5, 32.5, 47.5, 52.5, 52.5, 52.5, 57.5, 57.5, 62.5, 77.5; mean = \$52,500; median = \$52,500; mode = \$52,500. All three measures increased by \$2500.
(c) New data set: 31.5, 31.5, 47.25, 52.5, 52.5, 52.5, 57.75, 57.75, 63, 78.75; mean = \$52,500; median = \$52,500; mode = \$52,500. All three measures increased by 5%.
(d) New data set: 30, 30, 45, 50, 50, 50, 55, 55, 60, 100; mean = \$52,500; median = \$50,000; mode = \$50,000.
The mean increased by \$2500, but the median and the mode remained at \$50,000.

39. The trimmed mean is 0.875. The trimmed mean is resistant only if the extreme observation is the highest or lowest value.

41. **(a)** $\mu = 3.775$ miles; $M = 1.85$ miles; Because the mean is greater than the median, we expect the distribution of tornado lengths to be skewed right. The histogram confirms this.
(b) Texas: $\mu = 2.721$ miles; $M = 1.385$ miles; Georgia: $\mu = 5.346$ miles; $M = 3.26$ miles; Georgia has longer tornadoes by both measures of central tendency.
(c) $\mu = \$454,949$; $M = \$8000$; The histogram appears to show only observations between \$0 and \$25,000,000. However, there are a few tornadoes that had extreme property loss. In particular, a tornado in Georgia on January 22 had property loss of \$310,300,000!



- (d)** It does not make sense to determine the mean F scale because F scale is a qualitative variable.

43. **(a)** The mean will be less than 723 because in left-skewed distributions, the mean is usually less than the median.
(b) 50%

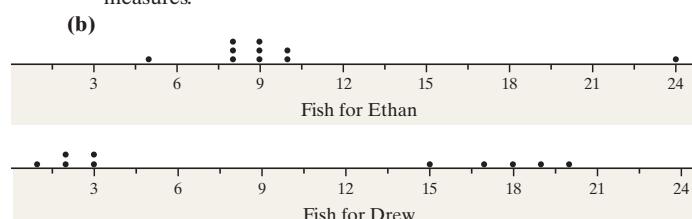
45. No. Each state has a different population size. This must be taken into account.

47. The salary distribution is skewed right, so the players' negotiator would want to use the median salary; the owners' negotiator would use the mean salary to refute the players' claim.

49. **(a)** Median; the distribution of home prices tends to be skewed right.
(b) Mode; the data are qualitative.
(c) Mean; the data are quantitative and symmetric.
(d) Median; the data are quantitative and skewed right.
(e) Median; NFL salaries are skewed right.
(f) Mode; the data are qualitative.
(g) Mode; jersey numbers categorize a player.

3.2 Assess Your Understanding (page 132)

1. zero
 3. True
 5. $s^2 = 36; s = 6$
 7. $\sigma^2 = 16; \sigma = 4$
 9. $s^2 = 196; s = 14$
 11. $R = 13.4 \text{ mpg}; s^2 = 26.35 \text{ mpg}^2; s = 5.13 \text{ mpg}$
 13. $R = 1150 \text{ psi}; s = 459.6 \text{ psi}$
 15. (b) has the higher standard deviation because the data go from 30 to 75; in (a) they are clustered between 40 and 60.
 17. (a) Set 1(a) has the higher standard deviation because the observations 6 and 8 create more dispersion than the corresponding observations 7 and 7. Set 1(a): $s = 2.2$; Set 2(a): $s = 2.1$
(b) Set 2(b) has the higher standard deviation because each observation is farther away from the mean by a factor of 10. Set 2(a): $s = 3.96$; Set 2(b): $s = 39.6$
(c) The standard deviation is the same because each corresponding observation is the same number of units from the mean. Set 3(a): $s = 4.3$; Set 3(b): $s = 4.3$
 19. (a) Traditional (29.1 vs. 28.4)
(b) Traditional (8.84 vs. 7.85)
(c) $R = 541.8; s = 145.88$; neither measure is resistant.
 21. (a) $\sigma = 7.7$ beats per minute (b), (c) Answers will vary.
 23. (a) Ethan: $\mu = 10$ fish, $R = 19$ fish; Drew: $\mu = 10$ fish, $R = 19$ fish.
There is no difference between Ethan and Drew based on these



Ethan appears to be more consistent.

- (c) Ethan: $\sigma = 4.9$ fish, Drew: $\sigma = 7.9$ fish; Ethan has the more consistent record.

(d) Answers will vary.

25. (a) $s = 0.036$ gram

(b) The histogram is approximately symmetric, so the Empirical Rule is applicable.

(c) 95% of the M&Ms should weigh between 0.803 and 0.947 gram.

(d) 96% of the M&Ms actually weigh between 0.803 and 0.947 gram.

(e) 16% of the M&Ms should weigh more than 0.911 gram.

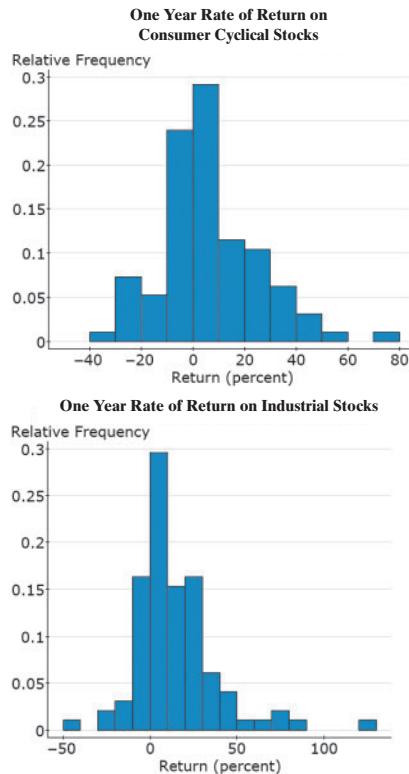
(f) 12% of the M&Ms actually weigh more than 0.911 gram.

27.

	Car 1	Car 2
Sample mean	$\bar{x} = 223.5 \text{ mi}$	$\bar{x} = 237.2 \text{ mi}$
Median	$M = 223 \text{ mi}$	$M = 230 \text{ mi}$
Mode	None	None
Range	$R = 93 \text{ mi}$	$R = 166 \text{ mi}$
Sample variance	$s^2 = 475.1 \text{ mi}^2$	$s^2 = 2406.9 \text{ mi}^2$
Sample standard deviation	$s = 21.8 \text{ mi}$	$s = 49.1 \text{ mi}$

Answers will vary.

29. (a)



It appears that industrial stocks have more dispersion.

(b) Consumer Cyclical: $\bar{x} = 6.595\%$; $M = 3.915\%$; Industrial: $\bar{x} = 14.425\%$; $M = 9.595\%$. Industrial stocks have the higher mean and median rate of return.

(c) Consumer Cyclical: $s = 19.078\%$; Industrial: $s = 23.851\%$. The sector with the higher rate of return also has the higher risk. Answers may vary regarding the determination as to whether it is worth the cost.

31. (a) 95% of people have an IQ score between 70 and 130.

(b) 5% of people have an IQ score either less than 70 or greater than 130.

(c) 2.5% of people have an IQ score greater than 130.

33. (a) 95% of pairs of kidneys weigh between 265 and 385 grams.

(b) 99.7% of pairs of kidneys weigh between 235 and 415 grams.

(c) 0.3% of pairs of kidneys weigh either less than 235 or more than 415 grams.

(d) 81.5% of pairs of kidneys weigh between 295 and 385 grams.

35. It depends. If you are a below average student (meaning you expect to have a mean score below 80%), you are better off with Professor Alpha, since about 97.5% of students will score 70% or better in the class. If you are an above average student, you would go with Professor Omega, since you likely want an A, and about 16% of the class will score 90% or higher.

37. (a) 88.9% of gas stations have prices within 3 standard deviations of the mean.

(b) 84% of gas stations have prices within 2.5 standard deviations of the mean. Gasoline priced from \$2.91 to \$3.21 is within 2.5 standard deviations of the mean.

(c) At least 75% of gas stations have prices between \$2.94 and \$3.18.

39. There is more variation among individuals than among means.

41. Sample of size 5: correct $s = 5.3$, incorrect $s = 27.9$

Sample of size 12: correct $s = 14.7$, incorrect $s = 22.7$

Sample of size 30: correct $s = 15.9$, incorrect $s = 19.2$

As the sample size increases, the effect of the misrecorded observation on standard deviation decreases.

43. (a) Coupe: $\bar{x} = \$17,592$; $s = \$4742.0$

Camaro: $\bar{x} = \$19,697.2$; $s = \$3206.0$

(b) Camaro's tend to cost more than the typical 2-door vehicle. The standard deviation is lower because there is less variability in prices of a specific vehicle.

45. MAD = 4.22 mpg, $s = 5.13$ mpg

47. (a) Average rate of return = 14.9%, risk level = 14.7%

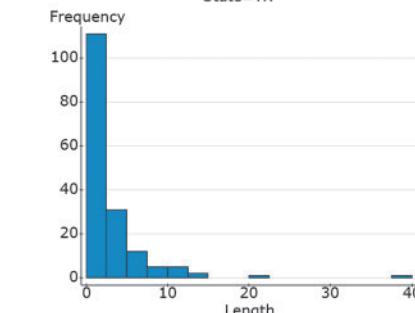
(b) To minimize risk, invest 30% in foreign stock.

(c) Answers will vary.

(d) At least 75% of returns are between -12.8% and 44.4% . At least 88.9% of returns are between -27.1% and 58.7% . Negative returns are not unusual.

49. Answers will vary.

51. (a) Range = 82.53 mi; $\sigma = 5.981$ mi



Texas appears to have more dispersion in length because the values are more spread out in the histogram.

(c) TX: Range = 39.7 mi, $\sigma = 4.349$ mi; OK: Range = 17.91 mi; $\sigma = 3.937$ mi

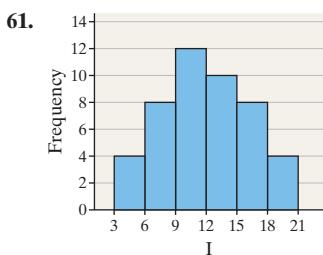
(d) The standard deviation is not defined because there is only one observation. At least two observations are needed to determine the value of the standard deviation.

53. Degrees of freedom refers to the number of data values that are free to be any value while requiring the entire data set to have a specified mean. For example, if a data set is composed of 10 observations, 9 of the observations are free to be any value, while the tenth must be a specific value to obtain a specific mean.

55. A statistic is biased if it consistently under- or overestimates the value of the corresponding parameter.

57. There is more spread among heights when gender is not accounted for. Think about the range of heights from the shortest female to the tallest male (in general) versus the range of heights from the shortest female to the tallest female and the shortest male to the tallest male.

59. The IQ of residents in your home town would have a higher standard deviation because the college campus likely has certain admission requirements which make the individuals more homogeneous as far as intellect goes.



Histogram I has the larger standard deviation.

3.3 Assess Your Understanding (page 143)

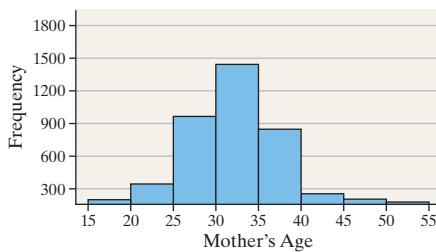
1. $\bar{x} = \$23,221.0$; $s = \$22,484.5$

3. $\bar{x} = 70.6^\circ\text{F}$, $s = 3.5^\circ\text{F}$

5. (a) $\mu = 32.0$ years, $\sigma = 5.7$ years

(b)

Number of Multiple Births in 2017



(c) 95% of mothers of multiple births are between 20.6 and 43.4 years of age.

7. Grouped data: $\bar{x} \approx 102.96$ mph, $s \approx 4.259$ mph;
raw data: $\bar{x} = 103.088$ mph, $s = 4.132$ mph

9. GPA = 3.27

11. Cost per pound = \$2.97

13. (a) Males: $\mu = 38.1$ years, $\sigma = 22.5$ years

(b) Females: $\mu = 38.8$ years, $\sigma = 23.1$ years

(c) Females have the higher mean age.

(d) Females have more dispersion in age.

15. $M = 2298.3$ square feet

17. Modal class: \$0–\$19,999

3.4 Assess Your Understanding (page 152)

1. z-score 3. Quartiles

5. -0.30 ; -0.43 ; the 40-week gestation baby weighs less relative to the gestation period.

7. The man is relatively taller.

9. Jacob deGrom had the better year because his ERA was 2.47 standard deviations below the National League mean ERA, while Snell's ERA was only 2.07 standard deviations below the American League's mean ERA.

11. Ryan is 3.39 standard deviations below the mean in the 100 m back and 3.05 standard deviations below the mean in the 200 m back. So, Ryan is better at the 100 m race.

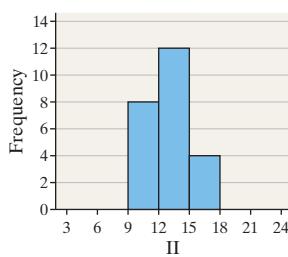
13. 239

15. (a) 15% of 3- to 5-month-old males have a head circumference that is 41.0 cm or less, and 85% of 3- to 5-month-old males have a head circumference that is greater than 41.0 cm.

(b) 90% of 2-year-old females have a waist circumference that is 52.7 cm or less, and 10% of 2-year-old females have a waist circumference that is more than 52.7 cm.

(c) The heights at each percentile decrease as the age increases. This implies that adult males are getting taller.

17. (a) 25% of the states have a violent crime rate that is 244.8 crimes per 100,000 population or less, and 75% of the states have a violent crime rate more than 244.8. 50% of the states have a violent crime rate that is 357.6 crimes per 100,000 population or less, while 50% of the states have a violent crime rate more than 357.6. 75% of the



states have a violent crime rate that is 454.8 crimes per 100,000 population or less, and 25% of the states have a violent crime rate more than 454.8.

(b) 210.0 crimes per 100,000 population; the middle 50% of all observations have a range of 210.0 crimes per 100,000 population.

(c) Yes

(d) Answers may vary. The difference between Q_1 and Q_2 is slightly more than the difference between Q_2 and Q_3 . However, the outlier in the right tail of the distribution implies that the distribution is skewed right.

19. (a) 50th percentile

(b) 90th percentile

(c) 105

21. (a) $z = -0.73$. The individual whose car got 36.3 miles per gallon was 0.73 standard deviation below the mean.

(b) By hand, TI-83 or 84, StatCrunch: $Q_1 = 36.85$ mpg, $Q_2 = 38.35$ mpg, $Q_3 = 41.0$ mpg; MINITAB: $Q_1 = 36.575$ mpg, $Q_2 = 38.35$ mpg, $Q_3 = 41.2$ mpg

(c) By hand, TI-83 or 84, StatCrunch:

IQR = 4.15 mpg; MINITAB: IQR = 4.625 mpg

(d) By hand, TI-83 or 84, StatCrunch: lower fence = 30.625 mpg, upper fence = 47.225 mpg.

Yes, 47.5 mpg is an outlier, MINITAB: lower fence = 29.6375 mpg, upper fence = 48.1375 mpg. There are no outliers using MINITAB's quartiles.

23. (a) By Hand, TI-83/84, StatCrunch: $Q_1: 5$, $Q_2: 8.5$, $Q_3: 18.5$;

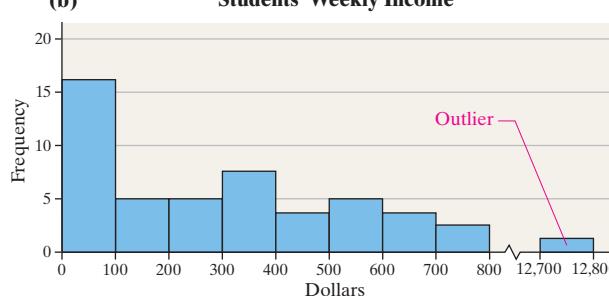
Minitab: $Q_1: 5$, $Q_2: 8.5$, $Q_3: 19.75$. Using the by-hand quartiles, 25% of the wait times are 5 minutes or less, and about 75% of the wait times exceed 5 minutes; 50% of the wait times are 8.5 minutes or less, and about 50% of the wait times exceed 8.5 minutes; 75% of the wait times are 18.5 minutes or less, and about 25% of the wait times exceed 18.5 minutes.

(b) 39, 44, and 52 are outliers

25. By hand, TI-83/84, StatCrunch: The cutoff point is 574 minutes. MINITAB: The cutoff point is 578 minutes. If more minutes are used, the customer is contacted.

27. (a) \$12,777 is an outlier.

Students' Weekly Income



(c) Answers will vary. One explanation: Student may have reported annual income.

29. (a) 0.5 mi; 1.85 mi; 4.7 mi

(b) IA: 5.07 mi; KS: 1.825 mi; Iowa

31. (a) $s = 58.0$ minutes; by hand, TI-83/84, StatCrunch: IQR = 56.5 minutes; MINITAB: IQR = 58.8 minutes

(b) $s = 115.0$ minutes; by hand, TI-83/84, StatCrunch: IQR = 56.5 minutes; MINITAB: IQR = 58.8 minutes. The standard deviation almost doubles in value, while the interquartile range is not affected. The standard deviation is not resistant to extreme observations, but the interquartile range is resistant.

33. Percentiles divide a data set into 100 parts. If there are more than 100 observations in a data set, it is possible for more than one observation to be in the top 1%.

35. No; when an outlier is discovered, it should be investigated to find its cause.

37. Comparing z -scores allows a unitless comparison of the number of standard deviations an observation is from the mean.

39. The first quartile is the 25th percentile, which means that 25% of the observations are less than or equal to the value and 75% of the observations are greater than the value. The second quartile is the 50th percentile, which means that 50% of the observations are less than or equal to the value and 50% of the observations are greater than the value. The third quartile is the 75th percentile, which means that 75% of the observations are less than or equal to the value and 25% of the observations are greater than the value.

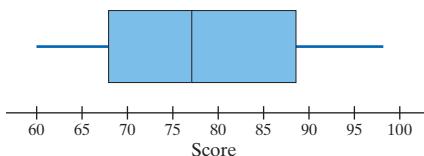
3.5 Assess Your Understanding (page 160)

1. The five-number summary consists of the minimum value in the data set, the first quartile, median, third quartile, and the maximum value in the data set.

- 3. (a)** Skewed right
(b) 0, 1, 3, 6, 16

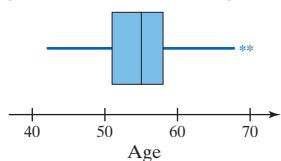
- 5. (a)** 40
(b) 52
(c) y
(d) Symmetric
(e) Skewed right

7. Statistics Exams Scores



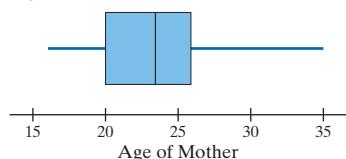
- 9. (a)** By hand, TI-83 or 84, StatCrunch: 42, 51, 55, 58, 70; MINITAB: 42, 50.5, 55, 59, 70

(b) Age of Presidents at Inauguration



- (c)** Symmetric with two outliers. Note: There are no outliers using Minitab's quantiles.

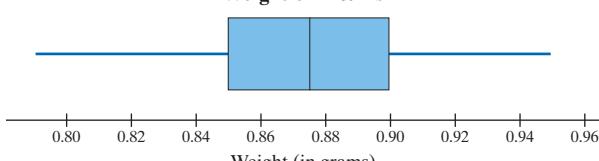
11. (a) Age of Mother at Time of First Birth



- (b)** Slightly skewed right

13.

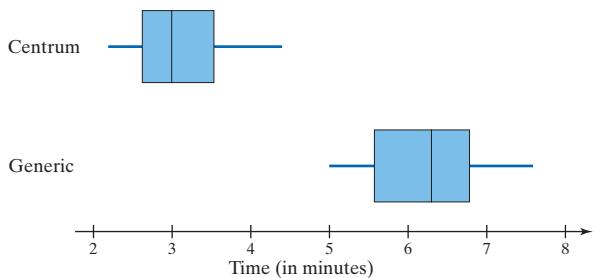
Weight of M&Ms



Since the range of the data between the minimum value and the median is roughly the same as the range between the median and the maximum value, and because the range of the data between the first quartile and median is the same as the range of the data between the median and third quartile, the distribution is symmetric.

15. (a)

Dissolving Time of Vitamins

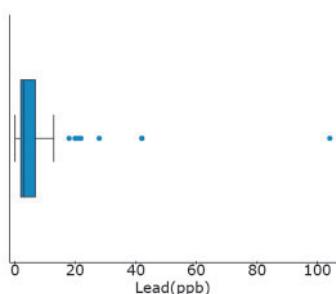


(b) Generic

(c) Centrum

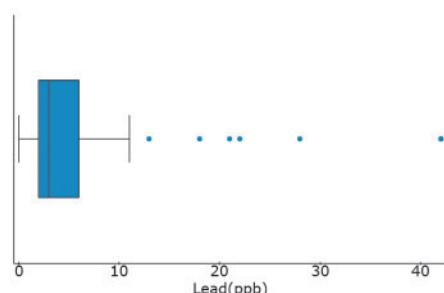
17. (a)

Lead Content in Water



- (b)** 8 out of 71 properties had lead readings above 15 ppb, which is 11.2% of the properties.

(c) Lead Content in Water (Two Observations Removed)



- (d)** 6 out of 69 properties had lead readings above 15 ppb, which is 8.7% of the properties.

(e) Answers may vary. Even though the removal of the two observations may be warranted from a purely statistical point of view (one had a filtration system and one was not a residential property), it is clear that investigation into the water supply was warranted, especially in light of the fact that the water source was changed just prior to the testing.

- 19. (a)** Answers may vary. Either a relative frequency bar graph or pie chart would answer the question.

(b) A relative histogram using a lower class limit of the first class of 0 and a class width of 10

(c) A frequency histogram with a lower class limit of the first class of 0 and a class width of 500

(d) A side-by-side boxplot by payment method

(e) Cash: $\bar{x} = \$11.216$; $M = \$8.25$; Credit card:

$$\bar{x} = \$17.203; M = \$8.375$$

(f) Cash: $s = \$8.514$; $IQR = \$5$; Credit card: $s = \$14.939$;

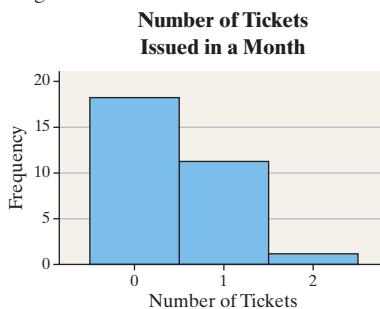
$IQR = \$22.625$. By both measures of dispersion, paying by credit card has more dispersion.

- 21.** (a) Completely randomized design
 (b) 30 subjects
 (c) Number of advantageous cards; quantitative; discrete
 (d) Hunger, impulsivity, BMI, age; Hunger; 2
 (e) In an attempt to create groups that were similar in terms of hunger, impulsivity, body mass index, and age, the subjects were randomly assigned to one of two treatment groups. In addition, the expectation is that randomization “evens out” the effect of any other explanatory variables not considered. The researchers verified that the impulsivity, ages, and body mass index of each group were not significantly different.
 (f) $\bar{x}_1 = 25.86$; $\bar{x}_2 = 33.36$
 (g) Hunger improves advantageous decision making.
- 23. Using the boxplot:** If the median is left of center in the box, and the right whisker is longer than the left whisker, the distribution is skewed right. If the median is in the center of the box, and the left and right whiskers are roughly the same length, the distribution is symmetric. If the median is right of center in the box, and the left whisker is longer than the right whisker, the distribution is skewed left.

Using the quartiles: If the distance from the median to the first quartile is less than the distance from the median to the third quartile, or the distance from the median to the minimum value in the data set is less than the distance from the median to the maximum value in the data set, then the distribution is skewed right. If the distance from the median to the first quartile is the same as the distance from the median to the third quartile, or the distance from the median to the minimum value in the data set is the same as the distance from the median to the maximum value in the data set, the distribution is symmetric. If the distance from the median to the first quartile is more than the distance from the median to the third quartile, or the distance from the median to the minimum value in the data set is more than the distance from the median to the maximum value in the data set, the distribution is skewed left.

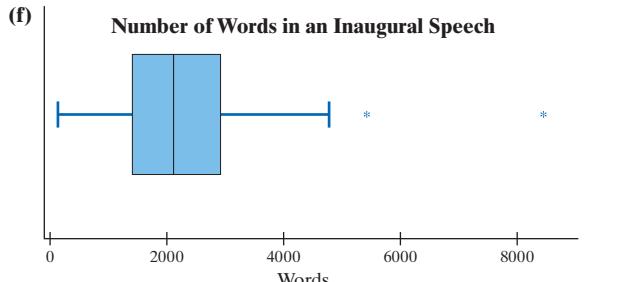
Chapter 3 Review Exercises (page 166)

- 1.** (a) Mean = 792.51 m/sec, median = 792.40 m/sec
 (b) Range = 4.8 m/sec, $s^2 = 2.03$, $s = 1.42$ m/sec
- 2.** (a) $\bar{x} = \$10,178.9$, $M = \$9980$
 (b) Range: $R = \$8550$, $s = \$3074.9$, IQR = $\$5954.5$; using MINITAB: IQR = $\$5954$; using StatCrunch: IQR = $\$5110$
 (c) $\bar{x} = \$13,178.9$, $M = \$9980$; range: $R = \$35,550$, $s = \$10,797.5$, IQR = $\$5954.5$ [using MINITAB: $\$5954$]. The median and interquartile range are resistant.
- 3.** (a) $\mu = 57.8$ years, $M = 58.0$ years, bimodal: 56 and 62
 (b) Range = 25 years, $\sigma = 7.0$ years
 (c) Answers will vary.
- 4.** (a) Skewed right



- (b)** Mean greater than median
 (c) $\bar{x} = 0.4$, $M = 0$
 (d) 0
- 5.** (a) 441; 759
 (b) 95% of the light bulbs have a life between 494 and 706 hours.
 (c) 81.5% of the light bulbs have a life between 547 and 706 hours.
 (d) The firm can expect to replace 0.15% of the light bulbs.
 (e) At least 84% of the light bulbs have a page count within 2.5 standard deviations of the mean.

- (f)** At least 75% of the light bulbs have a life between 494 and 706 hours.
- 6.** (a) $\bar{x} = 27.7$ minutes
 (b) $s = 19.1$ minutes
 7. 3.33
 8. Female, since her weight is more standard deviations above the mean
- 9.** (a) Two-seam fastball
 (b) Two-seam fastball
 (c) Four-seam fastball
 (d) 88 mph
 (e) Symmetric
 (f) Skewed right
- 10.** (a) $\mu = 2339$ words; $M = 2133.5$ words (Minitab: $M = 2134$ words)
 (b) $Q_1 = 1425$ words; $Q_2 = M = 2133.5$ words; $Q_3 = 2906$ words; MINITAB: $Q_1 = 1408$ words; $Q_2 = M = 2134$ words; $Q_3 = 2924$ words
 (c) 135, 1425, 2133.5, 2906, 8445; MINITAB: 135, 1408, 2134, 2924, 8445
 (d) $\sigma = 1374.5$ words; IQR = 1481 words [MINITAB: 1516 words]
 (e) Lower Fence = -796.5 words [MINITAB: -866]; Upper Fence = 5127.5 words [MINITAB: 5198]; Outliers: 5433 and 8445

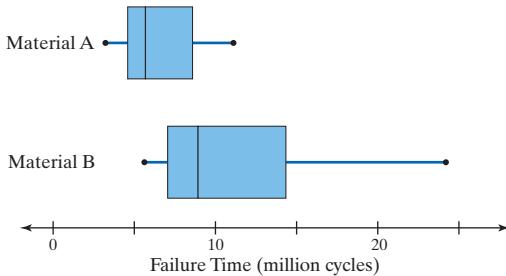


- (g)** Slightly skewed right since the right whisker is longer than the left whisker and considering the outliers.
(h) The median is the better measure since the outliers inflate the value of the mean.
(i) The interquartile range is the better measure since the outliers inflate the value of the standard deviation.
- 11.** 85% of 19-year-old females have a height that is 67.1 inches or less, and 15% of 19-year-old females have a height that is more than 67.1 inches.
- 12.** The median is used for three measures since it is likely the case that one of the three measures is extreme relative to the other two, thus substantially affecting the value of the mean. Since the median is resistant to extreme values, it is the better measure of central tendency.

Chapter 3 Test (page 167)

- 1.** (a) $\bar{x} = 80$ min
 (b) $M = 79.5$ min
 (c) $\bar{x} = 192.5$ min, $M = 79.5$ min; the median is resistant.
- 2.** From motor vehicles
3. 63 min
4. (a) 23.8 min
 (b) 41 min; the middle 50% of all study times has a range of 41 min.
 (c) The interquartile range is resistant; the standard deviation is not resistant.
- 5.** (a) 3282; 5322
 (b) 95% of the cartridges print between 3622 and 4982 pages.
 (c) The firm can expect to replace 2.5% of the cartridges.
 (d) At least 55.6% of the cartridges have a page count within 1.5 standard deviations of the mean.
 (e) At least 88.9% of the cartridges print between 3282 and 5322 pages.
- 6.** (a) 74.9 minutes
 (b) 14.7 minutes
7. \$2.17

8. (a) Material A: $\bar{x} = 6.404$ million cycles
Material B: $\bar{x} = 11.332$ million cycles
(b) Material A: $M = 5.785$ million cycles
Material B: $M = 8.925$ million cycles
(c) Material A: $s = 2.626$ million cycles
Material B: $s = 5.900$ million cycles
Material B is more dispersed.
(d) Material A (in million cycles): 3.17, 4.52, 5.785, 8.01, 11.92
Material B (in million cycles): 5.78, 6.84, 8.925, 14.71, 24.37

(e) **Bearing Failures**

Answers will vary.

(f) The distributions are skewed right.

9. Quartiles by hand, TI-83/84 (in grams): $Q_1 = 5.58$, $Q_2 = 5.60$, $Q_3 = 5.66$; quartiles by MINITAB (in grams): $Q_1 = 5.58$, $Q_2 = 5.60$, $Q_3 = 5.6625$. Using the by-hand quartiles: 25% of quarters have a weight that is 5.58 grams or less, 75% of the quarters have a weight more than 5.58 grams; 50% of the quarters have a weight that is 5.60 grams or less, 50% of the quarters have a weight more than 5.60 grams; 75% of the quarters have a weight that is 5.66 grams or less, 25% of the quarters have a weight that is more than 5.66 grams. The quarter whose weight is 5.84 grams is an outlier.

10. Armando should report his ACT math score since it is more standard deviations above the mean.

11. 15% of 10-year-old males have a height that is 53.5 inches or less, and 85% of 10-year-old males have a height that is more than 53.5 inches.

12. The median will be less than the mean for income data, so you should report the median.

13. (a) Report the mean since the distribution is symmetric.
(b) Histogram I has more dispersion. The range of classes is larger.

14. The standard deviation can be thought of as a measure of the deviations from the mean. We find the standard deviation by squaring the deviations about the mean because, otherwise, the deviations below the mean would offset deviations above the mean. The further a particular observation is from the mean, the higher the squared deviation—this is what causes data sets with more dispersion to have a higher standard deviation. Of course, to “undo” the squaring of the deviations, we need to take the square root after we add the squared deviations and divide by N (for population standard deviation) or $n - 1$ (for sample standard deviations).

CHAPTER 4 Describing the Relation between Two Variables**4.1 Assess Your Understanding (page 180)**

1. Univariate data measure the value of a single variable for each individual in the study. Bivariate data measure values of two variables for each individual.
3. scatter diagram
5. 1. The linear correlation coefficient is always between -1 and 1 , inclusive. That is, $-1 \leq r \leq 1$.
2. If $r = +1$, then a perfect positive linear relation exists between the two variables.
3. If $r = -1$, then a perfect negative linear relation exists between the two variables.
4. The closer r is to $+1$, the stronger is the evidence of positive association between the two variables.

5. The closer r is to -1 , the stronger is the evidence of negative association between the two variables.
6. If r is close to 0 , then little or no evidence exists of a linear relation between the two variables. So, r close to 0 does not imply no relation, just no linear relation.
7. The linear correlation coefficient is a unitless measure of association. So, the unit of measure for x and y plays no role in the interpretation of r .
8. The correlation coefficient is not resistant. Therefore, an observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.

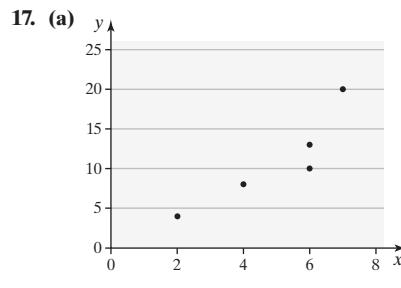
7. lurking

9. Nonlinear

11. Linear, positive

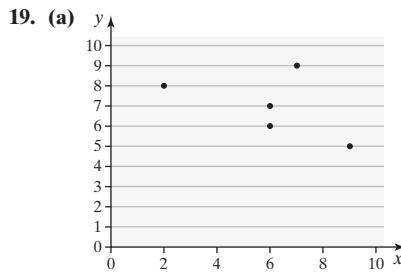
13. (a) III (b) IV
(c) II (d) I

15. (a) Linear; positive association
(b) The point $(59, 91000)$ appears to stick out. Reasons may vary. One explanation is the high concentration of government jobs that require a bachelor's degree, and the high proportion of lobbyists and attorneys who live in D.C.
(c) Yes, because $|0.760| > 0.361$ (critical value from Table II with $n = 30$).



- (b) $r = 0.896$

- (c) Linear relation



- (b) $r = -0.496$

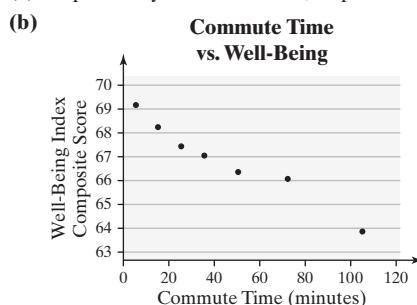
- (c) No linear relation

21. (a) Positive (b) Negative
(c) Negative (d) Negative
(e) No correlation

23. 0.137, 0.339, -0.431 , 0.869, -0.903

25. Yes; $0.79 > 0.361$ (critical r for $n = 30$), so countries in which students answered a greater percentage of items in the background questionnaire tended to have higher mean scores on the TIMSS exam.

27. (a) Explanatory: commute time; response: well-being score

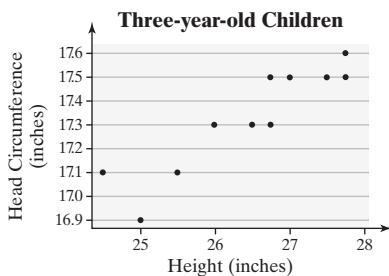


- (c) -0.981

(d) $| -0.981 | = 0.981 > 0.754$ (critical r from Appendix Table II), so a negative association exists between commute time and well-being index score.

29. (a) Explanatory: height; response: head circumference

(b)



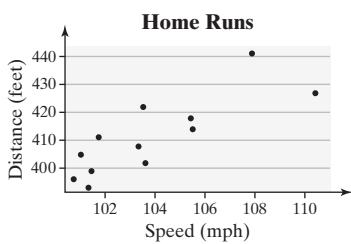
(c) $r = 0.911$

(d) Yes, positive association because $0.911 > 0.602$ (critical r from Table II).

(e) Converting to cm has no effect on the linear correlation coefficient.

31. (a) Speed

(b)

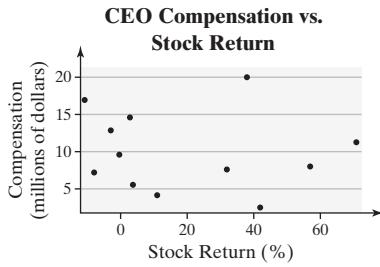


(c) $r = 0.823$

(d) Yes, because $|0.823| > 0.576$ (Table II), there is a linear relation between speed and distance.

33. (a) Stock return

(b)



(c) -0.103

(d) No, because $| -0.103 | < 0.576$, there is no linear relation between stock return and compensation.

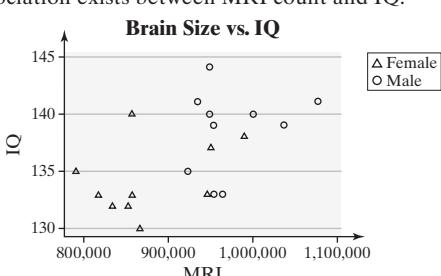
35. (a)

(b)



(b) $r = 0.548$. Because $0.548 > 0.444$ (Table II), a positive association exists between MRI count and IQ.

(c)

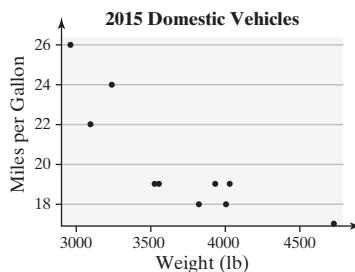


The females have lower MRI counts. Looking at each gender's plot, we see that the relation that appeared to exist between brain size and IQ disappears.

(d) Females: $r = 0.359$; males: $r = 0.236$. No linear relation exists between brain size and IQ.

37. (a) Explanatory: weight; response: miles per gallon

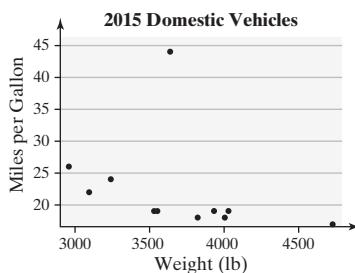
(b)



(c) $r = -0.842$

(d) Yes; $| -0.842 | > 0.632$ (Table II). There is a negative association between weight of a car and miles per gallon.

(e)

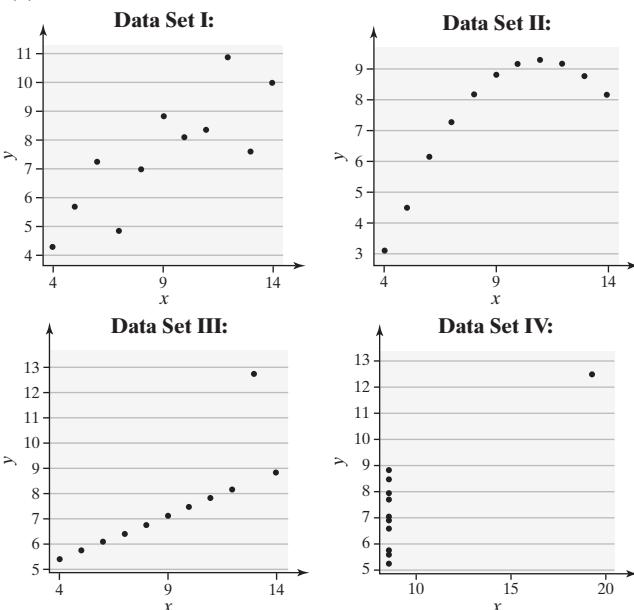


(f) Correlation coefficient (with Fusion included): $r = -0.331$

(g) The Fusion is a hybrid car; the other cars are not hybrids.

39. (a) 1: 0.816; 2: 0.817; 3: 0.816; 4: 0.817

(b)

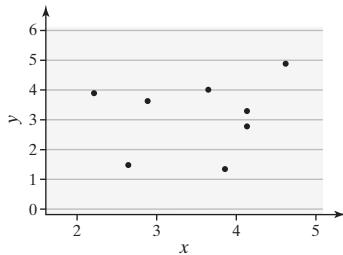


41. First Energy and Walt Disney have the lowest correlation (0.008). First Energy and Cisco Systems are negatively correlated (-0.182).

43. $r = 0.599$ implies that a positive linear relation exists between the number of television stations and life expectancy, but this is correlation, not causation. The more television stations a country has, the more affluent it is. The more affluent, the better the health care, so wealth is a likely lurking variable.

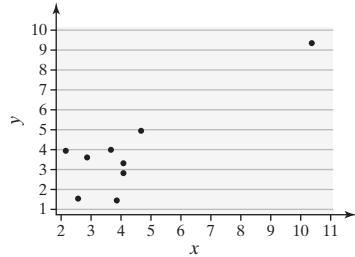
45. No; a likely lurking variable is the economy. In a strong economy, crime rates tend to decrease, and consumers are better able to afford cell phones.

47. (a)



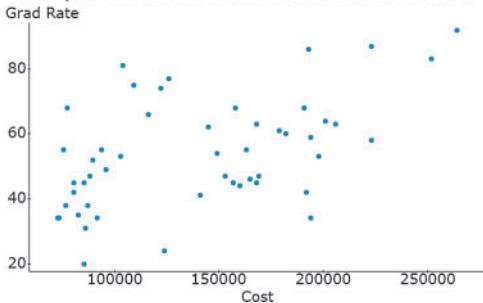
$$r = 0.228$$

(b)

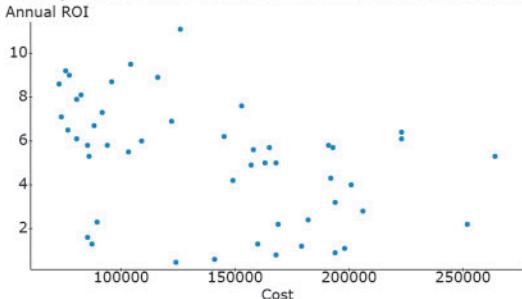


$r = 0.860$; A single point may affect the value of the correlation coefficient making it seem as if a strong linear relation exists between two variables, when it does not.

49. (a) The correlation between Cost and Graduation Rate is 0.513. The scatter diagram is in the next column. There is a positive association between cost and graduation rates for United States colleges and universities—as costs go up, so do graduation rates.

Four-year Cost and Graduation Rate of U.S. Universities


- (b) The correlation between Cost and Return on Investment is -0.420 . The scatter diagram is below. There is a negative association between cost and return on investment for United States colleges and universities—as costs go up, return on investment goes down.

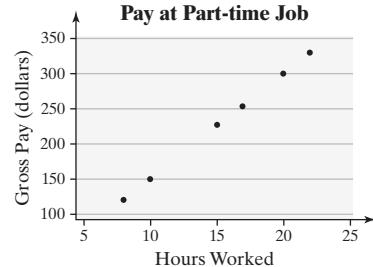
Four-year Cost and Return on Investment of U.S. Universities


51. If the correlation coefficient equals 1, then a perfect positive linear relation exists between the variables, and the points of the scatter diagram lie exactly on a straight line with a positive slope.

53. The linear correlation coefficient can only be calculated from bivariate quantitative data, and the gender of a driver is a qualitative variable. A possible better way to write the sentence: Gender is associated with the rate of automobile accidents.

55. Correlation describes a relation between two variables in an observational study. Causation describes a conditional (if—then) relation in an experimental study.

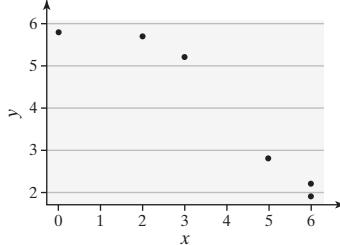
57. Deterministic


4.2 Assess Your Understanding (page 195)

1. residual

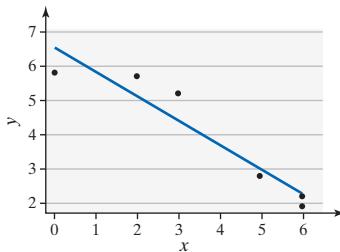
3. (a), (c), (e), (h), (i)

5. (a)

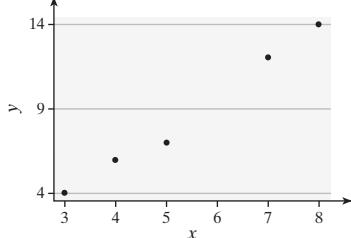


$$(b) \hat{y} = -0.7136x + 6.5499$$

(c)

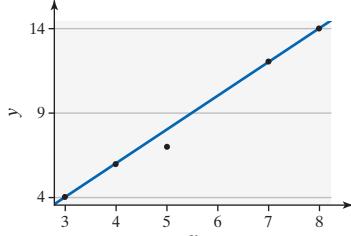


7. (a)



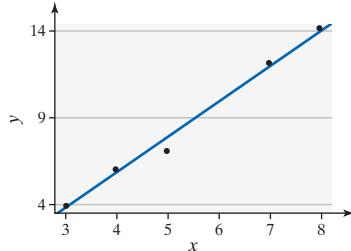
$$(b) \text{Using points } (3, 4) \text{ and } (8, 14): \hat{y} = 2x - 2$$

(c)



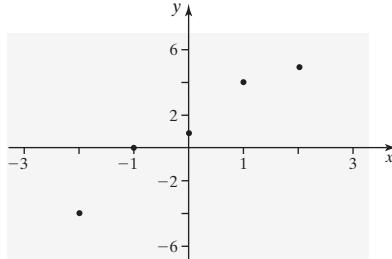
$$(d) \hat{y} = 2.0233x - 2.3256$$

(e)



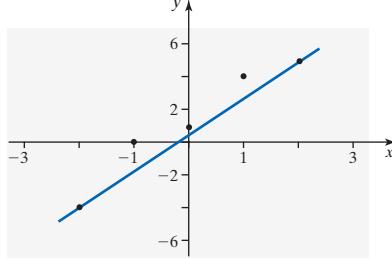
- (f) Sum of squared residuals (computed line): 1
 (g) Sum of squared residuals (least-squares line): 0.7907
 (h) Answers will vary.

9. (a)



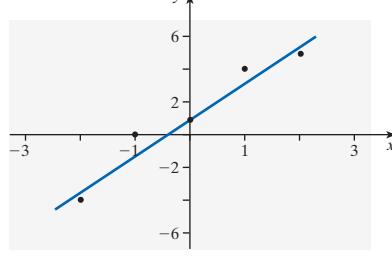
(b) Using points $(-2, -4)$ and $(2, 5)$: $\hat{y} = \frac{9}{4}x + \frac{1}{2} = 2.25x + 0.5$

(c)



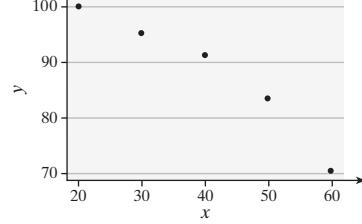
(d) $\hat{y} = 2.2x + 1.2$

(e)



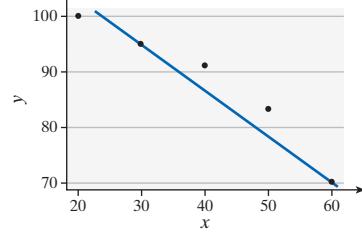
- (f) Sum of squared residuals (computed line): 4.875
 (g) Sum of squared residuals from least-squares line: 2.4
 (h) Answers will vary.

11. (a)



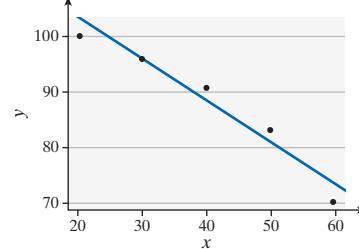
(b) Using points $(30, 95)$ and $(60, 70)$: $\hat{y} = -\frac{5}{6}x + 120$

(c)



(d) $\hat{y} = -0.72x + 116.6$

(e)



- (f) Sum of squared residuals (computed line): 51.6667
 (g) Sum of squared residuals from least-squares line: 32.4
 (h) Answers will vary.

13. (a) \$59,530

(b) Higher because the predicted income for a state with 28% of adults with at least a bachelor's degree is \$62,839 and North Dakota's median income is \$66,321.

(c) If the percentage of the population with at least a bachelor's degree increases 1%, median income increases \$1103, on average.
 (d) There are no observations near 0%, so this is outside the scope of the model.

15. (a) If literacy rate increases 1%, the age difference decreases 0.0527 year, on average.

(b) No. A 0% literacy rate is outside the scope of the model.

(c) 5.8 years

(d) No. Outside the scope of the model.

(e) The average age difference for a literacy rate of 99% is 1.88 years, so the United States has an age difference slightly above average.

17. (a) $\hat{y} = -0.0479x + 69.0296$

(b) Slope: for each 1-minute increase in commute time, the index score decreases by 0.0479, on average; y-intercept: The average index score for a person with a 0-minute commute is 69.0296. A person who works from home will have a 0-minute commute, so the y-intercept has a meaningful interpretation.

(c) 67.6

(d) Barbara's score is less than the mean score for people with a 20-minute commute, 68.1, so she is less well off.

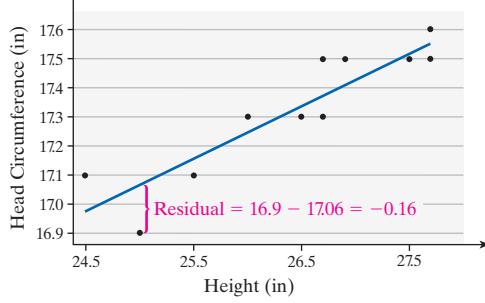
19. (a) $\hat{y} = 0.1827x + 12.4932$

(b) If height increases by 1 inch, head circumference increases by about 0.1827 inch, on average. It is not appropriate to interpret the y-intercept. It is outside the scope of the model.

(c) $\hat{y} = 17.06$ inches

(d) Residual = -0.16 inch; below

(e)



- (f) For children who are 26.75 inches tall, head circumference varies.
 (g) No; 32 inches is outside the scope of the model.

21. (a) $\hat{y} = 3.8170x + 15.0942$

(b) If the speed at which a ball is hit increases 1 mile per hour, the distance of a home run increases by 3.8170 feet, on average. It does not make sense to interpret the y-intercept because a home run hit with a speed of 0 mph makes no sense.

(c) 415.9 feet

(d) 415.9 feet

(e) We would expect the ball to travel 420.5 feet, so Yelich's home run was shorter than expected.

(f) No. This is outside the scope of the model.

23. (a) $\hat{y} = -0.0029x + 0.8861$

(b) For each additional cola consumed per week, bone mineral density will decrease by 0.0029 g/cm^2 , on average.
 (c) For a woman who does not drink cola, the mean bone mineral density will be 0.8861 g/cm^2 .

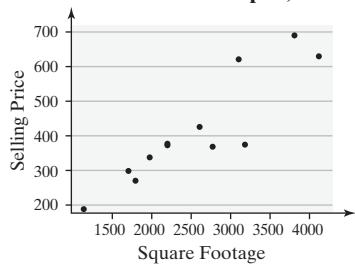
(d) The predicted bone mineral density is 0.8745 g/cm^2 .
 (e) This bone mineral density is below average for women who consume 4 cans of cola per week.
 (f) No; 2 cans of cola per day equates to 14 cans per week, which is outside the scope of the model.

25. In Problem 33 from Section 4.1, we determined that the stock return and CEO performance are not linearly related. Therefore, it does not make sense to find a least-squares regression model of the form $\hat{y} = b_1x + b_0$.

27. (a) Males: $\hat{y} = 0.3428x + 998.4488$ Females: $\hat{y} = 0.1045x + 514.1520$

(b) Males: If the number of licensed drivers increases by 1 (thousand), then the number of fatal crashes increases by 0.3428, on average. Females: If the number of licensed drivers increases by 1 (thousand), then the number of fatal crashes increases by 0.1045, on average. Since females tend to be involved in fewer fatal crashes, an insurance company may use this information to argue for higher rates for male customers.
 (c) Above average; above average; below average. An insurance company may use this information to argue for higher rates for younger drivers and lower rates for older drivers. The same relationship holds for females.

29. (a) Square footage

(b) **Homes Sold in Naples, FL**(c) $r = 0.903$ (d) Yes, because $0.903 > 0.576$ (Table II).(e) $\hat{y} = 0.1563x + 14.8740$

(f) For each square foot added to the area, the selling price of the house will increase by \$156.3 (that is, 0.1563 thousand dollars), on average.

(g) It is not reasonable to interpret the intercept.

(h) Above average; factors that could affect the price include location or number of bedrooms.

31. A residual is the difference between the observed value of y and the predicted value of y . If the residual is positive, the observed value is above average for the given value of the explanatory variable.

32. Values of the explanatory variable that are much larger or much smaller than those observed are considered *outside the scope of the model*. It is dangerous to make such predictions because we do not know the behavior of the data for which we have no observations.

33. Answers will vary.

35. Answers may vary. Discussions should involve the scope of the model and the idea of being “outside the scope.”

4.3 Assess Your Understanding (page 204)

1. coefficient of determination

3. (a) III (b) II (c) IV (d) I

5. 83.0% of the variation in the length of eruption is explained by the least-squares regression equation.

7. (a) $R^2 = 96.1\%$

(b) 96.1% of the variation in the well-being index composite score can be explained by the least-squares regression equation. The linear model appears to be appropriate, based on the residual plot.

9. (a) $R^2 = 83.0\%$

(b) 83.0% of the variation in head circumference is explained by the least-squares regression equation. The linear model appears to be appropriate, based on the residual plot.

11. (a) 67.7%

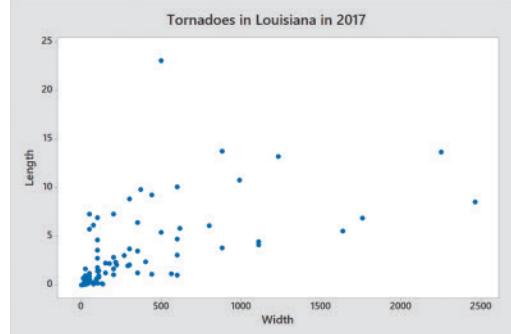
(b) 67.7% of the variability in home run distance is explained by the least-squares regression model (or speed). The residual plot does not show a discernable pattern and there are not any outliers or potentially influential observations.

13. The coefficient of determination with the Viper included is $R^2 = 23.0\%$. Adding the Viper reduces the amount of variability explained by the model by approximately 47.8%.

15. (a) Width

(b) For each tornado, two variables are measured: width and length. Both of these variables are quantitative.

(c)



(d) 0.579

(e) Yes, because $0.579 > 0.361$ (for $n = 30$ in Table II).(f) $\hat{y} = 0.0049x + 1.628$

(g) 4.078 miles

(h) Yes, it was expected to last 4.568 miles.

(i) If the width of the tornado increases by 1 yard, the distance for which the tornado is on the ground is expected to increase by 0.0049 mile.

(j) It does not make sense to talk about a tornado whose width is 0 yards.

(k) 33.5%

17. Answers will vary. The discussion should include a scatter diagram of the data, the correlation coefficient, and least-squares regression. In addition, an interpretation of the slope would be helpful because it shows the impact the cigarette taxes have on smuggling. In addition, the discussion should include the state of New Hampshire as an outlier. This is likely due to its proximity to high-tax states of New York and Massachusetts.

4.4 Assess Your Understanding (page 215)

1. A marginal distribution is a frequency or relative frequency distribution of either the row or column variable in a contingency table. A conditional distribution is the relative frequency of each category of one variable, given a specific value of the other variable in a contingency table.

3. Correlation is used with quantitative variables; association is used with categorical variables.

5. (a)

	x_1	x_2	x_3	Marginal Distribution
y_1	20	25	30	75
y_2	30	25	50	105
Marginal Distribution	50	50	80	180

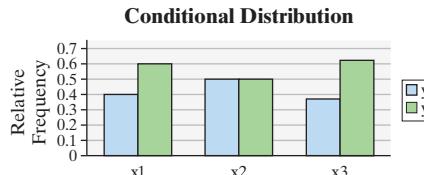
(b)

	x_1	x_2	x_3	Relative Frequency Marginal Distribution
y_1	20	25	30	0.417
y_2	30	25	50	0.583
Relative Frequency Marginal Distribution	0.278	0.278	0.444	1

(c)

	x_1	x_2	x_3
y_1	0.400	0.500	0.375
y_2	0.600	0.500	0.625
Total	1	1	1

(d)



7. (a) 2160 adult Americans were surveyed; 536 were 55 and older.

(b)

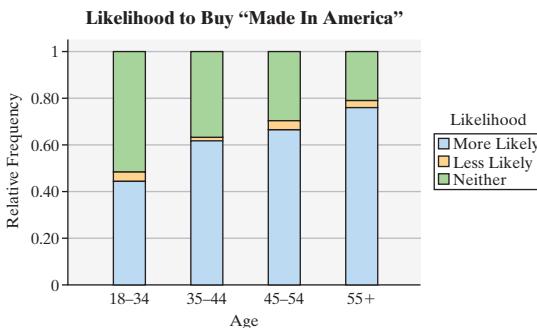
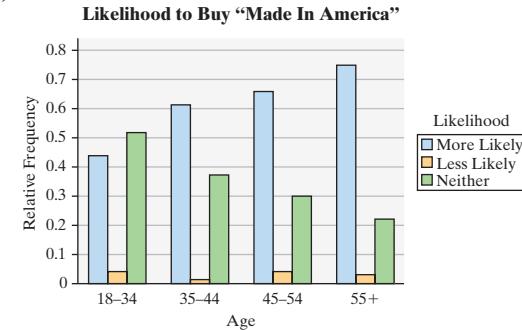
Likely to Buy	Age				Relative Frequency Marginal Distribution
	18–34	35–44	45–54	55 +	
More likely	238	329	360	402	0.615
Less likely	22	6	22	16	0.031
Neither	282	201	164	118	0.354
Relative Frequency Marginal Distribution	0.251	0.248	0.253	0.248	1

(c) 0.615

(d)

Likely to Buy	Age			
	18–34	35–44	45–54	55 +
More likely	0.439	0.614	0.659	0.750
Less likely	0.041	0.011	0.040	0.030
Neither	0.520	0.375	0.300	0.220
Total	1	1	1	1

(e)



(f) The number of people more likely to buy a product because it is made in America increases with age. On the other hand, age does not seem to be a significant factor in whether a person is less likely to buy a product because it is made in America.

9. (a)

Household Income in College	Bottom Quartile	2nd Quartile	3rd Quartile	Top Quartile	Totals
Bottom Quartile	72	62	46	40	220
2nd or 3rd Quartile	70	78	75	67	290
Top Quartile	32	47	52	84	215
Totals	174	187	173	191	725

(b)

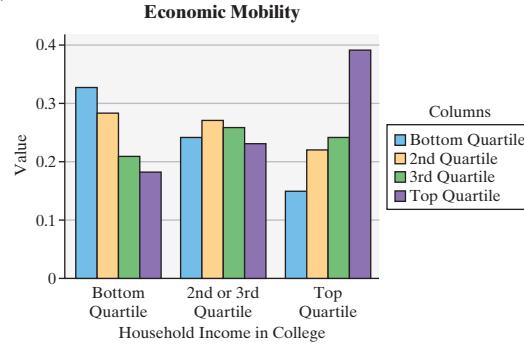
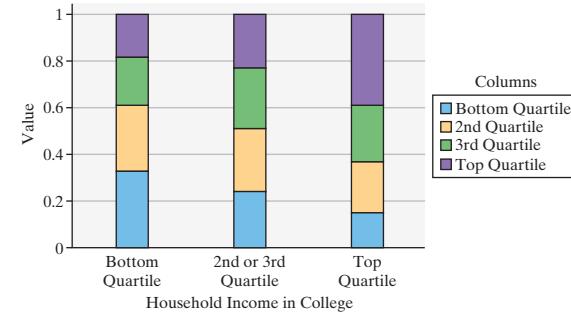
Household Income in College	Bottom Quartile	2nd Quartile	3rd Quartile	Top Quartile	Totals
Bottom Quartile	72	62	46	40	0.303
2nd or 3rd Quartile	70	78	75	67	0.400
Top Quartile	32	47	52	84	0.297
Totals	0.240	0.258	0.239	0.263	1

(c) 0.258

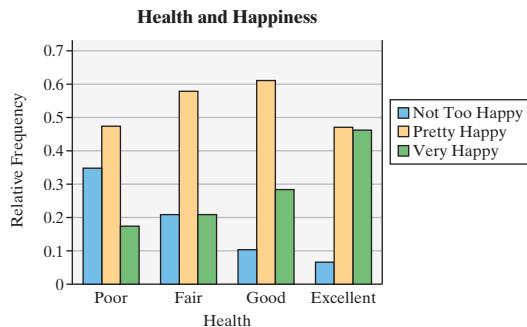
(d)

Household Income in College	Bottom Quartile	2nd Quartile	3rd Quartile	Top Quartile
Bottom Quartile	0.327	0.282	0.209	0.182
2nd or 3rd Quartile	0.241	0.269	0.259	0.231
Top Quartile	0.149	0.219	0.242	0.391

(e)

**Economic Mobility**

(f) If an individual is in the bottom quartile, he or she is more likely to remain in that quartile; if an individual is in the top quartile, he or she is more likely to remain in that quartile.

11.


Based on the conditional distribution by health, we can see that healthier people tend to be happier. As health increases, the percent who are very happy increases, while the percent who are not happy decreases.

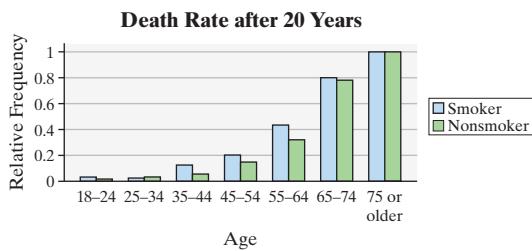
- 13. (a)** Smokers: 0.239; nonsmokers: 0.314; this implies that it is healthier to smoke.

(b) Smokers: 0.036; Nonsmokers: 0.016

(c) Proportion deceased after 20 years by smoking status and age

Age							
							75 or Older
	18–24	25–34	35–44	45–54	55–64	65–74	
Smoker	0.036	0.024	0.128	0.208	0.443	0.806	1
Nonsmoker	0.016	0.032	0.058	0.154	0.331	0.783	1

(d)

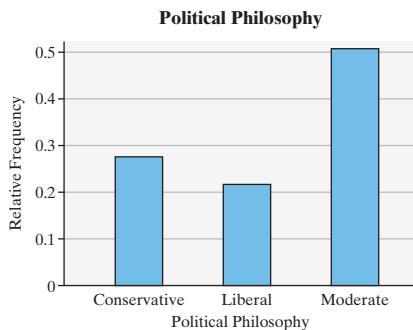


(e) Answers may vary. When taking age into account, the direction of association changed. In almost all age groups, smokers had a higher death rate than nonsmokers. The most notable exception is for the 25 to 34 age group, the largest age group for the nonsmokers. A possible explanation could be rigorous physical activity (e.g., rock climbing) that nonsmokers are more likely to participate in than smokers.

- 15. (a)** Moderates make up 50.75% of respondents.

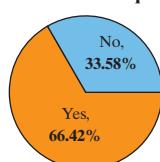
Political Philosophy	Relative Frequency
Conservative	0.2761
Liberal	0.2164
Moderate	0.5075

(b)



- (c)** It appears that a majority of the respondents believe there is gender income inequality.

Do You Believe There is Gender Income Inequality?



(d)

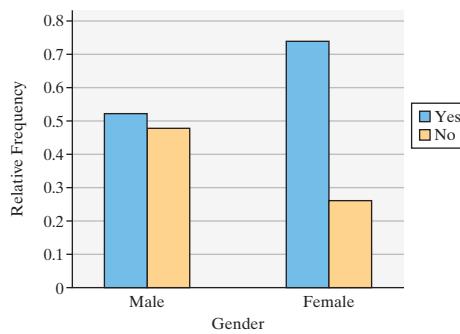
Gender	Yes	No	Relative Frequency Marginal Distribution
Male	24	22	0.343
Female	65	23	0.657
Relative Frequency Marginal Distribution	0.664	0.336	

- (e)** The conditional distribution by gender shows there is a gender gap when it comes to the belief there is gender income inequality.

Gender	Yes	No
Male	0.522	0.478
Female	0.739	0.261

(f)

Do You Believe There Is an Income Inequality Discrepancy between Males and Females?



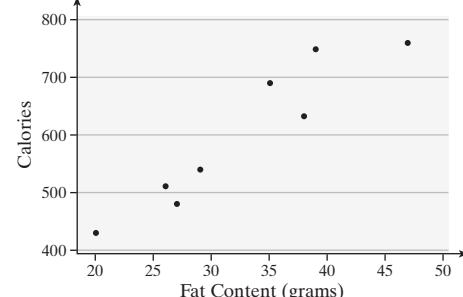
Chapter 4 Review Exercises (page 218)

- (a)** If the home team is favored by 3 points, the winning margin predicted by the regression equation is 3.009 points.
(b) If the visiting team is favored by 7 points, the winning margin (for the visiting team) predicted by the regression equation is 7.061 points.
(c) For each 1-point increase in the spread, the winning margin increases by 1.007 points, on average.
(d) If the spread is 0, the home team is expected to lose by 0.012 point, on average.
(e) 39% of the variation in winning margins can be explained by the least-squares regression equation.

- 2. (a)** Fat content

(b)

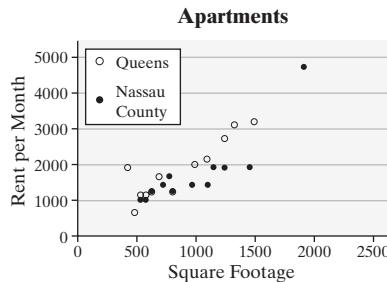
Calories vs. Fat Content



(c) $r = 0.944$

(d) Yes; a strong linear relation exists between fat content and calories in fast-food restaurant sandwiches because $0.944 > 0.707$ (Table II).

3. (a)



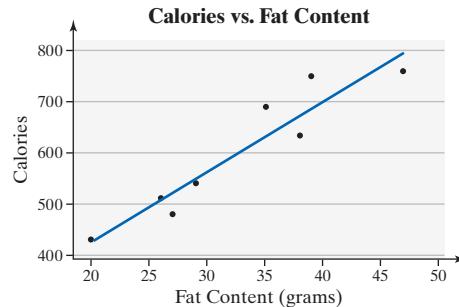
(b) Queens: $r = 0.909$; Nassau County: $r = 0.867$

(c) Both locations appear to have a positive linear association between square footage and monthly rent because $r > 0.576$ (Table II) for each location.

(d) For small apartments (those less than 1000 square feet in area), there seems to be no difference in rent between Queens and Nassau County. In larger apartments, Queens seems to have higher rents than Nassau County.

4. (a) $\hat{y} = 13.7334x + 150.9469$

(b)



(c) The slope indicates that each additional gram of fat in a sandwich adds approximately 13.73 calories, on average. The y -intercept indicates that a sandwich with no fat will contain about 151 calories.

(d) 562.9 calories

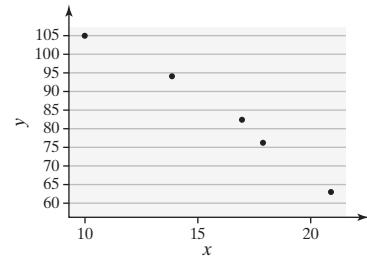
(e) Below average (the average value is 727.7 [Tech: 727.8] calories)

5. (a) $\hat{y} = 2.2091x - 34.3148$

(b) The slope of the least-squares regression line indicates that, for each additional square foot of floor area, the rent increases by \$2.21, on average. It is not appropriate to interpret the y -intercept since it is not possible to have an apartment with 0 square footage.

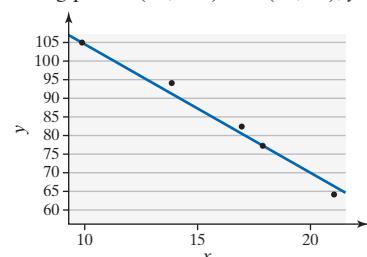
(c) This apartment's rent is below average.

6. (a)



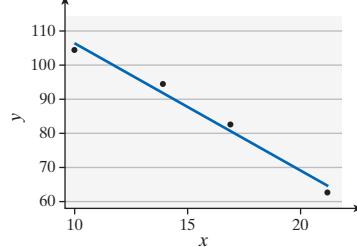
(b) Using points $(10, 105)$ and $(18, 76)$, $\hat{y} = -\frac{29}{8}x + \frac{565}{4}$.

(c)



(d) $\hat{y} = -3.8429x + 145.4857$

(e)



(f) Computed line: sum of squared residuals = 22.406

(g) Least-squares line: sum of squared residuals = 16.271

(h) Answers will vary.

7. $R^2 = 89.1\%$; 89.1% of the variation in calories is explained by the least-squares regression line.

8. $R^2 = 82.7\%$; 82.7% of the variance in rent is explained by the least-squares regression model.

9. No; correlation does not imply causation. Florida has a large number of tourists in warmer months, times when more people will be in the water to cool off. The larger number of people in the water splashing around leads to a larger number of shark attacks.

10. (a) 203

(b)

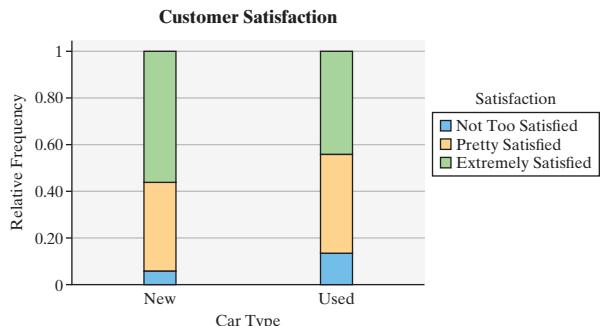
		Car Type		Relative Frequency Marginal Distribution
Satisfaction		New	Used	
Not too satisfied		11	25	0.091
Pretty satisfied		78	79	0.396
Extremely satisfied		118	85	0.513
Relative Frequency Marginal Distribution		0.523	0.477	1

(c) 0.513

(d)

		Car Type		Relative Frequency Marginal Distribution
Satisfaction		New	Used	
Not too satisfied		0.053	0.132	
Pretty satisfied		0.377	0.418	
Extremely satisfied		0.570	0.450	
Total		1	1	

(e)



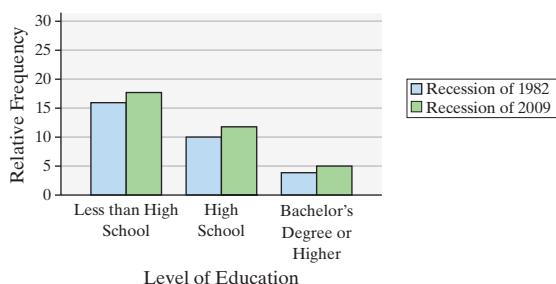
(f) There appears to be some association. Buyers of used cars are more likely to be dissatisfied and are less likely to be extremely satisfied than buyers of new cars.

11. (a) In 1982, the unemployment rate was approximately 10.3%. In 2009, the unemployment rate was approximately 10.0%. The 1982 recession appears to be worse.

- (b) The unemployment rates are displayed in the table.

Recession	Level of Education		
	Less than High School	High School	Bachelor's Degree or Higher
Recession of 1982	16.1%	10.2%	3.8%
Recession of 2009	16.7%	11.8%	4.8%

(c) Unemployment by Level of Education



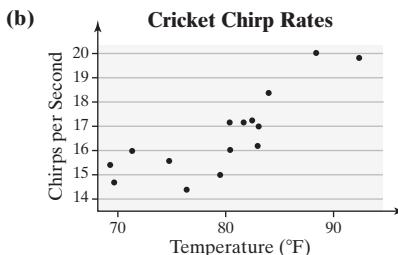
(d) Answer will vary. The discussion should include the observation that, although the overall unemployment rate was higher in 1982, the unemployment rate within each level of education was higher in 2009.

12. (a) A positive linear relation appears to exist between number of marriages and number unemployed.
 (b) Population is highly correlated with both the number of marriages and the number unemployed. The size of the population affects both variables.
 (c) No association exists between the two variables.
 (d) Answers may vary. A strong correlation between two variables may be due to a third variable that is highly correlated with the two original variables.
13. The eight properties of a linear correlation coefficient are the following:
- The linear correlation coefficient is always between -1 and 1 , inclusive. That is, $-1 \leq r \leq 1$.
 - If $r = +1$, there is a perfect positive linear relation between the two variables. See Figure 4(a) on page 174.
 - If $r = -1$, there is a perfect negative linear relation between the two variables. See Figure 4(d).
 - The closer r is to $+1$, the stronger is the evidence of positive association between the two variables. See Figures 4(b) and 4(c).
 - The closer r is to -1 , the stronger is the evidence of negative association between the two variables. See Figures 4(e) and 4(f).
 - If r is close to 0 , there is no evidence of a *linear* relation between the two variables. Because the linear correlation coefficient is a measure of the strength of the linear relation, r close to 0 does not imply no relation, just no linear relation. See Figures 4(g) and 4(h).
 - The linear correlation coefficient is a unitless measure of association. So the unit of measure for x and y plays no role in the interpretation of r .
 - The correlation coefficient is not resistant.

14. (a) Answers will vary.
 (b) The slope can be interpreted as “the school day decreases by 0.01 hour for each 1% increase in percent low income, on average.” The y -intercept can be interpreted as the length of the school day when 0% of the population is low income.
 (c) $\hat{y} = 6.91$ hours
 (d) – (i) Answers will vary.
15. (a) The twins are the individuals.
 (b) For each set of twins, the IQ is measured. This represents the bivariate quantitative data where $x = \text{IQ of twin 1}$ and $y = \text{IQ of twin 2}$.
 (c) 0.7225

Chapter 4 Test (page 222)

1. (a) Temperature



(c) $r = 0.835$

(d) Yes, a positive linear relation exists between temperature and chirps per second because $0.835 > 0.514$ (Table II).

2. (a) $\hat{y} = 0.2119x - 0.3091$

(b) If the temperature increases 1°F , the number of chirps per second increases by 0.2119, on average. Since there are no observations near 0°F , it does not make sense to interpret the y -intercept.

(c) At 83.3°F , $\hat{y} = 17.3$ chirps per second.

(d) Below average

(e) No; outside the scope of the model

3. $R^2 = 69.7\%$; 69.7% of the variation in number of chirps per second is explained by the least-squares regression line.

4. Correlation does not imply causation. It is possible that a lurking variable, such as income level or educational level, is affecting both the explanatory and response variables.

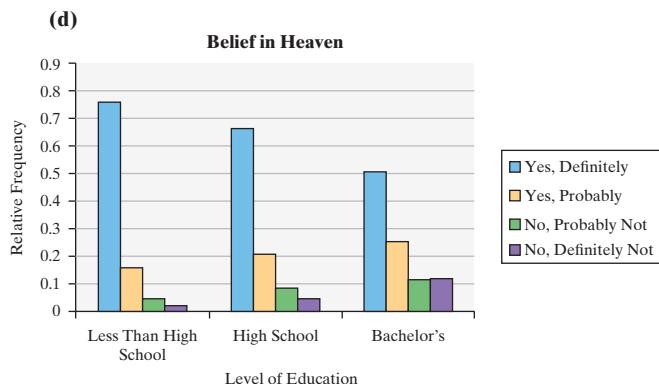
5. (a)

Education Level	Belief				Relative Frequency Marginal Distribution
	Yes, Definitely	Yes, Probably	No, Probably Not	No, Definitely Not	
Less than high school	316	66	21	9	0.173
High school	956	296	122	65	0.606
Bachelor's	267	131	62	64	0.221
Relative Frequency Marginal Distribution	0.648	0.208	0.086	0.058	1

(b) 0.648

(c)

Education Level	Belief				Total
	Yes, Definitely	Yes, Probably	No, Probably Not	No, Definitely Not	
Less than high school	0.767	0.160	0.051	0.022	1
High school	0.664	0.206	0.085	0.045	1
Bachelor's	0.510	0.250	0.118	0.122	1



(e) Yes; as education level increases, the percent who definitely believe in Heaven decreases (i.e., doubt in Heaven increases).

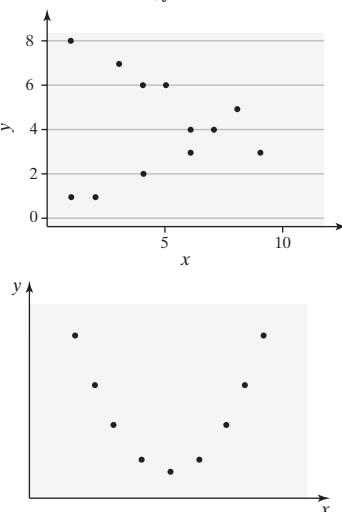
6. (a)

College Applicants		
Accepted	Denied	Total
Male	0.158	0.842
Female	0.310	0.690

- (b)** 0.158 of the males who applied was accepted. 0.310 of the females who applied was accepted.
(c) Conclusion: a higher proportion of females was accepted.
(d) 0.150 of the males applying to the business school was accepted. 0.143 of the females applying to the business school was accepted.
(e) 0.40 of the males applying to the social work school was accepted. 0.364 of the females applying to the social work school was accepted.
(f) Answers will vary. A larger number of males applied to the business school, which has an overall lower acceptance rate than the social work school, so more male applicants were declined.

7. If the slope of the least-squares regression line is negative, the correlation between the explanatory and response variables is also negative.

8. If a linear correlation coefficient is close to zero, then there is no linear relation between the explanatory and response variable. This does not mean there is no relation, just no *linear* relation.



9. The data would have a perfect negative linear relation. A scatter diagram would show all the observations being collinear (falling on the same line) with a negative slope.

CHAPTER 5 Probability

5.1 Assess Your Understanding (page 236)

- 1. (a)** The measure of the likelihood of a random phenomenon or chance behavior occurring.
 - (b)** Any process with uncertain results that can be repeated.
 - (c)** Any collection of outcomes from a probability experiment.
 - (d)** The collection of all possible outcomes from a probability experiment.
 - (e)** Each outcome has the same probability of occurring.
 - (f)** An event whose probability is 0.
 - (g)** An event that has a low probability of occurring.
 - 3.** Rule 1: All probabilities in the model are greater than or equal to zero and less than or equal to one.
- Rule 2: The sum of the probabilities in the model is 1.
- The outcome “blue” is an impossible event since $P(\text{blue}) = 0$.
- 5.** This is not a probability model because $P(\text{green}) < 0$.
 - 7.** The numbers 0, 0.01, 0.35, and 1 could be probabilities.
 - 9.** In 100 elections, where a senate candidate is winning his/her election with a 5% lead in the average of polls with a week until the election, we would expect the leading candidate to win about 89 of the elections.
 - 11.** The empirical probability is 0.95.
 - 13.** $P(2) \neq \frac{1}{11}$ since the outcomes are not equally likely.
 - 15.** $S = \{1H, 1T, 2H, 2T, 3H, 3T, 4H, 4T, 5H, 5T, 6H, 6T\}$
 - 17.** 0.428 **19.** $P(E) = \frac{3}{10} = 0.3$
 - 21.** $P(E) = \frac{2}{5} = 0.4$
 - 23. (a)** $P(\text{plays sports}) = \frac{288}{500} = 0.576$
 - (b)** If we sampled 1000 high school students, we would expect that about 576 of the students play organized sports.
 - 25. (a)** $P(\text{caught}) = \frac{85}{1000} = 0.085$
 - (b)** $P(\text{dropped}) = \frac{296}{1000} = 0.296$
 - (c)** $P(\text{hat}) = \frac{2}{85} \approx 0.024$; If 1000 caught homeruns are randomly selected, we would expect about 24 to be caught in a hat.
 - (d)** $P(\text{failed hat}) = \frac{8}{296} \approx 0.027$; If 1000 dropped homeruns are randomly selected, we would expect about 27 to be a failed hat attempt.
 - 27. (a)** $S = \{0, 00, 1, 2, 3, 4, \dots, 35, 36\}$
 - (b)** $P(8) = \frac{1}{38} = 0.0263$, if we spun the wheel 1000 times, we would expect about 26 of those times to result in the ball landing in slot 8.
 - (c)** $P(\text{odd}) = 9/19 = 0.4737$, if we spun the wheel 100 times, we would expect about 47 of the spins to result in an odd number.
 - 29. (a)** $\{SS, Ss, ss\}$; Note: There are two instances of *Ss*.
 - (b)** $P(ss) = \frac{1}{4} = 0.25$; In 100 instances where both the mother and father have one dominant normal-cell allele (*S*) and one recessive sickle-cell allele (*s*), we would expect 25 of the offspring to have genotype *ss*.
 - (c)** $P(\text{carrier}) = \frac{1}{2} = 0.5$; In 100 instances where both the mother and father have one dominant normal-cell allele (*S*) and one recessive sickle-cell allele (*s*), we would expect 50 of the offspring will be a carrier of, but will not have, sickle-cell anemia.

31. (a)	Response	Probability
Never	0.026	
Rarely	0.068	
Sometimes	0.116	
Most of the time	0.263	
Always	0.527	

(b) It would be unusual to randomly find a college student who never wears a seatbelt when riding in a car driven by someone else, because $P(\text{never}) < 0.05$.

33. (a)	Type of Larceny Theft	Probability
Pocket picking	0.006	
Purse snatching	0.009	
Shoplifting	0.207	
From motor vehicles	0.341	
Motor vehicle accessories	0.140	
Bicycles	0.065	
From buildings	0.223	
From coin-operated machines	0.008	

(b) Purse snatching larcenies are unusual.
(c) Bicycle larcenies are not unusual.

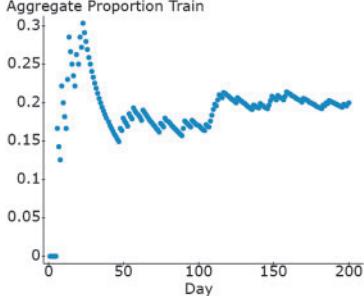
35. (a)	Movement	Probability
Up	0.433	
Down	0.5	
No change	0.067	

(b) Empirical
(c) $P(\text{up}) = 0.433$
(d) No, the probability is greater than 0.05.
(e) Yes, because of the Law of Large Numbers

37. (a) Getting stuck by a train is a random process because the outcome of any particular trial (arriving at the train tracks) is unknown. That is, you don't know ahead of time whether a train is present, or not.

(b) 0.125
(c) 0.25
(d) No

(e) **Random Process of Railroad Crossings**



(f) 0.2

39. A, B, C, and F are consistent with the definition of a probability model.

41. Use model B if the coin is known to always come up tails.

43. (a) $S = \{\text{JR, JC, JD, JM, RC, RD, RM, CD, CM, DM}\}$

(b) $P(\text{Clarice and Dominique attend}) = \frac{1}{10} = 0.10$
(c) $P(\text{Clarice attends}) = \frac{4}{10} = 0.40$
(d) $P(\text{John stays home}) = \frac{6}{10} = 0.60$

45. (a) Classical method is used.
(b) Empirical method is used.
(c) Subjective method is used.
(d) Empirical method is used.

47. Answers will vary.

49. (a)	Month	Probability
1 (January)	0.094	
2 (February)	0.050	
3 (March)	0.133	
4 (April)	0.154	
5 (May)	0.200	
6 (June)	0.105	
7 (July)	0.057	
8 (August)	0.079	
9 (September)	0.035	
10 (October)	0.054	
11 (November)	0.030	
12 (December)	0.010	

(b) $P(\text{April}) = 0.154$
(c) Yes, $P(\text{December}) = 0.010 < 0.05$. We would expect a tornado in December in about 1 of every 100 tornadoes.

51. (d)	F Scale	Probability
-9	0.043	
0	0.432	
1	0.412	
2	0.099	
3	0.011	
4	0.002	

(e) Yes, we would only expect an F scale of 4 in about 2 of every 1000 tornadoes.

51. (a) Row 1: No Hit, No Hit, No Hit, No Hit; Row 5: Hit; Hit, No Hit, No Hit

(b) Proportion of fifth at-bats resulting in a hit after first four at-bats are "no hit" is 0.301.

53. (a) This is to control other variables that may affect the response variable at a fixed level.
(b) The Law of Large Numbers

55. The Law of Large Numbers states that, as the number of repetitions of a probability experiment increases (long term), the proportion with which a certain outcome is observed (relative frequency) gets closer to the probability of the outcome. The games at a gambling casino are designed to benefit the casino in the long run; the risk to the casino is minimal because of the large number of gamblers.

57. An event is unusual if it has a low probability of occurring. The same cutoff should not always be used to identify unusual events. Selecting a cutoff is subjective and should take into account the consequences of incorrectly identifying an event as unusual.

59. Empirical probability is based on the outcomes of a probability experiment and is the relative frequency of the event. Classical probability is based on counting techniques and is equal to the ratio of the number of ways an event can occur to the number of possible outcomes of the experiment. Classical probability requires equally likely outcomes.

61. The smaller hospital due to the Law of Large Numbers. The larger hospital likely has more births so it is less likely that the births would deviate from the expected proportion of girls.

5.2 Assess Your Understanding (page 249)

1. Two events are disjoint (mutually exclusive) if they have no outcomes in common.
3. $P(E) + P(F) - P(E \text{ and } F)$
5. E and $F = \{5, 6, 7\}$; E and F are not mutually exclusive.
7. F or $G = \{5, 6, 7, 8, 9, 10, 11, 12\}$; $P(F \text{ or } G) = \frac{2}{3}$
9. There are no outcomes in event " E and G ." Events E and G are mutually exclusive.
11. $E^c = \{1, 8, 9, 10, 11, 12\}$; $P(E^c) = \frac{1}{2}$
13. $P(E \text{ or } F) = 0.55$
15. $P(E \text{ or } F) = 0.70$
17. $P(E^c) = 0.75$
19. $P(F) = 0.30$
21. $P(\text{Titleist or Maxfli}) = \frac{17}{20} = 0.85$
23. $P(\text{not a Titleist}) = \frac{11}{20} = 0.55$
25. (a) All the probabilities are nonnegative; the sum of the probabilities equals 1.
(b) $P(\text{head or face}) = 0.44$. If 100 injuries of youth baseball players ages 5–14 are randomly selected, we would expect 44 to be injuries of the head or face.
(c) $P(\text{head, face, or wrist}) = 0.49$. If 100 injuries of youth baseball players ages 5–14 are randomly selected, we would expect 49 to be injuries of the head, face, or wrist.
(d) $P(\text{something other than face}) = 0.67$. If 100 injuries of youth baseball players ages 5–14 are randomly selected, we would expect 67 to be injuries of something other than the face.
27. No; for example, on one draw of a card from a standard deck, let $E = \text{diamond}$, $F = \text{club}$, and $G = \text{red card}$.
29. (a) $P(5 - 5.9) = \frac{23}{125} = 0.184$
(b) $P(\text{not between } 5 - 5.9) = 1 - 0.184 = 0.816$
(c) $P(< 9) = 0.96$
(d) $P(\text{reduced payments}) = 0.08$. If 100 hospitals in Illinois are randomly selected, we would expect eight to have received reduced Medicare payments. Given the fact that reduced Medicare payments result due to a poor patient track record, we would like the proportion of hospitals receiving lower payments from Medicare to be close to zero. Therefore, this result is not unusual enough (unfortunately).
31. (a) $P(\text{heart or club}) = \frac{1}{2} = 0.5$
(b) $P(\text{heart or club or diamond}) = \frac{3}{4} = 0.75$
(c) $P(\text{ace or heart}) = \frac{4}{13} = 0.308$
33. (a) $P(\text{birthday is not November 8}) = \frac{364}{365} = 0.997$
(b) $P(\text{birthday is not 1st of month}) = \frac{353}{365} = 0.967$
(c) $P(\text{birthday is not 31st of month}) = \frac{358}{365} = 0.981$
(d) $P(\text{birthday is not in December}) = \frac{334}{365} = 0.915$
35. No; some people have both vision and hearing problems, but we do not know the proportion.
37. (a) $P(\text{only English or only Spanish is spoken}) = 0.907$
(b) $P(\text{language other than only English or only Spanish}) = 0.093$
(c) $P(\text{not only English is spoken}) = 0.216$
(d) No; the sum of the probabilities would be greater than 1, and there would be no probability model.
39. (a) $P(\text{died from cancer}) = 0.007$
(b) $P(\text{current smoker}) = 0.057$

- (c) $P(\text{died from cancer and was current cigar smoker}) = 0.001$
(d) $P(\text{died from cancer or was current cigar smoker}) = 0.063$

41. (a) $P(\text{placebo}) = \frac{2}{5} = 0.40$
(b) $P(\text{headache went away}) = \frac{94}{125} = 0.752$
(c) $P(\text{placebo and headache went away}) = \frac{28}{125} = 0.224$
(d) $P(\text{placebo or headache went away}) = \frac{116}{125} = 0.928$

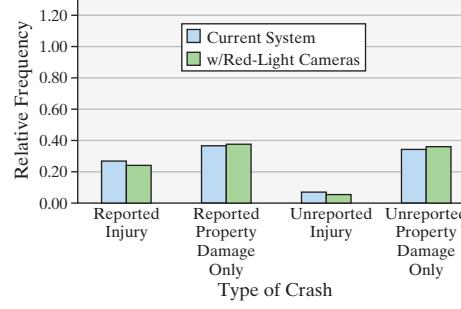
43. (a) $P(\text{normal weather}) = 0.859$
(b) $P(\text{daylight}) = 0.474$
(c) $P(\text{normal weather and daylight}) = 0.415$
(d) $P(\text{normal weather or daylight}) = 0.918$
(e) Yes. The probability of a fatality while it is dark outside (without light) and raining is 0.020. This is misleading. The real question is, "Among the drivers on the road when it is dark and raining, what proportion result in a fatality?"

45. (a) Crash type, system, and number of crashes
(b) Crash type: qualitative; system: qualitative; number of crashes: quantitative, discrete

(c)

Crash Type	Current System	Red-Light Cameras
Reported injury	0.26	0.24
Reported property damage only	0.35	0.36
Unreported injury	0.07	0.07
Unreported property damage only	0.32	0.33
Total	1	1

(d)

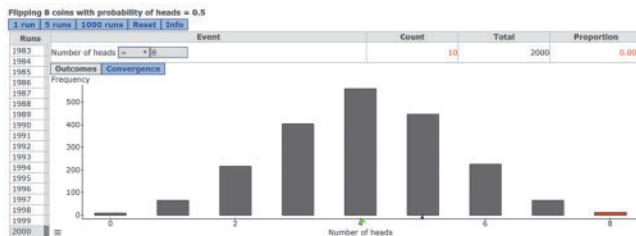
Crashes at Intersections

- (e) Current system: $\frac{1121}{13} \approx 86.2$ crashes
With cameras: $\frac{922}{13} \approx 70.9$ crashes

- (f) Not possible; we would need to know the number of crashes at each intersection.
(g) Since the mean number of crashes is less with the cameras, it appears that the program will be beneficial.
(h) $P(\text{reported injuries}) = \frac{289}{1121} = 0.258$
(i) $P(\text{only property damage}) = \frac{641}{922} = 0.695$
(j) When accounting for the cause of the crash (rear end vs. red-light running), the camera system does not reduce all types of accidents. Under the camera system, red-light running crashes decreased, but rear-end crashes increased.
(k) Recommendations may vary. The benefits of the decrease in red-light running crashes must be weighed against the negative of increased rear-end crashes. Seriousness of injuries and amount of property damage may need to be considered.

5.3 Assess Your Understanding (page 257)

1. independent 3. Addition 5. $P(E) \cdot P(F)$
7. (a) Dependent
(b) Dependent
(c) Independent
9. $P(E \text{ and } F) = 0.18$
11. $P(5 \text{ heads}) = \frac{1}{32} = 0.03125$; if we flipped a coin 5 times, one hundred different times, we would expect to observe 5 heads in a row about 3 times.
13. $P(\text{two left-handed people are chosen}) = 0.0169$
 $P(\text{at least one person chosen is right-handed}) = 0.9831$
15. (a) $P(\text{all five are negative}) = 0.9752$
(b) $P(\text{at least one is positive}) = 0.0248$
17. (a) $P(\text{two will live to 41 years}) = 0.99515$
(b) $P(\text{five will live to 41 years}) = 0.98791$
(c) $P(\text{at least one of five dies}) = 0.01209$; it would be unusual that at least one of the five dies before age 41.
19. (a) $P(\text{not default}) = 0.99$
(b) $P(\text{five not default}) = 0.951$
(c) $P(\text{worthless}) = 0.049$
(d) No. National economic conditions (such as recession) will impact all mortgages. So, if one mortgage defaults, the likelihood of a second mortgage defaulting increases.
21. (a) $P(\text{all three fail}) = 2.16 \times 10^{-7} = 0.000000216$
(b) $P(\text{at least one does not fail}) = 0.999999784$
23. (a) $P(\text{system does not fail}) = 0.999973$
(b) Six components
25. (a) $P(\text{two strikes in a row}) = 0.09$
(b) $P(\text{turkey}) = 0.027$
(c) $P(\text{3 strikes followed by nonstrike}) = 0.0189$
27. $P(\text{at least 1}) = 0.9999$; answers will vary.
29. $P(\text{girl and in excess of 40 pounds}) = 0.099$
31. (a) $P(\text{male and bets on professional sports}) = 0.0823$
(b) $P(\text{male or bets on professional sports}) = 0.5717$
(c) The independence assumption is not correct.
(d) $P(\text{male or bets on professional sports}) = 0.548$
33. No, assuming gender of children for different births are independent events.
35. (a) 0.5
(b) 0.0039
(c) The likelihood of obtaining a head with a fair coin is 0.5, which is the same as the likelihood of correctly guessing whether milk was put in the tea first or second. Flip eight fair coins many times and determine the proportion of times all heads occurs.
(d) Answers will vary. A sample result is shown below. Here $P(\text{eight heads}) = 0.005$.



- (e) If the lady tasting tea was guessing, we would expect her to guess correctly on all eight cups of tea in about 5 out of 1000 trials. It is highly unlikely that she was just guessing.

5.4 Assess Your Understanding (page 266)

1. F occurring; E has occurred
3. $P(F|E) = 0.75$
5. $P(F|E) = 0.568$
7. $P(E \text{ and } F) = 0.32$

9. The events are not independent.

11. $P(\text{club}) = \frac{1}{4}$; $P(\text{club} | \text{black card}) = \frac{1}{2}$

13. $P(\text{marriage lasts at least 20 years} | \text{bachelor's}) = \frac{0.27}{0.35} = 0.771$

15. $P(\text{unemployed} | \text{dropout}) = \frac{0.021}{0.080} = 0.263$

17. (a) $P(35-44 | \text{more likely}) = \frac{329}{1329} = 0.248$

(b) $P(\text{more likely} | 35-44) = \frac{329}{536} = 0.614$

(c) No; they are less likely to buy American; 0.439 for 18- to 34-year-olds compared to 0.615 in general.

19. (a) $P(\text{dawn/dusk} | \text{normal}) = 0.040$

(b) $P(\text{normal} | \text{dawn/dusk}) = 0.846$

(c) Rain is more dangerous because $P(\text{dark} | \text{normal}) = 0.276$ and $P(\text{dark} | \text{rain}) = 0.317$.

21. $P(\text{both TVs work}) = 0.4$; $P(\text{at least one TV does not work}) = 0.6$

23. (a) $P(\text{first card is a king and the second card is a king}) = \frac{1}{221} = 0.005$

(b) $P(\text{first card is a king and the second card is a king}) = \frac{1}{169} = 0.006$

25. $P(\text{Dave and then Neta are chosen}) = \frac{1}{20} = 0.05$

27. (a) $P(\text{like both songs}) = \frac{5}{39} = 0.128$; it is not unusual to like both songs.

(b) $P(\text{like neither song}) = \frac{14}{39} = 0.359$

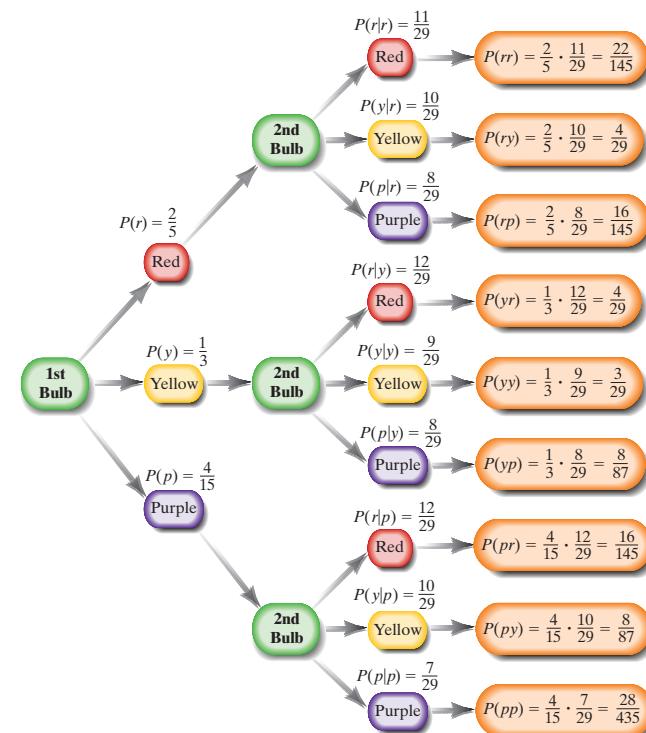
(c) $P(\text{like exactly one song}) = \frac{20}{39} = 0.513$

(d) $P(\text{like both songs}) = \frac{25}{169} = 0.148$;

$P(\text{like neither song}) = \frac{64}{169} = 0.379$;

$P(\text{like exactly one song}) = \frac{80}{169} = 0.473$

29.



- (a) $P(\text{two red bulbs}) = \frac{22}{145} = 0.152$
(b) $P(\text{first a red and then a yellow bulb}) = \frac{4}{29} = 0.138$
(c) $P(\text{first a yellow and then a red bulb}) = \frac{4}{29} = 0.138$
(d) $P(\text{a red bulb and a yellow bulb}) = \frac{8}{29} = 0.276$
31. $P(\text{female and smokes}) = 0.090$; it would not be unusual to randomly select a female who smokes.
33. (a) $P(10 \text{ people each have a different birthday}) = 0.883$
(b) $P(\text{at least two of the ten people have the same birthday}) = 0.117$
35. (a) Yes; $P(\text{male}) = \frac{1}{2} = P(\text{male} | 0 \text{ activities}) = \frac{1}{2}$
(b) No; $P(\text{female}) = \frac{1}{2} \neq P(\text{female} | 5+ \text{ activities}) = \frac{71}{109} = 0.651$
(c) Yes; $P(1-2 \text{ activities and } 3-4 \text{ activities}) = 0$
(d) No; $P(\text{male and } 1-2 \text{ activities}) = 0.2025 \neq 0$
37. (a) $P(\text{positive} | \text{no cancer}) = 0.08$
(b)

	Positive Mammogram	Negative Mammogram	Total
Cancer	160,000	40,000	200,000
No Cancer	3,184,000	36,616,000	39,800,000
Total	3,344,000	36,656,000	40,000,000

- (c) $P(\text{cancer} | \text{positive mammogram}) = 0.048$
39. (a) $P(\text{two defective chips}) = 0.0000245$
(b) Assuming independence: $P(\text{two defective chips}) = 0.000025$
41. $P(45-54 \text{ years old}) = \frac{546}{2160} = 0.253$; $P(45-54 \text{ years old} | \text{more likely}) = \frac{360}{1329} = 0.271$; the events are not independent.
43. (a) This is a cohort study because it follows a group of young millennial adults from teenage years to early adulthood. No attempt is made to influence variables that may play a role in "success."
(b) $P(\text{below poverty} | \text{failed to complete path to success}) = 0.53$
(c) $P(\text{below poverty} | \text{completed high school}) = 0.31$; $P(\text{below poverty} | \text{HS diploma & full-time job}) = 0.16$; $P(\text{below poverty} | \text{HS diploma & full-time job & marriage before children}) = 0.03$
(d) $P(\text{middle or upper class} | \text{followed path to success}) = 0.80$; $P(\text{middle or upper class} | \text{missed at least one stage in the path to success}) = 0.44$
(e) They are not independent. If an individual follows the path to success, he or she has a much higher likelihood of moving out of the lower-income class (80% versus 28%).

5.5 Assess Your Understanding (page 279)

- permutation
- True
- $5! = 120$
- $10! = 3,628,800$
- $0! = 1$
- ${}_6P_2 = 30$
- ${}_4P_4 = 24$
- ${}_5P_0 = 1$
- ${}_8P_3 = 336$
- ${}_8C_3 = 56$
- ${}_{10}C_2 = 45$
- ${}_{52}C_1 = 52$
- ${}_{48}C_3 = 17,296$

27. $ab, ac, ad, ae, ba, bc, bd, be, ca, cb, cd, ce, da, db, dc, de, ea, eb, ec, ed$; ${}_5P_2 = 20$
29. $ab, ac, ad, ae, bc, bd, be, cd, ce, de; {}_5C_2 = 10$
31. He can wear 24 different shirt-and-tie combinations.
33. Dan can arrange the songs $12! = 479,001,600$ ways.
35. The salesperson can take $8! = 40,320$ different routes.
37. At most 18,278 companies can be listed on the NYSE.
39. (a) 10,000 different codes are possible.
(b) $P(\text{correct code is guessed}) = \frac{1}{10,000}$
41. $26^8 \approx 2.09 \times 10^{11}$ different user names are possible.
43. (a) There are $50^3 = 125,000$ lock combinations.
(b) $P(\text{guessing the correct combination}) = \frac{1}{50^3} = \frac{1}{125,000}$
45. The top three cars can finish in ${}_{40}P_3 = 59,280$ ways.
47. The officers can be chosen in ${}_{20}P_4 = 116,280$ ways.
49. There are ${}_{25}P_4 = 303,600$ possible outcomes.
51. There are ${}_{50}C_5 = 2,118,760$ possible simple random samples of size 5.
53. ${}_6C_2 = 15$ different birth and gender orders are possible.
55. $\frac{10!}{3! \cdot 2! \cdot 2! \cdot 3!} = 25,200$ different sequences are possible.
57. The trees can be planted $\frac{11!}{4! \cdot 5! \cdot 2!} = 6930$ different ways.
59. $P(\text{winning}) = \frac{1}{{}_{39}C_5} = \frac{1}{575,757}$
61. (a) $P(\text{jury has only students}) = \frac{1}{153} = 0.0065$
(b) $P(\text{jury has only faculty}) = \frac{1}{34} = 0.0294$
(c) $P(\text{jury has two students and three faculty}) = \frac{20}{51} = 0.3922$
63. $P(\text{shipment is rejected}) = 0.1283$
65. (a) $P(\text{you like 2 of 4 songs}) = 0.3916$
(b) $P(\text{you like 3 of 4 songs}) = 0.1119$
(c) $P(\text{you like all 4 songs}) = 0.0070$
67. (a) Five cards can be selected from a deck ${}_{52}C_5 = 2,598,960$ ways.
(b) Three of the same card can be chosen $13 \cdot {}_4C_3 = 52$ ways.
(c) The remaining two cards can be chosen ${}_{12}C_2 \cdot {}_4C_1 \cdot {}_4C_1 = 1056$ ways.
(d) $P(\text{three of a kind}) = \frac{52 \cdot 1056}{2,598,960} = 0.0211$
69. $P(\text{all 4 modems tested work}) = 0.4912$
71. (a) 282 billion
(b) 272 billion
(c) 1 billion

5.6 Assess Your Understanding (page 286)

- (a) Day 188; Yes
(b) $P(\text{at least two people share a birthday}) = 461/2000 = 0.2305$
(c) Answers will vary, but results should be close to those in part (b).
- (a) Use a random-number generator to generate integers from 1 to 20. Because the outcome of each spin is a multiple of 5, multiply the integers by 5. Repeat this for the second spin. If the first spin results in a value of 85 or higher, the player wins. If the first spin results in a value of 75 or less, the player spins again. Compute the sum of the values of Spin 1 and Spin 2. If this sum is at least 85 but no more than 100, the player wins. If the sum is greater than 100, the player loses. If the player spins 80, the player must decide whether to spin again or not.
(b) $P(\text{win on first spin}) = 0.1915$
(c) $P(\text{tie on first spin}) = 0.0465$
(d) Player 2 spins a 65, so Player 2 must spin again. Second spin is a 25, which results in a sum of 90, so Player 2 wins. Player 2 spins an 80, so Player 2 decides to spin again. Second spin is 100, so Player 2 loses because the sum of the rolls is over 100.

(e) $P(\text{Player 2 wins}) = 0.3825$

(f) Player 2 spins a 20, so Player 2 must spin again. Second spin is 60, which results in a sum of 80, so there is a tie. In the playoff, Player 1 spins 80 and Player 2 spins 30, so Player 2 loses.

(g) $P(\text{Player 2 wins}) = 0.39$; If we assume the 8 ties are equally split between Player 1 and Player 2, $P(\text{Player 2 wins}) = 0.392$.

(h) Accepting the tie if a player spins 80 on the first spin is the better strategy.

5. (a) Use a random-number generator to generate two columns of integers from 1 to 6. For each row, compute the sum of the entries to represent the simulated sum of the dice. Count how many rolls are required before observing a 12, and record this result for all the simulated rolls.

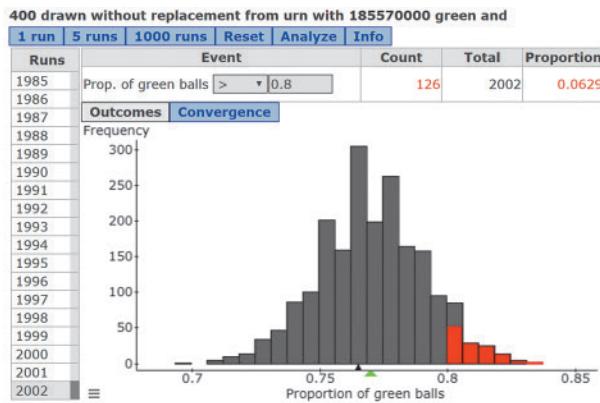
(b) Answers will vary.

(c) Answers will vary, but the result should be close to 6.

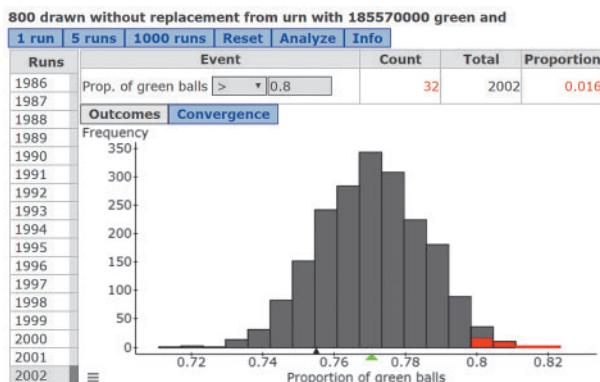
7. (a) Answers will vary.

(b) Answers will vary.

(c) Answers will vary. Sample results are shown. In this simulation, the probability of getting a sample proportion greater than 0.8 is 0.0629. Therefore, this would not be considered an unusual result (using the cut-off point of 0.05 to designate unusual events).



(d) Answers will vary. Sample results are shown. In this simulation, the probability of getting a sample proportion greater than 0.8 is 0.016. Therefore, this would be considered an unusual result (using the cut-off point of 0.05 to designate unusual events). This illustrates the Law of Large Numbers.



5.7 Assess Your Understanding (page 291)

- A permutation is an arrangement in which order matters; a combination is an arrangement in which order does not matter.
- AND is generally associated with multiplication; OR is generally associated with addition.

5. $P(E) = \frac{4}{10} = 0.4$

7. $abc, acb, abd, adb, abe, aeb, acd, adc, ace, aec, ade, aed, bac, bca, bad, bda, bae, bea, bcd, bdc, bce, bec, bde, bed, cbe, ceb, cab, cba, cad, cda, cae, cea, cbd, cdb, cde, ced, dab, dba, dac, dca, dae, dea, dbc, dcb,dbe, deb, dce, dec, eab, eba, eac, eca, ead, eda, ebc, ecb, ebd, ebd, ecd, edc$

9. $P(E \text{ or } F) = 0.7 + 0.2 - 0.15 = 0.75$

11. ${}_7P_3 = 210$

13. $P(E \text{ and } F) = P(E) \cdot P(F) = 0.4$

15. $P(E \text{ and } F) = P(E) \cdot P(F|E) = 0.27$

17. $P(\text{plays soccer}) = \frac{22}{500} = 0.044$

19. There are $3 \cdot 3 \cdot 2 \cdot 2 \cdot 1 \cdot 1 = 36$ ways to arrange the three men and three women.

21. (a) $P(\text{survived}) = \frac{711}{2224} = 0.320$

(b) $P(\text{female}) = \frac{425}{2224} = 0.191$

(c) $P(\text{female or child}) = \frac{534}{2224} = 0.240$

(d) $P(\text{female and survived}) = \frac{316}{2224} = 0.142$

(e) $P(\text{female or survived}) = \frac{820}{2224} = 0.369$

(f) $P(\text{survived}|\text{female}) = \frac{316}{425} = 0.744$

(g) $P(\text{survived}|\text{child}) = \frac{57}{109} = 0.523$

(h) $P(\text{survived}|\text{male}) = \frac{338}{1690} = 0.2$

(i) Yes, the survival rate was much higher for women and children.

(j) $P(\text{both females survived}) = \frac{316}{425} \cdot \frac{315}{424} = 0.552$

23. Shawn can fill out the report in ${}_{12}C_3 = 220$ different ways.

25. (a) $P(\text{win both}) = \frac{1}{5.2 \times 10^6} \cdot \frac{1}{705,600} = 0.0000000000027$

(b) $P(\text{win Jubilee twice}) = \left(\frac{1}{705,600} \right)^2 = 0.000000000002$

27. (a) No $P(\text{Bachelor's}) \neq P(\text{Bachelor's}|\text{never married})$

(b) $P(\text{Bachelor's and Never Married}) = P(\text{Bachelor's}) \cdot P(\text{Never Married}|\text{Bachelor's}) = (0.202)(0.186) = 0.038$

29. ${}_{12}C_8 = 495$ different sets of questions can be answered.

31. $2 \cdot 2 \cdot 3 \cdot 8 \cdot 2 = 192$ different cars are possible.

Chapter 5 Review Exercises (page 295)

1. (a) Possible probabilities are 0, 0.75, 0.41.

(b) Possible probabilities are $\frac{2}{5}, \frac{1}{3}, \frac{6}{7}$.

2. $P(E) = \frac{1}{5} = 0.2$

3. $P(F) = \frac{2}{5} = 0.4$

4. $P(E) = \frac{3}{5} = 0.6$

5. $P(E^c) = \frac{4}{5} = 0.8$

6. $P(E \text{ or } F) = 0.89$

7. $P(E \text{ or } F) = 0.48$

8. $P(E \text{ and } F) = 0.09$

9. Events E and F are not independent because $P(E \text{ and } F) \neq P(E) \cdot P(F)$.

10. $P(E \text{ and } F) = 0.2655$

11. $P(E|F) = 0.5$

12. (a) $7! = 5040$

(d) ${}_{10}C_3 = 120$

(b) $0! = 1$

(e) ${}_9P_2 = 72$

(c) ${}_9C_4 = 126$

(f) ${}_{12}P_4 = 11,880$

13. (a) $P(\text{green}) = \frac{4}{19} = 0.0526$; If the wheel is spun 100 times, we would expect about 5 spins to end with the ball in a green slot.

- (b) $P(\text{green or red}) = \frac{10}{19} = 0.5263$; If the wheel is spun 100 times, we would expect about 53 spins to end with the ball in either a green slot or a red slot.
- (c) $P(00 \text{ or red}) = \frac{19}{38} = 0.5$; If the wheel is spun 100 times, we would expect about 50 spins to end with the ball either in 00 or in a red slot.
- (d) $P(31 \text{ and black}) = 0$; this is called an impossible event.
14. (a) $P(\text{fatality was alcohol related}) = \frac{9477}{34,439} = 0.275$
(b) $P(\text{fatality was not alcohol related}) = 1 - 0.275 = 0.725$
(c) $P(\text{both fatalities were alcohol related}) = 0.076$
(d) $P(\text{neither fatality was alcohol related}) = 0.525$
(e) $P(\text{at least one fatality was alcohol related}) = 1 - 0.525 = 0.475$
15. (a)
- | Age | Probability |
|--------------|-------------|
| 18–79 | 0.121 |
| 80–89 | 0.252 |
| 90–99 | 0.253 |
| 100 or older | 0.373 |
- (b) It is not unusual for an individual to want to live between 18 and 79 years.
16. (a) $P(\text{baby was postterm}) = 0.057$
(b) $P(\text{baby weighed 3000–3999 grams}) = 0.651$
(c) $P(\text{baby weighed 3000–3999 grams and was postterm}) = 0.040$
(d) $P(\text{baby weighed 3000–3999 grams or was postterm}) = 0.668$
(e) $P(\text{baby weighed } < 1000 \text{ grams and was postterm}) = 0.000006$; this event is not impossible.
(f) $P(\text{baby weighed 3000–3999 grams} | \text{baby was postterm}) = 0.708$
(g) The events “postterm baby” and “weighs 3000–3999 grams” are not independent.
17. (a) $P(\text{trusts}) = 0.18$
(b) $P(\text{does not trust}) = 0.82$
(c) Yes, $P(\text{all three trust}) = 0.006$
(d) $P(\text{at least one of three does not trust}) = 0.994$
(e) No, $P(\text{all five do not trust}) = 0.371$
(f) $P(\text{at least one of five trusts}) = 0.629$
18. $P(\text{matching the three winning numbers}) = 0.001$
19. $P(\text{matching the four winning numbers}) = 0.0001$
20. $P(\text{drawing three aces}) = 0.00018$
21. $26^2 \cdot 10^4 = 6,760,000$ different license plates can be formed.
22. The students can be seated in ${}_{10}P_4 = 5040$ different arrangements.
23. There are $\frac{10!}{4! \cdot 3! \cdot 2!} = 12,600$ different vertical arrangements of the flags.
24. There are ${}_{55}C_8 = 1,217,566,350$ possible simple random samples.
25. $P(\text{winning Pick 5}) = \frac{1}{35}C_5 = \frac{1}{324,632} = 0.000003$
26. (a) $P(\text{three Merlot}) = 0.0455$
(b) $P(\text{two Merlot}) = 0.3182$
(c) $P(\text{no Merlot}) = 0.1591$
27. Answers will vary, but should be reasonably close to $\frac{1}{38}$ for part (a) and $\frac{1}{19}$ for part (b).
28. Answers will vary. Subjective probability is based on personal experience or intuition (e.g., “There is a 70% chance that the Packers will make it to the NFL playoffs next season.”)
29. (a) There are 13 clubs in the deck.
(b) Thirty-seven cards remain in the deck. Forty-one cards are not known by you. Eight of the unknown cards are clubs.
(c) $P(\text{next card dealt is a club}) = \frac{8}{41}$

- (d) $P(\text{two clubs in a row}) = 0.0341$
(e) No
30. (a) $P(\text{home run to left field}) = \frac{34}{70} = 0.486$; If 100 Mark McGwire home runs are randomly selected, we would expect 49 of them to be hit to left field.
(b) $P(\text{home run to right field}) = 0$
(c) No; it was not impossible for McGwire to hit a home run to right field. He just never did it.
31. Someone winning a lottery twice is not that unlikely considering millions of people play lotteries who have already won a lottery (sometimes more than one) each week, and many lotteries have multiple drawings each week.
32. (a) $P(\text{Bryce}) = \frac{119}{1009} = 0.118$; not unusual
(b) $P(\text{Gourmet}) = \frac{264}{1009} = 0.262$
(c) $P(\text{Mallory} | \text{Single Cup}) = \frac{3}{25} = 0.12$
(d) $P(\text{Bryce} | \text{Gourmet}) = \frac{3}{88} = 0.034$; yes, this is unusual.
(e) While it is not unusual for Bryce to sell a case, it is unusual for him to sell a Gourmet case.
(f) No; $P(\text{Mallory}) = \frac{186}{1009} \neq P(\text{Mallory} | \text{Filter}) = \frac{1}{3}$.
(g) No; $P(\text{Paige and Gourmet}) = \frac{42}{1009} \neq 0$.
33. (a) Patti and John both conducted a probability experiment using a random process. The randomness of winning or not is what causes different results.
(b) 1000 games would be better due to the Law of Large Numbers.
- Chapter 5 Test (page 297)**
- Possible probabilities are 0.23, 0, $\frac{3}{4}$.
 - $P(\text{Jason}) = \frac{1}{5}$
 - $P(\text{Chris or Elaine}) = \frac{2}{5}$
 - $P(E^c) = \frac{4}{5}$
 - (a) $P(E \text{ or } F) = P(E) + P(F) = 0.59$
(b) $P(E \text{ and } F) = P(E) \cdot P(F) = 0.0814$
 - (a) $P(E \text{ and } F) = P(E) \cdot P(F | E) = 0.105$
(b) $P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F) = 0.495$
(c) $P(E | F) = \frac{P(E \text{ and } F)}{P(F)} = \frac{0.105}{0.45} = 0.233$
(d) No; $P(E) = 0.15 \neq P(E | F) = 0.233$
 - (a) $8! = 40,320$
(b) ${}_{12}C_6 = 924$
(c) ${}_{14}P_8 = 121,080,960$
 - (a) $P(7 \text{ or } 11) = \frac{8}{36} = \frac{2}{9} = 0.222$; If the dice are thrown 100 times, we would expect the player will win on the first roll about 22 times.
(b) $P(2, 3, \text{ or } 12) = \frac{4}{36} = \frac{1}{9} = 0.111$; If the dice are thrown 100 times, we would expect that the player will lose on the first roll about 11 times.
 - (a) $P(\text{healthy}) = 0.26$; If we randomly selected 100 adult Americans, we would expect 26 to believe his/her diet is healthy.
(b) $P(\text{not healthy}) = 0.74$
 - (a) All the probabilities are greater than or equal to zero and less than or equal to one, and the sum of the probabilities is 1.
(b) $P(\text{a peanut butter cookie}) = 0.24$
(c) $P(\text{next box sold is mint, caramel, or shortbread}) = 0.53$
(d) $P(\text{next box sold is not mint}) = 0.75$
 - (a) $P(\text{ideal is 2}) = \frac{155}{297} = 0.522$

(b) $P(\text{female and ideal is 2}) = \frac{87}{297} = 0.293$

(c) $P(\text{female or ideal is 2}) = \frac{256}{297} = 0.862$

(d) $P(\text{ideal is 2} | \text{female}) = \frac{87}{188} = 0.463$

(e) $P(\text{male} | \text{ideal is 4}) = \frac{8}{36} = 0.222$

12. (a) $P(\text{win two in a row}) = 0.336$

(b) $P(\text{win seven in a row}) = 0.022$

(c) $P(\text{lose at least one of next seven}) = 1 - P(\text{win seven in a row}) = 0.978$

13. $P(\text{shipment accepted}) = 0.8$

14. There are $6! = 720$ different arrangements of the letters LINCEY.

15. $29C_5 = 118,755$ different subcommittees can be formed.

16. $P(\text{winning Cash 5}) = \frac{1}{43C_5} = \frac{1}{962,598} = 0.00000104$

17. $26 \cdot 36^7 \approx 2.04 \times 10^{12}$ different passwords are possible.

18. Subjective probability was used. It is not possible to repeat probability experiments to estimate the probability of life on Mars.

19. $\frac{15!}{2! \cdot 4! \cdot 4! \cdot 5!} = 9,459,450$ different sequences are possible.

20. $P(\text{guessing correctly}) = \frac{9}{40} = 0.225$

CHAPTER 6 Discrete Probability Distributions

6.1 Assess Your Understanding (page 308)

1. A random variable is a numerical measure of the outcome of a probability experiment.

3. Each probability must be between 0 and 1, inclusive, and the sum of the probabilities must equal 1.

5. (a) Discrete, $x = 0, 1, 2, \dots, 20$ (b) Continuous, $t > 0$
 (c) Discrete, $x = 0, 1, 2, \dots$ (d) Continuous, $s \geq 0$

7. (a) Continuous, $r \geq 0$ (b) Discrete, $x = 0, 1, 2, 3, \dots$
 (c) Discrete, $x = 0, 1, 2, 3, \dots$ (d) Continuous, $t > 0$

9. Yes, it is a probability distribution.

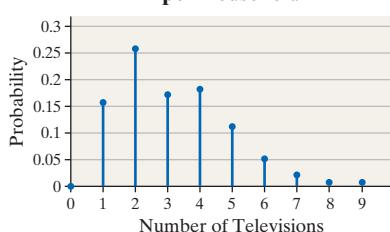
11. No, $P(50) < 0$.

13. No, $\sum P(x) \neq 1$

15. $P(4) = 0.3$

17. (a) Each probability is between 0 and 1, inclusive, and the sum of the probabilities equals 1.

(b) **Number of Televisions per Household**



The distribution is skewed right.

(c) $\mu_X = 3.2$ televisions; if we surveyed many households, we would expect the mean number of televisions per household to be 3.2.

(d) $\sigma_X = 1.7$ televisions

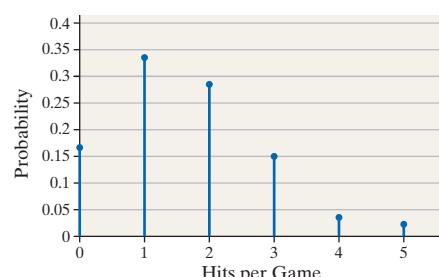
(e) $P(3) = 0.176$

(f) $P(3 \text{ or } 4) = 0.362$

(g) $P(0) = 0$; no, this is not an impossible event.

19. (a) Each probability is between 0 and 1, inclusive, and the sum of the probabilities equals 1.

(b) **Ichiro's Hit Parade**



The distribution is skewed right.

(c) $\mu_X = 1.6$ hits; over many games, Ichiro is expected to average about 1.6 hits per game.

(d) $\sigma_X = 1.2$ hits

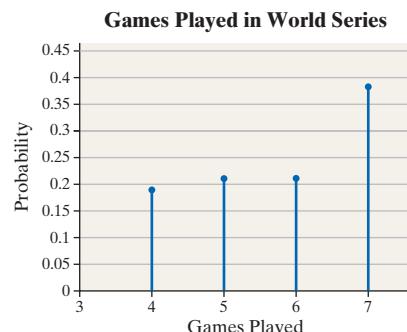
(e) $P(2) = 0.2857$

(f) $P(X > 1) = 0.4969$

21. (a)

x (games played)	P(x)
4	0.1895
5	0.2105
6	0.2105
7	0.3895

(b)



(c) $\mu_X = 5.8$ games; the World Series, if played many times, would be expected to last about 5.8 games, on average.

(d) $\sigma_X = 1.1$ games

23. (a) $P(4) = 0.074$

(b) $P(4 \text{ or } 5) = 0.103$

(c) $P(X \geq 6) = 0.027$

(d) $\mu_X = 2.2$ children; we would expect the mother to have had 2.2 children, on average.

25. $E(X) = \$86.00$; the insurance company expects to make an average profit of $\$86.00$ on every 20-year-old female it insures for one year.

27. $\sigma_x = \$5337$; The standard deviation is high because there is a wide range of possible outcomes.

29. (a) $-\$0.05$

(b) Lose $\$6$

31. $E(X) = -\$0.26$; if you played 1000 times, you would expect to lose about $\$260$.

33. (a) The expected cash prize is $\$0.30$. After paying $\$1.00$ to play, your expected profit is $-\$0.70$.

(b) The standard deviation is $\$1245$. This suggests there is a wide range of payouts.

(c) A grand prize of $\$118,000,000$ has an expected profit greater than zero.

(d) The size of the grand prize does not affect the chance of winning provided the probabilities remain constant.

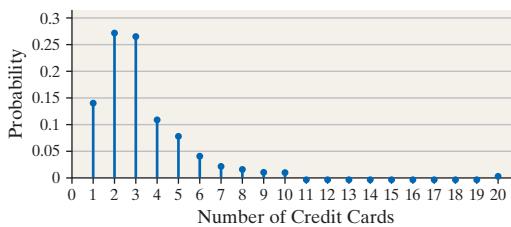
- 35.** (a) Until the 2016–2017 season, the expected value going for one point was 0.993, while the expected value of going for two points was 0.960. Because the expected value from one point is higher, it makes sense to go for one point.

(b) Starting with 2016–2017, the expected value of going for one point decreased to 0.959. Now, the two expected values are closer, so it is not as obvious which attempt should be made.

- 37.** (a) 3.3 credit cards (b) 2.3 credit cards

x (credit cards)	P(x)
1	0.14375
2	0.275
3	0.26875
4	0.1125
5	0.08125
6	0.04375
7	0.025
8	0.01875
9	0.0125
10	0.0125
20	0.00625

(d) Credit Cards



The distribution is skewed to the right.

(e) $\mu_X = 3.3; \sigma_X = 2.3$

(f) 0.05; this result is a little unusual.

(g) $P(\text{two with exactly two credit cards}) = 0.0744$. If we surveyed two individuals 100 different times, we would expect about seven of the surveys to result in two people with two credit cards.

6.2 Assess Your Understanding (page 324)

1. There are a fixed number of trials; the trials are independent; for each trial, there are two mutually exclusive outcomes called success or failure; the probability of success is the same for each trial.

3. True 5. np

7. Not binomial, because the random variable is continuous

9. Binomial

11. Not binomial, because the trials are not independent

13. Not binomial, because the number of trials is not fixed

15. Binomial

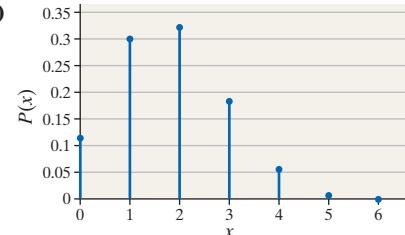
17. $P(3) = 0.2150$ 19. $P(38) = 0.0532$

21. $P(3) = 0.2786$ 23. $P(X \leq 3) = 0.9144$

25. $P(X > 3) = 0.5$ 27. $P(X \leq 4) = 0.5833$

x	P(x)	x	P(x)
0	0.1176	4	0.0595
1	0.3025	5	0.0102
2	0.3241	6	0.0007
3	0.1852		

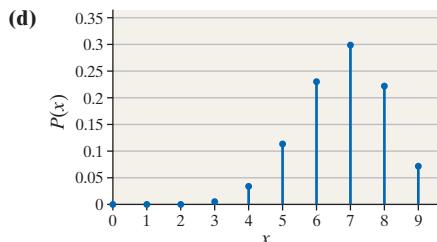
- (b) $\mu_X = 1.8; \sigma_X = 1.1$ (c) $\mu_X = 1.8; \sigma_X = 1.1$



The distribution is skewed right.

x	P(x)	x	P(x)
0	0.0000	5	0.1168
1	0.0001	6	0.2336
2	0.0012	7	0.3003
3	0.0087	8	0.2253
4	0.0389	9	0.0751

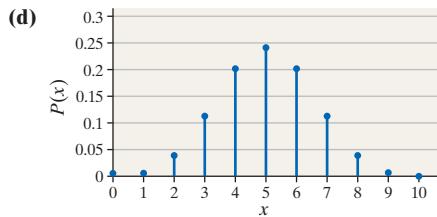
(b) $\mu_X = 6.75; \sigma_X = 1.3$



The distribution is skewed left.

x	P(x)	x	P(x)
0	0.0010	6	0.2051
1	0.0098	7	0.1172
2	0.0439	8	0.0439
3	0.1172	9	0.0098
4	0.2051	10	0.0010
5	0.2461		

(b) $\mu_X = 5; \sigma_X = 1.6$



The distribution is symmetric.

35. (a) This is a binomial experiment because:

1. It is performed a fixed number of times, $n = 15$.

2. The trials are independent.

3. For each trial, there are two possible mutually exclusive outcomes: on time and not on time.

4. The probability of “on time” is fixed at $p = 0.80$.

(b) $n = 15; p = 0.8$

(c) $P(10) = 0.1032$; in 100 trials of this study, we expect about 10 trials to result in exactly 10 flights being on time.

(d) $P(X < 10) = 0.0611$; in 100 trials of this study, we expect about 6 trials to result in fewer than 10 flights being on time.

(e) $P(X \geq 10) = 0.9389$; in 100 trials of this study, we expect about 94 trials to result in at least 10 flights being on time.

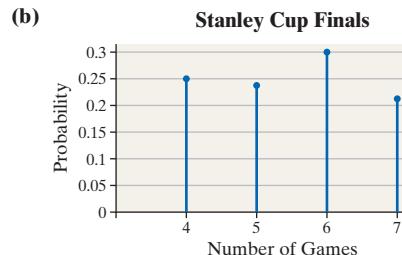
- (f) $P(8 \leq X \leq 10) = 0.16$; in 100 trials of this study, we expect about 16 trials to result in between 8 and 10 flights, inclusive, being on time.
- 37.** (a) This is a binomial experiment because:
1. It is performed a fixed number of times, $n = 20$.
 2. The trials are independent.
 3. For each trial there are two mutually exclusive outcomes: either the individual flushes a public toilet with their foot, or not.
 4. The probability an individual flushes a public toilet with their foot is fixed at $p = 0.64$.
- (b) $P(12) = 0.1678$. In 100 different surveys of 20 adult Americans, we would expect 17 of the surveys to result in exactly 12 who flush a public toilet with their foot.
- (c) $P(X \geq 16) = 0.1011$. In 100 different surveys of 20 adult Americans, we would expect 10 of the surveys to result in at least 16 who flush a public toilet with their foot.
- (d) $P(9 \leq X \leq 11) = 0.2436$. In 100 different surveys of 20 adult Americans, we would expect 24 of the surveys to result in 9 to 11, inclusive, who flush a public toilet with their foot.
- (e) Yes, $P(X > 17) = 0.0096$ under the assumption the proportion of adult Americans who flush a public toilet with their foot is 0.64. We would expect these results in only 1 of every 100 surveys.
- 39.** (a) $P(4) = 0.1655$ (b) $P(X < 3) = 0.4752$
 (c) Yes; the probability is 0.0272.
- 41.** (a) 0.1667 (b) $P(X \leq 2) = 0.0421$
 (c) The number of Hispanics on the jury is unusually low, given the composition of the population from which it came.
- 43.** (a) $\mu_X = 80$ flights; $\sigma_X = 4$ flights
 (b) We expect that, in a random sample of 100 flights from Dallas to Chicago, 80 will be on time.
 (c) It would not be unusual to have 75 on-time flights because 75 is within 2 standard deviations of the mean.
- 45.** (a) $E(X) = \mu_X = 320$
 (b) Yes, because 280 is more than 2 standard deviations below what we would expect.
- 47.** We would expect 824 out of 1030 parents to spank their children. The results suggest that parents' attitudes have changed since $781 < \mu_X - 2\sigma_X = 824 - 2(12.8) = 798.4$.
- 49.** We would expect $500,000(0.56) = 280,000$ of the stops to be pedestrians who are nonwhite. Because $500,000(0.56)(1 - 0.56) = 123,200 \geq 10$, we can use the Empirical Rule to identify cutoff points for unusual results. The standard deviation number of stops is 351. If the number of stops of nonwhite pedestrians exceeds $280,000 + 2(351) = 280,702$, we would say the result is unusual. The actual number of stops is $500,000(0.89) = 445,000$, which is definitely unusual. A potential criticism of this analysis is the use of 0.44 as the proportion of whites, since the actual proportion of whites may be different due to individuals commuting back and forth to the city.
- 51.** (a) $P(55 \text{ or } 56) = P(55) + P(56) = 0.0203$
 (b) $P(\text{bumped}) = P(X \geq 55) = 0.4423$
 (c) 266 tickets
- 53.** (a) $P(3) = 0.1187$
 (b)
- | x | $P(x)$ | x | $P(x)$ |
|-----|--------|-----|--------|
| 1 | 0.5240 | 6 | 0.0128 |
| 2 | 0.2494 | 7 | 0.0061 |
| 3 | 0.1187 | 8 | 0.0029 |
| 4 | 0.0565 | 9 | 0.0014 |
| 5 | 0.0269 | 10 | 0.0007 |
- (c) $\mu_X = 1.9$ free throws
 (d) $\mu_X = 1.9$ free throws; 1.9 free throws
- 55.** (a) A completely randomized design
 (b) Death, or not. This is qualitative with two possible outcomes.
 (c) 22
- (d) Neither the subject nor the individual administering the treatment knew which treatment group the subject belonged to.
 (e) The subjects were assigned to either the group receiving the experimental drug or placebo group by chance.
 (f) The probability of 10 or more subjects out of 12 dying using a probability of death of 0.3 (from the placebo group) is 0.0002. These results are unlikely to occur due to chance, which suggests the experimental drug may have contributed to the death of the subjects.
- 57.** Success means obtaining the outcome you are measuring.
- 59.** When n is small, the shape of the binomial distribution is determined by p . If p is close to zero, the distribution is skewed right; if p is close to 0.5, the distribution is approximately symmetric; and if p is close to one, the distribution is skewed left.

Chapter 6 Review Exercises (page 329)

- 1.** (a) Discrete, $s = 0, 1, 2, \dots$
 (b) Continuous, $s \geq 0$
 (c) Continuous, $h \geq 0$
 (d) Discrete, $x = 0, 1, 2, \dots$
- 2.** (a) It is not a probability distribution because $\sum P(x) \neq 1$.
 (b) It is a probability distribution.

3. (a)

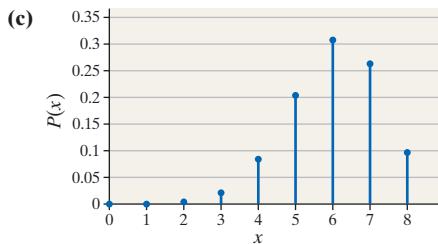
x	$P(x)$
4	0.25
5	0.2375
6	0.3
7	0.2125



- 5.** (a) Binomial experiment
 (b) Not a binomial experiment because the number of trials is not fixed.
- 6.** (a) $P(1) = 0.3151$; If we randomly selected 10 visitors to the ER 100 different times, we would expect about 32 of those times to result in exactly 1 of 10 visitors dying within one year.
 (b) $P(X < 2) = 0.6424$ (c) $P(X \geq 2) = 0.3576$
 (d) $P(X \leq 2) = 0.9885$ (e) No; $P(X > 3) = 0.0608$
 (f) $\mu_X = 50$; $\sigma_X = 6.9$
 (g) When $n = 800$, $\mu_X = 40$ and $\sigma_X = 6.2$, so 51 lies within two standard deviations of the mean. The hospital should not be investigated.
- 7.** (a) $P(10) = 0.1859$
 (b) $P(X < 5) = 0.0093$
 (c) $P(X \geq 5) = 0.9907$
 (d) $P(7 \leq X \leq 12) = 0.8779$ [Tech: 0.8778]
 (e) $\mu_X = 120$ women; $\sigma_X = 6.9$ women
 (f) It would not be unusual to have 110 females in a sample of 200 believe that the driving age should be 18, because 110 is within 2 standard deviations of the mean.

8. (a)	x	$P(x)$	x	$P(x)$
	0	0.00002	5	0.20764
	1	0.00037	6	0.31146
	2	0.00385	7	0.26697
	3	0.02307	8	0.10011
	4	0.08652		

(b) $\mu_X = 6; \sigma_X = 1.2$



The distribution is skewed left.

9. As a rule of thumb, if X is binomially distributed, the Empirical Rule can be used when $np(1 - p) \geq 10$.

10. We can sample without replacement and use the binomial probability distribution to approximate probabilities when the sample size is small in relation to the population size. As a rule of thumb, if the sample size is less than 5% of the population size, the trials can be considered nearly independent.

11. $P(X \geq 12) = 0.03$, so the result of the survey is unusual. This suggests that emotional abuse may be a factor that increases the likelihood of self-injurious behavior.

Chapter 6 Test (page 330)

1. (a) Discrete; $r = 0, 1, 2, \dots, 365$

- (b) Continuous; $m > 0$

- (c) Discrete; $x = 0, 1, 2, \dots$

- (d) Continuous; $w > 0$

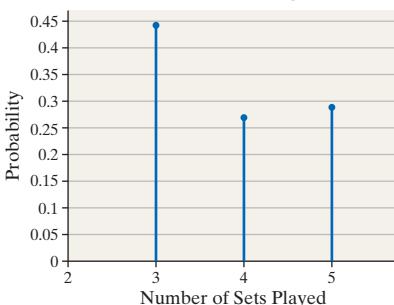
2. (a) It is a probability distribution.

- (b) It is not a probability distribution because $P(4) = -0.11$, which is negative.

3. (a)

x	$P(x)$
3	0.4423
4	0.2692
5	0.2885

(b) **Wimbledon Men's Singles Finals**



- (c) $\mu_X = 3.8$ sets. On average, we expect the Wimbledon men's singles final match for the championship to require 3.8 sets of play.

- (d) $\sigma_X = 0.7$ set

4. $E(X) = 72.50$; the insurance company expects to earn an average profit of \$72.50 for each 35-year-old male it insures for one year for \$100,000.

5. An experiment is binomial provided:

1. The experiment consists of a fixed number, n , of trials.
2. The trials are independent.

3. Each trial has two possible mutually exclusive outcomes: success and failure.
4. The probability of success, p , remains constant for each trial of the experiment.

6. (a) Not a binomial experiment because the trials are not independent and the number of trials is not fixed

- (b) Binomial experiment.

7. (a) $P(15) = 0.1746$; in 100 trials of this experiment, we expect about 17 trials to result in exactly 15 married people who hide purchases from their mates.

- (b) $P(X \geq 19) = 0.0692$; in 100 trials of this experiment, we expect about 7 trials to result in at least 19 married people who hide purchases from their mates.

- (c) $P(X < 19) = 0.9308$; in 100 trials of this experiment, we expect about 93 trials to result in fewer than 19 married people who hide purchases from their mates.

- (d) $P(15 \leq X \leq 17) = 0.5981$; in 100 trials of this experiment, we expect about 60 trials to result in between 15 and 17 married people, inclusive, who hide purchases from their mates.

8. (a) We would expect 600 out of 1200 adult Americans to say they pay too much tax.

- (b) We can use the Empirical Rule to identify unusual events since $np(1 - p) = 300 \geq 10$.

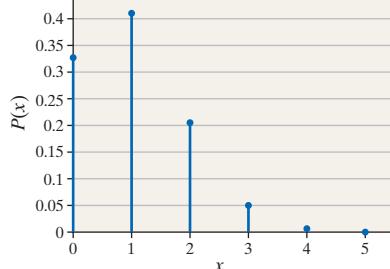
- (c) Yes; these results do contradict the belief since 640 is more than two standard deviations above the mean.

9. (a)

x	$P(x)$
0	0.3277
1	0.4096
2	0.2048
3	0.0512
4	0.0064
5	0.0003

- (b) $\mu_X = 1; \sigma_X = 0.9$

- (c)



The distribution is skewed right.

CHAPTER 7 The Normal Probability Distribution

7.1 Assess Your Understanding (page 339)

1. • The total area under the graph of the equation over all possible values of the random variable must equal 1.

- The height of the graph of the equation must be greater than or equal to 0 for all possible values of the random variable.

3. • The normal curve is symmetric about its mean.

- The normal curve has a single peak at its mean.

- The normal curve has inflection points at $x = \mu - \sigma$ and $x = \mu + \sigma$.

- The area under the normal curve is 1.

- The area under the normal curve to the right of the mean equals $\frac{1}{2}$ the area under the curve to the left of the mean, which equals $\frac{1}{2}$.

ANS-42 ANSWERS 7.2 Assess Your Understanding

- As x increases without bound (gets larger and larger), the graph approaches, but never reaches, the horizontal axis. As x decreases without bound (gets more and more negative), the graph approaches, but never reaches, the horizontal axis.
- The normal curve conforms to the Empirical Rule—that is, approximately 68% of the observations lie within one standard deviation of the mean; approximately 95% of the observations lie within two standard deviations of the mean; approximately 99.7% of the observations lie within three standard deviations of the mean.

5. This graph is not symmetric; it cannot represent a normal density function.

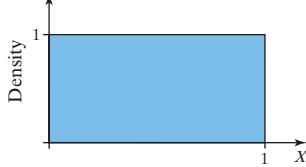
7. This graph is not always greater than zero; it cannot represent a normal density function.

9. The graph can represent a normal density function.

11. (a) $P(5 \leq X \leq 10) = \frac{1}{6}$ (b) 12

13. $P(X \geq 20) = \frac{1}{3}$

15. (a)



(b) $P(0 \leq X \leq 0.2) = 0.2$
(d) $P(X > 0.95) = 0.05$

(c) $P(0.25 \leq X \leq 0.6) = 0.35$
(e) Answers will vary.

17. Normal

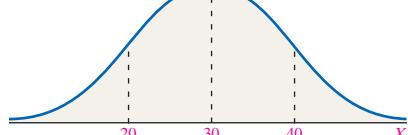
19. Not normal

21. Graph A: $\mu = 10$, $\sigma = 2$; graph B: $\mu = 10$, $\sigma = 3$. A larger standard deviation makes the graph lower and more spread out.

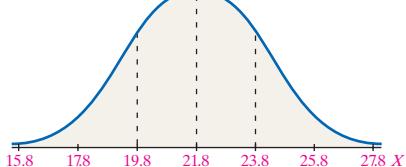
23. $\mu = 2$, $\sigma = 3$

25. $\mu = 100$, $\sigma = 15$

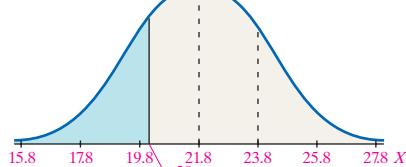
27.



29. (a)



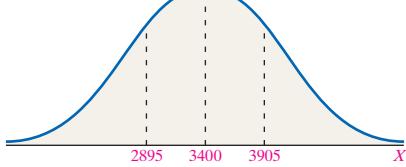
(b)



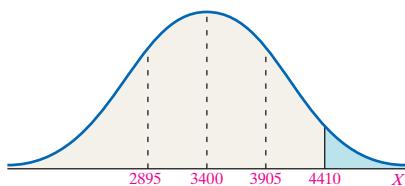
(c) (i) The proportion of 4-year-old males whose upper arm length is less than 20 cm is 0.1841.

(ii) The probability a randomly selected 4-year-old male will have an upper arm length less than 20 cm is 0.1841. That is, if we randomly selected 100 4-year-old males, we would expect 18 of them to have an upper arm length less than 20 cm.

31. (a)



(b)



(c) (i) 2.28% of all full-term babies have a birth weight of more than 4410 grams.

(ii) The probability is 0.0228 that the birth weight of a randomly chosen full-term baby is more than 4410 grams.

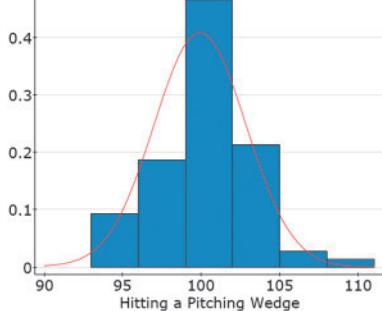
33. (a) (i) The proportion of human pregnancies that last more than 280 days is 0.1908.

(ii) The probability that a randomly selected human pregnancy lasts more than 280 days is 0.1908.

(b) (i) The proportion of human pregnancies that last between 230 and 260 days is 0.3416.

(ii) The probability that a randomly selected human pregnancy lasts between 230 and 260 days is 0.3416.

35. (a) Hitting a Pitching Wedge
Relative Frequency



(b) Answers will vary.

37. (a) This is an observational study in which the data are collected over time.

(b) Explanatory variable: hypothermia or not. It is qualitative.

(c) Survival status (qualitative); ICU stay (quantitative); time on ventilator (quantitative)

(d) Statistic

(e) Male patients who have an out-of-hospital cardiac arrest

(f) $P(\text{survive} \mid \text{hypothermia}) = 37/52 = 0.712$;
 $P(\text{survive} \mid \text{no hypothermia}) = 43/74 = 0.581$

7.2 Assess Your Understanding (page 351)

1. standard normal distribution

3. 0.3085

5. (a) Area = 0.0071

(b) Area = 0.3336

(c) Area = 0.9115

(d) Area = 0.9998

7. (a) Area = 0.9987

(b) Area = 0.9441

(c) Area = 0.0375

(d) Area = 0.0009

9. (a) Area = 0.9586

(b) Area = 0.2088

(c) Area = 0.8479

11. (a) Area = 0.0456 [Tech: 0.0455]

(b) Area = 0.0646

(c) Area = 0.5203 [Tech: 0.5202]

13. $z = -1.28$

15. $z = 0.67$

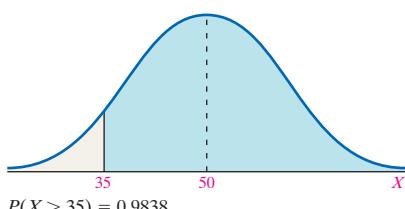
17. $z_1 = -2.575$ [Tech: -2.576];

$z_2 = 2.575$ [Tech: 2.576]

19. $z_{0.01} = 2.33$

21. $z_{0.025} = 1.96$

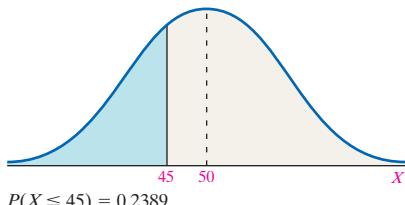
23.



$$P(X > 35) = 0.9838$$

Tech: $P(X > 35) = 0.9839$

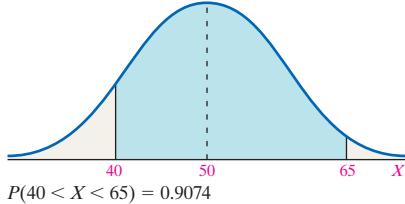
25.



$$P(X \leq 45) = 0.2389$$

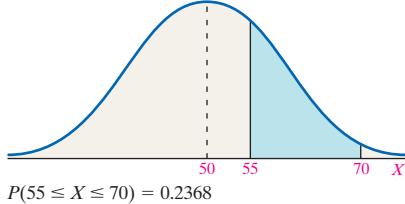
Tech: $P(X \leq 45) = 0.2375$

27.



$$P(40 < X < 65) = 0.9074$$

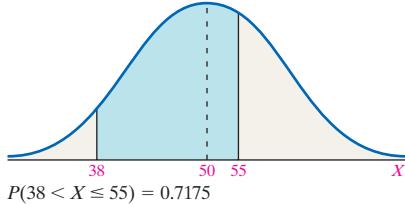
29.



$$P(50 \leq X \leq 70) = 0.2368$$

Tech: $P(50 \leq X \leq 70) = 0.2354$

31.



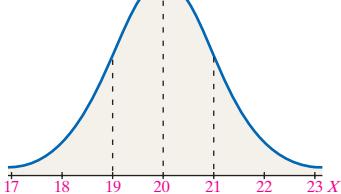
$$P(38 < X \leq 55) = 0.7175$$

Tech: $P(38 < X \leq 55) = 0.7192$

33. $x = 40.62$ [Tech: 40.61] is at the 9th percentile.

35. $x = 56.16$ [Tech: 56.15] is at the 81st percentile.

37. (a)



(b) $P(X < 20) = 0.1587$. If 100 eggs are randomly selected, we would expect 16 to incubate in less than 20 days.

(c) $P(X > 22) = 0.1587$. If 100 eggs are randomly selected, we would expect 16 to incubate in more than 22 days.

(d) $P(19 \leq X \leq 21) = 0.4772$. If 100 eggs are randomly selected, we would expect 48 to incubate in between 19 and 21 days.

(e) Yes; $P(X < 18) = 0.0013$. The model suggests that about 1 egg in 1000 incubates in less than 18 days.

39. (a) $P(1000 \leq X \leq 1400) = 0.8658$ [Tech: 0.8657]

(b) $P(X < 1000) = 0.0132$

(c) 0.7019 [Tech: 0.7004] of the bags have more than 1200 chocolate chips.

(d) 0.1230 [Tech: 0.1228] of the bags have fewer than 1125 chocolate chips.

(e) A bag that contains 1475 chocolate chips is at the 96th percentile.

(f) A bag that contains 1050 chocolate chips is at the 4th percentile.

41. (a) 0.4013 of pregnancies last more than 270 days.

(b) 0.1587 of pregnancies last fewer than 250 days.

(c) 0.7590 [Tech: 0.7571] of pregnancies last between 240 and 280 days.

(d) $P(X > 280) = 0.1894$ [Tech: 0.1908]

(e) $P(X \leq 245) = 0.0951$ [Tech: 0.0947]

(f) Yes; 0.0043 of births are very preterm. So about 4 births in 1000 births are very preterm.

43. (a) 0.0764 [Tech: 0.0766] of the rods have a length of less than 24.9 cm.

(b) 0.0324 [Tech: 0.0321] of the rods will be discarded.

(c) The plant manager expects to discard 162 [Tech: 161] of the 5000 rods manufactured.

(d) To meet the order, the plant manager should manufacture 11,804 [Tech: 11,808] rods.

45. (a) The favored team is equally likely to win or lose relative to the spread. Yes; a mean of 0 implies the spreads are accurate.

(b) $P(X \geq 5) = 0.3228$ [Tech: 0.3232]

(c) $P(X \leq -2) = 0.4286$ [Tech: 0.4272]

47. (a) The 17th percentile for incubation time is 20 days.

(b) From 19 to 23 days make up the middle 95% of incubation times of the eggs.

49. (a) The 30th percentile for the number of chips in an 18-ounce bag is 1201 [Tech: 1200] chips.

(b) The middle 99% of the bags contain between 958 and 1566 chocolate chips.

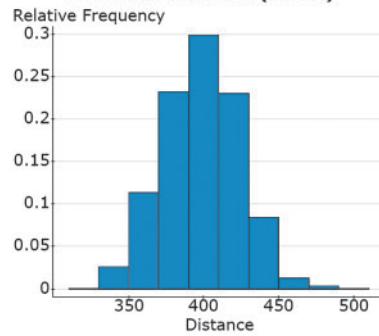
(c) $Q_1 = 1183$ [Tech: 1182]; $Q_3 = 1341$ [Tech: 1342]; $IQR = 158$ [Tech: 160]

51. (a) 11.51% of customers receive the service for half-price.

(b) So that no more than 3% receives the discount, the guaranteed time limit should be 22 minutes.

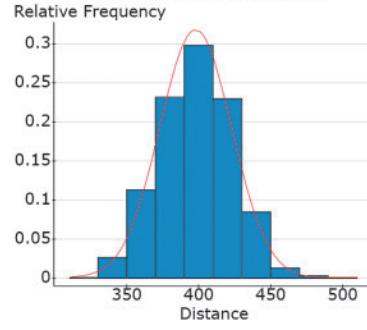
53. (a) The distribution is bell shaped.

Home Run Distances (in feet)



(b) $\mu = 397.6$ feet; $\sigma = 25.1$ feet

Home Run Distances (in feet)



(d) The area under the normal curve between 350 and 370 feet is 0.1068. The relative frequency with which a home run distance is between 350 and 370 feet is 0.1135. These values are very close, which suggests the normal model does a good job describing the variable "distance".

(e) The area under the normal curve to the right of 410 feet is 0.3106. The relative frequency with which a home run distance is at least 410 feet is 0.3293.

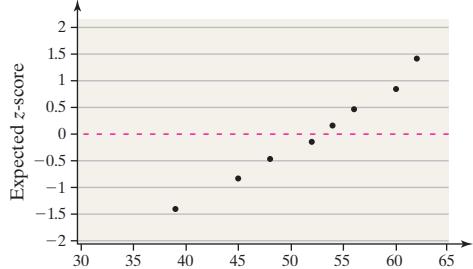
55. Reporting the probability as < 0.0001 accurately describes the event as possible but highly unlikely. Reporting the probability as 0.0000 might be incorrectly interpreted to mean that the event is impossible.

57. You did better on the SAT. On the ACT, you scored in the 83rd percentile. On the SAT, you scored in the 85th percentile.

7.3 Assess Your Understanding (page 358)

1. normal probability plot

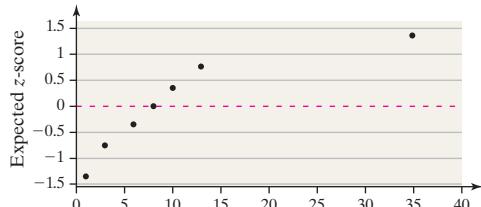
3. (a)



(b) 0.991

(c) Critical value: 0.906; the sample data come from a population that is normally distributed because $0.991 > 0.906$.

5. (a)



(b) 0.873

(c) Critical value: 0.898; the sample data do not come from a population that is normally distributed because $0.873 < 0.898$.

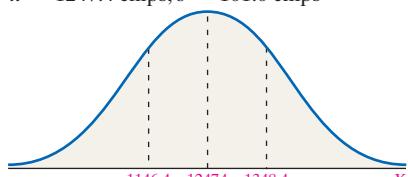
7. Because $0.979 > 0.951$ the sample data could come from a normally distributed population.

9. Because $0.88 < 0.951$ the sample data do not come from a normally distributed population.

11. (a) The sample data come from a normally distributed population.

(b) $\bar{x} = 1247.4$ chips, $s = 101.0$ chips

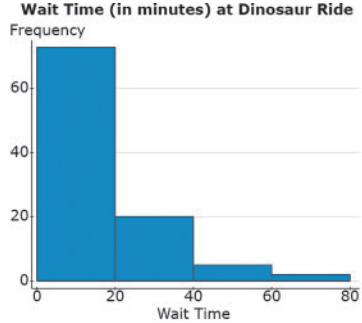
(c)



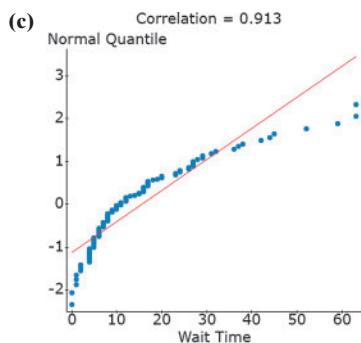
(d) $P(X \geq 1000) = 0.9929$ [Tech: 0.9928]

(e) $P(1200 \leq X \leq 1400) = 0.6153$ [Tech: 0.6152]

13. (a)



(b) Skewed right



Because $0.913 < 0.960$ (Table VI with $n = 30$), the data do not appear to come from a population that is approximately normal.

7.4 Assess Your Understanding (page 363)

1. $np(1-p) \geq 10; np; \sqrt{np(1-p)}$

3. $P(X \leq 4.5)$

5. Area under the normal curve to the right of $x = 39.5$

7. Area under the normal curve between $x = 7.5$ and $x = 8.5$

9. Area under the normal curve between $x = 17.5$ and $x = 24.5$

11. Area under the normal curve to the right of $x = 20.5$

13. Area under the normal curve to the left of $x = 500.5$

15. Using the binomial formula, $P(20) = 0.0616; np(1-p) = 14.4$, the normal distribution can be used. Approximate probability is 0.0618. [Tech: 0.0603]

17. Using the binomial formula, $P(30) < 0.0001; np(1-p) = 7.5$, the normal distribution cannot be used.

19. Using the binomial formula, $P(60) = 0.0677; np(1-p) = 14.1$, the normal distribution can be used. Approximate probability is 0.0630. [Tech: 0.0645]

21. (a) $P(130) \approx 0.0444$ [Tech: 0.0431]

(b) $P(X \geq 130) \approx 0.9332$ [Tech: 0.9328]

(c) $P(X < 125) \approx 0.0021$

(d) $P(125 \leq X \leq 135) \approx 0.5536$ [Tech: 0.5520]

23. (a) $P(490) \approx 0.0127$ [Tech: 0.0139]

(b) $P(X \leq 490) \approx 0.9015$ [Tech: 0.9022]

(c) $P(X \geq 503) \approx 0.0136$ [Tech: 0.0134]; the result suggests the proportion of adult women 18–24 years of age is higher than 0.64.

25. (a) $P(X \geq 130) \approx 0.0028$

(b) The result is unusual. Less than 3 samples in 1000 will result in 130 or more living at home if the true percentage is 55%. Perhaps a higher percentage of males are now living at home.

27. (a) $P(X \geq 103) \approx 0.004$

(b) About 4 samples in 1000 will result in 103 or more respondents having a positive view of socialism if the true percentage is 42%. This suggests that a higher proportion of students at this school have a positive view of socialism than adult Americans.

Chapter 7 Review Exercises (page 365)

1. (a) $\mu = 60$

(b) $\sigma = 10$

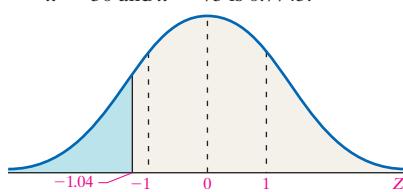
(c) (i) The proportion of values for the random variable to the right of $x = 75$ is 0.0668.

(ii) The probability that a randomly selected value is greater than $x = 75$ is 0.0668.

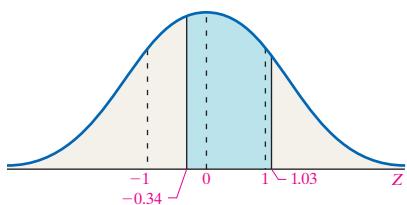
(d) (i) The proportion of values for the random variable between $x = 50$ and $x = 75$ is 0.7745.

(ii) The probability that a randomly selected value is between $x = 50$ and $x = 75$ is 0.7745.

2.



Area under the normal curve to the left of -1.04 is 0.1492.

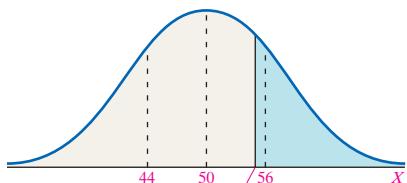
3.

Area under the normal curve between -0.34 and 1.03 is 0.4816 .

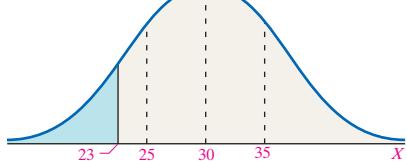
4. $z = 0.04$

5. $z_1 = -1.75$ and $z_2 = 1.75$

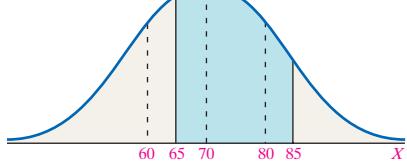
6. $z_{0.20} = 0.84$

7.

$$P(X > 55) = 0.2033 \text{ [Tech: 0.2023]}$$

8.

$$P(X \leq 23) = 0.0808$$

9.

$$P(65 < X < 85) = 0.6247$$

10. (a) 0.1271 [Tech: 0.1279] of the tires last at least 75,000 miles.

(b) 0.0116 [Tech: 0.0115] of the tires last at most 60,000 miles.

(c) $P(65,000 \leq X \leq 80,000) = 0.8613$ [Tech: 0.8606]

(d) The company should advertise 60,980 [Tech: 60,964] miles as its mileage warranty.

11. (a) The probability is 0.0475 [Tech: 0.0478] that the test taker scores above 125.

(b) The probability is 0.2514 [Tech: 0.2525] that the test taker scores below 90.

(c) 0.2476 [Tech: 0.2487] of the test takers score between 110 and 140.

(d) $P(X > 150) = 0.0004$

(e) A score of 131 places a child at the 98th percentile.

(f) Normal children score between 71 and 129 points.

12. (a) 0.0119 [Tech: 0.0120] of the baseballs produced are too heavy for use.

(b) 0.0384 [Tech: 0.0380] of the baseballs produced are too light to use.

(c) 0.9497 [Tech: 0.9500] of the baseballs produced are within acceptable weight limits.

(d) 8424 [Tech: 8421] should be manufactured.

13. (a) $np(1-p) = 62.1 > 10$, so the normal distribution can be used to approximate the binomial probabilities.

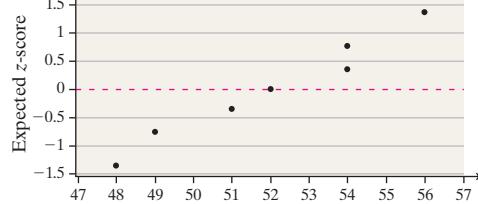
(b) $P(125) \approx 0.0213$ [Tech: 0.0226] Interpretation: Approximately 2 of every 100 random samples of 250 adult Americans will result in exactly 125 who state that they have read at least 6 books within the past year.

(c) $P(X < 120) \approx 0.7157$ [Tech: 0.7160] Interpretation: Approximately 72 of every 100 random samples of 250 adult

Americans will result in fewer than 120 who state that they have read at least 6 books within the past year.

(d) $P(X \geq 140) \approx 0.0009$ Interpretation: Approximately 1 of every 1000 random samples of 250 adult Americans will result in 140 or more who state that they have read at least 6 books within the past year.

(e) $P(100 \leq X \leq 120) \approx 0.7336$ [Tech: 0.7328] Interpretation: Approximately 73 of every 100 random samples of 250 adult Americans will result in between 100 and 120, inclusive, who state that they have read at least six books within the past year.

14. (a)

(b) 0.986

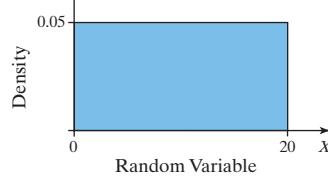
(c) Because $0.986 > 0.898$ (Table VI), the sample data could come from a population that is normally distributed.

15. Because $0.914 < 0.951$ (Table VI), the sample data do not come from a population that is normally distributed.

16. Because $0.873 < 0.959$ (Table VI with $n = 25$), the sample data do not come from a population that is normally distributed.

17. (a) $P(X \leq 30) = 0.0010$

(b) Yes; this contradicts the *USA Today* "Snapshot." About 1 sample in 1000 will result in 30 or fewer who do their most creative thinking while driving, if the true percentage is 20%.

18. (a)

(b) $P(0 \leq X \leq 5) = 0.25$

(c) $P(10 \leq X \leq 18) = 0.4$

19. A normal curve has the following properties:

1. It is symmetric about its mean μ .

2. Its highest point occurs at μ .

3. It has inflection points at $\mu - \sigma$ and $\mu + \sigma$.

4. The area under the curve is 1.

5. The area under the curve to the right of μ equals the area under the curve to the left of μ . Both equal $\frac{1}{2}$.

6. As the value of X increases, the graph approaches but never equals zero. When the value of X decreases, the graph approaches but never equals zero.

7. The Empirical Rule: Approximately 68% of the area under the standard normal curve is between $\mu - \sigma$ and $\mu + \sigma$.

Approximately 95% of the area under the standard normal curve is between $\mu - 2\sigma$ and $\mu + 2\sigma$. Approximately 99.7% of the area under the standard normal curve is between $\mu - 3\sigma$ and $\mu + 3\sigma$.

20. The graph plots actual observations against expected z -scores, assuming that the data are normal. If the plot is not linear, then we have evidence that the data are not normal.

Chapter 7 Test (page 367)

1. (a) $\mu = 7$

(b) $\sigma = 2$

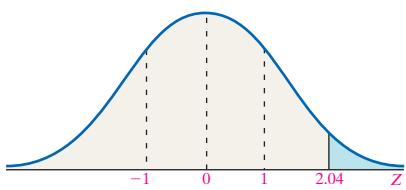
(c) (i) The proportion of values for the random variable to the left of $x = 10$ is 0.9332.

(ii) The probability that a randomly selected value is less than $x = 10$ is 0.9332.

(d) (i) The proportion of values for the random variable between $x = 5$ and $x = 8$ is 0.5328.

(ii) The probability that a randomly selected value is between $x = 5$ and $x = 8$ is 0.5328.

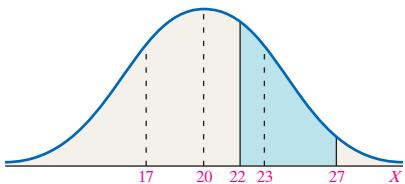
2.



Area under the normal curve to the right of 2.04 is 0.0207.

3. $z_1 = -1.555; z_2 = 1.555$

5. (a)



(b) $P(22 \leq X \leq 27) = 0.2415$ [Tech: 0.2427]

6. (a) 0.8944 of the time, the iPhone will last at least 6 hours.

(b) There is a 0.0062 probability that the iPhone will last less than 5 hours. This is an unusual result.

(c) About 8.3 hours would be the cutoff for the top 5% of all talk times.

(d) Yes; $P(X > 9) = 0.0062$. About 6 out of every 1000 full charges will result in the iPhone lasting more than 9 hours.

7. (a) The proportion of 20- to 29-year-old males whose waist circumference is less than 100 cm is 0.7088 [Tech: 0.7080].

(b) $P(80 \leq X \leq 100) = 0.5274$ [Tech: 0.5272]

(c) Waist circumferences between 70 and 115 cm make up the middle 90% of all waist circumferences.

(d) A waist circumference of 75 cm is at the 10th percentile.

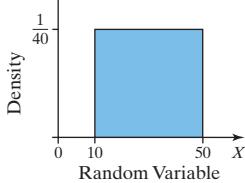
8. A: ≥ 76.4 ; B: 70.7–76.3; C: 57.3–70.6; D: 51.6–57.2; F: < 51.6

9. (a) $P(100) \approx 0.0025$

(b) $P(X < 60) \approx 0.0062$

10. The data likely come from a population that is normally distributed because $0.974 > 0.939$ (Table VI).

11. (a)



(b) $P(20 \leq X \leq 30) = 0.25$

(c) $P(X < 15) = 0.125$

CHAPTER 8 Sampling Distributions

8.1 Assess Your Understanding (page 379)

1. sampling distribution

3. standard error; mean

5. False

7. The sampling distribution of \bar{x} is approximately normal with $\mu_{\bar{x}} = 30$ and $\sigma_{\bar{x}} = \frac{8}{\sqrt{10}} \approx 2.530$.

9. $\mu_{\bar{x}} = 80, \sigma_{\bar{x}} = 2$

11. $\mu_{\bar{x}} = 52, \sigma_{\bar{x}} = \frac{10}{\sqrt{21}} \approx 2.182$

13. (a) $\mu_{\bar{x}} = 500$

(b) $\sigma_{\bar{x}} = 20$

(c) The population must be normally distributed.

(d) $\sigma = 80$

15. (a) \bar{x} is approximately normal with $\mu_{\bar{x}} = 80, \sigma_{\bar{x}} = 2$.

(b) $P(\bar{x} > 83) = 0.0668$. If we take 100 simple random samples of size $n = 49$ from a population with $\mu = 80$ and $\sigma = 14$, then about 7 of the samples will result in a mean that is greater than 83.

(c) $P(\bar{x} \leq 75.8) = 0.0179$. If we take 100 simple random samples of size $n = 49$ from a population with $\mu = 80$ and $\sigma = 14$, then about 2 of the samples will result in a mean that is less than or equal to 75.8.

(d) $P(78.3 < \bar{x} < 85.1) = 0.7969$ [Tech: 0.7970]. If we take 100 simple random samples of size $n = 49$ from a population with $\mu = 80$ and $\sigma = 14$, then about 80 of the samples will result in a mean that is between 78.3 and 85.1.

17. (a) The population must be normally distributed to compute probabilities involving the sample mean. If the population is normally distributed, then the sampling distribution of \bar{x} is approximately normal with $\mu_{\bar{x}} = 64$ and $\sigma_{\bar{x}} = \frac{17}{\sqrt{12}} \approx 4.907$.

(b) $P(\bar{x} < 67.3) = 0.7486$ [Tech: 0.7493]. If we take 100 simple random samples of size $n = 12$ from a population that is normally distributed with $\mu = 64$ and $\sigma = 17$, then about 75 of the samples will result in a mean that is less than 67.3.

(c) $P(\bar{x} \geq 65.2) = 0.4052$ [Tech: 0.4034]. If we take 100 simple random samples of size $n = 12$ from a population that is normally distributed with $\mu = 64$ and $\sigma = 17$, then about 40 or 41 of the samples will result in a mean that is greater than or equal to 65.2.

19. (a) $P(X < 260) = 0.3520$ [Tech: 0.3538]. If we randomly select 100 human pregnancies, then about 35 of the pregnancies will last less than 260 days.

(b) The sampling distribution of \bar{x} is normal with $\mu_{\bar{x}} = 266$ and $\sigma_{\bar{x}} = \frac{16}{\sqrt{20}} \approx 3.578$.

(c) $P(\bar{x} \leq 260) = 0.0465$ [Tech: 0.0468]. If we take 100 simple random samples of size $n = 20$ human pregnancies, then about 5 of the samples will result in a mean gestation period of 260 days or less.

(d) $P(\bar{x} \leq 260) = 0.0040$. If we take 1000 simple random samples of size $n = 50$ human pregnancies, then about 4 of the samples will result in a mean gestation period of 260 days or less.

(e) This result would be unusual, so the sample likely came from a population whose mean gestation period is less than 266 days.

(f) $P(256 \leq \bar{x} \leq 276) = 0.9844$ [Tech: 0.9845]. If we take 100 simple random samples of size $n = 15$ human pregnancies, then about 98 of the samples will result in a mean gestation period between 256 and 276 days, inclusive.

21. (a) $P(X > 95) = 0.3085$. If we select a simple random sample of $n = 100$ second-grade students, then about 31 of the students will read more than 95 words per minute.

(b) $P(\bar{x} > 95) = 0.0418$ [Tech: 0.0416]. If we take 100 simple random samples of size $n = 12$ second-grade students, then about 4 of the samples will result in a mean reading rate that is more than 95 words per minute.

(c) $P(\bar{x} > 95) = 0.0071$ [Tech: 0.0072]. If we take 1000 simple random samples of size $n = 24$ second-grade students, then about 7 of the samples will result in a mean reading rate that is more than 95 words per minute.

(d) Increasing the sample size decreases $P(\bar{x} > 95)$. This happens because $\sigma_{\bar{x}}$ decreases as n increases.

(e) A mean reading rate of 92.8 wpm is not unusual since $P(\bar{x} \geq 92.8) = 0.1056$ [Tech: 0.1052]. This means that the new reading program is not abundantly more effective than the old program.

(f) There is a 5% chance that the mean reading speed of a random sample of 20 second-grade students will exceed 93.7 words per minute.

23. (a) $P(X > 0) = 0.5675$ [Tech: 0.5694]. If we select a simple random sample of $n = 100$ months, then about 57 of the months will have positive rates of return.

(b) $P(\bar{x} > 0) = 0.7291$ [Tech: 0.7277]. If we take 100 simple random samples of size $n = 12$ months, then about 73 of the samples will result in a mean monthly rate of return that is positive.

(c) $P(\bar{x} > 0) = 0.8051$ [Tech: 0.8043]. If we take 100 simple random samples of size $n = 24$ months, then about 81 of the samples will result in a mean monthly rate of return that is positive.

- (d) $P(\bar{x} > 0) = 0.8531$ [Tech: 0.8530]. If we take 100 simple random samples of size $n = 36$ months, then about 85 of the samples will result in a mean monthly rate of return that is positive.
 (e) The likelihood of earning a positive rate of return increases as the investment time horizon increases.

25. (a) A sample size of at least 30 is needed to compute the probabilities.
 (b) $P(\bar{x} < 10) = 0.0028$. If we take 1000 simple random samples of size $n = 40$ oil changes, then about 3 of the samples will result in a mean time of less than 10 minutes.
 (c) There is a 10% chance of being at or below a mean oil-change time of 10.8 minutes.
 27. (a) Since we have a large sample ($n = 50 \geq 30$) the Central Limit Theorem allows us to say that the sampling distribution of the mean is approximately normal.

$$(b) \mu_{\bar{x}} = 3, \sigma_{\bar{x}} = \frac{\sqrt{3}}{\sqrt{50}} = \sqrt{\frac{3}{50}} \approx 0.245$$

- (c) $P(\bar{x} \geq 3.6) = 0.0071$ [Tech: 0.0072]. If we take 1000 simple random samples of size $n = 50$ ten-gram portions of peanut butter, then about 7 of the samples will result in a mean of at least 3.6 insect fragments. This result is unusual. We might conclude that the sample comes from a population with a mean higher than 3 insect fragments per ten-gram portion.

29. (a) No; the variable “weekly time spent watching television” is likely skewed right.
 (b) \bar{x} is approximately normal with $\mu_{\bar{x}} = 2.35$ and
 $\sigma_{\bar{x}} = \frac{1.93}{\sqrt{40}} \approx 0.305$.
 (c) $P(2 \leq \bar{x} \leq 3) = 0.8583$ [Tech: 0.8577]. If we take 100 simple random samples of size $n = 40$ adult Americans, then about 86 of the samples will result in a mean time between 2 and 3 hours watching television on a weekday.
 (d) $P(\bar{x} \leq 1.89) = 0.0793$. If we take 100 simple random samples of size $n = 35$ adult Americans who consider themselves to be avid Internet users, then about 8 of the samples will result in a mean time of 1.89 hours or less watching television on a weekday. This result is not unusual, so this evidence is insufficient to conclude that avid Internet users watch less television.

31. (a) $\mu = 45.5$ years
 (b) $(55, 44); (55, 33); (55, 41); (55, 41); (55, 59); (44, 33); (44, 41); (44, 41); (44, 59); (33, 41); (33, 41); (33, 59); (41, 41); (41, 59); (41, 59)$

(c)		\bar{x}	Probability	\bar{x}	Probability
37			$\frac{2}{15}$	48	$\frac{2}{15}$
38.5			$\frac{1}{15}$	49.5	$\frac{1}{15}$
41			$\frac{1}{15}$	50	$\frac{2}{15}$
42.5			$\frac{2}{15}$	51.5	$\frac{1}{15}$
44			$\frac{1}{15}$	57	$\frac{1}{15}$
46			$\frac{1}{15}$		

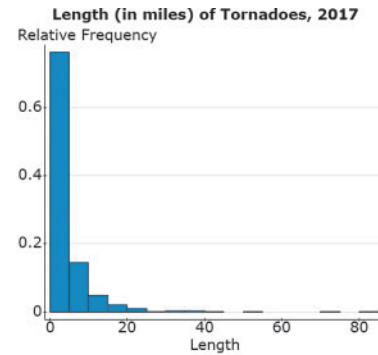
- (d) $\mu_{\bar{x}} = 45.5$ years old
 (e) $P(42.5 \leq \bar{x} \leq 48.5) = \frac{6}{15} = 0.4$

(f)		Samples		Sampling Distribution	
		\bar{x}	Probability	\bar{x}	Probability
55, 44, 33	44, 33, 41	38.3	0.05	39.3	0.10
55, 44, 41	44, 33, 41	42	0.05	43	0.10
55, 44, 41	44, 33, 59	44	0.05	44.3	0.10
55, 44, 59	44, 41, 41	45.3	0.05	45.7	0.05
55, 33, 41	44, 41, 59	46.7	0.10	47	0.05
55, 33, 59	33, 41, 41	48	0.10	49	0.05
55, 41, 41	33, 41, 59	51.7	0.10	52.7	0.05
55, 41, 59	33, 41, 59				
55, 41, 59	41, 41, 59				

$$\mu_{\bar{x}} = 45.5 \text{ years old}; \\ P(42.5 \leq \bar{x} \leq 48.5) = 0.6$$

As the sample size increases, the probability of obtaining a sample mean age within three years of the population mean age increases.

33. (a) Skewed right



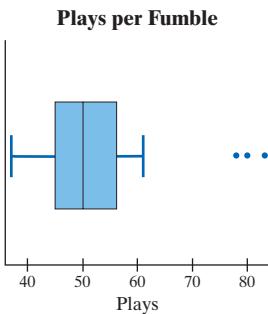
- (b) $\mu = 3.775$ miles; $\sigma = 5.981$ miles
 (c) Because the shape of the distribution is skewed right, the sample size, n , must be large.
 (d) The distribution will be approximately normal with
 $\mu_{\bar{x}} = 3.775$ miles; $\sigma_{\bar{x}} = \frac{5.981}{\sqrt{35}}$ miles.
 (e) $\bar{x} = 2.289$ miles (f) $P(\bar{x} < 2.289) = 0.0708$
 (g) If we obtained 100 different random samples of 35 tornadoes from 2017, we would expect approximately 7 to result in a sample mean less than 2.289 miles.

35. (a)	
x	$P(x)$
35	0.0263
-1	0.9737

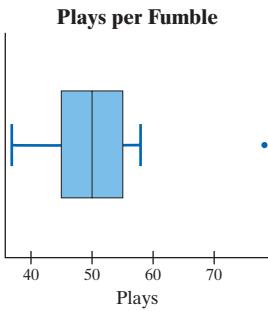
- (b) $\mu = -\$0.05$, $\sigma = \$5.76$
 (c) \bar{x} is approximately normal with $\mu_{\bar{x}} = -\$0.05$ and
 $\sigma_{\bar{x}} = \frac{5.76}{\sqrt{100}} = \0.576 .
 (d) $P(\bar{x} > 0) = 0.4641$ [Tech: 0.4654]
 (e) $P(\bar{x} > 0) = 0.4522$ [Tech: 0.4511]
 (f) $P(\bar{x} > 0) = 0.3936$ [Tech: 0.3918]
 (g) The probability of being ahead decreases as the number of games played increases.

- 37.** The Central Limit Theorem states that, regardless of the distribution of the population, the sampling distribution of the sample means becomes approximately normal as the sample size, n , increases.
- 39.** (b) The distribution for the larger sample size has the smaller standard deviation.
- 41.** (a) We would expect that Jack's distribution would be skewed left but not as much as the original distribution. Diane's distribution should be bell shaped and symmetric, that is, approximately normal.
- (b) We would expect both distributions to have a mean of 50.
- (c) We expect Jack's distribution to have standard deviation $\frac{10}{\sqrt{3}} \approx 5.8$. We expect Diane's distribution to have standard deviation $\frac{10}{\sqrt{30}} \approx 1.8$.
- 43.** (a) The population of interest is full-time students at your college. The sample is the ten students in the simple random sample.
- (b) The number of hours of sleep is a random variable because different people sleep different amounts.
- (c) Because the mean is based on a sample of ten randomly selected students, the mean is a sample mean, and therefore, a statistic.
- (d) The sample mean from part (c) is a random variable because its value will change depending on the ten individuals in the sample. In this study, there are two sources of variation. First, there is person-to-person variability since different people have different sleeping habits. Second, there is variability of sleeping habits of an individual day-to-day (this variability was described in the answer to Problem 42(d)).
- 8.2 Assess Your Understanding (page 389)**
1. 0.44
 3. False
 5. The sampling distribution of \hat{p} is approximately normal when $n \leq 0.05N$ and $np(1 - p) \geq 10$.
 7. The sampling distribution of \hat{p} is approximately normal with $\mu_{\hat{p}} = 0.4$ and $\sigma_{\hat{p}} = \sqrt{\frac{0.4(0.6)}{500}} \approx 0.022$.
 9. The sampling distribution of \hat{p} is approximately normal with $\mu_{\hat{p}} = 0.103$ and $\sigma_{\hat{p}} = \sqrt{\frac{0.103(0.897)}{1000}} \approx 0.010$.
 11. (a) The sampling distribution of \hat{p} is approximately normal with $\mu_{\hat{p}} = 0.8$ and $\sigma_{\hat{p}} = \sqrt{\frac{0.8(0.2)}{75}} \approx 0.046$.
 - (b) $P(\hat{p} \geq 0.84) = 0.1922$ [Tech: 0.1932]. About 19 out of 100 random samples of size $n = 75$ will result in 63 or more individuals (that is, 84% or more) with the characteristic.
 - (c) $P(\hat{p} \leq 0.68) = 0.0047$. About 5 out of 1000 random samples of size $n = 75$ will result in 51 or fewer individuals (that is, 68% or less) with the characteristic.
 13. (a) The sampling distribution of \hat{p} is approximately normal with $\mu_{\hat{p}} = 0.35$ and $\sigma_{\hat{p}} = \sqrt{\frac{0.35(0.65)}{1000}} \approx 0.015$.
 - (b) $P(\hat{p} \geq 0.39) = 0.0040$. About 4 out of 1000 random samples of size $n = 1000$ will result in 390 or more individuals (that is, 39% or more) with the characteristic.
 - (c) $P(\hat{p} \leq 0.32) = 0.0233$ [Tech: 0.0234]. About 2 out of 100 random samples of size $n = 1000$ will result in 320 or fewer individuals (that is, 32% or less) with the characteristic.
 15. (a) Qualitative with two possible outcomes—order a meal in a foreign language, or not.
 - (b) The source of the variability is the individuals in the survey and their ability to order a meal in a foreign language.
 - (c) The sampling distribution of \hat{p} is approximately normal with $\mu_{\hat{p}} = 0.47$ and $\sigma_{\hat{p}} = \sqrt{\frac{0.47(0.53)}{200}} \approx 0.035$.
- (d)** $P(\hat{p} > 0.5) = 0.1977$ [Tech: 0.1976]. About 20 out of 100 random samples of size $n = 200$ Americans will result in more than 100 individuals (that is, more than 50%) who can order a meal in a foreign language.
- (e)** $P(\hat{p} \leq 0.4) = 0.0239$ [Tech: 0.0237]. About 2 out of 100 random samples of size $n = 200$ Americans will result in 80 or fewer individuals (that is, 40% or less) who can order a meal in a foreign language. This result is unusual.
- 17.** (a) The sampling distribution of \hat{p} is approximately normal with $\mu_{\hat{p}} = 0.39$ and $\sigma_{\hat{p}} = \sqrt{\frac{0.39(0.61)}{500}} \approx 0.022$.
- (b) $P(\hat{p} < 0.38) = 0.3228$ [Tech: 0.3233]. About 32 out of 100 random samples of size $n = 500$ adult Americans will result in fewer than 190 individuals (that is, less than 38%) who believe that marriage is obsolete.
- (c) $P(0.40 < \hat{p} < 0.45) = 0.3198$ [Tech: 0.3203]. About 32 out of 100 random samples of size $n = 500$ adult Americans will result in between 200 and 225 individuals (that is, between 40% and 45%) who believe that marriage is obsolete.
- (d) $P(\hat{p} \geq 0.42) = 0.0838$ [Tech: 0.0845]. About 8 out of 100 random samples of size $n = 500$ adult Americans will result in 210 or more individuals (that is, 42% or more) who believe that marriage is obsolete. This result is not unusual.
- 19.** $P(X \geq 121) = P(\hat{p} \geq 0.11) = 0.1335$ [Tech: 0.1345]. This result is not unusual, so this evidence is insufficient to conclude that the proportion of Americans who are afraid to fly has increased above 0.10. In other words, the results obtained could be due to sampling error and the true proportion is still 0.10.
- 21.** $P(X \geq 164) = P(\hat{p} \geq 0.529) = 0.0853$ [Tech: 0.0846]. When the final election results will show 49% of voters supporting an increase in funding for education, approximately 9 out of 100 random samples of 310 voters will result in 52.9% or more who support the increase. This result is not unusual, so it would not be unusual for a wrong call to be made in an election if exit polling alone was considered. The exit polling could be biased in favor of an increase in funding for education, since a voter who voted against it in the privacy of the voting booth might not want to admit it to a pollster.
- 23.** (a) 62 more adult Americans must be polled to make $np(1 - p) \geq 10$.
- (b) 13 more adult Americans must be polled if $p = 0.2$.
- 25.** (a) Qualitative with two outcomes—believe in reincarnation, or not.
- (b) There is variability in the sample proportion due to sampling variability. Bob likely has different individuals from Alicia in his sample.
- (c) Randomly selecting individuals is important so that the results of the survey are representative of the population. This allows the researcher to generalize the results from the sample to the population.
- (d) $E(X) = np = 100(0.21) = 21$
- (e) The sample size is not large enough for the distribution of the sample proportion to be approximately normal. In fact, $np(1 - p) = 20(0.21)(1 - 0.21) = 3.318 < 10$.
- (f) Solve $n(0.21)(1 - 0.21) > 10$ and find that $n > 60.28$. Therefore, the sample size must be at least 61 to use the normal model to describe the distribution of the sample proportion.
- 27.** (a) $P(X \geq 25) = P\left(\hat{p} \geq \frac{25}{290}\right) = P(\hat{p} \geq 0.0862)$
 $= 0.7764$ [Tech: 0.7753]
- (b) Let X represent the number who show up so that $p = 0.9005$.
 $P(X \leq 300) = P\left(\hat{p} \leq \frac{300}{320}\right) = P(\hat{p} \leq 0.9375)$
 $= 0.9864$ [Tech: 0.9865]
- (c) Let X represent the number who do not show up. Here, $p = 0.04$. We need at least 15 to not show up.
 $P(X \geq 15) = P\left(\hat{p} \geq \frac{15}{300}\right) = P(\hat{p} \geq 0.05)$
 $= 0.1894$ [Tech: 0.1884]

- 29. (a)** The distribution is slightly skewed right. There are three outliers (Atlanta Falcons, New Orleans Saints, New England Patriots).



- (b)** The only outlier is the New England Patriots.



Chapter 8 Review Exercises (page 392)

- A sampling distribution is a probability distribution for all possible values of a statistic computed from a sample of size n .
- The sampling distribution of \bar{x} is approximately normal when the underlying population distribution is normal. The sampling distribution of \bar{x} is approximately normal when the sample size is large, usually greater than 30, regardless of how the population is distributed.
- The sampling distribution of \hat{p} is approximately normal when $np(1-p) \geq 10$, provided that $n \leq 0.05N$.

4. $\mu_{\bar{x}} = \mu$, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, and $\mu_{\hat{p}} = p$, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

5. (a) Quantitative

(b) $P(x > 2625) = 0.3085$. If we select a simple random sample of $n = 100$ pregnant women, then about 31 will have energy needs of more than 2625 kcal/day. This result is not unusual.

(c) \bar{x} is normal with $\mu_{\bar{x}} = 2600$ kcal and $\sigma_{\bar{x}} = \frac{50}{\sqrt{20}} = \frac{25}{\sqrt{5}} \approx 11.180$ kcal. Mother-to-mother variation.

(d) $P(\bar{x} > 2625) = 0.0125$ [Tech: 0.0127]. If we take 100 simple random samples of size $n = 20$ pregnant women, then about 1 of the samples will result in a mean energy need of more than 2625 kcal/day. This result is unusual.

6. (a) \bar{x} is approximately normal with $\mu_{\bar{x}} = 0.75$ inch and

$$\sigma_{\bar{x}} = \frac{0.004}{\sqrt{30}} \approx 0.001 \text{ inch.}$$

(b) $P(\bar{x} < 0.748) + P(\bar{x} > 0.752) = 0.0062$. The inspector will conclude that the machine needs adjustment when, in fact, it does not need adjustment in about 6 of every 1000 repetitions of this study.

7. (a) No, the variable “number of televisions” is likely skewed right.

(b) $\bar{x} = 2.55$ televisions

(c) $P(\bar{x} \geq 2.55) = 0.0778$ [Tech: 0.0777]. If we take 100 simple random samples of size $n = 40$ households, then about 8 of the samples will result in a mean of 2.55 televisions or more. This result is not unusual, so it does not contradict the results reported by A. C. Nielsen.

8. (a) Qualitative with two possible outcomes.

- (b)** The sampling distribution of \hat{p} is approximately normal with $\mu_{\hat{p}} = 0.72$ and $\sigma_{\hat{p}} = \sqrt{\frac{0.72(0.28)}{600}} \approx 0.018$. Person-to-person variability.

- (c) $P(\hat{p} \leq 0.70) = 0.1379$ [Tech: 0.1376]. About 14 out of 100 random samples of size $n = 600$ 18- to 29-year-olds will result in no more than 70% who would prefer to start their own business.

- (d) Yes, it would be a little unusual for 450 of the 600 randomly selected 18- to 29-year-olds to prefer to start their own business.

$$P(X \geq 450) = P(\hat{p} \geq 0.75) = 0.0505$$
 [Tech: 0.0509].

9. $P(X \geq 60) = P(\hat{p} \geq 0.12) = 0.0681$ [Tech: 0.0680]. This result is not that unusual. There is some evidence to suggest that the proportion of adults 25 years of age or older with advanced degrees has increased above 10%.

10. (a) The sampling distribution of \hat{p} is approximately normal with

$$\mu_{\hat{p}} = 0.280 \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{0.28(0.72)}{500}} \approx 0.020.$$

- (b) $P(\hat{p} \geq 0.310) = 0.0681$ [Tech: 0.0676]. It would not be unusual for a career 0.280 hitter to have a season in which he hits 0.310.

- (c) $P(\hat{p} \leq 0.255) = 0.1056$ [Tech: 0.1066]. It would not be unusual for a career 0.280 hitter to have a season in which he hits 0.255.

- (d) Batting averages between 0.260 and 0.300 lie within 1 standard deviation of the mean of the sampling distribution.

- (e) It is unlikely that a career 0.280 hitter hits 0.280 each season, so he will have seasons where he bats below 0.280 and seasons where he bats above 0.280, and batting averages as low as 0.260 and as high as 0.300 are not unusual [$P(\hat{p} \leq 0.260) \approx 0.16$ and $P(\hat{p} \geq 0.300) \approx 0.16$]. Based on a single season, we cannot conclude that a player who hit 0.260 is worse than a player who hit 0.300 because neither result would be unusual for players who had identical career batting averages of 0.280.

Chapter 8 Test (page 393)

- Regardless of the shape of the population, the sampling distribution of \bar{x} becomes approximately normal as the sample size n increases.

2. $\mu_{\bar{x}} = 50$, $\sigma_{\bar{x}} = \frac{24}{\sqrt{36}} = 4$

3. (a) $P(X > 100) = 0.3859$ [Tech: 0.3875]. If we select a simple random sample of $n = 100$ batteries of this type, then about 39 batteries would last more than 100 minutes. This result is not unusual.

- (b) \bar{x} is normally distributed with $\mu_{\bar{x}} = 90$ minutes and

$$\sigma_{\bar{x}} = \frac{35}{\sqrt{10}} \approx 11.068 \text{ minutes.}$$

- (c) $P(\bar{x} > 100) = 0.1841$ [Tech: 0.1831]. If we take 100 simple random samples of size $n = 10$ batteries of this type, then about 18 of the samples will result in a mean charge life of more than 100 minutes. This result is not unusual.

- (d) $P(\bar{x} > 100) = 0.0764$ [Tech: 0.0766]. If we take 100 simple random samples of size $n = 25$ batteries of this type, then about 8 of the samples will result in a mean charge life of more than 100 minutes.

- (e) The probabilities are different because a change in n causes a change in $\sigma_{\bar{x}}$.

4. (a) \bar{x} is approximately normally distributed with $\mu_{\bar{x}} = 2.0$ liters and $\sigma_{\bar{x}} = \frac{0.05}{\sqrt{45}} \approx 0.007$ liter.

- (b) $P(\bar{x} < 1.98) + P(\bar{x} > 2.02) = 0.0074$ [Tech: 0.0073]. The quality-control manager will shut down the machine even though it is correctly calibrated in about 7 of every 1000 repetitions of this study.

5. (a) The sampling distribution of \hat{p} is approximately normal with

$$\mu_{\hat{p}} = 0.224 \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{0.224(0.776)}{300}} \approx 0.024.$$

- (b) $P(X \geq 50) = P(\hat{p} \geq 0.1667) = 0.9913$ [Tech: 0.9914]. About 99 out of 100 random samples of size $n = 300$ adults will result in at least 50 adults (that is, at least 16.7%) who are smokers.
- (c) Yes; it would be unusual for 18% or less of 300 randomly selected adults to be smokers. $P(\hat{p} \leq 0.18) = 0.0336$ [Tech: 0.0338].
6. (a) For the sample proportion to be normal, the sample size must be large enough to meet the condition $np(1 - p) \geq 10$. Since $p = 0.01$, we have

$$\begin{aligned} n(0.01)(1 - 0.01) &\geq 10 \\ 0.0099n &\geq 10 \\ n &\geq \frac{10}{0.0099} \approx 1010.1 \end{aligned}$$

Thus, the sample size must be at least 1011 to satisfy the condition.

(b) No; it would not be unusual for a random sample of 1500 Americans to result in fewer than 10 with peanut or tree nut allergies. $P(X \leq 9) = P(\hat{p} \leq 0.006) = 0.0594$ [Tech: 0.0597].

7. $P(X \geq 82) = P(\hat{p} \geq 0.082) = 0.0681$ [Tech: 0.0685]. This result is not unusual, so this evidence is insufficient to conclude that the proportion of households with a net worth in excess of \$1 million has increased above 7%.

CHAPTER 9 Estimating the Value of a Parameter

9.1 Assess Your Understanding (page 406)

1. (a) A point estimate is the value of a statistic that estimates the value of a parameter.
- (b) A confidence interval for an unknown parameter consists of an interval of numbers based on a point estimate.
- (c) The level of confidence represents the expected proportion of intervals that will contain the parameter if a large number of different samples is obtained.
- (d) The margin of error determines the width of a confidence interval. It represents the number of standard errors below and above the point estimate the lower and upper bounds will be.

3. (b) < (d) < (a) < (c) 5. False

7. $z_{\alpha/2} = z_{0.05} = 1.645$

9. $z_{\alpha/2} = z_{0.01} = 2.33$

11. $\hat{p} = 0.225, E = 0.024, x = 270$

13. $\hat{p} = 0.4855, E = 0.0235, x = 816$

15. Lower bound: 0.146, upper bound: 0.254

17. Lower bound: 0.191, upper bound: 0.289

19. Lower bound: 0.759 [Tech: 0.758], upper bound: 0.805

21. (a) Flawed; no interval has been provided about the population proportion.

- (b) Flawed; this interpretation indicates that the level of confidence is varying.

- (c) Correct

- (d) Flawed; this interpretation suggests that this interval sets the standard for all the other intervals, which is not true.

23. We are 95% confident that the population proportion of adult Americans who dread Valentine's Day is between 0.135 and 0.225.

25. (a) $\hat{p} = 0.181$

- (b) The sample is a simple random sample, $n\hat{p}(1 - \hat{p}) = 341.8 \geq 10$, and the sample is less than 5% of the population.

- (c) Lower bound: 0.168, upper bound: 0.194

- (d) We are 90% confident that the proportion of adult Americans 18 years and older who have donated blood in the past two years is between 0.168 and 0.194.

27. (a) $\hat{p} = 0.519$

- (b) The sample is a simple random sample, $n\hat{p}(1 - \hat{p}) = 250.39 \geq 10$, and the sample is less than 5% of the population.

- (c) Lower bound: 0.488, upper bound: 0.550 [Tech: (0.489, 0.550)]

- (d) Yes; it is possible that the population proportion is more than 60%, because it is possible that the true proportion is not captured in the confidence interval. It is not likely.

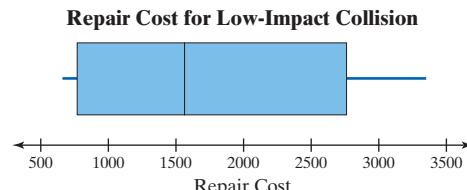
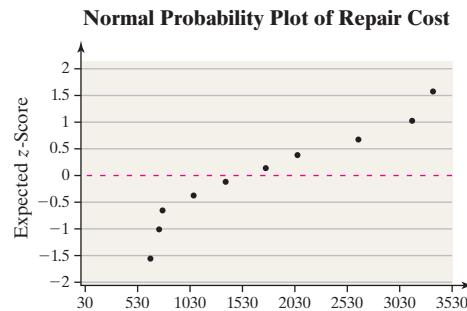
- (e) Lower bound: 0.450, upper bound: 0.512 [Tech: (0.450, 0.511)]

29. (a) Lower bound: 0.071; upper bound: 0.151
 (b) Lower bound: 0.058; upper bound: 0.164
 (c) As the level of confidence increases, the margin of error increases.
31. (a) The sample is the 1000 adults 19 years of age or older; the population is all adults 19 years of age or older.
 (b) The variable of interest is whether the individual brings and uses his or her cell phone every trip to the bathroom. It is qualitative with two possible outcomes—either the individual brings and uses his or her cell phone every trip to the bathroom, or not.
 (c) 0.241
 (d) The point estimate in part (c) is a statistic because its value is based on a sample. The point estimate is a random variable because its value may change depending on the individuals in the survey. The main source of variability are the individuals selected to be in the study. That is, there is person-to-person variability.
 (e) Lower bound: 0.214; upper bound: 0.268. We are 95% confident that the proportion of adults 19 years of age or older who bring and use their cell phone every trip to the bathroom is between 0.214 and 0.268.
 (f) To generalize these results to the population, the sample must be representative of the population. Random sampling is required to ensure the individuals in the sample are representative of the population.
33. (a) $p = 0.432$
 (b) $\hat{p} = 0.42$
 (c) The tornadoes were obtained through a simple random sample, $n\hat{p}(1 - \hat{p}) = 12.18 \geq 10$, and the sample size (50) is less than 5% of the population size (1473).
 (d) Lower bound: 0.305; Upper bound: 0.535
 (e) The 90% confidence interval from part (d) does include the population proportion, 0.432.
 (f) 0.90
 (g) If the sample proportion was in the tails of the distribution of the sample proportion (more than 1.645 standard errors from the population proportion), then the confidence interval would not include the population proportion.
 (h) Lower bound: 0.465; Upper bound: 0.695; this confidence interval does not include the population proportion because the random sample happened to result in a high proportion of F0 tornadoes.
35. (a) Using $\hat{p} = 0.635, n = 1708$. [Tech: 1726]
 (b) Using $\hat{p} = 0.5, n = 1842$. [Tech: 1844]
37. (a) Using $\hat{p} = 0.15, n = 1731$. [Tech: 1726]
 (b) Using $\hat{p} = 0.5, n = 3394$. [Tech: 3383]
39. (a) Using $\hat{p} = 0.53, n = 1064$.
 (b) Using $\hat{p} = 0.5, n = 1068$.
 (c) The results are close because $0.53(1 - 0.53) = 0.2491$ is very close to 0.25.
41. At least $n = 984$ people were surveyed.
43. The difference between the two point estimates is within the margin of error.
45. Lower bound: 0.016, upper bound: 0.313
47. (a) The variable of interest in the observational study is whether the individual says they wash their hands in a public restroom, or not. It is qualitative with two possible outcomes.
 (b) The sample is the 1001 adults interviewed. The population is all adults.
 (c) $n\hat{p}(1 - \hat{p}) = 73.7 \geq 10$ and the sample size is less than 5% of the population size.
 (d) Lower bound: 0.903; upper bound: 0.937
 (e) The variable of interest in the observational study is whether the individual is observed washing their hands in a public restroom, or not. It is qualitative with two possible outcomes.
 (f) Randomness could be achieved by using systematic sampling. For example, select every 10th individual who enters the bathroom. Also, to be able to generalize the results, we would want about half the individuals to be female, and half male.

- (g) Assume the sample is based on a random sample.
 $np(1 - \hat{p}) = 1076.1 \geq 10$ and the sample size is less than 5% of the population size.
- (h) Lower bound: 0.759; upper bound: 0.781
- (i) The proportion who say that they wash their hands is greater than the proportion who actually do. Explanations as to why may vary. One possibility is that people lie about their handwashing habits out of embarrassment.
- (j) In the telephone survey, there is certainly person-to-person variability. That is, different samples will result in different individuals surveyed, and therefore, different results. In the observational study, there is also person-to-person variability. There may also be individual variability. Perhaps an individual who does not always wash his or her hands is in the study, but during this particular observation does wash his or her hands. Another source of variability might be the type of public restroom the observation takes place in. For example, do people wash their hands more often in an airport restroom than they do at a sporting event?
49. Data must be qualitative with two possible outcomes to construct confidence intervals for a proportion.
51. They use the margin of error formula with $\hat{p} = 0.5$.
53. Mariya's interval is wrong. The upper and lower bounds are not the same distance from the point estimate.
- ### 9.2 Assess Your Understanding (page 419)
1. decreases
 3. (i) The t -distribution is different for different degrees of freedom.
 - (ii) The t -distribution is centered at 0 and is symmetric about 0.
 - (iii) The area under the curve is 1. The area under the curve to the right of 0 equals the area under the curve to the left of 0, which equals 1/2.
 - (iv) As t increases or decreases without bound, the graph approaches, but never equals, zero.
 - (v) The area in the tails of the t -distribution is a little greater than the area in the tails of the standard normal distribution, because we are using s as an estimate of σ , thereby introducing further variability into the t -statistic.
 - (vi) As the sample size n increases, the density curve of t gets closer to the standard normal density curve. This result occurs because, as the sample size increases, the values of s get closer to the value of σ , by the Law of Large Numbers.
 5. False; the population does not need to be normally distributed if the sample size is sufficiently large.
 7. (a) $t_{0.10} = 1.316$
 (b) $t_{0.05} = 1.697$
 (c) $t_{0.99} = -2.552$
 (d) $t_{0.05} = 1.725$
 9. Yes, $0.987 > 0.928$ (Table VI) and there are no outliers.
 11. No, there are outliers.
 13. $\bar{x} = 21, E = 3$
 15. $\bar{x} = 14, E = 9$
 17. (a) Lower bound: 103.7, upper bound: 112.3
 (b) Lower bound: 100.4, upper bound: 115.6; decreasing the sample size increases the margin of error.
 (c) Lower bound: 104.6, upper bound: 111.4; decreasing the level of confidence decreases the margin of error.
 (d) No; the sample sizes were too small.
 19. (a) Lower bound: 16.85, upper bound: 19.95
 (b) Lower bound: 17.12, upper bound: 19.68; increasing the sample size decreases the margin of error.
 (c) Lower bound: 16.32, upper bound: 20.48 [Tech: (16.33, 20.48)]; increasing the level of confidence increases the margin of error.
 (d) If $n = 15$, the population must be normal.
 21. (a) Flawed; this interpretation implies that the population mean varies rather than the interval.

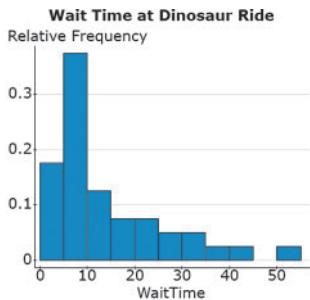
- (b) Correct
 (c) Flawed; this interpretation makes an implication about individuals rather than the mean.
 (d) Flawed; the interpretation should be about the mean number of hours worked by adult Americans, not about adults in Idaho.
23. (a) 163.1 minutes
 (b) 1.6 minutes
 (c) We are 90% confident the mean drive-thru service time at Taco Bell is between 161.5 seconds and 164.7 seconds.
25. (1) Increase the sample size, and (2) decrease the level of confidence to narrow the confidence interval.
27. (a) Since the distribution of blood alcohol concentrations is not normally distributed (highly skewed right), the sample must be large so that the distribution of the sample mean will be approximately normal.
 (b) The sample size is less than 5% of the population.
 (c) Lower bound: 0.1647, upper bound: 0.1693; we are 90% confident that the mean BAC in fatal crashes where the driver had a positive BAC is between 0.1647 and 0.1693 g/dL.
 (d) Yes; it is possible that the mean BAC is less than 0.08 g/dL, because it is possible that the true mean is not captured in the confidence interval, but it is not likely.

29. Lower bound: 317.63 licks [Tech: 317.64]; upper bound: 394.57 licks [Tech: 394.56]; We are 95% confident the mean number of licks to the center of a Tootsie pop is between 317.63 and 394.57.
31. (a) 4.893
 (b) Lower bound: 4.690, upper bound: 5.096 [Tech: 5.095]; We are 95% confident the mean pH of rain water in Tucker County, West Virginia, is between 4.690 and 5.096.
 (c) Lower bound: 4.607 [Tech: 4.606], upper bound: 5.179; We are 99% confident the mean pH of rain water in Tucker County, West Virginia, is between 4.607 and 5.179.
 (d) As the level of confidence increases, the margin of error also increases.
33. (a) Option 1: 0.961 [Tech: 0.966] > 0.918 (Table VI), so it is reasonable to conclude the data come from a normal population. The boxplot shows no outliers.
 Option 2: The boxplot shows no outliers and is roughly symmetric.

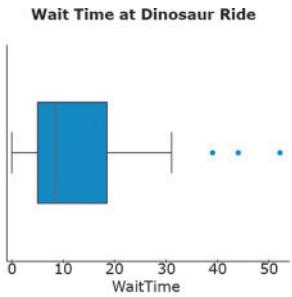


- (b) Lower bound: \$1035.8, upper bound: \$2477.2; We are 95% confident the mean repair cost for a low-impact collision involving mini- or micro-vehicles is between \$1035.8 and \$2477.2.
 (c) The 95% confidence interval would likely be narrower because there is less variability in the data because variability associated with make of the vehicle has been removed.

35. (a) The distribution is skewed right.



- (b) There are three outliers.

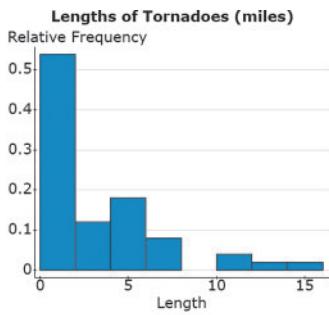


(c) Because the sample data are highly skewed to the right and there are three outliers, a large sample size is needed to be able to use Student's *t*-distribution to construct a confidence interval about the population mean.

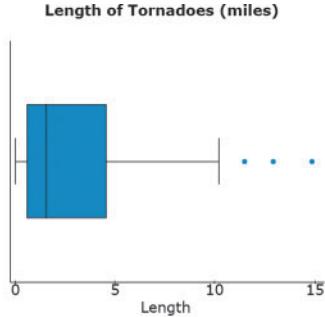
(d) Lower bound: 9.5 minutes; Upper bound: 17.4 minutes. We are 95% confident the mean wait time at Disney's Dinosaur Ride is between 9.5 minutes and 17.4 minutes.

37. (a) $\mu = 3.775$ miles

- (b) The distribution is skewed right.



- (c) There are three outliers.



(c) Because the sample data are highly skewed to the right and there are three outliers, a large sample size is needed to be able to use Student's *t*-distribution to construct a confidence interval about the population mean.

(d) Lower bound: 2.216 miles; Upper bound: 4.194 miles. We are 95% confident the mean length of a tornado in the United States in 2017 is between 2.216 miles and 4.194 miles. This confidence

interval does include the population mean. We would expect the proportion of intervals that capture the population mean to be 0.95.

39. For a 99% confidence level, $n = 298$ patients; for a 95% confidence level, $n = 173$ patients. Decreasing the confidence level decreases the sample size needed.

41. (a) To estimate within four books with 95% confidence, $n = 67$ subjects are needed.

(b) To estimate within two books with 95% confidence, $n = 265$ subjects are needed.

(c) Doubling the required accuracy quadruples the sample size.

(d) To estimate within four books with 99% confidence, $n = 115$ subjects are needed. Increasing the level of confidence increases the sample size. For a fixed margin of error, greater confidence can be achieved with a larger sample size.

43. (a) Set I: $\bar{x} \approx 99.1$; set II: $\bar{x} \approx 99.1$; set III: $\bar{x} \approx 99.0$

(b) Set I: Lower bound: 82.6, upper bound: 115.6 [Tech: 115.7]

Set II: Lower bound: 91.7, upper bound: 106.5

Set III: Lower bound: 93.5, upper bound: 104.5 [Tech: 104.6]

(c) As the size of the sample increases, the width of the confidence interval decreases.

(d) Set I: Lower bound: 58.6, upper bound: 117.2

Set II: Lower bound: 83.2, upper bound: 106.0

Set III: Lower bound: 88.1, upper bound: 103.9

(e) Each interval contains the population mean. The procedure for constructing the confidence interval is robust. This also illustrates the Law of Large Numbers, and Central Limit Theorem.

45. (a) Answers will vary. (b) Answers will vary.

(c) Answers will vary.

(d) Answers will vary. Expect 95% of the intervals to contain the population mean.

47. (a) Completely randomized design

(b) The treatment is the smoking cessation program. There are 2 levels.

(c) The response variable is whether or not the smoker had 'even a puff' from a cigarette in the past 7 days.

(d) The statistics reported are 22.3% of participants in the experimental group reported abstinence and 13.1% of participants in the control group reported abstinence.

$$(e) \frac{p(1-q)}{q(1-p)} = \frac{0.223(1-0.131)}{0.131(1-0.223)} \approx 1.90; \text{ this means that}$$

reported abstinence is almost twice as likely in the experimental group than in the control group.

(f) The authors are 95% confident that the population odds ratio is between 1.12 and 3.26.

(g) Answers will vary. One possibility: smoking cessation is more likely when the Happy Ending Intervention program is used rather than the control method.

49. The *t*-distribution has less spread as the degrees of freedom increase because as n increases s becomes closer to σ by the Law of Large Numbers.

51. The degrees of freedom are the number of data values that are free to vary.

53. We expect that the margin of error for population A will be smaller since there should be less variation in the ages of college students than in the ages of the residents of a town, resulting in a smaller standard deviation for population A.

9.3 Assess Your Understanding (page 426)

1. Confidence intervals for a population proportion are constructed on qualitative variables for which there are two possible outcomes.

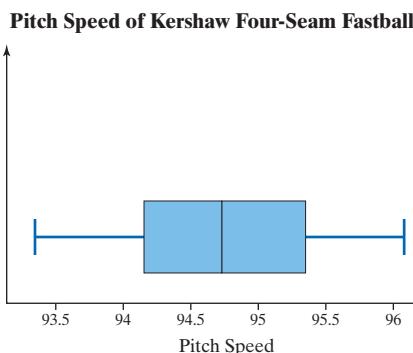
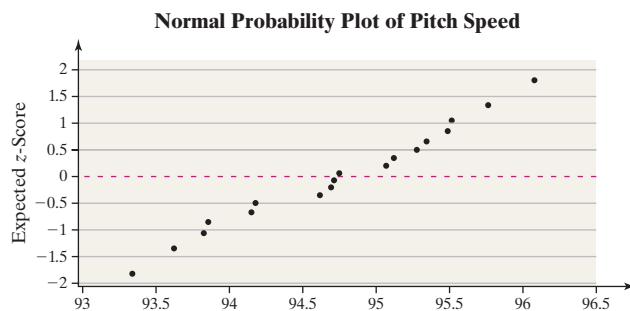
3. (1) data from a simple random sample or the result of a randomized experiment; (2) $n\hat{p}(1 - \hat{p}) \geq 10$; (3) sample size no more than 5% of the population size.

5. Lower bound: 0.069, upper bound: 0.165 [Tech: 0.164]

7. Lower bound: 37.74, upper bound: 52.26

9. Lower bound: 114.98, upper bound: 126.02
11. (a) Because the sample data are highly skewed to the right a large sample size is needed to be able to use Student's t -distribution to construct a confidence interval about the population mean.
(b) By hand (using 100 df)—Lower bound: 0.772 [Tech: 0.773]; Upper bound: 0.908 [Tech: 0.907]. We are 90% confident the mean reaction time is between 0.772 second and 0.908 second.
13. Lower bound: 2992.2, [Tech: 2992.1] upper bound: 3849.9; the Internal Revenue Service can be 90% confident that the mean additional tax owed is between \$2992.2 and \$3849.9.
15. Lower bound: 0.496, upper bound: 0.548; the Gallup organization can be 90% confident that the proportion of adult Americans who are worried about having enough money for retirement is between 0.496 and 0.548.

17. (a) Pitch speed is quantitative. This is important to know because confidence intervals for a mean are constructed on quantitative data, while confidence intervals for a proportion are constructed on qualitative data with two possible outcomes.
(b) **Option 1:** Correlation between raw data and normal score is 0.990 [Tech: 0.992], which is greater than the critical value from Table VI for $n = 18$ (0.946). Therefore, it is reasonable to conclude pitch speed comes from a population that is normally distributed. The boxplot shows no outliers.



Option 2: The boxplot shows the sample data is symmetric with no outliers. So, it is reasonable to construct a confidence interval for the mean.

- Because it is reasonable to conclude the data come from a population that is normally distributed, there are no outliers, and the data is the result of a random sample of pitches, it is reasonable to construct a confidence interval for the mean pitch speed.
(c) Lower bound: 94.358 mph; upper bound: 95.136 mph. We are 95% confident the mean pitch speed of a Clayton Kershaw four-seam fastball is between 94.358 mph and 95.136 mph.
(d) A 95% confidence interval for the mean pitch speed of all major league pitchers' four-seam fastballs would be wider because of the pitcher-to-pitcher variability in pitch speed is now part of the analysis.
19. (a) The study is cross-sectional because the data was obtained at a specific point in time (or over a very short period of time). It also means the study is an observational study.
(b) The variable of interest is whether an individual with sleep apnea has gum disease, or not. This is a qualitative variable because it categorizes the individual.

- (c) Lower bound: 0.546; upper bound: 0.654. We are 95% confident the proportion of individuals with sleep apnea who have gum disease is between 0.546 and 0.654.
21. (a) Qualitative with two possible outcomes.
(b) The sample is a simple random sample; $n\hat{p}(1 - \hat{p}) = 24.64 \geq 10$; and the sample size (100) is less than 5% of the population size (13,422). Construct a confidence interval for a population proportion. We are 95% confident the proportion of short-term rentals where the license is expired is between 0.343 and 0.537.
23. Confidence interval for a mean since the variable of interest is quantitative and we want to estimate the “typical” amount of time spent studying.
25. Confidence interval for a proportion since the variable of interest is qualitative with two possible outcomes—either LDL cholesterol decreases, or not.
- Chapter 9 Review Exercises (page 430)**
1. (a) $t_{0.005} = 2.898$
(b) $t_{0.05} = 1.706$
 2. We would expect 95 of the 100 intervals to include the mean 100. Random chance in sampling causes a particular interval to not include the mean 100. Any sample statistic in the tails of the distribution will result in a confidence interval that does not include the parameter.
 3. In a 95% confidence interval, the 95% represents the proportion of intervals that would contain the parameter (e.g., the population mean, population proportion, or population standard deviation) if a large number of different samples is obtained.
 4. If a large number of different samples is obtained, a 90% confidence interval for a population mean will not capture the true population mean 10% of the time.
 5. Area to the left of $t = -1.56$ is 0.0681 because the t -distribution is symmetric about zero.
 6. There is more area under the t -distribution to the right of $t = 2.32$ than under the standard normal distribution to the right of $z = 2.32$, because the t -distribution uses s to approximate σ , making it more dispersed than the z -distribution.
 7. The properties of Student's t -distribution:
 1. It is symmetric around $t = 0$.
 2. It is different for different sample sizes.
 3. The area under the curve is 1; half the area is to the right of 0 and half the area is to the left of 0.
 4. As t gets extremely large, the graph approaches, but never equals, zero. Similarly, as t gets extremely small (negative), the graph approaches, but never equals, zero.
 5. The area in the tails of the t -distribution is greater than the area in the tails of the standard normal distribution.
 6. As the sample size n increases, the distribution (and the density curve) of the t -distribution becomes more like the standard normal distribution.
 8. (a) Lower bound: 51.54, upper bound: 58.06
(b) Lower bound: 52.34, upper bound: 57.26; increasing the sample size decreases the width of the interval.
(c) Lower bound: 49.52, upper bound: 60.08; increasing the level of confidence increases the width of the interval.
 9. (a) Lower bound: 97.07, upper bound: 111.53
(b) Lower bound: 98.86, upper bound: 109.74; increasing the sample size decreases the width of the interval.
(c) Lower bound: 95.49, upper bound: 113.11; increasing the level of confidence increases the width of the interval.
 10. (a) The distribution is skewed left, but according to the Central Limit Theorem, the sampling distribution of \bar{x} is approximately normal when the sample size is large.
(b) Lower bound: 82.58, upper bound: 93.22; we are 95% confident that the mean age adult Americans would like to live is between 82.58 and 93.22 years.

- (c) A sample size of $n = 231$ is required to estimate the mean age adult Americans would like to live within 2 years with 95% confidence.

11. (a) The distribution is skewed right.
(b) Lower bound: 8.86, upper bound: 11.94; we are 90% confident that the population mean number of e-mails sent per day is between 8.86 and 11.94 e-mails.

12. (a) The sample is probably small because of the difficulty and expense of gathering data.
(b) Lower bound: 201.5, upper bound: 234.5; the researchers can be 95% confident that the population “mean total work performed” for the sports-drink treatment is between 201.5 and 234.5 kilojoules.
(c) Yes; it is possible that the population “mean total work performed” for the sports-drink treatment is less than 198 kilojoules, since it is possible that the true mean is not captured in the confidence interval. It is not likely.
(d) Lower bound: 161.5, upper bound: 194.5; the researchers can be 95% confident that the population “mean total work performed” for the placebo treatment is between 161.5 and 194.5 kilojoules.
(e) Yes; it is possible that the population “mean total work performed” for the placebo treatment is more than 198 kilojoules, since it is possible that the true mean is not captured in the confidence interval. It is not likely.
(f) Yes; our findings support the researchers’ conclusion. The confidence intervals do not overlap, so we are confident that the mean for the sports-drink treatment is greater than the mean for the placebo treatment.

13. (a) From the Central Limit Theorem, when the sample is large, \bar{x} is approximately normally distributed.
(b) Lower bound: 1.95, upper bound: 2.59 [Tech: 2.58]; we can be 95% confident that couples who have been married for 7 years have a mean number of children between 1.95 and 2.59.
(c) Lower bound: 1.85, upper bound: 2.69; we can be 99% confident that couples who have been married for 7 years have a mean number of children between 1.85 and 2.69.

14. (a) $\bar{x} = 147.3$ cm, $s = 28.8$ cm
(b) **Option 1:** Because 0.982 [Tech: 0.985] > 0.928 (Table VI) and there are no outliers, the conditions are met.
Option 2: The boxplot is roughly symmetric, so the robustness of using Student’s t -distribution to construct a confidence interval for the mean suggests the conditions are satisfied.
(c) Lower bound: 129.0, upper bound: 165.6; we are 95% confident that the population mean diameter of a Douglas fir tree in the western Washington Cascades is between 129.0 and 165.6 cm.

15. (a) $\hat{p} = 0.086$
(b) Lower bound: 0.065, upper bound: 0.107 [Tech: (0.064, 0.107)]; the Centers for Disease Control is 95% confident that the proportion of adult males 20 to 34 years old who have hypertension is between 0.065 and 0.107.
(c) 336 subjects would be needed if we use the point estimate of the proportion found in part (a).
(d) 1068 subjects would be needed if no prior estimate is available.

(b) Yes; our result from part (a) indicates that the mean time to graduate is more than 4 years. The entire interval is above 4.

5. (a) $\bar{x} = 57.8$ inches; $s = 15.4$ inches
(b) **Option 1:** Yes; the conditions are met. The distribution is approximately normal (since 0.960 [Tech: 0.957] > 0.928 –Table VI), and there are no outliers.
Option 2: The boxplot is roughly symmetric, so the robustness of using Student’s t -distribution to construct a confidence interval for the mean suggests the conditions are satisfied.
(c) Lower bound: 48.0 [Tech: 47.9], upper bound: 67.6; the student is 95% confident that the mean depth of visibility of the Secchi disk is between 48.0 and 67.6 inches.
(d) Lower bound: 44.0 [Tech: 43.9], upper bound: 71.6; the student is 99% confident that the mean depth of visibility of the Secchi disk is between 44.0 and 71.6 inches.

6. (a) $\hat{p} = 0.948$
(b) Lower bound: 0.932, upper bound: 0.964 [Tech: 0.965]; the EPA is 99% confident that the proportion of Americans who live in neighborhoods with acceptable levels of carbon monoxide is between 0.932 and 0.964.
(c) 593 Americans must be sampled if the prior estimate of p is used.
(d) 3007 Americans must be sampled for the estimate to be within 1.5 percentage points with 90% confidence if no prior estimate is available.

7. (a) $\bar{x} = 133.4$ minutes
(b) Since the distribution of the lengths of matches is not normally distributed, the sample must be large so that the distribution of the sample mean will be approximately normal.
(c) Lower bound: 114.9, upper bound: 151.9; the tennis enthusiast is 99% confident that the population “mean length of matches at Wimbledon” is between 114.9 and 151.9 minutes.
(d) Lower bound: 119.6, upper bound: 147.2; the tennis enthusiast is 95% confident that the population “mean length of matches at Wimbledon” is between 119.6 and 147.2 minutes.
(e) Increasing the level of confidence increases the width of the interval.
(f) No; because the tennis enthusiast only sampled matches at Wimbledon, the results cannot be generalized to include other professional tennis tournaments.

Chapter 9 Test (page 432)

1. **(a)** $t_{0.02} = 2.167$
(b) $t_{0.01} = 2.567$
 2. $\bar{x} = 139.2$, $E = 13.4$
 3. **(a)** The distribution is skewed right.
(b) Lower bound: 1.151, upper bound: 1.289 [Tech: (1.152, 1.288)]; we are 99% confident that the population “mean number of family members in jail” is between 1.151 and 1.289 members.
 4. **(a)** Lower bound: 4.319, upper bound: 4.841; we are 90% confident that the population “mean time to graduate” is between 4.319 and 4.841 years.

CHAPTER 10 Hypothesis Tests Regarding a Parameter

10.1 Assess Your Understanding (page 441)

- (d) The sample evidence led the real estate broker to conclude that the mean price of an existing single-family home is not lower in her neighborhood, when in fact the mean price is lower.
- 17.** (a) Null hypothesis: The standard deviation in the pressure required to open a valve is 0.7 psi. Alternative hypothesis: The standard deviation in the pressure required to open a valve is less than 0.7 psi.
 (b) $H_0: \sigma = 0.7$ psi, $H_1: \sigma < 0.7$ psi
 (c) The quality-control manager rejects the hypothesis that the variability in the pressure required is 0.7 psi, when the true variability is 0.7 psi.
 (d) The quality-control manager fails to reject that the variability in the pressure required is 0.7 psi, when the variability is less than 0.7 psi.
- 19.** (a) Null hypothesis: The mean monthly revenue per cell phone is \$38.66. Alternative hypothesis: The mean monthly revenue per cell phone is different from \$38.66.
 (b) $H_0: \mu = \$38.66$, $H_1: \mu \neq \$38.66$
 (c) The sample evidence led the researcher to believe the mean monthly cell phone bill is different from \$38.66, when in fact the mean bill is \$38.66.
 (d) The sample evidence did not lead the researcher to believe the mean monthly cell phone bill is different from \$38.66, when in fact the mean bill is different from \$38.66.
- 21.** There is sufficient evidence to conclude the proportion of students who enroll at Joliet Junior College and earn a bachelor's degree within six years exceeds 0.236.
- 23.** There is not sufficient evidence to conclude that the mean price of an existing single-family home in the realtor's neighborhood is less than \$395,000.
- 25.** There is not sufficient evidence to conclude that the variability in pressure has been reduced.
- 27.** There is sufficient evidence to conclude that the mean monthly revenue per cell phone is different from \$38.66.
- 29.** There is not sufficient evidence to conclude the proportion of students who enroll at Joliet Junior College and earn a bachelor's degree within six years exceeds 0.236.
- 31.** There is sufficient evidence to conclude that the mean price of an existing single-family home in the realtor's neighborhood is less than \$395,000.
- 33.** (a) $H_0: \mu = 12$ oz.; $H_1: \mu \neq 12$ oz.
 (b) Based on the sample evidence, there is sufficient evidence to conclude that the machine is out of calibration.
 (c) A Type I error has been made since the sample evidence led the quality-control manager to reject the null hypothesis, when the null hypothesis (not out of calibration) is true.
 (d) The level of significance should be 0.01 because this makes the probability of a Type I error small.
- 35.** (a) $H_0: p = 0.208$; $H_1: p > 0.208$
 (b) There is not sufficient evidence to conclude the proportion of high school students exceeds 0.208 at this counselor's high school.
 (c) A Type II error was committed because the sample evidence led the counselor to conclude the proportion of e-cig users at her school was 0.208, when, in fact, the proportion is higher.
- 37.** (a) $H_0: \mu = 4$ hours, $H_1: \mu < 4$ hours
 (b) *Consumer Reports* might reject the null hypothesis and conclude that a car with Prolong will not run 4 hours without oil.
- 39.** (a) \$30,000
 (b) This is a binomial experiment because (a) there are a fixed number of trials, (b) there are two mutually exclusive outcomes, (c) the trials are independent because the sample size is small relative to the population size, and (d) the probability of success is fixed for all trials. Here, $n = 20$, $p = 0.16$.
 (c) $P(8) = 0.0067$
 (d) $P(X < 8) = 0.9912$
 (e) Assume that $p = 0.16$, so $np(1 - p) = 500(0.16)(1 - 0.16) = 67.2 > 10$ and the sample size is less than 5% of the population

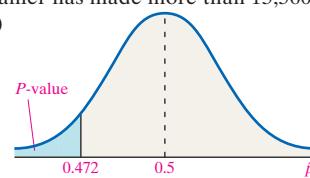
size assuming there are more than 10,000 60- to 75-year-olds in the population (which is reasonable). Therefore, the sample proportion is approximately normal with mean $500(0.16) = 80$ and standard deviation $\sqrt{(0.16)(1 - 0.16)/500} = 0.016$.

(f) $P(X \geq 100) = 0.0073$. This result is unusual. If the true proportion of 60- to 75-year-olds who believe it is safe to withdraw 6 to 8 percent of their savings annually is 0.16, we would expect about 7 of every 1000 samples of size 500 to result in at least 100 who believe this. We might conclude from this that the true proportion of 60- to 75-year-olds who believe it is safe to withdraw 6% to 8% of their savings annually is greater than 0.16.

41. As the level of significance, α , decreases, the probability of making a Type II error, β , increases. As we decrease the probability of rejecting a true null hypothesis, we increase the probability of not rejecting the null hypothesis when the alternative hypothesis is true.

43. Answers will vary.

10.2 Assess Your Understanding (page 454)

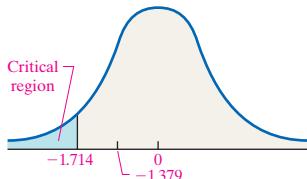
1. statistically significant
3. (a) < (c) < (b) < (e) < (d)
5. -1.28
7. $np_0(1 - p_0) = 42 > 10$
 - (a) Classical approach: $z_0 = 2.31 > z_{0.05} = 1.645$; reject the null hypothesis.
 - (b) *P*-value approach: *P*-value = 0.0104 [Tech: 0.0103] < $\alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to reject the null hypothesis.
9. $np_0(1 - p_0) = 37.1 > 10$
 - (a) Classical approach: $z_0 = -0.74 > -z_{0.10} = -1.28$; do not reject the null hypothesis.
 - (b) *P*-value approach: *P*-value = 0.2296 [Tech: 0.2301] > $\alpha = 0.10$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.10$ level of significance to reject the null hypothesis.
11. $np_0(1 - p_0) = 45 > 10$
 - (a) Classical approach: $z_0 = -1.49$ is between $-z_{0.025} = -1.96$ and $z_{0.025} = 1.96$; do not reject the null hypothesis.
 - (b) *P*-value approach: *P*-value = 0.1362 [Tech: 0.1360] > $\alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to reject the null hypothesis.
13. About 27 in 100 samples will give a sample proportion as high or higher than the one obtained if the population proportion really is 0.5. Because this probability is not small, we do not reject the null hypothesis. There is not sufficient evidence to conclude that the dart-picking strategy resulted in a majority of winners.
15. (a) $\hat{p} = 0.472$
 (b) $H_0: p = 0.5$ versus $H_1: p < 0.5$
 (c) The sample is a random sample; $np_0(1 - p_0) = 169.5 \geq 10$. The sample size is less than 5% of the population size, provided Cramer has made more than 13,560 predictions (which is likely).
- (d)

- (e) $P\text{-value} = P(\hat{p} \leq 0.472) = 0.0721$ [Tech: 0.0722]
- (f) If we obtained 100 different samples of size 678 from the population of Cramer predictions and the true proportion of correct predictions was 0.5, we would expect about seven of the samples to result in a sample proportion of correct predictions of 0.472 or less.
- (g) Because the *P*-value is greater than the level of significance, do not reject the null hypothesis. The sample data does not provide sufficient evidence to conclude that Cramer's predictions are correct less than half the time.

- 17.** The sample is a randomized trial; $np_0(1 - p_0) = 16.1 > 10$ and $n \leq 0.05N$. Hypotheses: $H_0: p = 0.019$, $H_1: p > 0.019$
- (a) Classical approach: $z_0 = 0.65 < z_{0.01} = 2.33$; do not reject the null hypothesis.
- (b) P -value approach: $P\text{-value} = 0.2578$ [Tech: 0.2582] $> \alpha = 0.01$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.01$ level of significance to conclude that more than 1.9% of Lipitor users experience flulike symptoms as a side effect.
- 19.** $np_0(1 - p_0) = 24.192 > 10$ and $n \leq 0.05N$. Hypotheses: $H_0: p = 0.36$, $H_1: p > 0.36$. Classical approach: $z_0 = 2.69 > z_{0.05} = 1.645$; reject the null hypothesis. P -value approach: $P\text{-value} = 0.0036 < \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude Hawaii has a higher proportion of traffic fatalities in which the driver has a positive BAC than in the United States.
- 21.** $H_0: p = 0.52$, $H_1: p \neq 0.52$; $np_0(1 - p_0) = 199.68 > 10$ and $n \leq 0.05N$. Classical approach: $z_0 = -11.32 < -z_{0.05/2} = -1.96$; reject the null hypothesis. P -value approach: $P\text{-value} < 0.0001 < \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence to conclude that the proportion of parents with children in high school who feel it was a serious problem that high school students were not being taught enough math and science has changed since 1994.
- 23.** $H_0: p = 0.47$, $H_1: p \neq 0.47$; $\hat{p} = 0.431$; $n\hat{p}(1 - \hat{p}) = 248.427 > 10$ and $n \leq 0.05N$. Lower bound: 0.401, upper bound: 0.461 [Tech: 0.462]. Since 0.47 is not contained in the interval, we reject the null hypothesis. There is sufficient evidence to conclude that parents' attitude toward the quality of education in the United States has changed since August 2002.
- 25.** 342
- 27.** (a) The researchers were testing $H_0: p = 0.5$ versus $H_1: p \neq 0.5$. The sample evidence indicated that there is not sufficient evidence to reject the statement in the null hypothesis. The researchers concluded that the patient could not distinguish black squares from circles on a white screen.
- (b) $H_0: p = 0.5$ vs. $H_1: p \neq 0.5$; data from a randomized trial; $np_0(1 - p_0) = 50 \geq 10$; the trials are independent; Classical approach: $z_0 = 2.55 > z_{0.05/2} = 1.96$, so reject the null hypothesis; $P\text{-value} = 0.0108$ [Tech: 0.0109], so reject H_0 ; The researchers are 95% confident the proportion of correct "guesses" is between 0.522 and 0.658. It would appear that the patient is not guessing (at least the results observed are very unusual if the patient is guessing).
- (c) $H_0: p = 0.5$ vs. $H_1: p \neq 0.5$; data from a randomized trial; $np_0(1 - p_0) = 50 \geq 10$; the trials are independent; Classical approach: $-z_{0.05/2} = 1.96 < z_0 = -1.56 < z_{0.05/2} = 1.96$, so do not reject the null hypothesis; $P\text{-value} = 0.1188$ [Tech: 0.1198], so do not reject H_0 ; The patient apparently is not able to identify other facial characteristics, such as gender.
- 29.** (a) $H_0: p = 0.5$ versus $H_1: p > 0.5$.
- (b) Treat the students as a random sample; $np_0(1 - p_0) = 4 < 10$. The normal model may not be used to describe the distribution of the sample proportion.
- (c) There are a fixed number of trials with two mutually exclusive outcomes (pass or not). The trials are independent and the probability of success is fixed at 0.5 for each trial.
- (d) $P\text{-value} = P(X \geq 11) = 0.1051$. If we taught the class 100 times with 16 students enrolled, we would expect 11 or more to pass in about 10 or 11 of the classes, assuming the probability of passing is 0.5.
- (e) Treat the students as a random sample; $np_0(1 - p_0) = 12 > 10$. Assuming there are over 960 students in the population (very likely), the trials are independent. The normal model may be used.
- (f) $P\text{-value} = 0.0045$ [Tech: 0.0047]. The P -value is less than the level of significance. There is sufficient evidence to support the belief that the blended course has a higher proportion who pass.
- (g) When there are small sample sizes, the evidence against the statement in the null hypothesis must be substantial. The moral is that you should be beware of studies that do not reject the null hypothesis when the test was conducted with a small sample size.
- 31.** (a) $H_0: p = 0.5$ versus $H_1: p > 0.5$.
- (b) $P\text{-value} = P(X \geq 28) = 0.000002$
- (c) Answers may vary. However, always be careful about drawing conclusions of causation from apparent associations.
- 33.** (a) $H_0: p = 0.465$ versus $H_1: p \neq 0.465$; Treat the data as a simple random sample; $np_0(1 - p_0) = 41.8 \geq 10$; assume the sample size is less than 5% of the population size; Classical approach: $-z_{0.025} = -1.96 < z_0 = 0.60 < z_{0.025} = 1.96$; do not reject the statement in the null hypothesis; P -value approach: $P\text{-value} = 0.5486$ [Tech: 0.5484]; do not reject the statement in the null hypothesis. There is not sufficient evidence to conclude that the proportion of F0 tornadoes in Texas is different than the proportion nationally.
- (b) $H_0: p = 0.465$ versus $H_1: p < 0.465$; Treat the data as a simple random sample; $np_0(1 - p_0) = 29.4 \geq 10$; assume the sample size is less than 5% of the population size; Classical approach: $z_0 = -2.20 < -z_{0.05} = -1.645$; Reject the statement in the null hypothesis; P -value approach: $P\text{-value} = 0.0139$ [Tech: 0.0142]; Reject the statement in the null hypothesis. There is sufficient evidence to conclude that the proportion of F0 tornadoes in Georgia is less than the proportion nationally. One might conclude from this that tornadoes in Georgia tend to have higher wind speeds.
- (c) We are 95% confident the proportion of F0 tornadoes in Georgia is between 0.277 [Tech: 0.278] and 0.451.
- 35.** Hypotheses: $H_0: p = 0.5$, $H_1: p \neq 0.5$. $np_0(1 - p_0) = 11.25 > 10$ and $n \leq 0.05N$. $P\text{-value} = 0.2938$ [Tech: 0.2967]; do not reject the null hypothesis. Yes, the data suggest that the spreads are accurate.
- 37.** (a) We do not reject the null hypothesis for values of p_0 between 0.44 and 0.62, inclusive. Each of these values of p_0 represents a possible value of the population proportion at the $\alpha = 0.05$ level of significance.
- (b) Lower bound: 0.432, upper bound: 0.628
- (c) At $\alpha = 0.01$, we do not reject the null hypothesis for any of the values of p_0 given in part (a), so that the range of values of p_0 for which we do not reject the null hypothesis increases. The lower value of α means we need more convincing evidence to reject the null hypothesis, so we would expect a larger range of possible values for the population proportion.
- 39.** (a) $H_0: p = 0.5$ vs. $H_1: p > 0.5$
- (b) Answers will vary.
- (c) Answers will vary.
- (d) There are a fixed number of trials with two mutually exclusive outcomes. The trials are independent and the probability of success is fixed at 0.5.
- (e) $P(X \geq 24) = 0.1341$
- (f) Answers will vary.
- (g) The sample evidence suggests that savants do not have the ability to predict card color as their results could easily be obtained by chance.
- 41.** (a) The randomness in the order in which the baby is exposed to the toys is important to avoid bias.
- (b) $H_0: p = 0.5$, $H_1: p > 0.5$
- (c) $P\text{-value} = 0.0021$; there is sufficient evidence to suggest the proportion of babies who choose the "helper" toy is greater than 0.5.
- (d) If the population proportion of babies who choose the helper toy is 0.5, a sample where all 12 babies choose the helper toy will occur in about 2 out of 10,000 samples of 12 babies.
- 43.** If the P -value for a particular test statistic is 0.23, we expect results at least as extreme as the test statistic in about 23 of 100 samples if the null hypothesis is true. Since this event is not unusual, we do not reject the null hypothesis.
- 45.** For the Classical Approach, the calculation of the test statistic, is simple, but you need to be very careful when determining the rejection region to interpret the result. For the P -value method, the P -value is harder to calculate than the test statistic, but the decision to reject or not is the same, regardless of the test. Plus, evidence as to the strength of evidence against the null hypothesis is reported. Most software will calculate both the test statistic and the P -value, which eliminates the disadvantages of the P -value method. Since the P -value is easier to interpret, the P -value method is easier to use with technology.
- 47.** Statistical significance means that the result observed in a sample is unusual when the null hypothesis is assumed to be true.

10.3 Assess Your Understanding (page 465)

1. (a) $t_{0.01} = 2.602$
 (c) $\pm t_{0.025} = \pm 2.179$
3. (a) $t_0 = -1.379$
 (c)

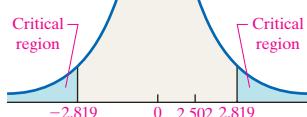
- (b) $-t_{0.05} = -1.729$
 (b) $-t_{0.05} = -1.714$



- (d) There is not enough evidence for the researcher to reject the null hypothesis because it is a left-tailed test and the test statistic is greater than the critical value ($-1.379 > -1.714$).

5. (a) $t_0 = 2.502$
 (c)

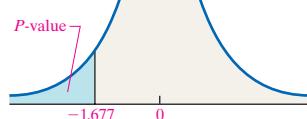
- (b) $-t_{0.005} = -2.819$; $t_{0.005} = 2.819$



- (d) There is not sufficient evidence for the researcher to reject the null hypothesis, since the test statistic is between the critical values ($-2.819 < 2.502 < 2.819$).
 (e) Lower bound: 99.39, upper bound: 110.21. Because the 99% confidence interval includes 100, we do not reject the statement in the null hypothesis.

7. (a) $t_0 = -1.677$

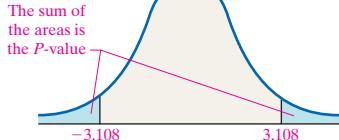
(b)



- (c) $0.05 < P\text{-value} < 0.10$ [Tech: $P = 0.0559$]. If we take 100 random samples of size 18, we would expect about six of the samples to result in a sample mean of 18.3 or less if $\mu = 20$.
 (d) The researcher will not reject the null hypothesis at the $\alpha = 0.05$ level of significance because the P -value is greater than the level of significance.

9. (a) No; $n \geq 30$

(c)



- (d) $0.002 < P\text{-value} < 0.005$ [Tech: $P\text{-value} = 0.0038$]. If we obtain 1000 random samples of size $n = 35$, we would expect about four samples to result in a mean as extreme or more extreme than the one observed if $\mu = 105$.

- (e) The researcher will reject the null hypothesis at the $\alpha = 0.01$ level of significance because the P -value is less than the level of significance ($0.0038 < 0.01$).

11. (a) $H_0: \mu = \$67$, $H_1: \mu > \$67$

- (b) There is a 0.02 probability of obtaining a sample mean of \$73 or higher from a population whose mean is \$67. So, if we obtained 100 simple random samples of size $n = 40$ from a population whose mean is \$67, we would expect about two of these samples to result in sample means of \$73 or higher.
 (c) Because the P -value is low ($P\text{-value} = 0.02 < \alpha = 0.05$), we reject the statement in the null hypothesis. There is sufficient evidence to conclude that the mean dollar amount withdrawn from a PayEase ATM is more than the mean amount from a standard ATM (that is, more than \$67).

13. (a) $H_0: \mu = 22$; $H_1: \mu > 22$

(b) The sample is random. The sample size is large, $n = 200 \geq 30$. We can reasonably assume that the sample is small relative to the population, so the scores are independent.

(c) Classical approach: $t_0 = 2.176 > t_{0.05} = 1.660$; reject the null hypothesis. P -value approach: $0.01 < P\text{-value} < 0.02$ [Tech: $P = 0.0154$]; reject the null hypothesis.

(d) There is sufficient evidence to conclude that students who complete the core curriculum are scoring above 22 on the math portion of the ACT.

15. Hypotheses: $H_0: \mu = 9.02 \text{ cm}^3$, $H_1: \mu < 9.02 \text{ cm}^3$. Classical approach: $t_0 = -4.553 < -t_{0.01} = -2.718 \text{ cm}^2$; reject the null hypothesis. P -value approach: $P\text{-value} < 0.0005$ [Tech: $P\text{-value} = 0.0004$] $< \alpha = 0.01$; reject the null hypothesis. There is sufficient evidence to conclude that the mean hippocampal volume in alcoholic adolescents is less than the normal mean volume of 9.02 cm^3 .

17. $H_0: \mu = 703.5$, $H_1: \mu > 703.5$. Classical approach: $t_0 = 0.813 < t_{0.05} = 1.685$ (39 degrees of freedom); do not reject the null hypothesis. P -value approach: $0.25 > P\text{-value} > 0.20$ [Tech: $P\text{-value} = 0.2105$] $> \alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence to conclude that the mean FICO score of high-income individuals is greater than that of the general population. In other words, it is not unlikely to obtain a mean credit score of 714.2 or higher even though the true population mean credit score is 703.5.

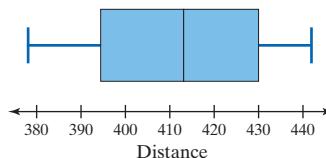
19. $H_0: \mu = 40.7$ years, $H_1: \mu \neq 40.7$ years; 95% confidence interval: Lower bound: 35.44 years, upper bound: 42.36 years. Because the interval includes 40.7 years, there is not significant evidence to conclude that the mean age of a death-row inmate has changed since 2002.

21. (a) Because $0.971 > 0.918$ (Table VI), it is reasonable to conclude the data come from a normal population. The boxplot does not show any outliers. Therefore, the conditions are satisfied. If using Option 2 to assess normality, the boxplot is "symmetric enough" to rely on the robustness of Student's t -distribution.

- (b) Hypotheses: $H_0: \mu = 84.3$ seconds, $H_1: \mu < 84.3$ seconds. Classical approach: $t_0 = -1.310 > -t_{0.10} = -1.383$ with 9 degrees of freedom; do not reject the null hypothesis. P -value approach: $0.15 > P\text{-value} > 0.10$ [Tech: $P\text{-value} = 0.1113$] $> \alpha = 0.10$; do not reject the null hypothesis. There is not sufficient evidence to conclude that the new system is effective.

23. (a) The shape of the distribution is roughly symmetric with no outliers. It is appropriate to use Student's t -distribution to determine the P -value.

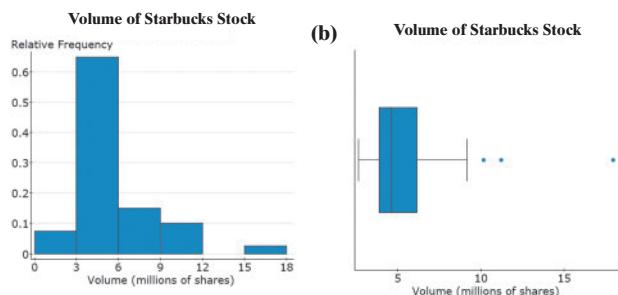
**Distance of Home Runs
in Coors Field**



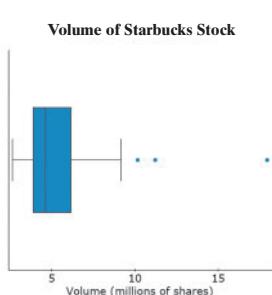
- (b) $H_0: \mu = 397.6$ feet versus $H_1: \mu > 397.6$ feet; Classical approach: $t_0 = 2.288 > t_{0.05} = 1.796$ with 11 degrees of freedom; reject the null hypothesis. P -value approach: $0.02 < P\text{-value} < 0.025$ [Tech: $P\text{-value} = 0.0215$] $> \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence to conclude a home run travels farther in Rockies stadium than in other Major League ballparks.

25. $H_0: \mu = 0.11 \text{ mg/L}$, $H_1: \mu \neq 0.11 \text{ mg/L}$. Classical approach: $t_0 = 1.707$ is between $-t_{0.025} = -2.262$ and $t_{0.025} = 2.262$; do not reject the null hypothesis. P -value approach: $0.1 < P\text{-value} < 0.2$ [Tech: $P = 0.122$]; do not reject the null hypothesis. Conclusion: There is not sufficient evidence to indicate that the calcium concentration in rainwater in Chautauqua, New York, has changed since 1990.

27. (a)



(b)



(c) The histogram indicates that the data is skewed right and the boxplot shows outliers.

(d) $H_0: \mu = 7.52$ million shares versus $H_1: \mu \neq 7.52$ million shares. Classical approach: $t_0 = -4.202$, $-t_{0.025} = -2.023$; since $t_0 < -t_{\alpha/2}$, reject H_0 . P-value approach: $P\text{-value} < 0.001$ [Tech: $P\text{-value} = 0.0001$], reject H_0 . There is sufficient evidence to conclude that the volume of Starbucks stock has changed since 2011.

29. $H_0: \mu = 0.11$ mg/L, $H_1: \mu \neq 0.11$ mg/L. Lower bound: 0.0948, upper bound: 0.2188. Because 0.11 is in the 95% confidence interval, we do not reject the statement in the null hypothesis. There is not sufficient evidence to indicate that the calcium concentration in rainwater in Chautauqua, New York, has changed since 1990.

31. $H_0: \mu = 7.52$, $H_1: \mu \neq 7.52$. Lower bound: 4.668, upper bound: 6.521. Because 7.52 is not in the 95% confidence interval, we reject the statement in the null hypothesis. The evidence suggests that the volume of Starbucks stock has changed since 2011.

33. (a) $H_0: \mu = 515$, $H_1: \mu > 515$

(b) Classical approach: $t_0 = 1.529 > t_{0.10} \approx 1.282$; reject the null hypothesis. P-value approach: $0.05 < P\text{-value} < 0.10$ [Tech: 0.0632] $< \alpha = 0.10$; reject the null hypothesis.

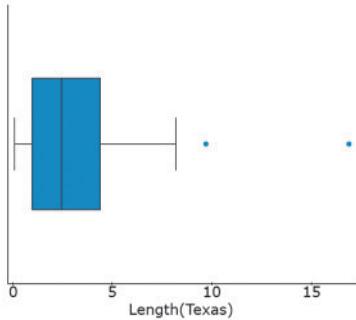
(c) Answers will vary.

(d) With $n = 400$ students: Classical approach: $t_0 = 0.721 < t_{0.10} \approx 1.29$; do not reject the null hypothesis. P-value approach: $0.20 < P\text{-value} < 0.25$ [Tech: 0.2358] $> \alpha = 0.10$; do not reject the null hypothesis.

35. (a) $\mu = 3.775$ feet

(b) The shape of the distribution is skewed right with two outliers.

Length of Tornadoes in Texas



(c) Because the sample data is skewed right with two outliers, a large sample size is needed to test a hypothesis about a population mean using Student's t -distribution.

(d) $H_0: \mu = 3.775$ feet versus $H_1: \mu \neq 3.775$ feet; Classical approach: $-t_{0.025} = -2.023 < t_0 = -0.962 < t_{0.025} = 2.023$ with 39 degrees of freedom; do not reject the null hypothesis. P-value approach: $0.30 < P\text{-value} < 0.40$ [Tech: $P\text{-value} = 0.3419$] $> \alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence to conclude the length of a tornado in Texas is different from the length of all tornadoes in the United States.

(e) $H_0: \mu = 3.775$ feet versus $H_1: \mu \neq 3.775$ feet; Classical approach: $t_0 = -5.037 < -t_{0.025} = -2.023$ with 39 degrees of freedom; reject the null hypothesis. P-value approach: $P\text{-value} < 0.0005$ [Tech: $P\text{-value} < 0.0001$] $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence to conclude the

length of a tornado in Texas is different from the length of all tornadoes in the United States.

(f) Different independent random samples can lead to different conclusions. The P -value is really “the P -value for a particular random sample.” Bottom line—the P -value is itself a random variable that varies from sample to sample. This is why low P -values mean the research should be validated by replicating the results.

37. (a) Answers will vary.

(b) We would expect five to result in a Type I error.

(c) Answers will vary.

(d) We know the population mean.

39. (a) To determine if artificial light from e-Readers disrupted sleep.

(b) Time to sleep; quantitative.

(c) e-Reader or book.

(d) This is a matched-pairs design where the individual is matched against himself or herself.

(e) The P -value of 0.009 means that if the null hypothesis of equal time to sleep for e-Readers and printed books was true, we would expect a mean time difference of 10 minutes or more in about 9 of every 1000 samples.

41. This means that minor departures from normality will not adversely affect the results of the test.

43. Answers will vary.

10.4 Assess Your Understanding (page 470)

1. Hypotheses: $H_0: \mu = 1$, $H_1: \mu < 1$. Classical approach: $t_0 = -2.179 > -t_{0.01} = -2.552$ with 18 degrees of freedom; do not reject the null hypothesis. P-value approach: $0.025 > P\text{-value} > 0.02$ [Tech: $P\text{-value} = 0.0214$] $> \alpha = 0.01$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.01$ level of significance to conclude that the mean is less than 1.

3. Hypotheses: $H_0: \mu = 25$, $H_1: \mu \neq 25$. Classical approach: $t_0 = -0.738$ is between $-t_{0.005} = -2.977$ and $t_{0.005} = 2.977$ with 14 degrees of freedom; do not reject the null hypothesis. P-value approach: $0.40 < P\text{-value} < 0.50$ [Tech: $P\text{-value} = 0.4729$] $> \alpha = 0.01$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.01$ level of significance to conclude that the population mean is different from 25.

5. Hypotheses: $H_0: \mu = 100$, $H_1: \mu > 100$. Classical approach: $t_0 = 3.003 > t_{0.05} = 1.685$; reject the null hypothesis. P-value approach: $0.001 < P\text{-value} < 0.0025$ [Tech: $P\text{-value} = 0.0023$] $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the population mean is greater than 100.

7. $H_0: \mu = 100$, $H_1: \mu > 100$. Classical approach: $t_0 = 1.278 < t_{0.05} = 1.729$; do not reject the null hypothesis. P-value approach: $0.10 < P\text{-value} < 0.15$ [Tech: 0.1084] $> \alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence to conclude that mothers who listen to Mozart have children with higher IQs.

9. (a) The variable of interest is whether the student passes, or not. It is qualitative with two outcomes.

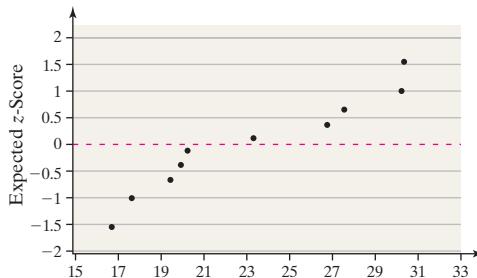
(b) $H_0: p = 0.526$, $H_1: p > 0.526$; $np_0(1 - p_0) = 119.7 > 10$.

Assuming there are over 9600 students eligible for this course (reasonable for large schools), $n \leq 0.05N$. Classical approach:

$z_0 = 1.32 < z_{0.01} = 2.33$; do not reject the null hypothesis.

P-value approach: $P\text{-value} = 0.0934$ [Tech: 0.0922] $> \alpha = 0.01$; do not reject the null hypothesis. There is not sufficient evidence to conclude that the proportion of students who pass is greater than 0.526.

(c) Changing the delivery method for a course for the entire campus would be extremely time-consuming. Plus, building labs would be expensive. The researchers would want to be sure the evidence is overwhelming against the statement in the null hypothesis.

11. Normal Probability Plot of Gas Mileage

The correlation between gas mileage and expected z-scores is 0.962 [Tech: 0.968] > 0.918 (Table VI). It is reasonable to conclude that the data come from a population that is normally distributed. A boxplot indicates that there are no outliers. $H_0: \mu = 28 \text{ mpg}$, $H_1: \mu \neq 28 \text{ mpg}$; Classical approach: $t_0 = -2.963 < -t_{0.05} = -1.833$, reject H_0 . P -value approach: $0.01 < P\text{-value} < 0.02$ [Tech: 0.0159] < $\alpha = 0.10$, reject H_0 . There is sufficient evidence to conclude that the individuals are getting different gas mileage than the EPA suggests.

13. Hypotheses: $H_0: p = 0.26$, $H_1: p \neq 0.26$; $np_0(1 - p_0) = 11.544 > 10$ and $n \leq 0.05N$. Classical approach: $z_0 = 1.589$ is between $-z_{0.025} = -1.96$ and $z_{0.025} = 1.96$; do not reject the null hypothesis. P -value approach: $P\text{-value} = 0.1118$ [Tech: 0.112] > $\alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the proportion of individuals who text while driving is different from 0.26. The data do not contradict the results from Toluna.

15. Assuming the congresswoman wants to avoid voting for the tax increase unless she is confident that the majority of her constituents are in favor of it, the value of α should be small, such as $\alpha = 0.01$ or $\alpha = 0.05$, to avoid a Type I error. Let p be the population proportion in favor of the tax increase. Hypotheses: $H_0: p = 0.5$, $H_1: p > 0.5$; $np_0(1 - p_0) = 2062.5 > 10$ and $n \leq 0.05N$. P -value = 0.0392 [Tech: 0.0391]. The P -value is greater than $\alpha = 0.01$ but less than $\alpha = 0.05$, so the conclusion depends on the value of α . Recommendations will vary.

17. Because the data is a simple random sample, $np_0(1 - p_0) = 49.73 \geq 10$, and the sample size n is less than 5% of the population size N (there have been over 10,000 trading days the past 50 years), we may use the normal model to test the hypothesis. $H_0: p = 0.537$ versus $H_1: p \neq 0.537$; Classical approach: Because $z_0 = 2.35 > z_{0.025} = 1.96$, reject the null hypothesis. P -value approach: $P\text{-value} = 0.0188$ [Tech: $P\text{-value} = 0.0186$] < $\alpha = 0.05$; reject the null hypothesis. There is sufficient evidence to conclude the number of days Amazon stock is up is different from the market as a whole. We are 95% confident the proportion of days Amazon stock is up is between 0.553 and 0.687.

19. **(a)** Do not reject H_0 for $\mu_0 = 104, 105, \dots, 110, 111$.
(b) Lower bound: 103.43, Upper bound: 111.17; The lower and upper bounds for the 95% confidence interval represent the cut-off points for which the null hypothesis is rejected.
(c) More values of μ_0 would lead to not rejecting the null hypothesis.

21. Confidence interval; mean

23. Hypothesis test; proportion; $H_0: p = 0.55$ vs. $H_1: p > 0.55$

25. Hypothesis test; mean; $H_0: \mu = 0$ vs. $H_1: \mu > 0$

Chapter 10 Review Exercises (page 474)

- 1.** **(a)** $H_0: \mu = 3173$; $H_1: \mu < 3173$
(b) If we reject the null hypothesis and the mean credit card debt of college undergraduates has not decreased, then we make a Type I error.
(c) If we do not reject the null hypothesis and the mean credit card debt of college undergraduates has decreased, then we make a Type II error.
(d) There is not sufficient evidence at the α level to support the researcher's belief that the mean credit card debt of college undergraduates has decreased.

- (e)** There is sufficient evidence at the α level to support the researcher's belief that the mean credit card debt of college undergraduates has decreased.
- 2.** **(a)** $H_0: p = 0.13$; $H_1: p \neq 0.13$
(b) If we reject the null hypothesis and the proportion of cards that result in default is not different today, we make a Type I error.
(c) If we do not reject the null hypothesis and the proportion of cards that result in default is different today, we make a Type II error.
(d) There is not sufficient evidence at the α level to support the credit analyst's belief that the proportion of cards that result in default is different today.
(e) There is sufficient evidence at the α level to support the credit analyst's belief that the proportion of cards that result in default is different today.
- 3.** The probability of a Type I error is 0.05.
- 4.** The probability of a Type II error is 0.113.
- 5.** **(a)** The sample size must be large to use the Central Limit Theorem because the distribution of the population may not be normal.
(b) Classical approach: $t_0 = 2.0515 > t_{0.05} = 1.691$ for 34 degrees of freedom; reject the null hypothesis. P -value approach: $0.02 < P\text{-value} < 0.025$ [Tech: $P = 0.0240$]; reject the null hypothesis.
- 6.** **(a)** The distribution must be normal because the sample size is small.
(b) Classical approach: $t_0 = -1.795$ is between $-t_{0.025} = -2.145$ and $t_{0.025} = 2.145$ for 14 degrees of freedom; do not reject the null hypothesis. P -value approach: $0.05 < P\text{-value} < 0.10$ [Tech: $P = 0.0943$]; do not reject the null hypothesis.
- 7.** Hypotheses: $H_0: p = 0.6$, $H_1: p > 0.6$; $np_0(1 - p_0) = 60 > 10$ and $n \leq 0.05N$. Classical approach: $z_0 = 1.94 > z_{0.05} = 1.645$; reject the null hypothesis. P -value approach: $P\text{-value} = 0.0262$ [Tech: 0.0264] < $\alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that $p > 0.6$.
- 8.** Hypotheses: $H_0: p = 0.35$, $H_1: p \neq 0.35$; $np_0(1 - p_0) = 95.6 > 10$ and $n \leq 0.05N$. Classical approach: $z_0 = -0.92$ is between $-z_{0.025} = -1.96$ and $z_{0.025} = 1.96$; do not reject the null hypothesis. P -value approach: $P\text{-value} = 0.3576$ [Tech: 0.3572] > $\alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that $p \neq 0.35$.
- 9.** **(a)** $H_0: p = 0.733$, $H_1: p \neq 0.733$.
(b) The sample is random. $np_0(1 - p_0) = 19.57 \geq 10$. We can reasonably assume that the sample is less than 5% of the population.
(c) The P -value is 0.2892 [Tech: 0.2881], so a result this far from the proportion stated in the null hypothesis will occur in about 29 of 100 samples of this size when the null hypothesis is true. We do not reject the null hypothesis at any commonly used level of significance α . Mary's sample evidence does not contradict Professor Wilson's findings.
- 10.** Hypotheses: $H_0: p = 0.05$, $H_1: p > 0.05$. $np_0(1 - p_0) = 11.875 \geq 10$. The P -value is 0.0951 [Tech: 0.0958]. We reject the null hypothesis at $\alpha = 0.10$, but we do not reject it for $\alpha = 0.01$ or $\alpha = 0.05$. The administrator should be a little concerned.
- 11.** **(a)** Data obtained via a random sample. Yes, we can reasonably assume that the distances between retaining rings is normally distributed. In addition, the sample size is large.
(b) $H_0: \mu = 0.875$, $H_1: \mu > 0.875$
(c) Recalibrating the machine could be very costly, so he wants to avoid the consequences of making a Type I error.
(d) Classical approach: $t_0 = 1.2 < t_{0.01} = 2.438$; do not reject the null hypothesis. P -value approach: $0.10 < P\text{-value} < 0.15$ [Tech: $P\text{-value} = 0.1191$]; do not reject the null hypothesis. No, the evidence does not suggest that the machine be recalibrated.
(e) The quality-control engineer would make a Type I error if he recalibrated the machine when it did not need to be recalibrated. He would make a Type II error if he did not recalibrate the machine when it actually did need to be recalibrated.

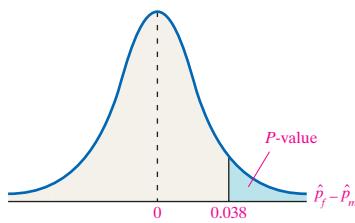
- 12.** Hypotheses: $H_0: \mu = 98.6$, $H_1: \mu < 98.6$. Classical approach: $t_0 = -15.119 < -t_{0.01} = -2.364$, using 100 degrees of freedom; reject the null hypothesis. P -value approach: P -value < 0.0005 [Tech: P -value < 0.0001]; reject the null hypothesis. The evidence suggests that the mean temperature of humans is less than 98.6°F.
- 13.** **(a)** Yes, because $0.951 > 0.928$ (Table VI), it is reasonable to conclude the data come from a population that is normally distributed and the boxplot shows no outliers.
- (b)** Hypotheses: $H_0: \mu = 1.68$, $H_1: \mu \neq 1.68$. Lower bound: 1.6781 [Tech: 1.6782], upper bound: 1.6839 [Tech: 1.6838]. The mean stated in the null hypothesis is included in the 95% confidence interval, so do not reject the null hypothesis. Conclusion: There is not sufficient evidence to conclude that the golf balls do not conform to USGA standards.
- 14.** Hypotheses: $H_0: \mu = 480$, $H_1: \mu < 480$. Classical approach: $t_0 = -1.577 > -t_{0.05} = -1.812$; do not reject the null hypothesis. P -value approach: $0.05 < P$ -value < 0.10 [Tech: P -value < 0.0730]; do not reject the null hypothesis. There is not sufficient evidence to suggest that the students are studying less than 480 minutes per week.
- 15.** $H_0: p = 0.5$, $H_1: p > 0.5$; $np_0(1 - p_0) = 150(0.5)(1 - 0.5) = 37.5 > 10$ and $n \leq 0.05N$. Classical approach: $z_0 = 0.98 < z_{0.05} = 1.645$; do not reject the null hypothesis; P -value approach: P -value = 0.1635 [Tech: 0.1636] $> \alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence to suggest that a majority of pregnant women nap at least twice each week.
- 16.** Hypotheses: $H_0: p = 0.52$, $H_1: p > 0.52$. The P -value is 0.0793 [Tech: 0.0788]. Reject the null hypothesis at the $\alpha = 0.10$ level of significance. The results are statistically significant. The new retention rate was about 53.3%, not a great increase over 52%, so the results might not be considered practically significant. The cost of the new policies would have to be weighed against such a small increase in the retention rate when considering whether other community colleges should implement the policies.
- 17.** Hypotheses: $H_0: p = 0.4$, $H_1: p > 0.4$; $np_0(1 - p_0) = 9.6 < 10$. P -value = 0.3115 $> \alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to support the researcher's notion that the proportion of adolescents who pray daily has increased.
- 18.** Hypotheses: $H_0: \mu = 73.2$, $H_1: \mu < 73.2$. Classical approach: $t_0 = -2.018 < -t_{0.05} = -1.645$; reject the null hypothesis. P -value approach: P -value = 0.0218 $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to support the coordinator's concern that the mean test scores decreased. The results do not illustrate any practical significance. The average score only declined 0.4 point.
- 19.** If we accept the null hypothesis, we are saying it is true. If we do not reject the null hypothesis, we are saying that we do not have enough evidence to conclude that it is not true. It is the difference between saying H_0 is true and saying we are not convinced that it is false.
- 20.** $H_0: \mu = 1.22$, $H_1: \mu < 1.22$. The P -value indicates that if the population mean is 1.22 hours as stated in the null hypothesis, then results as low as or lower than those obtained by the researcher would occur in about 3 of 100 similar samples.
- 21.** In the Classical Approach, a test statistic is calculated and a rejection region based on the level of significance α is determined within the appropriate distribution for the population parameter stated in the null hypothesis. The null hypothesis is rejected if the test statistic is in the rejection region.
- 22.** In the P -value approach, the probability of getting a result as extreme as the one obtained in the sample is calculated assuming the null hypothesis to be true. If the P -value is less than a predetermined level of significance α , we reject the null hypothesis.
- Chapter 10 Test (page 475)**
- 1.** **(a)** $H_0: \mu = 42.6$ minutes, $H_1: \mu > 42.6$ minutes
(b) There is sufficient evidence to conclude that the mean amount of daily time spent on phone calls and answering or writing emails has increased since 2006.
- (c)** We would reject the null hypothesis that the mean is 42.6 minutes, when, in fact, the mean amount of daily time spent on phone calls and answering or writing emails is 42.6 minutes.
- (d)** We would not reject the null hypothesis that the mean is 42.6 minutes, when, in fact, the mean amount of time spent on phone calls and answering or writing emails is greater than 42.6 minutes.
- 2.** **(a)** $H_0: \mu = 167.1$ seconds, $H_1: \mu < 167.1$ seconds
(b) By choosing a level of significance of 0.01, the probability of rejecting the null hypothesis that the mean is 167.1 seconds in favor of the alternative hypothesis that the mean is less than 167.1 seconds, when, in fact, the mean is 167.1 seconds, is small.
(c) Classical approach: $t_0 = -1.75 > -t_{0.01} = -2.381$ (using 70 degrees of freedom); do not reject the null hypothesis. P -value approach: $0.025 < P$ -value < 0.05 [Tech: P -value = 0.0423] is greater than $\alpha = 0.01$; do not reject the null hypothesis. There is not sufficient evidence to conclude that the drive-through service time has decreased.
- 3.** $H_0: \mu = 8$, $H_1: \mu < 8$. Classical approach: $t_0 = -1.755 < -t_{0.05} = -1.660$ (using 100 degrees of freedom); reject the null hypothesis; P -value approach: $0.025 < P$ -value < 0.05 [Tech: P -value = 0.0406] $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence to conclude that postpartum women get less than 8 hours of sleep each night.
- 4.** $H_0: \mu = 1.3825$ inches, $H_1: \mu \neq 1.3825$ inches. Lower bound: 1.3824 inches; upper bound: 1.3828 inches. Because the interval includes the mean stated in the null hypothesis, we do not reject the null hypothesis. There is not sufficient evidence to conclude that the mean is different from 1.3825 inches. Therefore, we presume that the part has been manufactured to specifications.
- 5.** $H_0: p = 0.6$, $H_1: p > 0.6$; Assume data collected through a random sample. $np_0(1 - p_0) = 1.561(0.6)(0.4) = 374.64 > 10$ and $n \leq 0.05N$. We will use an $\alpha = 0.05$ level of significance. Classical approach: $z_0 = 0.89 < z_{0.05} = 1.645$; do not reject the null hypothesis. P -value approach: P -value = 0.1867 [Tech: 0.1843] $> \alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence to conclude that a supermajority of Americans did not feel that the United States would need to fight Japan in their lifetimes. It is interesting, however, that significantly fewer than half of all Americans felt the United States would not have to fight Japan in their lifetimes. About 2.5 years after this survey, the Japanese attack on Pearl Harbor thrust the United States into World War II.
- 6.** $H_0: \mu = 0$, $H_1: \mu > 0$. Classical approach: $t_0 = 2.634 > t_{0.05} = 1.664$ (using 80 df); reject the null hypothesis. P -value approach: $0.0025 < P$ -value < 0.005 [Tech: P -value = 0.0051] $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence to suggest that the diet is effective. However, losing 1.6 kg of weight over the course of a year does not seem to have much practical significance.
- 7.** $H_0: p = 0.37$, $H_1: p > 0.37$, Assume data collected through a random sample. $np_0(1 - p_0) = 6.993 < 10$. P -value = $P(\hat{p} \geq 0.37) = P(X \geq 16) = 0.0501 < \alpha = 0.10$; reject the null hypothesis. There is sufficient evidence to conclude that the proportion of 20- to 24-year-olds who live on their own and do not have a landline is greater than 0.37.

CHAPTER 11 Inferences on Two Population Parameters

11.1 Assess Your Understanding (page 487)

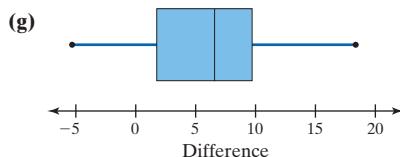
1. Independent
3. Dependent; quantitative
5. Independent; qualitative
7. Independent; quantitative
- 9.** **(a)** $H_0: p_1 = p_2$ versus $H_1: p_1 > p_2$
(b) $z_0 = 3.07$
(c) $z_{0.05} = 1.645$
(d) P -value = 0.0011 [Tech: 0.0010]. Because $z_0 > z_{0.05}$ (or P -value $< \alpha$), we reject the null hypothesis. There is sufficient evidence to support the claim that $p_1 > p_2$.

- 11.** (a) $H_0: p_1 = p_2$ versus $H_1: p_1 \neq p_2$
 (b) $z_0 = -0.34$
 (c) $-z_{0.025} = -1.96; z_{0.025} = 1.96$
 (d) $P\text{-value} = 0.7264$ [Tech: 0.7307]. Because $-z_{0.025} < z_0 < z_{0.025}$ (or $P\text{-value} > \alpha$), we do not reject the null hypothesis. There is not sufficient evidence to support the claim that $p_1 \neq p_2$.
- 13.** Lower bound: -0.075 , upper bound: 0.015
- 15.** Lower bound: -0.063 , upper bound: 0.043 [Tech: 0.044]
- 17.** Each sample is the result of a completely randomized design experiment; the response variable is qualitative with two outcomes. $n_1\hat{p}_1(1 - \hat{p}_1) = 91 \geq 10$ and $n_2\hat{p}_2(1 - \hat{p}_2) = 60 \geq 10$; and each sample is less than 5% of the population size. $H_0: p_1 = p_2; H_1: p_1 > p_2$. Classical approach: $z_0 = 2.20 > z_{0.05} = 1.645$; reject H_0 . $P\text{-value}$ approach: $P\text{-value} = 0.0139 < \alpha = 0.05$; reject H_0 . There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that a higher proportion of subjects in the treatment group (taking Prevnar) experienced fever as a side effect than in the control (placebo) group.
- 19.** $\hat{p}_{1947} = 0.37, \hat{p}_{\text{recent}} = 0.303$. Each independent sample is a simple random sample; the variable of interest is qualitative with two outcomes. $n_{1947}\hat{p}_{1947}(1 - \hat{p}_{1947}) \geq 10$ and $n_{\text{recent}}\hat{p}_{\text{recent}}(1 - \hat{p}_{\text{recent}}) \geq 10$. The sample size is less than 5% of the population for each sample. $H_0: p_{1947} = p_{\text{recent}}$ vs. $H_1: p_{1947} \neq p_{\text{recent}}$. Classical approach: $z_0 = 3.33$ is greater than $z_{0.025} = 1.96$; reject H_0 . $P\text{-value}$ approach: $P\text{-value} = 0.0008 < \alpha = 0.05$; reject H_0 . There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the proportion of adult Americans who were abstainers in 1947 is different from the recent proportion of abstainers.
- 21.** $H_0: p_m = p_f$ vs. $H_1: p_m \neq p_f$. Lower bound: -0.008 [Tech: -0.009], upper bound: 0.048 . We are 95% confident that the difference in the proportion of males and females that have at least one tattoo is between -0.008 and 0.048 . Because the interval includes zero, we do not reject the null hypothesis. There is no significant difference in the proportion of males and females that have tattoos.
- 23.** (a) The data were obtained from two independent simple random samples, $n\hat{p}(1 - \hat{p}) \geq 10$ for each independent sample, and the sample size is less than 5% of the population size for each sample.
 (b) $H_0: p_f = p_m$ vs. $H_1: p_f > p_m$
 (c) $\hat{p}_f - \hat{p}_m$ is approximately normal with $\mu_{\hat{p}_f - \hat{p}_m} = 0$ and $\sigma_{\hat{p}_f - \hat{p}_m} = \sqrt{0.349(1 - 0.349)}\left(\frac{1}{540} + \frac{1}{560}\right) \approx 0.029$.
- (d) $P\text{-value} = p(\hat{p}_f - \hat{p}_m > 0.038) = 0.0905$ [Tech: 0.0918]
 (e) If we obtain 1000 different simple random samples as described in the original problem, we would expect 93 to have a difference in sample proportion of 0.038 or higher if the population proportion difference was 0.
 (f) Since $P\text{-value} > \alpha$, do not reject H_0 . There is not sufficient evidence to conclude the proportion of females annoyed by people who repeatedly check their mobile phones while having an in-person conversation is greater than the proportion of males.
- 25.** (a) Completely randomized design.
 (b) The response variable is whether the subject experiences dry mouth, or not. It is qualitative with two possible outcomes.
 (c) The explanatory variable is type of drug. It has two levels—Claritin or placebo.
 (d) Double-blind means that neither the subject nor the individual monitoring the subject knows which treatment the subject is receiving (Claritin or placebo).



- 27.** (a) Completely randomized design
 (b) Death, or not; qualitative with two possible outcomes
 (c) The subjects were randomly assigned to one of two treatment groups: $n\hat{p}_{OPDIVO}(1 - \hat{p}_{OPDIVO}) = 210(0.214)(1 - 0.214) = 35.3 \geq 10; n\hat{p}_D(1 - \hat{p}_D) = 208(0.106)(1 - 0.106) = 19.7 \geq 10$; There are over 10,000 individuals with metastatic melanoma, so the sample size is less than 5% of the population size. We are 95% confident the difference in the proportion of subjects receiving OPDIVO versus dacarbazine who survived 12 months is between 0.039 and 0.177 [Tech: 0.178].
- 29.** (a) $H_0: p_{TX} = p_{GA}$ vs. $H_1: p_{TX} \neq p_{GA}$; we are treating the data as a simple random sample. The sample sizes are large and the sample sizes are small relative to the size of the populations (since we are treating this data as a sample of all tornadoes in each state). Classical approach: $z_0 = 2.08 > z_{0.025} = 1.96$. Reject H_0 ; $P\text{-value}$ approach: $P\text{-value} = 0.0375$ [Tech: 0.0379] < 0.05 (level of significance). Reject H_0 . There is sufficient evidence to suggest the proportion of F0 tornadoes in Texas is different from the proportion of F0 tornadoes in Georgia.
 (b) We are 95% confident the difference in the proportion of F0 tornadoes in Texas is between 0.009 and 0.239 higher than the proportion of F0 tornadoes in Georgia.
- 31.** (a) In sentence A, the verbs are “was having” and “was taking.” In sentence B, the verbs are “had” and “took.”
 (b) 98; 98
 (c) $\hat{p}_A = 0.276; \hat{p}_B = 0.5$
 (d) $H_0: p_A = p_B$ vs. $H_1: p_A \neq p_B$; Students were randomly assigned to one of two groups; $n\hat{p}(1 - \hat{p}) \geq 10$ for both sentence A and sentence B group; assume the size of the population of students is large enough so that the sample size is no more than 5% of the population size. Classical approach: $z_0 = -3.22 < -z_{0.025} = -1.96$ (using 0.05 level of significance). Reject H_0 . $P\text{-value}$ approach: $P\text{-value} = 0.0013 < 0.05$, so reject H_0 . There is sufficient evidence to conclude the proportion of students reading sentence A who think the politician will be re-elected is different from the proportion of students reading sentence B who think the politician will be re-elected.
 (e) Answers will vary. The wording in sentence A suggests that the actions were taking place over a period of time and habitual behavior may be continuing in the present or might be expected to continue at some time in the future. The wording in sentence B suggests events that are concluded and were possibly brief in duration or one-time occurrences.
- 33.** (a) $n = n_1 = n_2 = 1406$
 (b) $n = n_1 = n_2 = 2135$
- 35.** (a) The response variable is whether the fund manager outperforms the market, or not. The explanatory variable is whether it is a high-dispersion year or a low-dispersion year.
 (b) The individuals are the fund managers.
 (c) This study applies to the population of fund managers.

- (d) $H_0: p_{\text{high}} = p_{\text{low}}$ vs. $H_1: p_{\text{high}} > p_{\text{low}}$
- (e) If the null hypothesis were true and we conducted the study 100 times, we would expect to observe the results as extreme or more extreme than the results observed in about eight of the studies. There is some evidence to suggest that the proportion of fund managers who outperform the market in high-dispersion years is greater than the proportion of fund managers who outperform the market in low-dispersion years.
37. A pooled estimate of p is the best point estimate of the common population proportion p . However, when finding a confidence interval, the sample proportions are not pooled because no assumption about their equality is made.
- ### 11.2 Assess Your Understanding (page 497)
1. <
 3. (a)
- | Observation | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-------------|------|---|------|------|-----|------|------|
| d_i | −0.5 | 1 | −3.3 | −3.7 | 0.5 | −2.4 | −2.9 |
- (b) $\bar{d} = -1.614$; $s_d = 1.915$
- (c) Classical approach: $t_0 = -2.230 < -t_{0.05} = -1.943$; reject the null hypothesis. P-value approach: $0.025 < P\text{-value} < 0.05$ [Tech: $P\text{-value} = 0.0336$] $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to reject the null hypothesis that $\mu_d = 0$.
- (d) We can be 95% confident that the mean difference is between −3.39 and 0.16.
5. (a) $H_0: \mu_d = 0$ vs. $H_1: \mu_d > 0$
- (b) Classical approach: $t_0 = 2.606 > t_{0.05} = 1.753$; reject the null hypothesis. P-value approach: $0.005 < P\text{-value} < 0.01$ [Tech: $P\text{-value} = 0.0099$] $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance that a baby will watch the climber approach the hinderer toy for a longer time, on average, than the baby will watch the climber approach the helper toy.
- (c) Answers will vary. The fact that the babies watch the surprising behavior for a longer period of time suggests that they are curious about it.
7. (a) This is matched-pairs data because two measurements (A and B) are taken on the same round.
- (b) Hypotheses: $H_0: \mu_d = 0$, $H_1: \mu_d \neq 0$; $d_i = A_i - B_i$. Classical approach: $t_0 = 0.852$; do not reject the null hypothesis since $-t_{0.005} = -3.106 < 0.852 < t_{0.005} = 3.106$. P-value approach: $0.50 > P\text{-value} > 0.40$ [Tech: $P\text{-value} = 0.4125$] $> \alpha = 0.01$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.01$ level of significance to conclude that there is a difference in the measurements of velocity between device A and device B.
- (c) Lower bound: −0.309, upper bound: 0.542. We are 99% confident that the mean difference in measurement is between −0.31 and 0.54 feet per second.
- (d)
-
- Yes, the boxplot supports that there is no difference in measurements.
9. (a) These are matched pairs because the car and the SUV were involved in the same collision.
- (b)
-
- The median of the differences is well to the left of 0 suggesting that SUVs do have a lower repair cost.
- (c) $H_0: \mu_d = 0$ vs. $H_1: \mu_d < 0$. Classical approach: $t_0 = -2.685 < -t_{0.05} = -1.943$; reject the null hypothesis. P-value approach: $0.01 < P\text{-value} < 0.02$ [Tech: $P\text{-value} = 0.0181$] $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to suggest that the mean repair cost for the car is higher.
11. Hypotheses: $H_0: \mu_d = 0$, $H_1: \mu_d > 0$; $d_i = Y_i - X_i$. Classical approach: $t_0 = 0.392 < t_{0.10} = 1.356$; do not reject the null hypothesis. P-value approach: $P\text{-value} > 0.25$ [Tech: $P\text{-value} = 0.3508$] $> \alpha = 0.10$; do not reject the null hypothesis. No; there is not sufficient evidence at the $\alpha = 0.10$ level of significance to conclude that sons are taller than their fathers.
13. $H_0: \mu_d = 0$; $H_1: \mu_d \neq 0$; $d_i = \text{diamond} - \text{steel}$, $\bar{d} = 1.333$; $s_d = 1.5$, $t_{0.025} = 2.306$. Lower bound: 0.2; upper bound: 2.5. We are 95% confident that the difference in hardness reading is between 0.2 and 2.5. Because the interval does not include 0, we reject the null hypothesis. There is sufficient evidence to conclude that the two indenters produce different hardness readings.
15. (a) To control for any “learning” that may occur in using the simulator.
- (b) Hypotheses: $H_0: \mu_d = 0$, $H_1: \mu_d \neq 0$. Lower bound: 0.688, upper bound: 1.238. $d_i = Y_i - X_i$. We can be 95% confident that the mean difference in reaction time when teenagers are driving impaired from when driving normally is between 0.688 second and 1.238 seconds. Because the interval does not contain zero, we reject the null hypothesis. There is sufficient evidence to conclude there is a difference in braking time with impaired vision and normal vision.
17. (a) Use the same camera to eliminate variability due to camera location.
- (b) A large sample size is needed due to the outlier.
- Difference in Violations at Each Camera**
-
- (c) $H_0: \mu_d = 0$ vs. $H_1: \mu_d \neq 0$; Classical approach: $-t_{0.025} = -2.045 < t_0 = -0.650 < t_{0.025} = 2.045$; do not reject H_0 . P-value approach: $P\text{-value} > 0.5$ [Tech: $P\text{-value} = 0.5207$]; do not reject H_0 . There is not sufficient evidence to conclude there is a difference in the number of moving violations identified by the camera between Saturday and Wednesday.
- (d) Answers may vary. Weather is certainly a potential confounder.
19. (a) It is matched-pairs data because the flying conditions are similar for each windmilling/stopped-propeller matching.
- (b) The decision to windmill or stop the propeller first was determined randomly to control for any “learning” that may take place while flying without the aid of an engine.
- (c) The pilot can see whether the propeller is spinning or not.
- (d) Response: time to descend 800 feet; treatments: windmilling or stopped propeller.
- (e)
-
- An outlier is present.
- (f) The conditions of windmilling and stopped propeller were not the same, so the data are not matched by weather conditions.



Yes

(h) $H_0: \mu_d = 0$ vs. $H_1: \mu_d > 0$

(i) Classical: $t_0 = 5.49 > t_{0.05} = 1.708$, reject the null hypothesis; $P\text{-value} < 0.0005$ [Tech: < 0.0001], reject the null hypothesis.

There is sufficient evidence to conclude that stopping the propeller increases the mean time to fall 800 feet (thereby increasing the horizontal distance that the plane will travel).

(j) If you experience engine failure, you should stop the propeller from windmilling to increase the distance that the plane will glide.

11.3 Assess Your Understanding (page 508)

1. (a) $H_0: \mu_1 = \mu_2, H_1: \mu_1 \neq \mu_2$. Classical approach: $t_0 = 0.898$ is between $-t_{0.025} = -2.145$ and $t_{0.025} = 2.145$; do not reject H_0 . $P\text{-value}$ approach: $0.30 < P\text{-value} < 0.40$ [Tech: $P\text{-value} = 0.3767$] $> \alpha = 0.05$; do not reject H_0 . There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the population means are different.

(b) Lower bound: -1.53 [Tech: -1.41], upper bound: 3.73 [Tech: 3.61]

3. (a) $H_0: \mu_1 = \mu_2, H_1: \mu_1 > \mu_2$. Classical approach: $t_0 = 3.081 > t_{0.10} = 1.333$; reject H_0 . $P\text{-value}$ approach: $0.0025 < P\text{-value} < 0.005$ [Tech: $P\text{-value} = 0.0024$] $< \alpha = 0.10$; reject H_0 . There is sufficient evidence at the $\alpha = 0.10$ level of significance to conclude that $\mu_1 > \mu_2$.

(b) Lower bound: 3.57 [Tech: 3.67], upper bound: 12.83 [Tech: 12.73]

5. $H_0: \mu_1 = \mu_2; H_1: \mu_1 < \mu_2$. Classical approach: $t_0 = -3.158 < -t_{0.02} = -2.172$; reject H_0 . $P\text{-value}$ approach: $0.001 < P\text{-value} < 0.0025$ [Tech: $P\text{-value} = 0.0013$] $< \alpha = 0.02$; reject H_0 . There is sufficient evidence at the $\alpha = 0.02$ level of significance to conclude that $\mu_1 < \mu_2$.

7. (a) The response variable is time to graduate. The explanatory variable is community college first, or not.

(b) There are two independent groups: those who enroll directly in four-year institutions, and those who first enroll in community college and the response variable is quantitative. Treat each sample as a simple random sample. Each sample size is large, so each sample mean is approximately normal. Each population is small relative to its population size.

(c) $H_0: \mu_{CC} = \mu_{NT}$ vs. $H_1: \mu_{CC} > \mu_{NT}$. Classical approach: $t_0 = 12.977 > t_{0.01} \approx 2.364$ (using 100 df); reject H_0 . $P\text{-value}$ approach: $P\text{-value} < 0.0005$ [Tech: $P\text{-value} < 0.0001$]; reject H_0 . The evidence suggests that the mean time to graduate for students who first start in community college is longer than the mean time to graduate for those who do not transfer.

(d) Lower bound: 0.847 [Tech: 0.848], upper bound: 1.153 [Tech: 1.152]. We are 95% confident that the mean additional time to graduate for students who start in community college is between 0.847 year and 1.153 years.

(e) No; this is observational data. Community college students may be working more hours, which does not allow them to take additional classes.

9. (a) This is an observational study. The researcher did not influence the data.

(b) (1) Treat the data as a simple random sample, (2) the samples are obtained independently, (3) the sample sizes are large, and (4) each sample is small relative to the population size.

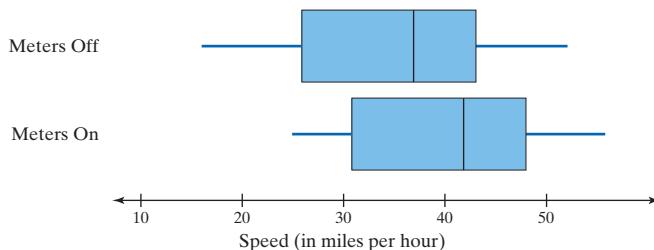
(c) $H_0: \mu_A = \mu_D, H_1: \mu_A \neq \mu_D$. Classical approach: $t_0 = 0.846$ is between $-t_{0.025} = -2.032$ and $t_{0.025} = 2.032$; do not

reject H_0 . $P\text{-value}$ approach: $0.50 > P\text{-value} > 0.40$ [Tech: $P\text{-value} = 0.4013$] $> \alpha = 0.05$; do not reject H_0 . There is not sufficient evidence at the $\alpha = 0.05$ level of significance to say that travelers walk at different speeds depending on whether they are arriving or departing an airport.

11. (a) Lower bound: 4.71 , upper bound: 6.29 , using 100 degrees of freedom. We are 95% confident that the mean difference in scores between students who think about being a professor and students who think about soccer hooligans is between 4.71 and 6.29 .

(b) Since the 95% confidence interval does not contain 0 , the results suggest that priming does have an effect on scores.

13. (a)



There are no outliers. The speed with meters on appears to be higher than with meters off.

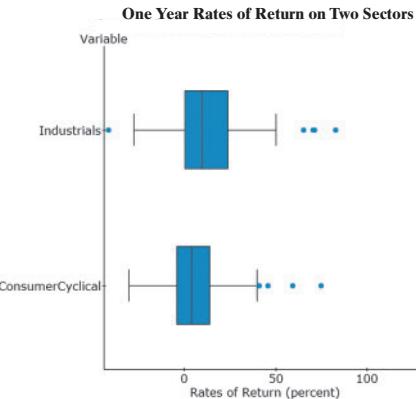
(b) $H_0: \mu_{on} = \mu_{off}, H_1: \mu_{on} > \mu_{off}$. Classical approach: $t_0 = 1.713 > t_{0.10} = 1.345$; reject the null hypothesis. $P\text{-value}$ approach: $0.05 < P\text{-value} < 0.10$ [Tech: $P\text{-value} = 0.0489$]; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.10$ level of significance that the ramp meters are effective in maintaining higher speed on the freeway.

15. $H_0: \mu_{carpet} = \mu_{no carpet}, H_1: \mu_{carpet} > \mu_{no carpet}$. Classical approach: $t_0 = 0.956 < t_{0.05} = 1.895$; do not reject H_0 . $P\text{-value}$ approach: $0.15 < P\text{-value} < 0.20$ [Tech: $P\text{-value} = 0.1780$] $> \alpha = 0.05$; do not reject H_0 . There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that carpeted rooms have more bacteria than uncarpeted rooms.

17. (a) $H_0: \mu_T = \mu_G$ vs. $H_1: \mu_T \neq \mu_G$. Classical approach: $t_0 = -3.336 < -t_{0.025} = -1.984$ (with 100 df); $P\text{-value}$ approach: $0.001 < P\text{-value} < 0.002$ [Tech: $P\text{-value} = 0.001$]. Reject H_0 . There is sufficient evidence to suggest the length of tornadoes in Texas is different from the length of tornadoes in Georgia.

(b) We are 95% confident the mean length of a tornado in Georgia is between 1.064 [Tech: 1.072] miles and 4.186 [Tech: 4.178] miles longer than the length of a tornado in Texas.

19. (a)



Industrial stocks appear to have a higher median rate of return.

(b) (1) Treat each sample as a simple random sample, (2) each sample is obtained independently of the other, (3) each sample size is large, and (4) each sample size is small relative to the size of its population. The response variable is quantitative and there are two groups to compare.

- (c)** $H_0: \mu_{cc} = \mu_I$ vs. $H_1: \mu_{cc} \neq \mu_I$. Classical approach:
 $t_0 = -2.528 < -t_{0.025} = -1.984$ (100 df). Reject H_0 . P -value approach: $0.01 < P\text{-value} < 0.02$ [Tech: $P\text{-value} = 0.0123$]. Reject H_0 . There is sufficient evidence to conclude that the mean rate of return on consumer cyclical stocks differs from the mean rate of return of industrial stocks.
- (d)** Lower bound: 1.684 [Tech: 1.719], Upper bound : 13.976 [Tech: 13.942]

We are 95% confident the mean difference in rate of return of industrial stocks versus consumer cyclical stocks is between 1.684% and 13.976%. This suggests that the one-year rate of return on industrial stocks was higher than consumer cyclical stocks by somewhere between 1.684% and 13.976% for this time period.

- 21.** Lower bound: 0.71 [Tech: 0.72], upper bound: 2.33 [Tech: 2.32]. We can be 90% confident that the mean difference in daily leisure time between adults without children and those with children is between 0.71 and 2.33 hours. Since the confidence interval does not include zero, we can conclude that there is a significant difference in the leisure time of adults without children and those with children.

- 23. (a)** Completely randomized design
(b) Final exam score; online versus traditional homework
(c) Teacher; location; time; text; syllabus; tests
(d) The assumption is that the students “randomly” enrolled in the course.

(e) $H_0: \mu_F = \mu_S$; $H_1: \mu_F > \mu_S$. Classical:
 $t_0 = 1.795 > t_{0.05} = 1.711$; reject H_0 . $0.025 < P\text{-value} < 0.05$ [Tech: $P\text{-value} = 0.0395$] $< \alpha = 0.05$; reject H_0 . There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the final exam scores in the fall semester were higher than the final exam scores in the spring semester. It would appear to be the case that the online homework system helps in raising final exam scores.

(f) One factor is the fact that the weather is pretty lousy at the end of the fall semester, but pretty nice at the end of the spring semester. If “spring fever” kicked in for the spring semester students, then they probably studied less for the final exam.

- 25.** The sampling method is independent (the freshman cannot be matched to the corresponding seniors). Therefore, the inferential method that may be applied is a two-sample t -test. This comparison, however, has major shortcomings. The goal of the CLA+ is to measure gains in critical thinking, analytical reasoning, and so on, as a result of four years of college. The logical design to measure this as a matched-pairs design where the exam is administered before and after college to the same student.

- 27.** The degrees of freedom obtained from Formula (2) are larger than the smaller of $n_1 - 1$ or $n_2 - 1$, and t_α decreases as the degrees of freedom increase. The larger the critical value is, the harder it is to reject the null hypothesis.

11.4 Assess Your Understanding (page 514)

- 1.** $H_0: p_1 = p_2$ vs. $H_1: p_1 \neq p_2$. All requirements to conduct the test are satisfied. Classical: $z_0 = -1.18$ is between $-z_{0.005} = -2.575$ and $z_{0.005} = 2.575$; do not reject the null hypothesis. $P\text{-value} = 0.238$ [Tech: 0.242] $> \alpha = 0.01$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that there is a difference in the proportions.
- 3.** $H_0: p_1 = p_2$ vs. $H_1: p_1 < p_2$. All requirements to conduct the test are satisfied. Classical: $z_0 = -1.84 < -z_{0.05} = -1.645$; reject the null hypothesis. $P\text{-value} = 0.0329$ [Tech: 0.0335] $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that proportion 1 is less than proportion 2.

- 5.** $H_0: \mu_d = 0$ vs. $H_1: \mu_d > 0$; $d_i = Y_i - X_i$. Classical:
 $t_0 = 1.324 < t_{0.05} = 2.132$; do not reject the null hypothesis.
 $0.10 < P\text{-value} < 0.15$ [Tech: $P\text{-value} = 0.128$] $> \alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence at the

$\alpha = 0.05$ level of significance to conclude that the treatment is effective.

- 7. (a)** There are a few very large collision claims relative to the majority of claims.
(b) $H_0: \mu_{30-59} = \mu_{20-24}$ vs. $H_1: \mu_{30-59} < \mu_{20-24}$. All requirements to conduct the test are satisfied; use $\alpha = 0.05$. Classical:
 $t_0 = -1.890 < -t_{0.05} = -1.685$; reject the null hypothesis.
 $0.025 < P\text{-value} < 0.05$ [Tech: $P\text{-value} = 0.0313$] $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the mean collision claim of a 30- to 59-year-old is less than the mean claim of a 20- to 24-year-old. Given that 20- to 24-year-olds tend to claim more for each accident, it makes sense to charge them more for coverage.
- 9. (a)** The response variable is age.
(b) The sampling method is dependent because each husband is matched with the wife.
(c) To estimate the mean difference, construct a 95% confidence interval for the mean difference of the data. First, verify the differences are approximately normal by drawing a normal probability plot and verify the differences contain no outliers by drawing a boxplot. Compute the differences as “Husband minus Wife.” Lower bound: -0.5; upper bound: 6.0. We are 95% confident the mean difference in age between a husband and wife is between -0.5 year and 6.0 years.
- 11.** This requires a single sample t -test for a population mean.
 $H_0: \mu = 92.0$ mph versus $H_1: \mu > 92.0$ mph. The sample size is small. A normal probability plot indicates the data could come from a population that is normally distributed. A boxplot indicates there are no outliers. Classical approach: $t_0 = 1.086 < t_{0.05} = 1.771$. Do not reject the null hypothesis. P -value approach: $0.10 < P\text{-value} < 0.15$ [Tech: $P\text{-value} = 0.1485$]. Do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude the pitcher has a fastball that exceeds 92.0 mph.
- 13.** $H_0: p = 0.5$ vs. $H_1: p > 0.5$. All requirements to conduct the test are satisfied. Classical approach: $z_0 = 9.217 > z_{0.05} = 1.645$; reject the null hypothesis. P -value approach:
 $P\text{-value} < 0.0002$ [Tech: $P\text{-value} < 0.0001$] $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to suggest that a quick 1-second view of a black and white photo represents enough information to judge the winner of an election.
- 15.** $H_0: p_{>100K} = p_{<100K}$ vs. $H_1: p_{>100K} \neq p_{<100K}$. Lower bound: 0.019 [Tech: 0.020], upper bound: 0.097. Because the confidence interval does not include 0, there is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that there is a difference in the proportions. It seems that a higher proportion of individuals who earn over \$100,000 per year feel it is morally wrong for unwed women to have children.
- 17. (a)** The response variable is score on the quiz. The explanatory variable is whether texting was required or the cell phone was turned off (no texting allowed).
(b) Students were randomly given instructions at the beginning of the class. Both groups received the same lecture.
(c) The sampling method is independent.
(d) $H_0: \mu_{text} = \mu_{cell\ off}$ versus $H_1: \mu_{text} < \mu_{cell\ off}$; the sample sizes are large enough. Assume this study applies to all students, so the sample size is less than 5% of the population size.
Classical approach: $t_0 = -6.141 < -t_{0.05} = -1.697$. Reject the null hypothesis. P -value approach: $P\text{-value} < 0.0001$. There is sufficient evidence to suggest the mean score in the texting group is less than the mean score in the cellphone off group. Apparently, students cannot multitask.
(e) Lower bound: -20.243 [Tech: -20.17], upper bound: -11.477 [Tech: -11.55]. We are 90% confident the texting group scored between 11.477 points and 20.243 points worse, on average, than the cell-phone off group.

19.

Variable	Hypothesis	P-value (Using Technology)	Conclusion
Age	$H_0: \mu_{\text{Readmit}} = \mu_{\text{Non-readmit}}$	0.0006	Reject the null hypothesis
	$H_1: \mu_{\text{Readmit}} > \mu_{\text{Non-readmit}}$		
Length of stay	$H_0: \mu_{\text{Readmit}} = \mu_{\text{Non-readmit}}$	< 0.0001	Reject the null hypothesis
	$H_1: \mu_{\text{Readmit}} > \mu_{\text{Non-readmit}}$		
Admission in previous calendar year	$H_0: p_{\text{Readmit}} = p_{\text{Non-readmit}}$	< 0.0001	Reject the null hypothesis
	$H_1: p_{\text{Readmit}} > p_{\text{Non-readmit}}$		
Season	$H_0: p_{\text{Readmit}} = p_{\text{Non-readmit}}$	0.0001	Reject the null hypothesis
	$H_1: p_{\text{Readmit}} > p_{\text{Non-readmit}}$		
Floor	$H_0: p_{\text{Readmit}} = p_{\text{Non-readmit}}$	0.3959	Do not reject the null hypothesis
	$H_1: p_{\text{Readmit}} > p_{\text{Non-readmit}}$		

From the analysis, the readmits were older and had a longer length of stay. A higher proportion of readmits were admitted in the previous calendar year and a higher proportion of readmits were discharged in the winter. The proportion of readmits who were on the cardiac floor is not significantly higher than the proportion of non-readmits who were on the cardiac floor.

- 21. (a)** The response variable is sale, or not; the explanatory variable is web page design.
(b) $H_0: p_1 = p_{II}$ vs. $H_1: p_1 \neq p_{II}$; Classical approach:
 $-1.96 = -z_{0.025} < z_0 = -0.91 < z_{0.025} = 1.96$; P-value approach:
 $P\text{-value} = 0.3628$ [Tech: $P\text{-value} = 0.3629$]; Do not reject H_0 .
 There is not sufficient evidence to conclude the proportion of sales in the two designs differs.

- 23.** The sample mean difference (computing “With – Without”) is the same for both samples: 0.078 meters per second. The standard error treating the data as an independent sample is

$$\sigma_{\bar{x}_W - \bar{x}_{WO}} = \sqrt{\frac{0.136^2}{12} + \frac{0.141^2}{12}} = 0.057$$

meter per second. The standard error for the dependent sample $s_d = 0.006$ meter per second. The standard error for the matched-pairs (dependent) data is smaller. By pairing the data, quite a bit of variability in speed was reduced. Accounting for other variables (such as swimmer) reduces variability.

- 25.** Hypothesis test for two proportions, independent sample (using a completely randomized design)
27. Confidence interval for a single mean
29. Two sample t -test of independent means
31. Matched-pairs design. Analyze using t -test for a dependent sample
33. Hypothesis test of two independent proportions

Chapter 11 Review Exercises (page 519)

1. Dependent; quantitative
 2. Independent; quantitative

3. (a)

Observation	1	2	3	4	5	6
$d_i = X_i - Y_i$	-0.7	0.6	0	-0.1	-0.4	-0.4

(b) $\bar{d} = -0.167$; $s_d = 0.450$

- (c)** Hypotheses: $H_0: \mu_d = 0$, $H_1: \mu_d < 0$. Classical approach:
 $t_0 = -0.909 > -t_{0.05} = -2.015$; do not reject the null

hypothesis. P-value approach: $0.20 < P\text{-value} < 0.25$ [Tech: $P\text{-value} = 0.2030$] $> \alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence to conclude that the mean difference is less than zero.

(d) Lower bound: -0.79, upper bound: 0.45

- 4. (a)** Hypotheses: $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$. Classical approach: $t_0 = 2.290 > t_{0.05} = 1.895$; reject the null hypothesis. P-value approach: $0.05 < P\text{-value} < 0.10$ [Tech: $P\text{-value} = 0.0351$] $< \alpha = 0.10$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.1$ level of significance to conclude that $\mu_1 \neq \mu_2$.

(b) Lower bound: 0.73 [Tech: 1.01], upper bound: 7.67 [Tech: 7.39]

- 5.** Hypotheses: $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 > \mu_2$. Classical approach: $t_0 = 1.472 < t_{0.01} = 2.423$; do not reject the null hypothesis. P-value approach: $0.05 < P\text{-value} < 0.10$ [Tech: $P\text{-value} = 0.0726$] $> \alpha = 0.01$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.01$ level of significance to conclude that the mean of population 1 is larger than the mean of population 2.

6. $H_0: p_1 = p_2$, $H_1: p_1 \neq p_2$. Classical: $z_0 = -1.68$ is between $-z_{0.025} = -1.96$ and $z_{0.025} = 1.96$; do not reject the null hypothesis. P-value = 0.0930 [Tech: 0.0895] $> \alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the proportion in population 1 is different from the proportion in population 2.

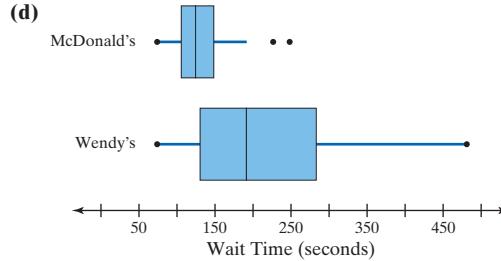
- 7. (a)** The sampling method is dependent because the same individual is used for both measurements.

(b) Hypotheses: $H_0: \mu_d = 0$, $H_1: \mu_d \neq 0$. Classical approach: $t_0 = 0.512$ is between $-t_{0.025} = -2.262$ and $t_{0.025} = 2.262$; do not reject the null hypothesis. P-value approach: $P\text{-value} > 0.50$ [Tech: $P\text{-value} = 0.6209$] $> \alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that arm span is different from height. The sample evidence does not contradict the belief that arm span and height are the same.

- 8. (a)** The sampling method is independent since the cars selected for the McDonald's sample had no bearing on the cars chosen for the Wendy's sample.

(b) Time in drive-through, which is quantitative.

(c) Hypotheses: $H_0: \mu_{\text{McD}} = \mu_W$, $H_1: \mu_{\text{McD}} \neq \mu_W$. Classical approach: $t_0 = -4.059 < -t_{0.05} = -1.706$; reject the null hypothesis. P-value approach: $P\text{-value} < 0.0005$ [Tech: $P\text{-value} = 0.0003$] $< \alpha = 0.10$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.10$ level of significance to conclude that the wait times in the drive-throughs of the two restaurants differ.

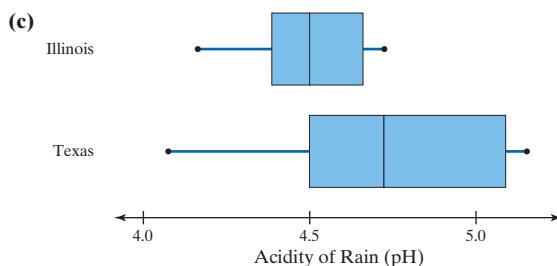


Based on the boxplots, it would appear to be the case that the wait time at McDonald's is less than the wait time at Wendy's.

- 9. (a)** This is a completely randomized design with two treatments: placebo and 5 mg of Actonel. The response variable is whether the subject has a bone fracture over the course of one year, or not. It is qualitative with two outcomes.

(b) A double-blind experiment is one in which neither the subject nor the individual administering the treatment knows which group (experimental or control) the subject is in.

- (c) Each sample is the result of a randomized experiment; $n_1\hat{p}_1(1 - \hat{p}_1) = 26 \geq 10$ and $n_2\hat{p}_2(1 - \hat{p}_2) = 45 \geq 10$; and each sample is less than 5% of the population size. Hypotheses: $H_0: p_{\text{exp}} = p_{\text{control}}$, $H_1: p_{\text{exp}} < p_{\text{control}}$. Classical approach: $z_0 = -2.68 < -z_{0.01} = -2.33$; reject the null hypothesis. P -value approach: P -value = 0.0037 [Tech: 0.0033] $< \alpha = 0.01$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.01$ level of significance to conclude that a lower proportion of women in the experimental group experienced a bone fracture than in the control group.
- (d) Lower bound: -0.06 , upper bound: -0.01 . We are 95% confident that the difference in the proportion of women who experienced a bone fracture between the experimental and control group is between -0.06 and -0.01 .
10. (a) $n = n_1 = n_2 = 2136$
 (b) $n = n_1 = n_2 = 3383$
11. Lower bound: -1.37 , upper bound: 2.17 . Since the interval includes zero, we conclude that there is not sufficient evidence at the $\alpha = 0.05$ level of significance to reject the claim that arm span and height are equal.
12. Lower bound: -128.868 [Tech: -128.42]; upper bound: -42.218 [Tech: -42.67]. A marketing campaign could be initiated by McDonald's touting the fact that wait times are up to 2 minutes less at McDonald's.
- Chapter 11 Test (page 520)**
1. Independent
 2. Dependent
 3. (a)
- | Observation | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-------------------|-----|------|-----|-----|------|------|------|
| $d_i = X_i - Y_i$ | 0.2 | -0.5 | 0.2 | 0.6 | -0.6 | -0.5 | -0.8 |
- (b) $\bar{d} = -0.2$; $s_d = 0.526$
 (c) Hypotheses: $H_0: \mu_d = 0$, $H_1: \mu_d \neq 0$. Classical approach: $t_0 = -1.006$ is between $-t_{0.005} = -3.707$ and $t_{0.005} = 3.707$; do not reject the null hypothesis. $0.30 < P\text{-value} < 0.40$ [Tech: $P\text{-value} = 0.3532$] $> \alpha = 0.01$; do not reject the null hypothesis. There is not sufficient evidence to conclude that the mean difference is different from zero.
 (d) Lower bound: -0.69 , upper bound: 0.29
4. (a) Hypotheses: $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$. Classical approach: $t_0 = -2.054 < -t_{0.05} = -1.714$. Reject the null hypothesis. $0.05 < P\text{-value} < 0.10$ [Tech: $P\text{-value} = 0.0464$] $< \alpha = 0.1$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.1$ level of significance to conclude that the means are different.
 (b) Lower bound: -12.44 [Tech: -12.3], upper bound: 0.04 [Tech: -0.10]
5. Hypotheses: $H_0: \mu_1 = \mu_2$, $H_1: \mu_1 < \mu_2$. Classical approach: $t_0 = -1.357 > -t_{0.05} = -1.895$; do not reject the null hypothesis. $0.10 < P\text{-value} < 0.15$ [Tech: $P\text{-value} = 0.0959$] $> \alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the mean of population 1 is less than the mean of population 2.
6. $H_0: p_1 = p_2$ vs. $H_1: p_1 < p_2$ Classical:
 $z_0 = -0.80 > -z_{0.05} = -1.645$; do not reject the null hypothesis. $P\text{-value} = 0.2119$ [Tech: 0.2124] $> \alpha = 0.05$; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the proportion in population 1 is less than the proportion in population 2.
7. (a) The sampling method is independent since the dates selected for the Texas sample have no bearing on the dates chosen for the Illinois sample.
 (b) Both samples must come from populations that are normally distributed.



The boxplots indicate the Chicago rain has a lower pH than Houston rain.

- (d) Hypotheses: $H_0: \mu_{\text{Texas}} = \mu_{\text{Illinois}}$, $H_1: \mu_{\text{Texas}} \neq \mu_{\text{Illinois}}$. Classical approach: $t_0 = 2.276 > t_{0.025} = 2.201$; reject the null hypothesis. $0.04 < P\text{-value} < 0.05$ [Tech: $P\text{-value} = 0.0387$] $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the acidity of the rain near Houston is different from its acidity near Chicago.
8. (a) The response variable is the student's GPA. The explanatory variable is whether the student has a sleep disorder or not.
 (b) Hypotheses: $H_0: \mu_{\text{sleep disorder}} = \mu_{\text{no sleep disorder}}$, $H_1: \mu_{\text{sleep disorder}} < \mu_{\text{no sleep disorder}}$. Classical approach: $t_0 = -3.784$ is less than $-t_{0.05} = -1.660$; reject H_0 . P -value approach: $P\text{-value} < 0.0005$ [Tech: $P\text{-value} < 0.0001$] $< \alpha = 0.05$; reject H_0 . There is sufficient evidence at the $\alpha = 0.05$ level of significance to suggest that sleep disorders adversely affect a student's GPA.
9. Hypotheses: $H_0: \mu_d = 0$ vs. $H_1: \mu_d < 0$; Classical approach: $t_0 = -1.927 < -t_{0.10} = -1.440$; reject the null hypothesis. P -value approach: $0.05 < P\text{-value} < 0.10$ [Tech: $P\text{-value} = 0.0511$] $< \alpha = 0.1$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.1$ level of significance to suggest that the repair cost for the car is higher.
10. (a) Completely randomized design.
 (b) Whether the subject gets dry mouth or not.
 (c) Each sample is the result of a randomized experiment. $n_1\hat{p}_1(1 - \hat{p}_1) = 66 \geq 10$ and $n_2\hat{p}_2(1 - \hat{p}_2) = 31 \geq 10$; and each sample is less than 5% of the population size.
 Hypotheses: $H_0: p_{\text{exp}} = p_{\text{control}}$, $H_1: p_{\text{exp}} > p_{\text{control}}$. Classical approach: $z_0 = 2.20 > z_{0.05} = 1.645$; reject the null hypothesis. $P\text{-value} = 0.0139$ [Tech: 0.0136] $< \alpha = 0.05$; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that a higher proportion of subjects in the experimental group experienced a dry mouth than in the control group.
11. $H_0: p_M = p_F$, $H_1: p_M \neq p_F$. Lower bound: 0.051 , upper bound: 0.089 . Because the confidence interval does not include 0, we reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.1$ level of significance to conclude that the proportion of males and females for which hypnotism led to quitting smoking is different.
12. (a) $n = n_1 = n_2 = 762$
 (b) $n = n_1 = n_2 = 1201$
13. $H_0: \mu_p = \mu_n$, $H_1: \mu_p \neq \mu_n$. Lower bound: 1.03 kg, upper bound: 1.17 kg. Because the confidence interval does not contain 0, we reject the null hypothesis. There is sufficient evidence to conclude that Naltrexone is effective in preventing weight gain among individuals who quit smoking. Answers will vary regarding practical significance, but one must ask, "Do I want to take a drug so that I can keep about 1 kg off?" Probably not.

CHAPTER 12 Inference on Categorical Data

12.1 Assess Your Understanding (page 533)

1. True
3. expected counts; np_i

5.

p_i	0.2	0.1	0.45	0.25
Expected counts	100	50	225	125

7. (a) $\chi^2_0 = 2.72$

(b) $df = 3$

(c) $\chi^2_{0.05} = 7.815$

(d) Do not reject H_0 , since $\chi^2_0 < \chi^2_{0.05}$. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that at least one of the proportions is different from the others.

9. (a) $\chi^2_0 = 12.56$

(b) $df = 4$

(c) $\chi^2_{0.05} = 9.488$

(d) Reject H_0 , since $\chi^2_0 > \chi^2_{0.05}$. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that X is not binomial with $n = 4, p = 0.8$.

11. H_0 : the distribution of colors is as stated by M&M; H_1 : the distribution is different from that stated by M&M. Classical approach: $\chi^2_0 = 6.744 < \chi^2_{0.05} = 11.070$; do not reject the null hypothesis. P -value approach: P -value $> 0.10 > \alpha = 0.05$ [Tech: P -value = 0.2404]; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the distribution of candies in a bag of M&Ms is different from 13% brown, 14% yellow, 13% red, 20% orange, 24% blue, and 16% green.

13. (a) Answers will vary depending on the desired probability of making a Type I error. One possible choice: $\alpha = 0.01$.

(b) H_0 : the digits follow Benford's Law; H_1 : the digits do not follow Benford's Law. Classical approach: $\chi^2_0 = 21.693$; compare χ^2_0 to your chosen χ^2_α . P -value approach: P -value is between 0.01 and 0.005 [Tech: P -value = 0.0055].

(c) Answers will vary. One possible answer: Since P -value < 0.01 , there is sufficient evidence to conclude that Benford's Law is not followed and the employee is guilty of embezzlement.

15. (a) Classical approach: $\chi^2_0 = 121.367 > \chi^2_{0.05} = 9.488$; reject the null hypothesis. P -value approach: P -value $< 0.005 < \alpha = 0.05$ [Tech: P -value < 0.0001]; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the distribution of fatal injuries for riders not wearing a helmet does not follow the distribution for all riders.

(b)

Location of Injury	Multiple Locations	Abdomen/Lumbar/Spine			
		Head	Neck	Thorax	Lumbar/Spine
Observed	1036	864	38	83	47
Expected	1178.76	641.08	62.04	124.08	62.04

We notice that the observed count for head injuries is much higher than expected, while the observed counts for all the other categories are lower. We might conclude that motorcycle fatalities from head injuries occur more frequently for riders not wearing a helmet.

17. (a) Group 1: 84; Group 2: 84; Group 3: 84; Group 4: 81. Classical approach: $\chi^2_0 = 0.084$ [Tech: 0.081] $< \chi^2_{0.05} = 7.815$; do not reject the null hypothesis. P -value approach: P -value $> 0.99 > \alpha = 0.05$ [Tech: P -value = 0.994]; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that there are differences among the groups in attendance patterns.

(b) Group 1: 84; Group 2: 81; Group 3: 78; Group 4: 76. Classical approach: $\chi^2_0 = 0.463$ [Tech: 0.461] $< \chi^2_{0.05} = 7.815$; do not reject the null hypothesis. P -value approach: P -value $> 0.90 > \alpha = 0.05$ [Tech: P -value = 0.9274]; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that there are differences among the groups in attendance patterns. It is curious that the farther a group's original position is located from the front of the room, the more the attendance rate for the group decreases.

(c) Group 1: 20; Group 2: 20; Group 3: 20; Group 4: 20. Classical approach: $\chi^2_0 = 2.60 < \chi^2_{0.05} = 7.815$; do not reject the null

hypothesis. P -value approach: P -value $> 0.10 > \alpha = 0.05$

[Tech: P -value = 0.4575]; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that there is a significant difference in the number of students in the top 20% of the class by group.

(d) Though not statistically significant, the group located in the front had both better attendance and a larger number of students in the top 20%. Choose the front.

19. Classical approach: $\chi^2_0 = 15.122 > \chi^2_{0.01} = 11.345$; reject the null hypothesis. P -value approach: P -value < 0.005 [Tech: P -value = 0.0017]; reject the null hypothesis. There is sufficient evidence to suggest that hockey players' birthdates are not evenly distributed throughout the year. More than expected are born in the early part of the year.

21. (a) $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$

(b) $H_0: p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$ versus H_1 : At least one of the proportions differs from the others; Classical approach: $\chi^2_0 = 8.593 < \chi^2_{0.05} = 11.070$; P -value approach: P -value > 0.10 [Tech: P -value = 0.1264] > 0.05 . Do not reject H_0 . There is not sufficient evidence to conclude the first roll frequency differs from $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$.

(c) $H_0: p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$ versus H_1 : At least one of the proportions differs from the others; Classical approach: $\chi^2_0 = 889.68 > \chi^2_{0.05} = 11.070$; P -value approach: P -value < 0.0005 [Tech: P -value < 0.0001] < 0.05 . Reject H_0 . There is sufficient evidence to conclude the first roll frequency differs from $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$.

(d) $H_0: p_1 = 0.083, p_2 = 0.139, p_3 = 0.194, p_4 = 0.25, p_5 = 0.306, p_6 = 0.028$ versus H_1 : The distribution does not follow the "better choice" distribution; Classical approach: $\chi^2_0 = 10.891 < \chi^2_{0.05} = 11.070$; P -value approach: $0.05 < P$ -value < 0.10 [Tech: P -value = 0.0536] > 0.05 . Do not reject H_0 . There is not sufficient evidence to conclude the first roll frequency differs from $p_1 = 0.083, p_2 = 0.139, p_3 = 0.194, p_4 = 0.25, p_5 = 0.306, p_6 = 0.028$.

(e) Although the sample data did not allow us to reject the statement in the null hypothesis at the 0.05 level of significance, the P -value is low enough to suggest the distribution differs slightly from that in the null hypothesis. For example, we observed 88 subjects who reported rolling a six, and we would expect 72 to roll a six under the better option theory. So, perhaps there are some honest folks in the study.

23. (a) 64 students

Grade Expected Number of Students

A	8.512
B	12.224
C	15.744
D	6.656
F	7.296
W	13.568

(b) Classical approach: $\chi^2_0 = 9.989 < \chi^2_{0.01} = 15.086$; Do not reject H_0 . P -value approach: $0.05 < P$ -value < 0.10 [Tech: P -value = 0.0756]; Do not reject H_0 . There is not sufficient evidence to conclude the grade distribution in the MRP program differs from traditional classes at the $\alpha = 0.01$ level of significance.

(c) It is a big change to adjust an entire curriculum, so we don't want to make a Type I error. That is, we don't want to conclude the grade distribution in the MRP program differs from traditional, when, in fact, it does not differ.

(d) Classical approach: $\chi^2_0 = 19.977 > \chi^2_{0.01} = 15.086$; reject H_0 . P -value approach: P -value < 0.005 [Tech: P -value = 0.0013] $< \alpha = 0.01$; reject H_0 . There is sufficient evidence to conclude the distribution of grades in the MRP program differs from that of the traditional classes. Looking at the observed versus expected values, there is a higher number

of Bs than expected, a lower number of Cs, and a higher number of Ds in the MRP program. With small sample sizes, the evidence against the null hypothesis must be overwhelming to be able to reject the statement in the null hypothesis. So, watch out for studies that suggest there is not significant evidence when the sample size is small.

- 25. (a)** Expected number with low birth weight: 17.04; expected number without low birth weight: 222.96.
(b) $H_0: p = 0.071$ vs. $H_1: p > 0.071$. Classical approach: $\chi^2_0 = 1.554 < \chi^2_{0.05} = 3.841$; do not reject the null hypothesis.
 P-value approach: $P\text{-value} > 0.10 > \alpha = 0.05$
 [Tech: $P\text{-value} = 0.2125$]; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that mothers between the ages of 35 and 39 have a higher percentage of low-birth-weight babies.
(c) $H_0: p = 0.71$ vs. $H_1: p > 0.71$; $np_0(1 - p_0) = 15.83 > 10$. Classical approach: $z_0 = 1.25 < z_{0.05} = 1.645$; do not reject the null hypothesis. P-value approach:
 $P\text{-value} = 0.1056 > \alpha = 0.05$ [Tech: $P\text{-value} = 0.1063$]; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that mothers between the ages of 35 and 39 have a higher percentage of low-birth-weight babies.
- 27. (a)** $\mu = 0.9323$ hits
(b) All expected frequencies will not be greater than or equal to 1. Also, more than 20% (37.5%) of the expected frequencies are less than 5.
(c), (d)
- | x | P(x) | Expected Number of Regions |
|-----------|--------|----------------------------|
| 0 | 0.3936 | 226.714 |
| 1 | 0.3670 | 211.392 |
| 2 | 0.1711 | 98.554 |
| 3 | 0.0532 | 30.643 |
| 4 or more | 0.0151 | 8.698 |
- (e)** Classical approach: $\chi^2_0 = 1.011 < \chi^2_{0.05} = 9.488$; do not reject the null hypothesis. P-value approach:
 $P\text{-value} > 0.90 > \alpha = 0.05$ [Tech: $P\text{-value} = 0.9081$]; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the distribution of rocket hits is different from the Poisson distribution. That is, the rocket hits do appear to be modeled by a Poisson random variable.
- 29. (a)** Quantitative
(b) $\bar{x} = \$41,130.9$, $M = \$41,215$
(c) $s = \$897.8$, IQR = \$1592
(d) The correlation between the raw data and normal scores is $0.991 > 0.939$ (Table VI), so it is reasonable to conclude that the data come from a population is normally distributed.
- (e)**
- Price of a 2019 A4**
-
- (f)** Lower bound: \$40,722.7 [Tech: \$40,772.6], upper bound: \$41,539.1 [Tech: \$41,539.2]; We are 90% confident the mean price of a new 2019 Audi A4 is between \$40,722.7 and \$41,539.1.
(g) Wider because there is more variability in the data. By removing the variability due to car type, the interval becomes more precise.
- 31.** The χ^2 goodness-of-fit tests are always right tailed because the numerator in the test statistic is squared, making every test statistic other than a perfect fit positive. So we are measuring if $\chi^2_0 > \chi^2_\alpha$.

- 33.** $H_0: p_0 = p_1 = \dots = p_9 = 0.1$ versus H_1 : at least one proportion differs from the others. To answer any of these questions, you would need to obtain a random sample of individuals from your population of interest. Be careful that the data collection allows for equal representation of each age. Run a chi-square test for goodness-of-fit by comparing expected counts to those actually observed. For example, for the marathon study, you could determine the age of all the registrants and count how many folks have each last digit.

12.2 Assess Your Understanding (page 549)

- True
- (a)** $\chi^2_0 = 1.701$
(b) Classical approach: $\chi^2_0 = 1.698 < \chi^2_{0.05} = 5.991$; do not reject the null hypothesis. P-value approach:
 $P\text{-value} > 0.10 > \alpha = 0.05$ [Tech: $P\text{-value} = 0.4278$]; do not reject the null hypothesis. There is evidence at the $\alpha = 0.05$ level of significance to conclude that X and Y are independent. We conclude that X and Y are not related.
- Classical approach: $\chi^2_0 = 1.989 < \chi^2_{0.01} = 9.210$; do not reject the null hypothesis. P-value approach: $P\text{-value} > 0.10 > \alpha = 0.01$ [Tech: $P\text{-value} = 0.3699$]; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.01$ level of significance to conclude that at least one of the proportions is different from the others.

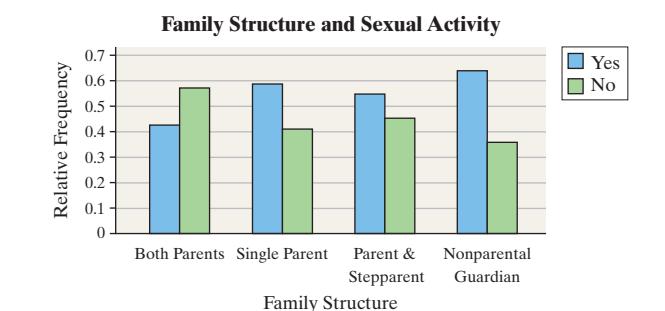
- 7. (a)**

Had Sexual Intercourse?	Both Biological/			
	Adoptive Parents	Single Parent	Parent and Stepparent	Nonparental Guardian
Yes	78.553	52.368	41.895	26.184
No	71.447	47.632	38.105	23.816

- (b)** (1) All expected frequencies are greater than or equal to 1, and (2) no more than 20% of the expected frequencies are less than 5.
(c) $\chi^2_0 = 10.357$
(d) H_0 : family structure and sexual activity are independent; H_1 : family structure and sexual activity are not independent. Classical approach: $\chi^2_0 = 10.357 > \chi^2_{0.05} = 7.815$; reject the null hypothesis. P-value approach:
 $P\text{-value} < 0.025 < \alpha = 0.05$ [Tech: $P\text{-value} = 0.0158$]; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that sexual activity and family structure are associated.
(e) The biggest difference between observed and expected occurs under the family structure in which both parents are present. Fewer females were sexually active than was expected when both parents were present. This means that having both parents present seems to have an effect on whether the child is sexually active.

- (f)**

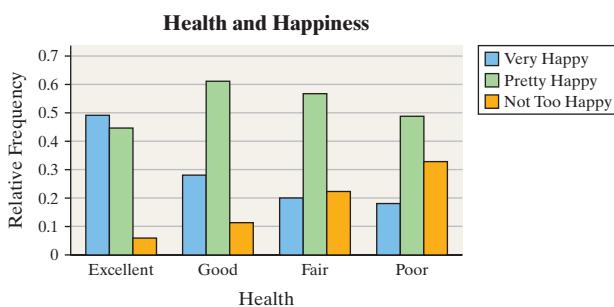
Had Sexual Intercourse?	Both Biological/			
	Adoptive Parents	Single Parent	Parent and Stepparent	Nonparental Guardian
Yes	0.427	0.59	0.55	0.64
No	0.573	0.41	0.45	0.36



- 9.** (a) H_0 : health and happiness are independent; H_1 : health and happiness are not independent. Classical approach: $\chi^2_0 = 182.174 > \chi^2_{0.05} = 12.592$; reject the null hypothesis. P -value approach: P -value < 0.005 < $\alpha = 0.05$ [Tech: P -value < 0.0001]; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that happiness and health are dependent. That is, happiness and health are related to each other.

(b)

Happiness	Health			
	Excellent	Good	Fair	Poor
Very Happy	0.492	0.280	0.202	0.183
Pretty Happy	0.448	0.609	0.570	0.486
Not Too Happy	0.060	0.111	0.227	0.330



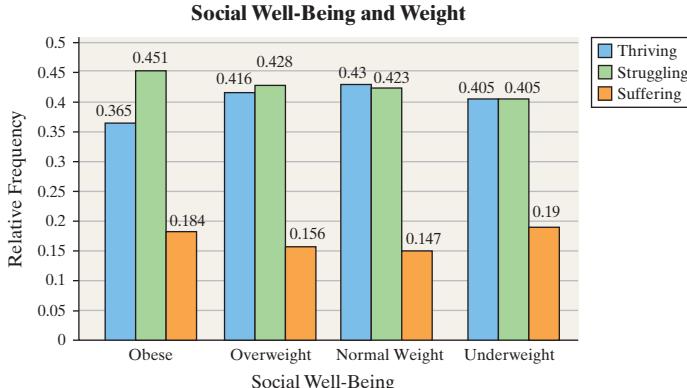
- (c) The proportion of individuals who are “very happy” is much higher for individuals in “excellent” health than any other health category. Further, the proportion of individuals who are “not too happy” is much lower for individuals in “excellent” health compared to the other health categories. Put simply, the level of happiness seems to decline as health status declines.

- 11.** (a) The data represent the measurement of two variables (weight classification and social well-being) on each individual in the study. Because two variables are measured on a single individual, use a chi-square test for independence.

- (b) Classical: $\chi^2_0 = 7.167 < \chi^2_{0.05} = 12.592$; do not reject H_0 . P -value approach: P -value > 0.10 [Tech: P -value = 0.306]; do not reject H_0 .

There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude there is an association between social well-being and weight classification.

(c)



- (d) Answers may vary. There is not enough sample evidence to suggest that one's social well-being is associated with one's weight classification. However, it is worth noting some differences in the relative frequencies. For example, both obese and underweight individuals have a higher relative frequency for “suffering.”

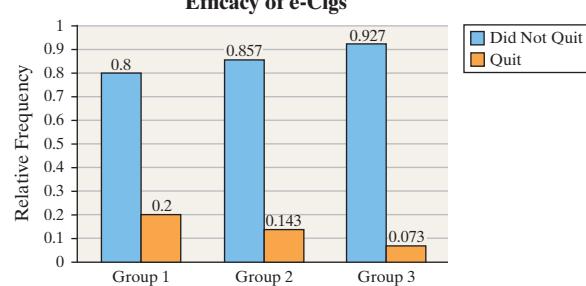
- 13.** (a) A completely randomized design with three levels of treatment.
(b) The response variable is whether the subject abstained from cigarette smoking, or not. It is qualitative with two possible outcomes.
(c) Current smokers

- (d) $H_0: p_1 = p_2 = p_3$

H_1 : At least one proportion differs from the others

- (e) Classical approach: $\chi^2_0 = 3.960 < \chi^2_{0.05} = 5.991$; do not reject H_0 . P -value approach: P -value > 0.1 [Tech: P -value = 0.1381] > $\alpha = 0.05$; do not reject H_0 .

(f)



- (g) There is not sufficient evidence at the 0.05 level of significance to suggest that the proportion of individuals who abstain from cigarette smoking in the three groups differs.

- 15.** (a) Because there are three distinct populations that are being surveyed (Democrat, Republican, Independent) and the response variable is qualitative with two outcomes (positive/negative), we analyze the data using homogeneity of proportions.

- (b) $H_0: p_D = p_I = p_R; H_1$: at least one proportion differs.

Classical approach: $\chi^2_0 = 96.733 > \chi^2_{0.05} = 5.991$; reject the null hypothesis. P -value approach: P -value < 0.005 [Tech: P -value < 0.0001]; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance that a different proportion of individuals within each political affiliation reacts positively to the word socialism.

(c)

	Democrat	Independent	Republican
Positive	0.4409	0.2599	0.1501
Negative	0.5591	0.7401	0.8499
Total	1	1	1

- (d) Independents and Republicans are far more likely to react negatively to the word *socialism* than Democrats are. However, it is important to note that a majority of Democrats in the sample did have a negative reaction, so the word *socialism* has a negative connotation among all groups.

17. (a)

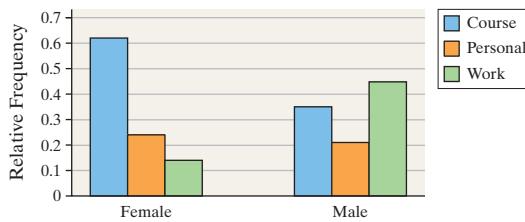
	Course	Personal	Work
Female	13	5	3
Male	10	6	13

- (b) Classical approach: $\chi^2_0 = 5.595 > \chi^2_{0.10} = 4.605$; reject the null hypothesis. P -value approach: P -value < 0.10 = α [Tech: P -value = 0.0609]; reject the null hypothesis. The evidence suggests a relation between gender and drop reason. Females are more likely to drop because of the course, while males are more likely to drop because of work.

(c)

	Reason		
	Course	Personal	Work
Female	0.619	0.238	0.143
Male	0.345	0.207	0.448

Why Did You Drop?



19. There are three different populations to study: (1) all capital letters, (2) all lower case letters, and (3) upper and lower case accurately. Obtain a random sample of applications from each population. Determine whether the loan corresponding to the application resulted in default. Conduct a hypothesis test for homogeneity of proportions to determine if the sample data suggest at least one population differs from the others.

21. (a) Completely randomized design

(b) Course is the treatment and it has three levels.

(c) Pass, or not; qualitative with two outcomes

(d) $H_0: p_1 = p_2 = p_3$ versus $H_1:$ at least one proportion differs from the others. Classical approach: $\chi^2_0 = 14.923 > \chi^2_{0.05} = 5.991$. P -value approach: P -value < 0.005 [Tech: P -value = 0.0006].

Reject H_0 . There is sufficient evidence at the 0.05 level of significance to conclude at least one proportion differs from the others. Based on the proportion of students passing each course, it would appear the proportion passing Statistics (0.561) is higher than the proportion passing Elementary Algebra (Course 1).

(e) Variables controlled and fixed: Semester course offered; university course offered at (CUNY). Time of day was not controlled (as far as we know). Instructor is not mentioned as being controlled.

(f) Teacher could be a confounding variable especially if Teacher C is a much better instructor than Teacher A. We would not know if the difference in pass rate was due to the course structure or the teacher.

(g) $H_0: p = 0.37$ versus $H_1: p > 0.37$. Classical approach:

$z_0 = 2.47 > z_{0.05} = 1.645$. P -value approach:

P -value = 0.0068 [Tech: 0.0066]. Reject H_0 . There is sufficient evidence to conclude the students in Course 2 had a higher pass rate than the historical Elementary Algebra pass rate.

23. (a) $H_0: \mu_C = \mu_T$ versus $H_1: \mu_C \neq \mu_T$. Classical approach: $t_0 = 2.197 > t_{0.025} = 1.994$ (using 70 df). P -value approach: $0.02 < P$ -value < 0.04 [Tech: P -value = 0.0298]. Reject H_0 . There is sufficient evidence to conclude there is a difference between the corequisite gain and traditional gain. We are 95% confident the gain in the corequisite course is between 0.96 [Tech: 1.04] and 19.84 [Tech: 19.76] points higher.

(b) $H_0:$ Course grade is independent of course type versus $H_1:$ Course grade is associated with course type. Classical approach: $\chi^2_0 = 12.677 > \chi^2_{0.05} = 11.070$ (using 5 degrees of freedom at the 0.05 level of significance). P -value approach: $0.025 < P$ -value < 0.05 [Tech: P -value = 0.0266]. Reject H_0 . There is sufficient evidence to conclude there is an association between grade and course type. Comparing expected counts to observed counts, it appears more students earn an A in the traditional course, but a higher number of Bs and Cs is observed than expected in the corequisite course.

(c) $\hat{p}_C = 0.896; \hat{p}_T = 0.837$

(d) $H_0: p_C = p_T$ versus $H_1: p_C \neq p_T$. Classical approach: $-z_{0.025} = -1.96 < z_0 = 1.10 < z_{0.025} = 1.96$. P -value approach: P -value = 0.2714 [Tech: 0.2719]. Do not reject H_0 . There is not sufficient evidence to conclude the pass rates between the two courses are different.

25. Answers may vary, but differences should include chi-square test for independence compares two characteristics from a single population, whereas the chi-square test for homogeneity of proportions compares a single characteristic from two (or more) populations. Similarities should include the procedures of the two tests and the assumptions of the two tests are the same.

12.3 Assess Your Understanding (page 565)

1. 82.7

3. 0; σ

5. (a) $\beta_0 \approx b_0 = -2.3256, \beta_1 \approx b_1 = 2.0233$

(b) $s_e = 0.5134$ is the point estimate for σ .

(c) $s_{b_1} = 0.1238$

(d) Because the P -value < 0.001 < $\alpha = 0.05$ (or $t_0 = 16.344$ [Tech: 16.344] > $t_{0.025} = 3.182$), we reject the null hypothesis and conclude that a linear relation exists between x and y .

7. (a) $\beta_0 \approx b_0 = 1.200, \beta_1 \approx b_1 = 2.200$

(b) $s_e = 0.8944$ is the point estimate for σ .

(c) $s_{b_1} = 0.2828$

(d) Because the P -value = 0.0044 < $\alpha = 0.05$ (or $t_0 = 7.778$ [Tech: 7.779] > $t_{0.025} = 3.182$), we reject the null hypothesis and conclude that a linear relation exists between x and y .

9. (a) $\beta_0 \approx b_0 = 116.600, \beta_1 \approx b_1 = -0.7200$

(b) $s_e = 3.2863$ is the point estimate for σ .

(c) $s_{b_1} = 0.1039$

(d) Because the P -value = 0.0062 < $\alpha = 0.05$ (or $t_0 = -6.929$ [Tech: -6.928] < $-t_{0.025} = -3.182$), we reject the null hypothesis and conclude that a linear relation exists between x and y .

11. (a) $\beta_0 \approx b_0 = 69.0296; \beta_1 \approx b_1 = -0.0479$

(b) $s_e = 0.3680$

(c) $s_{b_1} = 0.0043$

(d) Because the P -value < 0.001 [Tech: P -value = 0.0001] < $\alpha = 0.05$ (or $t_0 = -11.140$ [Tech: -11.157] < $-t_{0.025} = -2.571$), we reject the null hypothesis and conclude that a linear relation exists between commute time and score on a well-being survey.

(e) 95% confidence interval: lower bound: -0.0589, upper bound: -0.0369

13. (a) $\beta_0 \approx b_0 = 12.4932, \beta_1 \approx b_1 = 0.1827$

(b) $s_e = 0.0954$

(c) The correlation between the residuals and expected z -scores is 0.986. Because 0.986 [Tech: 0.987] > 0.923 (Table VI), conclude the residuals are approximately normally distributed.

(d) $s_{b_1} = 0.0276$

(e) Because the P -value < 0.001 < $\alpha = 0.01$ (or $t_0 = 6.62$ [Tech: 6.63] > $t_{0.005} = 3.250$), we reject the null hypothesis and conclude that a linear relation exists between a child's height and head circumference.

(f) 95% confidence interval: lower bound: 0.1204; upper bound: 0.2451

(g) A good estimate of the child's head circumference would be 17.33 inches.

15. (a) $\beta_0 \approx b_0 = 2675.6, \beta_1 \approx b_1 = 0.6764$

(b) $s_e = 271.04$

(c) $s_{b_1} = 0.2055$

(d) Because the P -value = 0.011 < $\alpha = 0.05$ (or $t_0 = 3.291 > t_{0.025} = 2.306$), we reject the null hypothesis and conclude that a linear relation exists between 7-day strength and 28-day strength.

(e) 95% confidence interval: lower bound: 0.2025; upper bound: 1.1503

(f) The mean 28-day strength of this concrete if the 7-day strength is 3000 psi is 4704.8 psi.

17. (a) $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 < 0$. $t_0 = -3.210 < -t_{0.01} \approx -2.4$, therefore reject H_0 . P -value 0.0012 < $\alpha = 0.01$, therefore reject H_0 . Conclude that a linear relation exists between cost and ROI.

(c) 90% confidence interval: lower bound: -0.000036; upper bound: -0.000008

(d) When the cost is \$180,000, the mean return on investment is 4.25%.

19. (a) $\beta_0 \approx b_0 = 24.9838; \beta_1 \approx b_1 = -0.5344$

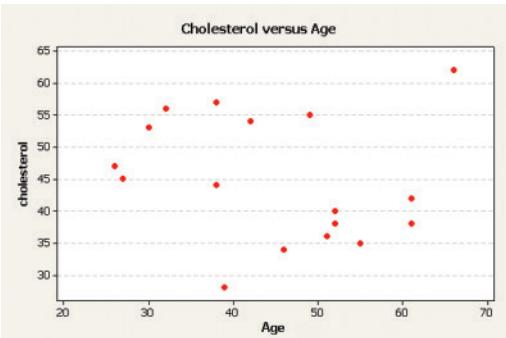
(b) Because the P -value > 0.5 [Tech: P -value = 0.751]

> $\alpha = 0.05$ (or $t_0 = -0.326 > -t_{0.025} = -2.228$), we do not reject the null hypothesis. There is not sufficient evidence to conclude that a linear relation exists between compensation and stock return.

(c) 95% confidence interval: lower bound: -4.1847, upper bound: 3.1159

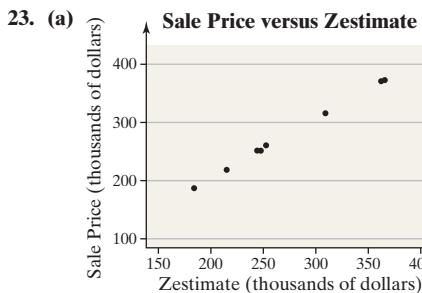
(d) No, the results do not indicate that a linear relation exists.

21. (a)



No linear relation appears to exist.

- (b) $\hat{y} = 50.7841 - 0.1298x$
 (c) Because the P -value = 0.530 > $\alpha = 0.01$ (or $t_0 = -0.642$ [Tech: $-0.643 > -t_{0.005} = -2.947$]), we do not reject the null hypothesis and conclude that a linear relation does not exist between the age and HDL levels.
 (d) 95% confidence interval: lower bound: -0.5605 ; upper bound: 0.3009
 (e) Do not recommend using the least-squares regression line to predict the HDL cholesterol levels since we did not reject the null hypothesis. A good estimate for the HDL cholesterol level would be $\bar{y} = 44.9$.



- (b) $\hat{y} = 1.0228x - 0.7590$; because the P -value $< 0.001 < \alpha = 0.05$ (or $t_0 = 105.426 > t_{0.025} = 2.447$), we reject the null hypothesis and conclude that there is a linear relation between the Zestimate and the sale price.
 (c) $\hat{y} = 0.5220x + 115.8094$; because $0.20 > P$ -value > 0.10 [Tech: P -value = 0.1927] $> \alpha = 0.05$ (or $t_0 = 1.441 < t_{0.025} = 2.365$), we do not reject the null hypothesis. There is not sufficient evidence to conclude that a linear relation exists between the Zestimate and the sale price. Yes, this observation is influential.

25. The y -coordinates on the least-squares regression line represent the mean value of the response variable for any given value of the explanatory variable.

27. We do not conduct inference on the linear correlation coefficient because a hypothesis test on the slope and a hypothesis test on the linear correlation coefficient yield the same conclusion. Moreover, the requirements for conducting inference on the linear correlation coefficient are very hard to verify.

12.4 Assess Your Understanding (page 572)

1. confidence; mean

3. (a) $\hat{y} = 11.8$
 (b) Lower bound: 10.8 [Tech: 10.9]; upper bound: 12.8
 (c) $\hat{y} = 11.8$
 (d) Lower bound: 9.9; upper bound: 13.7
 (e) The confidence interval is an interval estimate for the mean value of y at $x = 7$, whereas the prediction interval is an interval estimate for a single value of y at $x = 7$.
 5. (a) $\hat{y} = 4.3$
 (b) Lower bound: 2.5; upper bound: 6.1
 (c) $\hat{y} = 4.3$
 (d) Lower bound: 0.9; upper bound: 7.7
 7. (a) $\hat{y} = 68.07$
 (b) Lower bound: 67.723; upper bound: 68.420
 (c) $\hat{y} = 68.07$
 (d) Lower bound: 67.252; upper bound: 68.891
 (e) The prediction made in part (a) is an estimate of the mean well-being index composite score for all individuals whose commute time is 20 minutes. The prediction made in part (c) is an estimate of the well-being index composite score of one individual, Jane, whose commute time is 20 minutes.
 9. (a) $\hat{y} = 17.20$ inches
 (b) 95% confidence interval: lower bound: 17.12 inches; upper bound: 17.28 inches
 (c) $\hat{y} = 17.20$ inches
 (d) 95% prediction interval: lower bound: 16.97 inches; upper bound: 17.43 inches

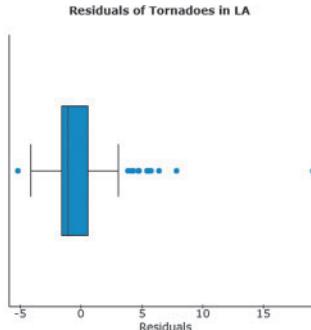
- (e) The confidence interval is an interval estimate for the mean head circumference of all children who are 25.75 inches tall. The prediction interval is an interval estimate for the head circumference of a single child who is 25.75 inches tall.

11. (a) $\hat{y} = 4400.4$ psi
 (b) 95% confidence interval: lower bound: 4147.8 psi; upper bound: 4635.1 psi
 (c) $\hat{y} = 4400.4$ psi
 (d) 95% prediction interval: lower bound: 3726.3 psi; upper bound: 5074.6 psi
 (e) The confidence interval is an interval estimate for the mean 28-day strength of all concrete cylinders that have a 7-day strength of 2550 psi. The prediction interval is an interval estimate for the 28-day strength of a single cylinder whose 7-day strength is 2550 psi.

13. (a) $\hat{y} = 4.25\%$
 (b) 95% confidence interval: lower bound: 3.35%; upper bound: 5.15%
 (c) $\hat{y} = 4.25\%$
 (d) 95% prediction interval: lower bound: -0.88% ; upper bound: 9.39%
 (e) Although the predicted return on investment in parts (a) and (c) are the same, the intervals are different because the distribution of the mean ROI, part (a), has less variability than the distribution of the individual ROI of a particular four-year school, part (c).

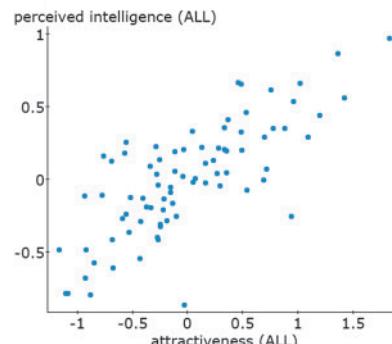
15. It does not make sense to construct either a confidence interval or prediction interval based on the least-squares regression equation because the evidence indicated that there is no linear relation between CEO compensation and stock return.

17. (a) $b_0 = 1.6280$; $b_1 = 0.0049$
 (b) There are many outliers.



- (c) $H_0: \beta_1 = 0$ versus $H_1: \beta_1 > 0$. Classical approach: $t_0 = 6.549 > t_{0.05} = 1.664$ (using 80 df); P -value approach: P -value < 0.0005 [Tech: P -value < 0.0001]. Reject H_0 . There is sufficient evidence to conclude a positive association exists between the width and length of a tornado in Louisiana.
 (d) With 95% confidence, the mean length of all tornadoes whose width is 500 yards is between 3.329 miles and 4.839 miles.
 (e) With 95% confidence, the mean length of a tornado whose width is 500 yards is between -2.561 (or 0) miles and 10.728 miles.
 (f) The large value of the standard error of the estimate causes the wide interval.

19. (a)



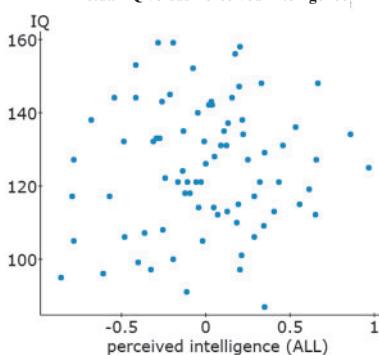
(b) $r = 0.762$; because $0.762 > 0.361$ (Table II with $n = 30$), we conclude there is a linear relation between attractiveness and perceived intelligence.

(c) $\hat{y} = 4.6008x + 0.0000$

(d) A person of average attractiveness is perceived to be of average intelligence.

(e) $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 > 0$, $P\text{-value} < 0.0001$. Reject H_0 .

(f)



$r = 0.086 < 0.361$, No linear relation between perceived intelligence and IQ.

(g) $\hat{y} = -1.8893x + 122.6038$. $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 > 0$. $P\text{-value} = 0.6329 > \alpha = 0.1$. Do not reject H_0 .

(h) $\hat{y} = 18.1100x + 128.8345$, $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 > 0$. $P\text{-value} = 0.0287 < \alpha = 0.1$. Reject H_0 .

(i) Lower bound: 126.1; upper bound: 177.9

(j) Lower bound: 106.3; upper bound: 197.7

Chapter 12 Review Exercises (page 576)

1. H_0 : the wheel is in balance; H_1 : the wheel is out of balance. Classical approach: $\chi^2_0 = 0.578 < \chi^2_{0.05} = 5.991$; do not reject the null hypothesis. $P\text{-value}$ approach: $P\text{-value} > 0.10 > \alpha = 0.05$ [Tech: $P\text{-value} = 0.7489$]; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the wheel is out of balance.

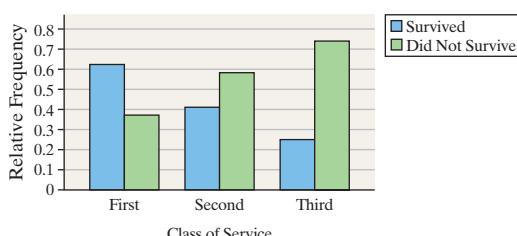
2. H_0 : the teams are evenly matched; H_1 : the teams are not evenly matched. Classical approach: $\chi^2_0 = 8.151 > \chi^2_{0.05} = 7.815$; reject the null hypothesis. $P\text{-value}$ approach: $0.025 < P\text{-value} < 0.05 = \alpha$ [Tech: $P\text{-value} = 0.043$]; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the teams playing in the World Series have not been evenly matched. The $P\text{-value}$ is suggestive of an issue. In particular, there are fewer six-game series than we would expect. Perhaps the team that is down “goes all out” in game 6, trying to force game 7.

3. (a) H_0 : class is independent of survival status; H_1 : class is associated with survival status. Classical approach: $\chi^2_0 = 133.052 > \chi^2_{0.05} = 5.991$; reject the null hypothesis. $P\text{-value}$ approach: $P\text{-value} < 0.005 < \alpha = 0.05$ [Tech: $P\text{-value} < 0.0001$]; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the class of service and survival rates are dependent.

(b)

Titanic	Class		
	First	Second	Third
Survived	0.625	0.414	0.252
Did not survive	0.375	0.586	0.748

Titanic Survival



This summary supports the existence of a relationship between class and survival rate. Individuals with higher-class tickets survived in greater proportions than individuals with lower-class tickets.

4. H_0 : gestational period is independent of degree;

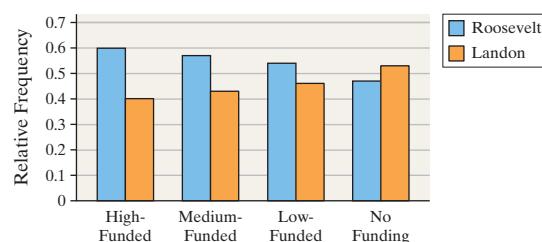
H_1 : gestational period is not independent of degree. Classical approach: $\chi^2_0 = 10.239 > \chi^2_{0.05} = 9.488$; reject the null hypothesis. $P\text{-value}$ approach: $P\text{-value} < 0.05$ [Tech: $P\text{-value} = 0.0366$]; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that length of the gestational period and completion of a high school diploma are dependent.

5. (a) Classical approach: $\chi^2_0 = 37.518 > \chi^2_{0.05} = 7.815$; reject the null hypothesis. $P\text{-value}$ approach: $P\text{-value} < 0.005$ [Tech: $P\text{-value} < 0.0001$]; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance that the level of funding received by the counties is associated with the candidate.

(b)

	High-Funded	Medium-Funded	Low-Funded	No Funding
Roosevelt	0.5994	0.5698	0.5400	0.4692
Landon	0.4006	0.4302	0.4600	0.5308
Total	1	1	1	1

Roosevelt versus Landon



6. $H_0: p_v = p_v$ vs. $H_1: p_v \neq p_v$. Classical: $\chi^2_0 = 3.636 < \chi^2_{0.05} = 3.841$; do not reject H_0 . $P\text{-value}: 0.05 < P\text{-value} < 0.10$ [Tech: 0.0565] $> \alpha = 0.05$; do not reject H_0 . There is not sufficient evidence at the 0.05 level of significance to suggest the proportion of individuals who believe voting is a civic duty differs from the proportion who believe jury duty is a civic duty.

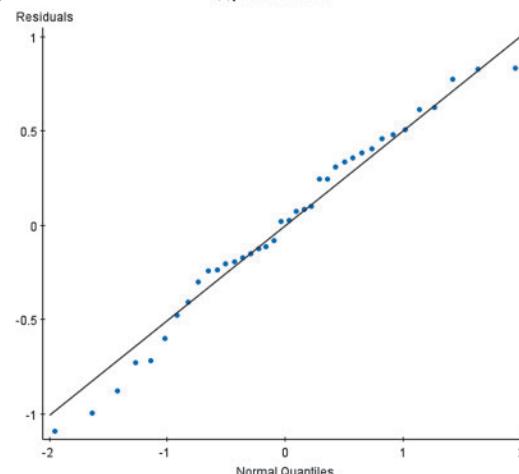
7. The least-squares regression model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The requirements to perform inference on the least-squares regression line are (1) for any particular values of the explanatory variable x , the mean of the corresponding responses in the population depends linearly on x , and (2) the response variables, y_i , are normally distributed with mean $\mu_{y|x} = \beta_0 + \beta_1 x$ and standard deviation σ . We verify these requirements by checking to see that the residuals are normally distributed, with mean 0 and constant variance σ^2 and that the residuals are independent. We do this by constructing residual plots and a normal probability plot of the residuals.

8. (a) $\beta_0 \approx b_0 = 3.8589$; $\beta_1 \approx b_1 = -0.1049$; $\hat{y} = 3.334$

(b) $s_e = 0.5102$

(c)

QQ plot of Residuals



The residuals are normally distributed.

(d) $s_{b_1} = 0.0366$

(e) Because the $P\text{-value} = 0.007 < \alpha = 0.05$ (or $t_0 = -2.866$ [Tech: $-2.868 < t_{0.025} = -2.028$]), we reject the null hypothesis and conclude that a linear relation exists between the row chosen by students on the first day of class and their cumulative GPAs.

(f) 95% confidence interval: lower bound: -0.1791 ; upper bound: -0.0307

(g) 95% confidence interval: lower bound: 3.159 ; upper bound: 3.509

(h) $\hat{y} = 3.334$

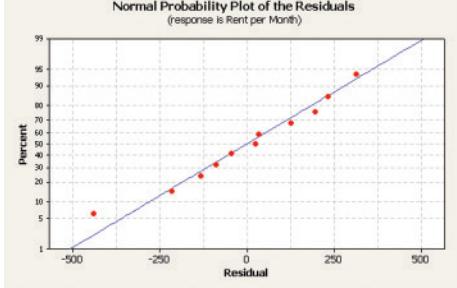
(i) 95% prediction interval: lower bound: 2.285 ; upper bound: 4.383 [Tech: 4.384]

(j) Although the predicted GPAs in parts (a) and (h) are the same, the intervals are different because the distribution of the mean GPAs, part (a), has less variability than the distribution of individual GPAs, part (h).

9. (a) $\beta_0 \approx b_0 = -399.2496$; $\beta_1 \approx b_1 = 2.5315$; $\hat{y} = 1879.1$, so the mean rent of a 900-square-foot apartment in Queens is \$1879.10.

(b) $s_e = 229.547$

(c)



The correlation between the residuals and expected z -scores is 0.985. Because $0.985 > 0.923$, the residuals are normally distributed.

(d) $s_{b_1} = 0.2166$

(e) There is evidence that a linear relation exists between the square footage of an apartment in Queens, New York, and the monthly rent.

(f) 95% confidence interval about the slope of the true least-squares regression line: lower bound: 2.0416 , upper bound: 3.0214 .

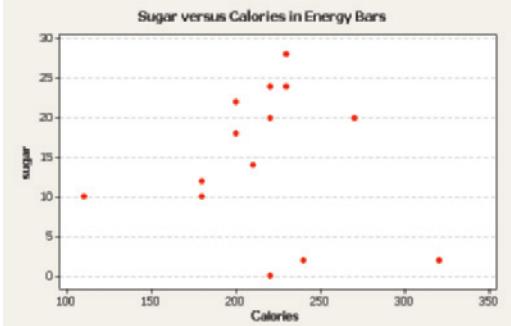
(g) 90% confidence interval about the mean rent of 900-square-foot apartments: lower bound: $\$1752.20$ [Tech: $\$1752.22$]; upper bound: $\$2006.00$ [Tech: $\$2005.96$].

(h) When an apartment has 900 square feet, $\hat{y} = \$1879.1$ [Tech: 1879.1].

(i) 90% prediction interval for the rent of a particular 900-square-foot apartment: lower bound: $\$1439.60$ [Tech: $\$1439.59$]; upper bound: $\$2318.60$ [Tech: $\$2318.59$].

(j) Although the predicted rents in parts (a) and (h) are the same, the intervals are different because the distribution of the means, part (a), has less variability than the distribution of the individuals, part (h).

10. (a)



No linear relation appears to exist.

(b) $\hat{y} = 17.8675 - 0.0146x$

(c) $s_e = 9.3749$

(d) A normal probability plot of the residuals shows that they are approximately normally distributed.

(e) $s_{b_1} = 0.0549$

(f) Because the $P\text{-value} = 0.795 > \alpha = 0.01$ (or $t_0 = -0.266$ [Tech: $-0.265 > -t_{0.005} = -3.055$]), we do not reject the null hypothesis. We conclude that a linear relation does not exist between the number of calories per serving and the number of grams of sugar per serving in high-protein and moderate-protein energy bars.

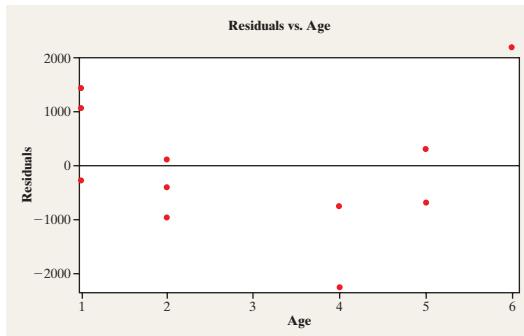
(g) 95% confidence interval: lower bound: -0.1342 [Tech: -0.1343]; upper bound: 0.1050 [Tech: 0.1051]

(h) Do not recommend using the least-squares regression line to predict the sugar content of the energy bars because we did not reject the null hypothesis. A good estimate for the sugar content is $\bar{y} = 14.7$ grams.

11. (a) $\hat{y} = 20,976.2 - 2054.6x$

(b) Reject H_0 . There is sufficient evidence at the $\alpha = 0.05$ level of significance that a linear relation exists between the age of Chevy Camaros and their selling price ($P\text{-value} < 0.001$).

(c)



(d) The residual plot shows a pattern indicating that a linear model is not appropriate. The moral is to perform graphical diagnostic tests along with the inferential procedures before drawing any conclusions regarding the model.

Chapter 12 Test (page 579)

1. H_0 : the dice are fair; H_1 : the dice are not fair. Classical approach: $\chi^2_0 = 4.940 < \chi^2_{0.01} = 23.209$; do not reject the null hypothesis. P -value approach: $P\text{-value} > 0.10 > \alpha = 0.01$ [Tech: $P\text{-value} = 0.8951$]; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.01$ level of significance to conclude that the dice are loaded.

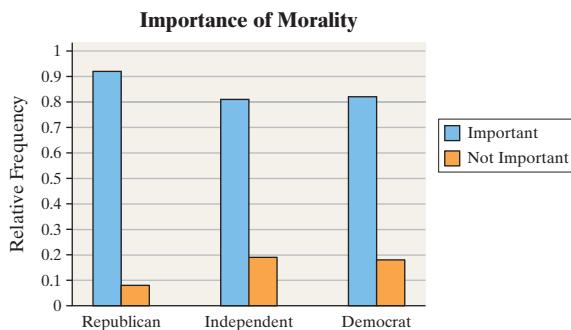
2. H_0 : educational attainment is the same today as 2000; H_1 : educational attainment has changed since 2000. Classical approach: $\chi^2_0 = 2.661 < \chi^2_{0.10} = 9.236$; do not reject the null hypothesis. P -value approach: $P\text{-value} > 0.10 = \alpha$ [Tech: $P\text{-value} = 0.7521$]; do not reject the null hypothesis. There is not sufficient evidence at the $\alpha = 0.10$ level of significance to conclude that the distribution of educational attainment of Americans today is different from the distribution of education attainment in 2000. That is, the evidence suggests that educational attainment has not changed.

3. (a) $H_0: p_R = p_I = p_D; H_1: \text{at least one proportion differs}$

Classical approach: $\chi^2_0 = 41.767 > \chi^2_{0.05} = 5.991$; reject the null hypothesis. P -value approach: $P\text{-value} < 0.005 < \alpha = 0.05$ [Tech: $P\text{-value} < 0.0001$]; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that at least one proportion is different from the others. That is, the evidence suggests that the proportion of adults who feel morality is important when deciding how to vote is different for at least one political affiliation.

(b)

	Republican	Independent	Democrat
Important	0.920	0.810	0.820
Not Important	0.080	0.190	0.180

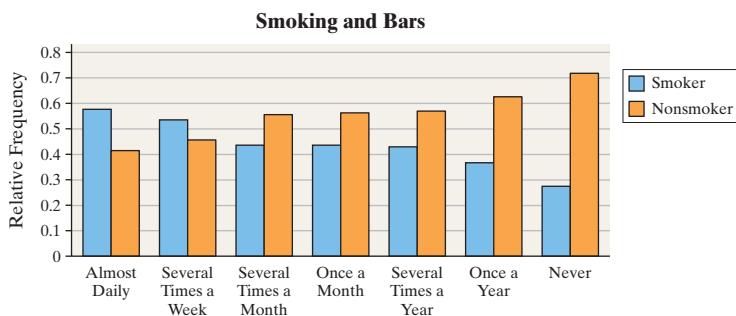


A higher proportion of Republicans appears to feel that morality is important when deciding how to vote than do Democrats or Independents.

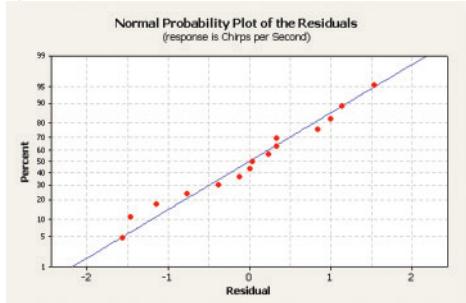
4. $H_0: P_{1970s} = P_{1980s} = P_{1990s} = P_{2000s}$ vs. H_1 : at least one proportion differs; Classical approach: $\chi^2_0 = 4155.584 > \chi^2_{0.05} = 7.815$; reject the null hypothesis. P -value approach: P -value < 0.005 [Tech: P -value < 0.0001]; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance that different proportions of 18- to 29-year-olds have been affiliated with religion in the past four decades.

5. Classical approach: $\chi^2_0 = 330.803 > \chi^2_{0.05} = 12.592$; reject the null hypothesis. P -value approach: P -value < 0.005 < $\alpha = 0.05$ [Tech: P -value < 0.0001]; reject the null hypothesis. There is sufficient evidence at the $\alpha = 0.05$ level of significance to conclude that the proportions for the time categories spent in bars differ between smokers and nonsmokers. From the conditional distribution and bar graph, it appears that a higher proportion of smokers spends more time in bars than nonsmokers.

	Almost Daily	Several Times a Week	Several Times a Month	Once a Month	Several Times a Year	Once a Year	Never
Smoker	0.584	0.539	0.437	0.435	0.430	0.372	0.277
Nonsmoker	0.416	0.461	0.563	0.565	0.570	0.628	0.723



6. (1) For any particular values of the explanatory variable x , the mean of the corresponding responses in the population depends linearly on x .
 (2) The response variable, y_i , is normally distributed with mean $\mu_{y|x} = \beta_0 + \beta_1 x$ and standard deviation σ .
 7. (a) $\beta_0 \approx b_0 = -0.3091$; $\beta_1 \approx b_1 = 0.2119$; $\hat{y} = 16.685$ [Tech: 16.687]; so the mean number of chirps when the temperature is 80.2°F is 16.69.
 (b) $s_e = 0.9715$



The residuals are normally distributed.

- (d) $s_{b_1} = 0.0387$
 (e) There is sufficient evidence that a linear relation exists between temperature and the cricket's chirps (P -value = 0.0001).
 (f) 95% confidence interval about the slope of the true least-squares regression line: lower bound: 0.1283; upper bound: 0.2955.
 (g) 90% confidence interval about the mean number of chirps at 80.2°F: lower bound: 16.25 [Tech: 16.24]; upper bound: 17.13.
 (h) When the temperature is 80.2°F, $\hat{y} = 16.69$ chirps.

- (i) 90% prediction interval for the number of chirps of a particular cricket at 80.2°F: lower bound: 14.91; upper bound: 18.47 [Tech: 18.46].
 (j) Although the predicted numbers of chirps in parts (a) and (h) are the same, the intervals are different because the distribution of the means, part (a), has less variability than the distribution of the individuals, part (h).
 8. (a) $\beta_0 \approx b_0 = 29.7048$; $\beta_1 \approx b_1 = 2.6351$; $\hat{y} = 48.15$; so the mean height of a 7-year-old boy is 48.15 inches.
 (b) $s_e = 2.45$
 (c) There is sufficient evidence that a linear relation exists between boys' ages and heights (P -value < 0.0001).
 (d) 95% confidence interval about the slope of the true least-squares regression line: lower bound: 2.2009; upper bound: 3.0692.
 (e) 90% confidence interval about the mean height of 7-year-old boys: lower bound: 47.11; upper bound: 49.19 inches.
 (f) The predicted height of a randomly chosen 7-year-old boy is 48.15 inches, $\hat{y} = 48.15$.
 (g) 90% prediction interval for the height of a particular 7-year-old boy: lower bound: 43.78; upper bound: 52.52 inches.
 (h) Although the predicted heights in parts (a) and (f) are the same, the intervals are different because the distribution of the means, part (a), has less variability than the distribution of the individuals, part (f).
 9. (a) $\beta_0 \approx b_0 = 67.388$; $\beta_1 \approx b_1 = -0.2632$
 (b) There is not sufficient evidence at the $\alpha = 0.05$ level of significance to support that a linear relation exists between a woman's age and grip strength (P -value = 0.138).
 (c) Based on the answer to (b), a good estimate of the grip strength of a 42-year-old female would be the mean strength of the population, 57 psi.

Subject Index

A

Absolute deviation, mean, 137
Acceptance sampling, 263–264
Addition Rule, 257
with contingency tables, 246–247
for Disjoint Events, 242–244, 246, 253
Benford's Law and, 243–244
General, 245–247
Agresti, A., 409
Alternative hypothesis, 435–438
definition of, 436
structuring, 437
American Time Use Survey, 370, 382, 394
Analysis of variance (ANOVA)
definition of, B12
F-test statistic, B16–B19
null hypothesis in, B12
one-way, B12–B23
conceptual understanding of, B15–B16
decision rule in, B19
equal population variances in, B14
normal distribution in, B14, B22
requirements of, B14–B15
robustness of, B14
testing hypothesis with, B15–B22
using technology, B20–B21, B22–B23
Type I error in, B13
Anecdotal claims, 3
ANOVA table, B19
Anticyclones, 192
Approach to problem solving, 7
Area
under normal curve, 336–339, 343–347
interpreting, 339
as probability, 338
as proportion, 338
as probability, 333
Ars Conjectandi (Bernoulli), 228, 313
Ars Magna (Cardano), 233
Associated variables, 172, 176
Association
among categorical data, conditional distribution to identify, 208–212
causation vs., 16–17
Simpson's Paradox and, 213
At-least probability, 256–257
Average, 107, 108. *See also* Mean(s)

B

Bar graph(s), 64–68, 69–70, 544–545
of conditional distribution, 211–212
frequency and relative frequency, 65–66
horizontal bars, 67–68
side-by-side, 66–67
stacked (segmented), 211, 214
technology to draw, 65–66, 69–70, 214
BeautifulSoup, 19
Before-after (pretest-posttest) experiments, 49
Behrens-Fisher problem, 502
Bell-shaped distribution, 82, 129–131
Empirical Rule to describe, 129–131
Benford, Frank, 243, 534
Benford's Law, 243–244, 534
Bernoulli, Jacob, 228, 313
Bernoulli random values, 286
Bernoulli trial, 313
Between-sample variability, B16

Bias, 38

in census, 41–42
in sampling, 38–42
frame and, 38
misrepresented answers, 40
ordering of questions or words, 40
sources of, 38–42

Biased statistic, 129

Bimodal data set, 115

Binomial experiment
computing probabilities of, 314–320
criteria for, 313, 360
identifying, 313–314

Binomial probability distribution, 312–327, 370, 452–453
binomial probability distribution function (pdf), 316–317
binomial table to compute binomial probabilities, 318–319
constructing, 314–316
cumulative, table of, A7–A10
graphs of, 321–322
mean and standard deviation of binomial random variable, 320
negative, 326–327
normal approximation to, 360–364
notation used in, 313
table, A3–A10
using technology, 319–320, 323–324

Binomial random variable, 313

normal approximation to, 362–363

Binomial tables, 318–319

Bivariate data, 170. *See also* Relation between two variables

Bivariate normal distribution, 564

Blinding, 45

Blocking, 137

Box, George, 507

Boxplots, 157–160
comparing two distributions using, 159–160
constructing, 157–158
distribution shape based on, 158–160
side-by-side, of ANOVA results, B20–B21
technology to draw, 160

Bureau of Labor Statistics, 370, 394

C

Callahan, Paul X., 159

Callbacks, 39

Cardano, Fazio, 233

Cardano, Girolamo, 233

Case-control studies, 18

Categorical data

contingency tables and association, 206–217
conditional distribution, 208–212, 544–545
marginal distribution of variable, 207–208
using technology, 214
goodness-of-fit test
technology for, 530–531, 532–533
tests for homogeneity of proportions, 545–548

Categorical variable, 7

Causation

association vs., 16–17
correlation vs., 178–179

Cdf (cumulative distribution function), 319

Cells, 207, 246

Census, 18–19

bias in, 41–42

definition of, 19

Census Bureau, 19, 35

Central Limit Theorem, 377–378, 458

- Central tendency, measures of, 108–121
 arithmetic mean. *See Mean(s)*
 from grouped data, 139–145
 mean, 115
 median, 108, 110–111, 115, 145, 150, 159
 approximating from grouped data, 145
 computing, 110–111, 115
 definition of, 110
 quartile and, 158–159
 shape of distribution identified using, 113–114
 mode, 108, 114–115
 bimodality, 115
 computing, 114–115
 definition of, 114
 multimodality, 115
 of qualitative data, 115
 of quantitative data, 114–115
 trimmed mean, 120
- Certainty, 230
- Chart(s)
 Pareto, 66
 pie, 68–70
- Chebyshev, Pafnuty, 131, 132
- Chebyshev's Inequality, 131–132
- Chicago Tribune*, 221
- Chi-square distribution, 524–526
 characteristics of, 525
 critical values for, 525–526
 definition of, 524
 table, A15
- Chi-square test, 176
 for homogeneity of proportions, 545–548
 definition of, 545
 steps in, 546–548
 for independence, 538–545
 definition of, 539
 expected counts in, 539–541
 expected frequencies in, 541
 steps in, 541–542
 test statistic for, 541
 using technology, 548, 549
- Claim, 46
- Class(es)
 data, 77–80
 midpoint of, 139–140
 modal, 145
 width of, 78
 determining, 80
- Classical approach to hypothesis testing
 in chi-square test
 for homogeneity of proportions, 547
 for independence, 542, 543
- of difference between two means using independent samples, 502–503
- of difference between two population proportions, McNemar's Test
 for, B8, B9
- of difference of two means using independent samples, 504–505
- in goodness-of-fit test, 528–529, 530, 532
- in least-squares regression model, 560
- about population mean, 459, 460–461, 462–463, 464
- about population proportion, 444–451
 from independent samples, 482–483
- to testing claims about matched-pairs data, 491–492, 494
- Classical method, 232–236
 empirical method compared to, 234–236
 logic of, 443–447
- Classical probability, 299
- Classroom survey, 41–42
- Closed question, 41
- Cluster sampling, 33–34, 36
- Coefficient of determination, 201–206
 computing, 201–204
 by hand, 204
 using technology, 204
 definition of, 201
- Coefficient of skewness, 137
- Coefficient of variation, 138
- Cohort studies, 18
 prospective, 178
- Collaborative Atorvastatin Diabetes Study (CARDS), 46
- Column variable, 207, 246
- Combinations, 278–279
 counting problems solved using, 274–276
 definition of, 274
 formula, 274–275
 listing, 274
 of simple random samples, 276
 technology to compute, 275, 279
- Complement of event, 247
- Complement Rule, 247–249
 at-least probabilities and, 256
- Completely randomized design, 48–49
- Computational formula, 123, 125
- Conceptual formula, 123, 125
- Conclusions, stating, 440–441
- Conditional distribution, 208–212
 bar graph of, 211–212
 constructing, 210–211, 544–545
 definition of, 209
- Conditional probability, 260–269
 definition of, 260
 independence and, 265–266
 using the General Multiplication Rule, 262–266
- Conditional Probability Rule, 261
- Confidence, level of, 5, 397, 401
 margin of error and, 403–404
 in method vs. interval, 401
 simulation to illustrate meaning of, 399
- Confidence interval(s)
 definition of, 397, 569
 determining appropriate procedure for constructing, 425–426
 for difference between two population proportions, 483–485
 using technology, 486–487
 about the difference of two means, 506–508
 hypothesis testing using, 451–452
 interpretation of, 400–401
 margin of error for 95%, 398–400, 418
 for mean response, 569–570, 571
 using technology, 571–572
 for population mean, 411–412, 415–418
 for population mean difference of matched-pairs data, 495–496
 for population proportion, 396–404
 constructing and interpreting, 396–404
 point estimate for population proportion, 396
 sample size determination, 404–405
 using technology, 403, 405–406
 probability and, 399
 for slope of regression line, 562–564
 constructing, 562–564
 definition, 562
- Confounding, 16
 experimental design and, 47
- Constants, 7
- Contingency (two-way) table(s), 206–217, 246–247. *See also Chi-square test*
 Addition Rule with contingency tables, 246–247
 conditional distribution, 208–212, 544–545
 definition of, 207
 marginal distribution of variable, 207–208
 using technology, 214

- Continuity, correction for, 361
 Continuous data, 8, 9
 histograms of, 80–81
 quantitative, mean for, 139–140
 in tables, 78–80
 Continuous random variable, 300–301
 probability density functions to find probabilities for, 333–334
 Control group, 45
 Convenience sampling, 6, 24, 34–35
 Correction for continuity, 361
 Correlation, causation vs., 178–179
 Correlation coefficient, 531
 critical values for, table of, A2
 linear, 173–178, 564
 computing and interpreting, 175–177
 definition of, 173
 properties of, 173–175
 technology to determine, 179–180
 Cortical blindness, 455
 Coull, B., 409
 Counting problems, 269–282
 appropriate technique to use, determining, 289–290
 combinations for, 274–276, 278–279
 Multiplication Rule for, 269–272
 permutations for, 272–274, 276–277, 279
 without repetition, 271
 Counts, expected, 527, 539
 Cramer, Jim, 454
 Critical *F*-value, B21
 Critical value(s), 400
 for chi-square distribution, 525–526
 for correlation coefficient table, A2
 determining linear relation between two variables using, 177–178
 F-distribution table, A16–A19
 Cross-sectional studies, 18
 Cumulative Binomial Probability Distribution table, A7–A10
 Cumulative distribution function (cdf), 319
 Current Population Survey, 35, 39, 371
Current Population Survey: Design and Methodology, The, 35
 Cutoff point, 230
- D**
- Darwin, Charles, 192
 Data, 3–4. *See also* Qualitative data; Quantitative data
 bivariate, 170
 categorical. *See* Categorical data
 classes of, 76, 78–80
 collection of, 6
 continuous, 8
 histograms of, 80–81
 quantitative, mean for, 139–140
 in tables, 78–80
 discrete, 8, 76–77
 downloading from Web, 20
 existing sources of, 18
 grouped, 139–145
 mean from, 139–140, 143
 standard deviation from, 141–143
 misuse of, 3
 polling, 6
 qualitative, 8
 quantitative, 8
 raw, 62, 80
 univariate, 170
 variability in, 4
 variables vs., 8–9
 through web scraping (data mining), 19–20
 Data-entry error, 41
- Data organization and summary, 62–106. *See also* Numerically summarizing data
 graphical misrepresentations, 92–100
 by manipulating vertical scale, 93–95
 qualitative data, 63–76
 bar graphs, 64–68, 69–70
 pie charts, 68–70
 tables, 63–64
 quantitative data, 76–92
 dot plots, 82
 histograms, 77–78, 80–81, 84–85
 shape of distribution, 82–83
 tables, 76–77, 78–80
 time-series graphs, 84–85
 Data sets
 comparing, 66–67, 121–122
 standard deviation of two, 128
 small, 355
 Decision making, informed, 60
 Degrees of freedom, 125, 412, 415, B17, B21
 inferences about two means, 502, 505–506
 de Moivre, Abraham, 337, 339, 360
 Density function(s)
 exponential, 394–395
 probability, 333–334
 uniform, 335
 Dependent events, 253
 Dependent (response) variable, 15, 46, 171, 209
 Dependent sample(s), 478–479, 490–496. *See also* Matched-pairs (dependent) design
 confidence interval for matched-pairs data, 495–496
 independent vs. dependent sampling, 478–479
 McNemar's Test to compare two proportions from, B6–B10
 testing hypotheses regarding matched-pairs data, 491–494
 Descriptive statistics, 5, 6
 Designed experiment, 14–23, 44–54. *See also* Experiment(s): design of
 completely randomized design, 48–49
 defined, 15
 matched-pairs design, 49–50
 observational study vs., 14–20
 steps in designing, 46–47
 Determination, coefficient of, 201–206
 computing, 201–204
 using technology, 204
 definition of, 201
 Deviation(s), 202
 explained, 202
 about the mean, 123
 total, 202
 unexplained, 202
 Dice problem, 236
 Die
 fair, 229
 loaded, 229
 Diophantus, 235
 Discrete data, 8
 histograms of, 77–78
 in tables, 76–77
 Discrete probability distributions, 299–339
 binomial, 312–327, 370, 452–453
 binomial probability distribution function (pdf), 316–317
 binomial table to compute binomial probabilities, 318–319
 constructing, 314–316
 graphs of, 321–322
 identifying binomial experiment, 313–314
 mean and standard deviation of binomial random variable, 320
 negative, 326–327
 normal approximation to, 360–364

- Discrete probability distributions (*Continued*)
 notation used in, 313
 using technology, 319–320, 323–324
 definition of, 301
 geometric, 326
 identifying, 301–302
 Poisson, 370
 rules for, 301
- Discrete random variable, 300–312
 continuous random variables distinguished from, 300–301
 mean of, 303–306, 307
 computing, 303–305
 defined, 304
 as an expected value, 305–306
 interpreting, 304–306
 using technology, 307–308
 variance and standard deviation of, 306–308
- Disjoint events, 242–244, 253
 independent events vs., 254
- Dispersion, measures of, 121–143
 Chebyshev's Inequality, 131–132
 Empirical Rule, 129–131
 from grouped data, 139–145
 mean absolute deviation, 137
 range, 122
 computing, 122
 definition of, 122
 interquartile (IQR), 122, 149–150, 157
 technology to determine, 132
- standard deviation, 123–128, 176
 of binomial random variable, 320
 of discrete random variables, 306–308
 from grouped data, 141–143
 interpretations of, 127
 outlier distortion of, 150
 population, 123–125, 556
 sample, 125–127, 557
 of sampling distribution of sample mean, 374
 technology to approximate, 143
 technology to determine, 132
 of two data sets, 128
- unusual results in binomial experiment and, 323
- variance, 128–129
 of discrete random variables, 307
 population, 128
 sample, 128–129
 technology to determine, 132
- Distribution function
 binomial probability (pdf), 316–317
 cumulative (cdf), 319
- Doctrine of Chance, The* (Leibniz), 337
- Dot plots, 82
 using technology, 85
- Double-blind experiment, 45
- E**
- EDA (exploratory data analysis), 155
Éléments de Géométrie (Legendre), 190
Elements of Sampling Theory and Methods (Govindarajulu), 35
 Empirical Method, 231–232, 234–236
 classical method compared to, 234–236
- Empirical Rule, 129–131, 337
 unusual results in binomial experiment and, 323
- Equally likely outcomes, 232, 234
- Error(s)
 input (data-entry), 41
 margin of, 6, 397, 398
 for constructing confidence intervals about a population mean, 415
 definition of, 402
- level of confidence and, 403–404
 sample size and, 404–405, 418
 mean square due to (MSE), B17, B19
 nonsampling vs. sampling, 41–42
 residual, 189
 round-off, 128
 sampling, interviewer error, 39–40
 standard, 558
 computing, 558
 definition of, 557
 of the mean, 374
 sum of squares due to, B17
- Type I, 438–440, 445, 506
 in ANOVA, B13
 probability of, 440, 445
- Type II, 438–440
 probability of, 440
- Estimates, pooled, 480
- Estimation, 395
- Estimator, biased, 129
- Ethics of web scraping (data mining), 19–20
- Event(s), 229
 certain, 230
 complement of, 247
 dependent, 253
 disjoint, 242–244, 253, 254
 impossible, 230
 independent, 253–256, 265–266
 Multiplication Rule for, 254–256, 266
- not so unusual, 230
- simple, 229
 unusual, 230
- Excel, 28
 area under the normal curve using, 351
 bar graph using, 65, 69
 binomial probabilities using, 323–324
 boxplots using, 160
 chi-square tests using, 549
 coefficient of determination using, 204
 combinations using, 275, 279
 confidence intervals using, 572
 for population mean, 419
 for population proportion, 406, 487
- correlation coefficient using, 180
- factorials on, 279
- goodness-of-fit test, 533
- histograms using, 85
- hypothesis testing using, for population mean, 464
- If/Else statements in, 285
- inference for two population proportions using, 487
- marginal and conditional distributions using, 214
- for population proportion, 453, 487
- least-squares regression line using, 191–192, 195
- least-squares regression model using, 564
- linear correlation coefficient using, 176–177
- mean using, 116, 308
- median using, 116
- normal probability plot using, 358
- one-way ANOVA using, B20, B23
- permutations using, 273, 279
- pie charts using, 70
- prediction intervals using, 572
- quartiles using, 152
- random integers using, 285
- scatter diagrams using, 176–177, 180
- standard deviation of discrete random variable using, 308
- standard error using, 558
- sum of two columns in, 285

- two-sample *t*-tests, dependent sampling using, 496
 two-sample *t*-tests, independent sampling using, 508
- Expected counts, 527–529
 in chi-square test, 539–541
 in goodness-of-fit test, 527
- Expected value, 305–306
- Experiment(s), 62, 229, 299
 characteristics of, 45–46
 definition, 45
 design of, 44–45
 completely randomized design, 48–49
 matched-pairs design, 49–50
 simple random sampling and, 50
 steps in, 46–47
- double-blind, 45
 with equally likely outcomes, 232
- observational study vs., 14–20
 probability, 231, 232
 single-blind, 45
- Experimental units (subjects), 45, 47
 in matched-pairs design, 49, 50
- Explained deviation, 202
- Explanatory (predictor or independent) variable, 171
- Explanatory variable, 15, 45, 209
- Exploratory data analysis (EDA), 155
Exploratory Data Analysis (Tukey), 155
- Exponential density function, 394–395
- Exponential probability distribution, 424
- F**
- Factor(s), 45, 47
- Factorial notation, 273
- Factorials, technology to compute, 279
- Factorial symbol, 272
- FADS2 gene, 4n
- Fair die, 229
- F*-distribution
 critical values of
 table, A16–A19
- Fences, 151
- Fermat, Pierre de, 235, 236, 306
- Fermat's Last Theorem, 235
- Fisher, Sir Ronald A., 45, 176, 443, 531, B13
- Five-number summary, 155–157
- Frame, 25, 32, 38
- Framingham Heart Study, 18
- Frequency, relative, 64
- Frequency bar graph, 65–66
- Frequency distribution(s). *See also* Relative frequency distribution
 based on boxplots, 158–160
 bell shaped, 82, 129–131
 Empirical Rule to describe, 129–131
 center of (average), 107
 characteristics of, 107
 chi-square, 526
 characteristics of, 525
 definition of, 524
 comparing, 159–160
 conditional, 208–212
 bar graph, 211–212
 constructing, 210–211, 544–545
 definition of, 209
 from continuous data, 78–80
 histograms, 80–81
 tables, 78–80
 critical value for, 400
 of discrete data, 76–78
 histograms, 77–78
 tables, 76–77
- marginal, 207–208
 relative frequency, 208
 mean of variable from, 139–140
 of qualitative data, 63–64
 relative, 64
 shape of, 82–83, 107, 113–114
 skewed, 112–113
 mean or median versus skewness, 112–113
 spread of, 107
- F*-test, 507
 computing, B16–B19
- Functional status, 16
- G**
- Gallup, George, 39
 Gallup Organization, 23, 40
 Galton, Sir Francis, 192, 259
 Gauss, Carl, 339
 General Addition Rule, 245–247
 General Multiplication Rule, conditional probability using, 262–266
 Geometric probability distribution, 326
 Good fit, 187
 Goodness-of-fit test, 526–538
 characteristics of chi-square distribution, 525
 definition of, 526
 expected counts, 527
 technology for, 529, 532–533
 testing hypotheses using, 528–532
 test statistic for, 528
- Gosset, William Sealey, 411–412
- Graph(s), 107
 binomial probability, 321–322
 characteristics of good, 97
 of discrete probability distributions, 302–303
 dot plots, 82
 using technology, 85
 histograms
 of continuous data, 80–81
 of discrete data, 77–78
 technology to draw, 84–85
 misleading or deceptive, 92–97
 of normal curve, 335–336
 polar area diagram, 80
 tables, 106, 107
 binomial, 318–319
 continuous data in, 78–80
 discrete data in, 76–77
 open-ended, 78
 qualitative data in, 63–64
 time-series, 83–84
 technology to draw, 85
- Greek letters, use of, 108
- Grouped data, 139–145
 mean from, 139–140
 standard deviation from, 141–143
- H**
- Histogram(s), 77–78
 of continuous data, 80–81
 of discrete data, 77–78
 relative frequency, 338
 technology to draw, 84–85
- Homeless, Census count of, 19
- Homogeneity of proportions, chi-square test for, 545–548
 definition of, 545
 steps in, 546–548
- Horizontal lines, B3–B4
- How to Lie with Statistics* (Huff), 97
- Huff, Darrell, 97

- Huygens, Christiaan, 306
- Hypothesis/hypotheses**
- alternative, 435–438
 - definition of, 436
 - structuring, 437
 - definition of, 436
 - forming, 437–438
 - null, 435–438, 526, 529
 - in ANOVA, B12
 - assumption of trueness of, 444
 - definition of, 436
 - not rejecting versus acceptance of, 440–441
 - structuring, 437
- Hypothesis testing, 395, 434–476. *See also* Inferences; One-tailed test(s); Two-tailed tests
- choosing method for, 470–472
 - definition of, 436
 - language of, 435–443
 - outcomes from, 438
 - for population mean, 458–469
 - classical approach to, 459, 460–461, 462–463
 - with large sample, 459–461, 464
 - P*-value approach to, 459, 460–461, 462–463
 - with small sample, 461–463
 - technology in, 464
 - for population proportion, 443–458
 - binomial probability distribution for, 452–453
 - classical approach to, 444–445, 446–451
 - left-tailed, 447
 - logic of, 443–447
 - P*-value approach to, 445–451
 - technology in, 453
 - two-tailed, 449–451
 - using confidence interval, 451–452
 - probability of Type I error, 440, 445
 - probability of Type II error, 440
 - small samples, 452–453
 - stating conclusions, 440–441
 - steps in, 436
 - for two population proportions from independent samples, 479–483
 - classical approach to, 482–483
 - P*-value approach to, 482–483
 - Type I and Type II errors, 438–440
 - using one-way ANOVA, B15–B22
- I**
- Impossible event, 230
 - Incentives, nonresponse and, 39
 - Independence**
 - chi-square test for, 538–545
 - definition of, 539
 - expected counts in, 539–541
 - expected frequencies in, 541
 - steps in, 541–542
 - test statistic for, 541
 - conditional probability and, 265–266
 - Independent events, 253–256, 265–266
 - disjoint events vs., 254
 - Multiplication Rule for, 254–256, 266
 - Independent (explanatory or predictor) variable, 171
 - Independent samples, 478–479
 - difference between two means from
 - confidence intervals regarding, 506–508
 - technology for, 505, 508
 - testing hypotheses regarding, 502–506
 - hypothesis testing regarding two population proportions from, 479–483
 - inferences about two means from, 502–512
 - Independent trials, 313–314

Individual, population vs., 5

Inequality, Chebyshev's, 131–132

Inferences, 371, 477–581. *See also* Least-squares regression model

 - contingency tables and association, 206–217
 - determining appropriate test to perform, 513
 - goodness-of-fit test, 526–538
 - tests for independence and the homogeneity of proportions, 538–553
 - about two means
 - dependent samples, 490–496
 - independent samples, 502–512
 - about two population proportions, 478–490
 - confidence intervals, 483–485
 - hypothesis testing from independent samples, 479–483
 - independent vs. dependent sampling, 478–479
 - McNemar's Test to compare two proportions from matched-pairs data, B6–B10
 - sample size requirements for, 485–486
 - using technology, 486–487
 - Inferential statistics, 5, 6, 47, 395, 434
 - Inflection points, 336
 - Input errors, 41
 - Integers, random, 285–286
 - Intercept, 190, 193–194
 - of least-squares regression model, inference on, 559
 - Internet surveys, 34
 - Interquartile range (IQR), 122, 149–150, 156, 157
 - Interval, for a population parameter, 397
 - Interval level of measurement, 10
 - Interviewer error, 39–40

J

 - Jefferson, Thomas, 19
 - Jointly normally distributed, 564

K

 - Kolmogorov, Andrei Nikolaevich, 265

L

 - Landon, Alfred M., 39
 - Laplace, Pierre Simon, 377
 - Law of Large Numbers, 229–230, 231, 299, 414
 - Least-squares regression line, 187–214
 - definition of, 189
 - diagnostics on
 - coefficient of determination, 201–206
 - equation of, 190
 - finding, 187–192
 - by hand, 190
 - using technology, 191–192, 195
 - interpretation of predicted values, 191
 - slope and *y*-intercept of, 193–194
 - interpreting, 193–194
 - rounding rule for, 190
 - sum of squared residuals, 194–195
 - Least-squares regression model, 554–568
 - confidence interval about slope of, 562–564
 - constructing, 562–564
 - definition of, 562
 - confidence intervals for mean response, 569, 570, 572
 - using technology, 571–572
 - definition of, 557
 - example of, 554–555
 - inference on, 555–557
 - inference on the slope and intercept, 559–562
 - using technology, 561–562, 564
 - normally distributed residuals in, 558–559
 - prediction intervals for an individual response, 570–572
 - requirements of, 555–557
 - robustness of, 560

- significance of, 554–564
 standard error of the estimate, 557–558
- Legendre, Adrien Marie, 189, 190
- Leibniz, Gottfried Wilhelm, 337
- Level of confidence, 5, 397, 401
 margin of error and, 403–404
 in method vs. interval, 401
 simulation to illustrate meaning of, 399
- Level of measurement of variable, 9–10
- Level of significance, 440
- Liber de Ludo Alaea* (Cardano), 233
- Life on the Mississippi* (Twain), 200
- Linear correlation coefficient, 173–178, 564
 absolute value of, determining linear relation between variables from, 177
 computing and interpreting, 175–177
 definition of, 173
 properties of, 173–175
 using technology, 179–180
- Linear relation between two variables, determining, 177–178
- Lines, B1–B6
 equation of, B4–B5
 point-slope form of, B3–B4
 slope of, B1–B3, B5
 calculating and interpreting, B1–B2
 graphing line using, B2
 y -intercept of, B5
- Literary Digest*, 39
- Loaded die, 229
- Lower class limit, 78
 of the first class, guidelines for determining, 80
- Lurking variable, 4, 16–17, 178, 212–213
- Lyell's syndrome (Toxic Epidermal Necrolysis), 327
- M**
- McNemar's Test, B6–B10
- Marginal distribution
 definition of, 207
 relative frequency, 208
- Margin of error, 6, 397, 398
 for constructing confidence intervals about a population mean, 415
 definition of, 402
 level of confidence and, 403–404
 sample size and, 404–405, 485–486
 sample size to estimate population mean given, 418
- Matched-pairs (dependent) design
 confidence intervals about population mean difference of, 495–496
 testing claims about
 using technology, 494, 496
 testing hypotheses regarding, 491–494
- Matched-pairs (dependent) samples, 478–479
 McNemar's Test to compare two proportions from, B6–B10
- Matched-pairs design, 49–50, 478–479
- Mathematics, statistics vs., 4
- Mean(s), 108–110, 115. *See also* Population mean; Sample mean
 of binomial random variable, 320
 as center of gravity, 110
 comparing median and, 112–113
 comparing three or more. *See* Analysis of variance (ANOVA)
 computing, 109–110
 using technology, 111, 116
 definition of, 108
 deviation about the, 123, 127
 of discrete random variable, 303–308
 computing, 303–305
 defined, 304
 as an expected value, 305–306
 interpreting, 304–306
 using technology, 307–308
- from grouped data, 139–140, 143
 least-squares regression model and, 555–556
 outlier distortion of, 150
 of sampling distribution of sample mean, 374
 shape of distribution from, 113–114
 standard error of the, 374
 technology to approximate, 143
 trimmed, 120
 unusual results in binomial experiment and, 323
 variability in individuals versus variability in, 374
 weighted, 140–141
- Mean absolute deviation, 137
- Mean response, confidence intervals for, 569–570, 571–572
- Mean squares, B17
 due to error (MSE), B17, B19
 due to treatment (MST), B17
- Méchanique céleste* (Laplace), 377
- Median, 108, 110–111, 115, 150, 155, 159
 approximating from grouped data, 145
 comparing mean and, 112–113
 computing, 110–111
 with even number of observations, 111
 with odd number of observations, 110–111
 using technology, 111, 116
 definition of, 110
 quartile and, 158–159
 sampling distribution of, 394–395
 shape of distribution from, 113–114
- Midpoint, class, 139–140
- Midrange, 120
- Minitab, 27–28
 area under the normal curve using, 351
 normal values corresponding to, 351
 bar graph using, 65, 69
 binomial probabilities using, 323
 boxplots using, 160
 chi-square tests using, 549
 coefficient of determination using, 204
 comparing mean and median using, 112
 confidence bands in, 357
 confidence intervals using, 571–572
 for population mean, 417, 419
 for population proportion, 405–406
 correlation coefficient using, 180
 dot plots using, 85
 five-number summary using, 156
 goodness-of-fit test, 532
 histograms using, 85
 hypothesis testing using
 about population mean, 464
 about population proportion, 453
 inference of two population proportions using, 486–487
 least-squares regression line using, 191–192, 195
 least-squares regression model using, 561–562, 564
 marginal and conditional distributions using, 214
 McNemar's Test, B10
 mean and median using, 116
 normality assessed using, 358
 normal probability plot using, 358
 one-way ANOVA using, B20, B22
 pie charts using, 69
 prediction intervals using, 571–572
 quartiles using, 149, 151
 sampling distribution of sample mean from normal population, 372–373
 sampling from a population that is not normal, 376
 scatter diagrams using, 179
 time-series plots using, 85
 two-sample *t*-tests using

- Minitab (*Continued*)
 dependent sampling, 496
 independent sampling, 508
- Misrepresentation of data, graphical, 92–100
 three-dimensional scale, 96–97
- Misrepresented answers, 40
- Modal class, 145
- Mode, 108, 114–115
 bimodality, 115
 computing, 114–115
 definition of, 114
 multimodality, 115
 of qualitative data, 115
 of quantitative data, 114–115
- Models, 335, 338
 probability, 230–231, 232
- Monte Carlo Method, 282
- Multimodal data set, 115
- Multimodal instruction, 511
- Multiplication Rule, 257, 539, 540
 for counting, 269–272
 General Multiplication Rule, 262–266
 for Independent Events, 254–256, 265–266
- Multistage sampling, 35
- Mutual fund, 476
- Mutually exclusive (disjoint) events, 242–244, 253
 independent events vs., 254
- N**
- Negative binomial probability distribution, 326–327
- Negatively associated variables, 172
- Newcomb, Simon, 243, 534
- Newton, Isaac, 337
- Nielsen Media Research, 35
- Nightingale, Florence, 80
- Nine-enders, 538
- Nomen, 9
- Nominal level of measurement, 9
- Nonparametric statistics, 418
- Nonresponse bias, 39, 41
- Nonsampling errors, 41
 data-entry errors, 41
 nonresponse bias, 39
 response bias, 39–41
 undercoverage, 38–39
- Normal curve, 335–337
 area under, 337–339, 343–347
 interpreting, 339
 as probability, 338
 as proportion, 338
 cautionary thoughts on, 347
 inflection points on, 336
- Normal distribution, standard, 343, A11–A12
 t -distribution and, 416
- Normal populations, sampling distribution of sample mean from, 371–375
- Normal probability density function, 338
- Normal probability distribution, 332–369, 370
 applications of, 343–354
 area under, 337–339
 interpreting, 339
 as proportion or probability, 338
 assessing normality, 355–358
 normal probability plots for, 355–358
- bivariate, 564
- cautionary thoughts on, 347
- graph of, 335–336
- jointly, 564
- least-squares regression model and, 555–556, 558–559
- normal approximation to the binomial probability distribution, 360–364
- in one-way ANOVA, B14, B22
- properties of, 333–343
 statement of, 336–337
- technology to find, 350–351
- uniform probability distribution, 332, 333–335
 definition of, 333
- Normal probability plot, 355–358
 drawing, 355–358
 linearity of, 355
 technology to assess, 358
- Normal random variable, 337–338
 probability of, 345–346
 standardizing, 343
 value of, 347–350
- Normal score, 355
- Not so unusual events, 230
- Nouvelles méthodes pour la détermination des orbites des comètes* (Legendre), 189
- Null hypothesis, 435–438, 526, 529
 in ANOVA, B12
 assumption of trueness of, 444
 definition of, 436
 not rejecting versus accepting, 440–441
 structuring, 437
- Numerically summarizing data, 107–170
 boxplots, 157–160
 comparing two distributions using, 159–160
 constructing, 157–158
 distribution shape based on, 158–160
 technology to draw, 160
- five-number summary, 155–157
- measures of central tendency, 108–121
 from grouped data, 139–145
 median, 108, 110–111, 112–113, 116, 145, 155, 159
 midrange, 120
 mode, 108, 114–115
 shape of distribution from mean and median, 113–114
 trimmed mean, 120
- measures of dispersion, 121–143, 156
 Chebyshev's Inequality, 131–132
 Empirical Rule, 129–131
 from grouped data, 139–145
 mean absolute deviation, 137
 range, 122, 132, 156
 variance, 128–129, 132, 307
- measures of position, 145–155
 percentiles, 147–150, 345
 quartiles, 147–150, 151–152
 z -scores, 146–147, 355–358, 411–412, 414
- resistant statistic, 112–114
- O**
- Observational studies, 14–23, 62
 defined, 15
 experiment vs., 14–20
 types of, 18
- Odds ratio, 424
- One-tailed test(s), 436–437, 502–503
 of difference between two means, 493
 of difference between two means: independent samples, 502–503
 in least-squares regression model, 560
- One-way ANOVA, B12–B23
 conceptual understanding of, B15–B16
 decision rule in, B19
 equal population variances in, B12, B13
 normal distribution in, B14, B22

- requirements of, B14–B15
- robustness of, B14
- testing hypothesis with, B15–B22
- using technology, B20, B22–B23
- Open-ended tables, 78
- Open question, 41
- Ordinal level of measurement, 9
- Oswiecimski, Paul, 423
- Outcomes, 226, 227, 228
 - equally likely, 232, 234
- Outliers, 107, 150–151
 - quartiles to check, 151
- P**
- Parameters, 5, 108
- Pareto chart, 66
- Parsing, 19
- Pascal, Blaise, 235, 236, 306
- Pearson, Karl, 137, 174, 176, 282, 336, 531
- People Meter, 35
- Percentile(s), 147–150
 - interpreting, 147
 - k th, 147
 - quartiles, 147–150
 - ranks by, 345
 - value of normal random variable corresponding to, 347–349
- Permutations, 272–274, 278–279
 - computing, 273
 - using technology, 273, 279
 - definition of, 272
 - of distinct items, 277
 - with nondistinct items, 276–277
- Phone-in polling, 34
- Pie charts, 68–70
 - constructing, 68–69
 - drawing, 69–70
 - three-dimensional, 96–97
- Placebo, 45
- Point estimate
 - definition of, 396
 - of population mean, 410–411
 - of population proportion, 396
 - of two population means, 495
- Points, problem of, 236
- Point-slope form of line, 187, B3–B4
- Poisson probability distribution, 370
- Polar area diagram, 80
- Politics, statistics in, 2
- Polling, phone-in, 34
- Polling data, 6
- Pooled estimate, 480
- Pooled t -statistic, 507–508
- Pooling, 507
- Population, 5
 - mean of (i), 108–110
 - portion of, 338
- Population mean, 108–110, 139
 - confidence interval about, 495–496
 - confidence interval for, 397, 411–412, 415–418
 - technology for, 419
 - forming hypothesis about, 437–438
 - hypothesis testing about, 458–469
 - classical approach to, 459, 460–461, 462–463, 464
 - with large sample, 459–461, 464
 - P -value approach to, 459, 460–461, 462–463, 464
 - with small sample, 461–463
 - point estimate of, 410–411
 - sample size to estimate, within given margin of error, 418
- Population proportion(s), 338
 - confidence interval for, 396–404
 - constructing and interpreting, 396–404
 - point estimate for population proportion, 396
 - technology for, 403, 405–406
 - difference between two
 - confidence intervals, 483–485
 - McNemar's Test to compare two proportions from matched-pairs data, B6–B10
 - sample size requirements for, 485–486
 - using technology, 486–487
 - forming hypothesis about, 437–438
 - hypothesis testing for, 443–458
 - binomial probability distribution for, 452–453
 - classical approach using, 444–445, 446–451
 - left-tailed, 447
 - logic of, 443–447
 - P -value approach using, 445–451
 - technology in, 453
 - two-tailed, 449–451
 - using confidence interval, 451–452
 - hypothesis testing regarding two (independent samples), 479–483
 - classical approach to, 482–483
 - P -value approach to, 482–483
 - point estimate for, 396
 - pooled estimate of, 480
 - sample size determination, within specified margin of error, 404–405
 - sampling distribution of, 397–398
- Population size, sample size and, 387
- Population standard deviation(s), 123–125, 141
 - forming hypothesis about, 438
 - least-squares regression model and, 556
- Population variance, 128
 - in one-way ANOVA, B13, B14
- Population z -score, 146
- Position, measures of, 145–155
 - outliers, 107, 150–151
 - percentiles, 147–150
 - quartiles, 147–150
 - z -scores, 146–147
- Positively associated variables, 172
- Practical significance
 - definition of, 463
 - statistical significance vs., 463–464
- Prediction interval(s)
 - definition of, 569
 - for an individual response, 570–572
- Predictor (independent or explanatory) variable, 171
- Pretest-posttest (before-after) experiments, 49
- Probability(ies), 226–298
 - Addition Rule
 - with contingency tables, 246–247
 - General, 245–247
 - Addition Rule for Disjoint Events, 242–244, 253
 - Benford's Law and, 243–244
 - area as, 333
 - area under normal curve as, 338
 - at-least, 256–257
 - classical, 232–236, 299
 - Complement Rule, 247–249, 256
 - conditional, 260–269
 - definition of, 260
 - independence and, 265–266
 - using the General Multiplication Rule, 262–266
 - confidence interval and, 399
 - counting problems, 269–282
 - combinations for, 274–276, 278–279
 - Multiplication Rule for, 269–272

- Probability(ies) (*Continued*)
 permutations for, 272–274, 276–277, 279
 without repetition, 271
 defined, 228
 Empirical Method to approximate, 231–232, 234–236
 events and the sample space of probability experiment, 229
 to identify unusual events, 230
 Multiplication Rule for Independent Events, 254–256, 266
 relative frequency to approximate, 231
 rules of, 227–241, 257
 application of, 230–231
 determining appropriate, 287–289
 random processes and the law of large numbers, 227–229
 simulation to approximate, 282–287
 technology in, 285–286
 subjective, 236
 value of normal random variable corresponding to, 347, 350
- Probability density function (pdf), 333–334, 338
 normal, 338
- Probability distribution, 370. *See also* Normal probability distribution
 binomial, 370, 452–453
 cumulative, A7–A10
 table, A3–A10
 discrete, 301–303
 exponential, 424
 geometric, 326
 negative binomial, 326–327
 Poisson, 370
 testing claims regarding. *See* Contingency (two-way) table(s);
 Goodness-of-fit test
- Probability experiment, 229, 231, 232, 299
 binomial. *See* Binomial experiment
- Probability histograms of discrete probability distributions, binomial, 321–322
- Probability model, 230–231, 299
 for random variables. *See* Discrete probability distributions
 from survey data, 232
- Proportion(s). *See also* Population proportion(s); Sample proportion
 area under normal curve as, 338
 homogeneity of, 545–548
 definition of, 545
 steps in, 546–548
 value of normal random variable corresponding to, 347
- Prospective cohort studies, 178
- Prospective studies, 18
- P*-value approach to hypothesis testing, 445–451
 in chi-square test
 for homogeneity of proportions, 547–548
 for independence, 542, 543
 definition of, 446
 of difference between two means using independent samples, 502–503,
 504–505
 of difference between two population proportions
 from independent samples, 482–483
 McNemar's Test for, B8, B9
 goodness-of-fit, 528–529, 530, 532
 in least-squares regression model, 560–562
 logic of, 445–447
 of matched-pairs data, 491–492, 494
 in one-way ANOVA, B20
 about population mean, 459, 460–461, 462–463, 464
 two-tailed, 452n
- Q**
- Qualitative data, 8, 63–76. *See also* Categorical data
 bar graphs of, 64–68, 69–70
 frequency distribution of, 63–64
 relative, 64
 mode of, 115
 pie charts of, 68–70
 tables of, 63–64
- Qualitative variable, 7
 nominal or ordinal, 9
- Quantitative data, 8, 76–92
 dot plots of, 82, 85
 histograms of, 77–78, 80–81, 84–85
 mode of, 114–115
 shape of distribution of, 82–83
 tables of, 76–77, 78–80
 time-series graphs of, 83–84
 technology to draw, 84–85
- Quantitative variable, 7
 interval or ratio, 10
- Quartiles, 147–150
 checking for outliers using, 151
 using technology, 149, 151–152
- Questionnaires, 60
 ordering of questions or words in, 40
- Questions
 ordering of, 40
 type of, 41
 wording of, 40, 60
- Questions and Answers in Attitude Surveys* (Schuman and Presser), 40
- R**
- Random assignment, 283
- Random digit dialing (RDD) telephone surveys, 39
- Randomization, 47
- Random number generator, 26–27, 282
- Random numbers, table of, 25–26, A1
- Random processes, 227–229
 definition of, 227, 282
- Random sampling, 23–30, 32, 33, 34, 36
 combinations of samples, 276
 definition of, 23
 illustrating, 24–25
 obtaining sample, 24–28
- Random selection, 283–285
- Random values, Bernoulli, 286
- Random variable(s), 299, 370
 binomial, 313
 normal approximation to, 362–363
 continuous, 300–301
 probability density functions to find probabilities for, 333–334
 definition of, 300
 discrete, 300–312
 continuous random variables distinguished from, 300–301
 definition of, 300
 mean of, 303–308
 variance and standard deviation of, 306–308
 normal, 337–338
 probability of, 345–346
 standardizing, 343
 value of, 347–350
 probability models for. *See* Discrete probability distributions
 statistics as, 370
- Range, 122, 132
 computing, 122
 definition of, 122
 interquartile (IQR), 149–150, 156, 157
 technology to determine, 132
- Ratio level of measurement, 10
- Raw data, 62
 continuous, 80
 mean of variable from, 108–110
 median of a variable from, 110–111
 range of variable from, 122

- standard deviation of variable from, 123–128
 variance of variable from, 128–129
- Reformatting processes, 19
- Regression analysis, 192
- Relation between two variables, 170–233
 contingency tables and association, 206–217
 conditional distribution to identify association among categorical data, 208–212
 marginal distribution of a variable, 207–208
 Simpson's Paradox, 212–213
 using technology, 214
- correlation versus causation, 178–179
- least-squares regression line, 187–214
 coefficient of determination, 201–206
 definition of, 189
 equation of, 190
 finding, 187–192
 interpreting the slope and y -intercept of, 193–194
 sum of squared residuals, 194–195
- linear, determining, 177–178
- linear correlation coefficient, 173–178, 564
 computing and interpreting, 175–177
 definition of, 173
 properties of, 173–175
 using technology, 179–180
- scatter diagrams, 171–172
 definition of, 171
 drawing, 171–172, 176–177, 179–180
- testing for linear relation, 559–562
- Relative frequency(ies)
 association between two categorical variables and, 208–209
 probability using, 231
 of qualitative data, 64
- Relative frequency bar graph, 65–66
 side-by-side, 66–67, 68
 using horizontal bars, 67–68
- Relative frequency distribution, 64
 from continuous data, 79–81
 of discrete data, 76–77
 histogram of, 338
- Relative frequency marginal distributions, 208
- Replication, 47
- Research objective, identification of, 6
- Residual(s), 189
 normally distributed, 558–559
 sum of squared, 194–195
- Resistant statistic, 112–114
- Response bias, 39–41
- Response (dependent) variable, 15, 45, 46, 171, 209
- Retrospective studies, 18
- Rewards, nonresponse and, 39
- Rise, B1
- Risk, 135
- Robustness, 416, 492
 of least-squares regression model, 560
 of one-way ANOVA, B14
- Roman letters, use of, 108
- Roosevelt, Franklin D., 39
- Rounding, 8, 80
 for slope and intercept, 190
- Round-off error, 128
- Row variable, 207, 246
- Run, B1
- S**
- Sample(s)
 convenience, 6
 correlation coefficient, 173
 defined, 5
- matched-pairs (dependent), 478–479
 confidence intervals for, 495–496
 hypotheses testing for a population mean regarding, 491–494
 McNemar's Test to compare two proportions from, B6–B10
 mean (x -bar), 108–109
 self-selected (voluntary response), 34
- Sample mean, 108–109, 139
 sample size and, 373–374
 sampling distribution of, 371–384, 395
 definition of, 371
 describing, 374–379
 mean and standard deviation of, 374
 from nonnormal populations, 375–379
 from normal populations, 371–375
 shape of, 374
 standard deviation of, 395
- Sample proportion
 computing, 385
 definition of, 385
 sampling distribution of, 384–391
 describing, 385–387
 probabilities of, 388–389
 simulation to describe distribution of, 385–387
- Sample size, 35
 for difference of two population proportions, 485–486
 distribution shape and, 377–378
 hypothesis testing and, 452–453, 459–463, 464
 margin of error and, 404–405
 for population mean within given margin of error, 418
 for population proportion estimation within specified margin of error, 404–405
 population size and, 387
 sampling variability and, 373–374
 shape of the t -distribution and, 414
- Sample space, 229
- Sample standard deviation, 125–127, 141, 557
- Sample variance, 128–129
- Sample z -score, 146
- Sampling, 23–44
 acceptance, 263–264
 bias in, 38–42
 frame and, 38
 misrepresented answers, 40
 nonresponse bias, 39, 41
 ordering of questions or words, 40
 response bias, 39–41
 sampling bias, 38–39
 wording of questions, 40
- cluster, 33–34, 36
- convenience, 34–35
- dependent, 478–479
- goal of, 30, 38
- independent, 478–479
- interviewer error in, 39–40
- multistage, 35
- with replacement, 25
- without replacement, 25
- sample size considerations, 35
- simple random, 23–30, 32, 33, 36
 combinations of, 276
 definition of, 24
 illustrating, 24–25
 obtaining sample, 24–28
- stratified, 30–32, 34, 36
- systematic, 32–33, 36
- Sampling distribution(s), 370–395
 of difference of two means, 502
 in least-squares regression model, 555
 of median, 394–395

- Sampling distribution(s) (*Continued*)
 of population proportion, 398
 of sample mean, 371–384, 395
 definition of, 371
 describing, 374–379
 mean and standard deviation of, 374
 from nonnormal populations, 375–379
 from normal populations, 371–375
 shape of, 374
 of sample proportion, 384–391
 describing, 385–387
 probabilities of, 388–389
- Sampling error, 41
- Scatter diagrams, 171–172
 definition of, 171
 drawing, 171–172, 176–177, 179–180
- Seed, 26–27
- Segmented (stacked) bar graphs, 211, 214
- Self-selected samples, 34
- Side-by-side bar graph, 66–67
 using horizontal bars, 67–68
- Sigma (Σ), 109
- Significance
 of least-squares regression model, 554–564
 level of, 440
 practical
 definition of, 463
 statistical vs., 463–464
 statistical, 444
 definition of, 444
 practical vs., 463–464
- Type I error and, 440
- Simple events, 229
- Simple random sample, 23–30, 32, 33, 36
 combinations of, 276
 definition of, 24
 designed experiment and, 50
 illustrating, 24–25
 obtaining, 24–28
- Simpson's Paradox, 212–213
- Simulation, 227, 282–287
 standard normal distribution compared to *t*-distribution using, 412–413
 using technology, 285–286
- Single-blind experiment, 45
- Skewed distributions, 82, 112–113
 mean or median versus skewness, 112–113
 quartile for, 158–159
- Skewness, 158–159
 coefficient of, 137
- Slope, 190, B1–B3, B5
 calculating and interpreting, B1–B2
 definition of, B1
 graphing line using, B2
 of least-squares regression model, 193–194
 confidence interval about, 562–564
 hypothesis test, 564
 of least-squares regression model, inference on, 559–562
- Slope–intercept form of an equation of a line, B5
- Sports, statistics in, 2
- Stacked (segmented) bar graphs, 211, 214
- Standard deviation, 123–128, 132, 176
 of binomial random variable, 320
 of discrete random variables, 306–308
 technology to find, 307–308
 from grouped data, 141–143
 interpretations of, 127
 outlier distortion of, 150
- population, 123–125, 141
 least-squares regression model and, 556
- sample, 125–127, 141, 557
 of sample mean, 395
 of sampling distribution of sample mean, 374
 of two data sets, 128
 unusual results in binomial experiment and, 323
 using technology, 126–127, 143
- Standard error, 558
 computing, 557–558
 definition of, 557
 of the mean, 374
- Standard normal probability distribution, 343
 table, A11–A12
- StatCrunch, 27, 28
 area under the normal curve using, 351
 normal values corresponding to, 351
 bar graph using, 70
 Bernoulli random values using, 286
 binomial probabilities using, 319–320, 324
 boxplots using, 160
 chi-square tests using, 549
 coefficient of determination using, 204
 combinations using, 279
 confidence intervals using, 572
 for population mean, 419
 for population proportion, 406
 for the slope of the true regression line using, 563
 correlation coefficient using, 180
 expression computation using, 286
 factorials using, 279
 goodness-of-fit test, 533
 in hypothesis testing
 about population mean, 464
 about population proportion, 453
 inference of two population proportions using, 487
 least-squares regression model using, 191–192, 195, 564
 marginal and conditional distributions using, 214
 McNemar's Test, B10
 mean using, 111, 143, 307
 sample size determination for, 419
 median using, 111, 143
 normal probability plot using, 357–358
 one-way ANOVA using, B20, B23
 permutations using, 279
 pie charts using, 70
 prediction intervals using, 572
 quartiles using, 152
 random integers using, 286
 scatter diagrams using, 180
 testing claims about matched-pairs data using, 494
 time-series plots using, 85
 two-sample *t*-tests using
 dependent sampling, 496
 independent sampling, 508
- Statistic, 5
 biased, 129
 defined, 5
 parameter versus, 5
 as random variable, 370
 resistant, 112–114
 sample mean as, 108
- Statistical Abstract of the United States*, 236
- Statistically significant, defined, 444
- Statistical significance, 444
 practical significance vs., 463–464
- Statistical spreadsheets. *See also* Excel
 pie charts on, 68
 time-series on, 84–85

- Statistical thinking, 3–4
 Statistics, 3–14
 - definition of, 3–4
 - descriptive, 5, 6
 - inferential, 5, 6, 47
 - mathematics vs., 4
 - process of, 4–6
 - roles in everyday life of, 2
 - variables in, 6–10
 - data vs., 8–9
 - discrete vs. continuous, 7–8
 - qualitative (categorical) vs. quantitative, 7
- Status quo statement, 436, 437
 Strata, 31
 Stratified sampling, 30–32, 36
 Subject (experimental unit), 45, 47
 - in matched-pairs design, 49, 50
- Subjective probability, 236
 Summers, Lawrence, 136
 Sum of squared residuals, 194–195
 Sum of squares, B17, B19
 - due to error, B17
 - total (SS (Total)), B17
 - due to treatment (SST), B17
- Survey data, probability model from, 232
 Surveys
 - American Time Use Survey, 370, 394
 - classroom, 41–42
 - Current Population Survey, 371
 - Internet, 34
 - random digit dialing (RDD) telephone, 39
 - random sample, 23–30
- Suzuki, Ichiro, 309
 Symmetric distributions, 82, 112
 Systematic sampling, 33–34, 36
- T**
- Tables, 106, 107
 - binomial, 318–319
 - continuous data in, 78–80
 - discrete data in, 76–77
 - open-ended, 78
 - qualitative data in, 63–64
- t*-distribution, 411–415, 458–459
 - finding values, 414–415
 - hypothesis testing and, 458–459
 - normality condition for, 416
 - properties of, 411–414
 - sample size and, 414
 - standard normal distribution compared to, 412–413
 - statement of, 412
 - table, A14, A20–A22
- Technology. *See also* Excel; Minitab; StatCrunch; Statistical spreadsheets; TI-83/84 Plus and TI-84 Plus CE graphing calculators
 - TI-83/84 Plus and TI-84 Plus CE graphing calculators
 - ANOVA using
 - one-way, B20–B21, B22–B23
 - binomial probabilities using, 319–320, 323–324
 - boxplots using, 160
 - chi-square tests using, 542, 543–544, 548, 549
 - coefficient of determination using, 204
 - combinations using, 275, 279
 - confidence intervals using, 571–572
 - for matched-pairs data, 495–496
 - for population mean, 417, 419
 - for population proportion, 403–404, 405–406
 - for slope of the true regression line, 563–564
 - difference between two means using, 505, 508
 - difference between two population proportions using, 486–487
 - exact *P*-values using, 530–531
 - factorials using, 279
 - five-number summary using, 156–157
 - goodness-of-fit test using, 530–531, 532–533
 - in hypothesis testing
 - about population mean, 464
 - about population proportion, 453
 - least-squares regression line using, 191–192
 - least-squares regression model using, 561–562, 564
 - linear correlation coefficient using, 176–177, 179–180
 - McNemar's Test, B10
 - mean and standard deviation using, 143, 307–308
 - normal probability distribution using, 350–351
 - normal probability plot using, 358
 - normal random variable using, 347–348
 - permutations using, 273, 279
 - prediction intervals using, 571–572
 - probabilities of a sample proportion using, 388
 - scatter diagram using, 176–177, 179–180
 - simple random sample using, 27–28
 - simulation using, 285–286
 - standard error using, 558
 - testing claims about matched-pairs data using, 494, 496
 - t*-value using, 415
 - two-sample *t*-tests, independent sampling, 508
 - TI-83/84 Plus and TI-84 Plus CE graphing calculators, 27
 - area under normal curve using, 350
 - binomcdf* command, 362
 - binomial probabilities using, 323
 - boxplots using, 160
 - chi-square tests using, 549
 - coefficient of determination using, 204
 - combinations using, 275, 279
 - confidence intervals using, 572
 - for population mean, 419
 - for population proportion, 405
 - correlation coefficient using, 179
 - factorials using, 279
 - goodness-of-fit test, 532
 - in hypothesis testing
 - about population mean, 464
 - about population proportion, 453
 - inference between two population proportions using, 486
 - invT feature, 415
 - least-squares regression line using, 191–192, 195
 - least-squares regression model using, 564
 - McNemar's Test, B10
 - mean and median on, 116
 - mean and standard deviation using
 - approximation, 143
 - from grouped data, 143
 - normal probability plot using, 358
 - normal random variable using, 347–348
 - one-way ANOVA using, B20, B22
 - permutations using, 273, 279
 - prediction intervals using, 572
 - quartiles using, 149, 151
 - random integers using, 285
 - role of level of confidence on margin of error using, 403
 - scatter diagrams using, 179
 - standard deviation using, 126–127, 307
 - time-series plots using, 84
 - two-sample *t*-tests using
 - dependent sampling, 496
 - independent sampling, 508
 - z*-value for area under the standard normal curve using, 350
 - Time-series graphs, 83–84
 - technology to draw, 84
 - t*-interval, 415
 - Total deviation, 202

Total sum of squares (SS (Total)), B17

Toxic Epidermal Necrolysis (TEN), 327

Treatment, 45

Tree diagram, 235

Trials, 313, 360

Trimmed mean, 120

t-statistic, 412

pooled, 507–508

two-sample, 507–508

Welch's approximate, 502

Tufte, Edward, 97

Tukey, John, 155

Twain, Mark, 200

Two-tailed tests, 436

of difference between two means, 491–492

of difference of two means: independent samples, 502–503

in least-squares regression model, 559–562

Type I error, 438–440, 445, 506

in ANOVA, B13

probability of, 440, 445

Type II error, 438–440

probability of, 440

U

Ulam, Stanislas, 282

Undercoverage, 38

Unexplained deviation, 202

Uniform density function, 335

Uniform probability distribution, 82, 332, 333–335

definition of, 333

Unimodal instruction, 511

United States Census, 19

Univariate data, 170

Unusual events, 230

Upper class limit, 78

V

Value, expected, 305–306

Variability, 137

between-sample, B16

within-sample, B16

Variable(s), 6–10. *See also* Random variable(s)

associated, 172, 176

column, 207, 246

data vs., 8–9

defined, 6

dependent (response), 171

discrete vs. continuous, 7–8

explanatory, 15, 45, 209

independent (explanatory or predictor), 171

level of measurement of, 9–10

linear relation between two, determining, 177–178

lurking, 4, 16–17, 178, 212–213

marginal distribution of, 207–208

modal class of, 145

relation between two. *See* Relation between two variables

response, 15, 45, 46

row, 207, 246

Variance, 128–129, 132

of discrete random variables, 307

population, 128

sample, 128–129

technology to determine, 132

Variation, coefficient of, 138

Venn diagram, 245, 248

Venn diagrams, 242

Visual Display of Quantitative Information, The (Tufte), 97

Voluntary response samples, 34

von Neumann, John, 282

Vos Savant, Marilyn, 241

W

Web scraping (data mining), 19–20

Weighted mean, 140–141

Welch, Bernard Lewis, 502

Welch's approximate *t*, 502

Whiskers, 157

Wiles, Andrew, 235

Within-sample variability, B16

Wording of questions, 40

X

XLSTAT, 69

Y

y-intercept, 190, 193–194, B5

Z

z-score, 146–147, 411–412, 414

comparing, 146–147

expected, 355–358

population, 146

sample, 146

for specified area to the right, 350

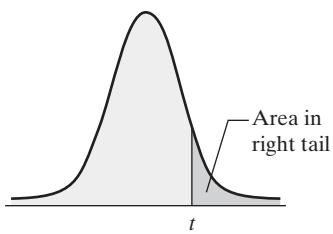


Table VII

df	t-Distribution Area in Right Tail											
	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.894	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
31	0.682	0.853	1.054	1.309	1.696	2.040	2.144	2.453	2.744	3.022	3.375	3.633
32	0.682	0.853	1.054	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.365	3.622
33	0.682	0.853	1.053	1.308	1.692	2.035	2.138	2.445	2.733	3.008	3.356	3.611
34	0.682	0.852	1.052	1.307	1.691	2.032	2.136	2.441	2.728	3.002	3.348	3.601
35	0.682	0.852	1.052	1.306	1.690	2.030	2.133	2.438	2.724	2.996	3.340	3.591
36	0.681	0.852	1.052	1.306	1.688	2.028	2.131	2.434	2.719	2.990	3.333	3.582
37	0.681	0.851	1.051	1.305	1.687	2.026	2.129	2.431	2.715	2.985	3.326	3.574
38	0.681	0.851	1.051	1.304	1.686	2.024	2.127	2.429	2.712	2.980	3.319	3.566
39	0.681	0.851	1.050	1.304	1.685	2.023	2.125	2.426	2.708	2.976	3.313	3.558
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
70	0.678	0.847	1.044	1.294	1.667	1.994	2.093	2.381	2.648	2.899	3.211	3.435
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
90	0.677	0.846	1.042	1.291	1.662	1.987	2.084	2.368	2.632	2.878	3.183	3.402
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
<i>z</i>	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.291

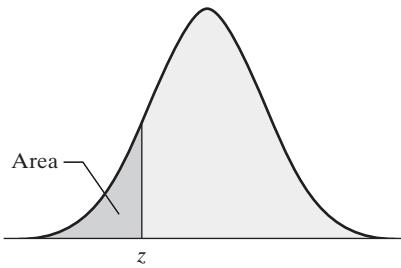


Table V

<i>z</i>	Standard Normal Distribution									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-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

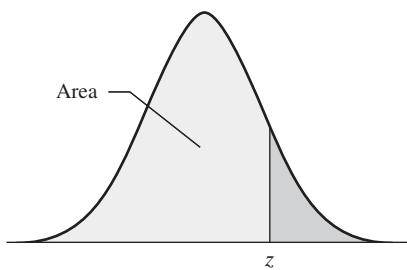


Table V (continued)

Chapter 2 Organizing and Summarizing Data

- Relative frequency = $\frac{\text{frequency}}{\text{sum of all frequencies}}$
- Class midpoint: The sum of consecutive lower class limits divided by 2.

Chapter 3 Numerically Summarizing Data

- Population Mean: $\mu = \frac{\sum x_i}{N}$
- Sample Mean: $\bar{x} = \frac{\sum x_i}{n}$
- Range = Largest Data Value – Smallest Data Value
- Population Standard Deviation:
$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{N}}{N}}$$
- Sample Standard Deviation
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}}$$
- Population Variance: σ^2
- Sample Variance: s^2
- Empirical Rule:** If the shape of the distribution is bell-shaped, then
 - Approximately 68% of the data lie within 1 standard deviation of the mean
 - Approximately 95% of the data lie within 2 standard deviations of the mean
 - Approximately 99.7% of the data lie within 3 standard deviations of the mean
- Population Standard Deviation from Grouped Data:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2 f_i}{\sum f_i}} = \sqrt{\frac{\sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{\sum f_i}}{\sum f_i}}$$
- Sample Standard Deviation from Grouped Data:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{(\sum f_i) - 1}} = \sqrt{\frac{\sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{\sum f_i}}{\sum f_i - 1}}$$
- Population z -score: $z = \frac{x - \mu}{\sigma}$
- Sample z -score: $z = \frac{x - \bar{x}}{s}$
- Interquartile Range: $IQR = Q_3 - Q_1$
- Lower and Upper Fences: Lower fence = $Q_1 - 1.5(IQR)$
Upper fence = $Q_3 + 1.5(IQR)$
- Five-Number Summary

Minimum, Q_1 , M , Q_3 , Maximum

Chapter 4 Describing the Relation between Two Variables

- Correlation Coefficient: $r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)}{n-1}$
- The equation of the least-squares regression line is

$$\hat{y} = b_1 x + b_0$$
, where \hat{y} is the predicted value, $b_1 = r \cdot \frac{s_y}{s_x}$
 is the slope, and $b_0 = \bar{y} - b_1 \bar{x}$ is the intercept.
- Residual = observed y – predicted y = $y - \hat{y}$
- $R^2 = r^2$ for the least-squares regression model

$$\hat{y} = b_1 x + b_0$$
- The coefficient of determination, R^2 , measures the proportion of total variation in the response variable that is explained by the least-squares regression line.

Chapter 5 Probability

- Empirical Probability
$$P(E) \approx \frac{\text{frequency of } E}{\text{number of trials of experiment}}$$
- Classical Probability
$$P(E) = \frac{\text{number of ways that } E \text{ can occur}}{\text{number of possible outcomes}} = \frac{N(E)}{N(S)}$$
- Addition Rule for Disjoint Events

$$P(E \text{ or } F) = P(E) + P(F)$$
- Addition Rule for n Disjoint Events

$$P(E \text{ or } F \text{ or } G \text{ or } \dots) = P(E) + P(F) + P(G) + \dots$$
- General Addition Rule

$$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$$

- Complement Rule

$$P(E^c) = 1 - P(E)$$

- Multiplication Rule for Independent Events

$$P(E \text{ and } F) = P(E) \cdot P(F)$$

- Multiplication Rule for n Independent Events

$$P(E \text{ and } F \text{ and } G \dots) = P(E) \cdot P(F) \cdot P(G) \cdot \dots$$

- Conditional Probability Rule

$$P(F|E) = \frac{P(E \text{ and } F)}{P(E)} = \frac{N(E \text{ and } F)}{N(E)}$$

- General Multiplication Rule

$$P(E \text{ and } F) = P(E) \cdot P(F|E)$$

Chapter 6 Discrete Probability Distributions

- Mean (Expected Value) of a Discrete Random Variable

$$\mu_X = \sum x \cdot P(x)$$

- Standard Deviation of a Discrete Random Variable

$$\sigma_X = \sqrt{\sum (x - \mu)^2 \cdot P(x)} = \sqrt{\sum [x^2 P(x)] - \mu_X^2}$$

- Factorial

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 3 \cdot 2 \cdot 1$$

- Permutation of n objects taken r at a time:

$${}_nP_r = \frac{n!}{(n-r)!}$$

- Combination of n objects taken r at a time:

$${}_nC_r = \frac{n!}{r!(n-r)!}$$

- Permutations with Repetition:

$$\frac{n!}{n_1! \cdot n_2! \cdot \dots \cdot n_k!}$$

Chapter 7 The Normal Distribution

- Standardizing a Normal Random Variable

$$z = \frac{x - \mu}{\sigma}$$

- Finding the Score: $x = \mu + z\sigma$

Chapter 8 Sampling Distributions

- Mean and Standard Deviation of the Sampling Distribution of \bar{x}

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Sample Proportion: $\hat{p} = \frac{x}{n}$

- Mean and Standard Deviation of the Sampling Distribution of \hat{p}

$$\mu_{\hat{p}} = p \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Chapter 9 Estimating the Value of a Parameter

Confidence Intervals

- A $(1 - \alpha) \cdot 100\%$ confidence interval about p is

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- A $(1 - \alpha) \cdot 100\%$ confidence interval about μ is

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Note: $t_{\alpha/2}$ is computed using $n - 1$ degrees of freedom.

Sample Size

- To estimate the population proportion with a margin of error E at a $(1 - \alpha) \cdot 100\%$ level of confidence:

$$n = \hat{p}(1 - \hat{p}) \left(\frac{z_{\alpha/2}}{E} \right)^2 \text{ rounded up to the next integer,}$$

where \hat{p} is a prior estimate of the population proportion, or $n = 0.25 \left(\frac{z_{\alpha/2}}{E} \right)^2$ rounded up to the next integer when no prior estimate of p is available.

- To estimate the population mean with a margin of error E at a $(1 - \alpha) \cdot 100\%$ level of confidence: $n = \left(\frac{z_{\alpha/2} \cdot s}{E} \right)^2$ rounded up to the next integer.

Chapter 10 Hypothesis Tests Regarding a Parameter

Test Statistics

- $$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- $$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Chapter 11 Inferences on Two Samples

- Test Statistic Comparing Two Population Proportions (Independent Samples)

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{where } \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

- Confidence Interval for the Difference of Two Proportions (Independent Samples)

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- Test Statistic for Matched-Pairs Data

$$t_0 = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$$

where \bar{d} is the mean and s_d is the standard deviation of the differenced data.

- Confidence Interval for Matched-Pairs Data

$$\bar{d} \pm t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}}$$

Note: $t_{\alpha/2}$ is found using $n - 1$ degrees of freedom.

- Test Statistic Comparing Two Means (Independent Sampling)

$$t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Confidence Interval for the Difference of Two Means (Independent Samples)

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Note: $t_{\alpha/2}$ is found using the smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom.

Chapter 12 Additional Inferential Procedures

- Expected Counts (when testing for goodness of fit)

$$E_i = \mu_i = np_i \quad \text{for } i = 1, 2, \dots, k$$

- Expected Frequencies (when testing for independence or homogeneity of proportions)

$$\text{Expected frequency} = \frac{(\text{row total})(\text{column total})}{\text{table total}}$$

- Chi-Square Test Statistic

$$\chi_0^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$i = 1, 2, \dots, k$$

All $E_i \geq 1$ and no more than 20% less than 5.

- Standard Error of the Estimate

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum \text{residuals}^2}{n-2}}$$

- Standard error of b_1

$$s_{b_1} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- Test Statistic for the Slope of the Least-Squares Regression Line

$$t_0 = \frac{b_1 - \beta_1}{s_e / \sqrt{\sum (x_i - \bar{x})^2}} = \frac{b_1 - \beta_1}{s_{b_1}}$$

- Confidence Interval for the Slope of the Regression Line

$$b_1 \pm t_{\alpha/2} \cdot \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

where $t_{\alpha/2}$ is computed with $n - 2$ degrees of freedom.

- Confidence Interval about the Mean Response of y , \hat{y}

$$\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

where x^* is the given value of the explanatory variable and $t_{\alpha/2}$ is the critical value with $n - 2$ degrees of freedom.

- Prediction Interval about an Individual Response, \hat{y}

$$\hat{y} \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

where x^* is the given value of the explanatory variable and $t_{\alpha/2}$ is the critical value with $n - 2$ degrees of freedom.

This page intentionally left blank

Table I

Row Number	Random Numbers									
	Column Number									
01–05	06–10	11–15	16–20	21–25	26–30	31–35	36–40	41–45	46–50	
01	89392	23212	74483	36590	25956	36544	68518	40805	09980	00467
02	61458	17639	96252	95649	73727	33912	72896	66218	52341	97141
03	11452	74197	81962	48443	90360	26480	73231	37740	26628	44690
04	27575	04429	31308	02241	01698	19191	18948	78871	36030	23980
05	36829	59109	88976	46845	28329	47460	88944	08264	00843	84592
06	81902	93458	42161	26099	09419	89073	82849	09160	61845	40906
07	59761	55212	33360	68751	86737	79743	85262	31887	37879	17525
08	46827	25906	64708	20307	78423	15910	86548	08763	47050	18513
09	24040	66449	32353	83668	13874	86741	81312	54185	78824	00718
10	98144	96372	50277	15571	82261	66628	31457	00377	63423	55141
11	14228	17930	30118	00438	49666	65189	62869	31304	17117	71489
12	55366	51057	90065	14791	62426	02957	85518	28822	30588	32798
13	96101	30646	35526	90389	73634	79304	96635	06626	94683	16696
14	38152	55474	30153	26525	83647	31988	82182	98377	33802	80471
15	85007	18416	24661	95581	45868	15662	28906	36392	07617	50248
16	85544	15890	80011	18160	33468	84106	40603	01315	74664	20553
17	10446	20699	98370	17684	16932	80449	92654	02084	19985	59321
18	67237	45509	17638	65115	29757	80705	82686	48565	72612	61760
19	23026	89817	05403	82209	30573	47501	00135	33955	50250	72592
20	67411	58542	18678	46491	13219	84084	27783	34508	55158	78742

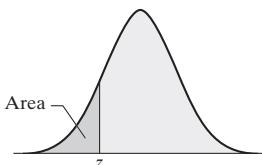
Table II

Critical Values (CV) for Correlation Coefficient							
<i>n</i>	CV	<i>n</i>	CV	<i>n</i>	CV	<i>n</i>	CV
3	0.997	10	0.632	17	0.482	24	0.404
4	0.950	11	0.602	18	0.468	25	0.396
5	0.878	12	0.576	19	0.456	26	0.388
6	0.811	13	0.553	20	0.444	27	0.381
7	0.754	14	0.532	21	0.433	28	0.374
8	0.707	15	0.514	22	0.423	29	0.367
9	0.666	16	0.497	23	0.413	30	0.361

Table VI

Critical Values for Normal Probability Plots					
Sample Size, <i>n</i>	Critical Value	Sample Size, <i>n</i>	Critical Value	Sample Size, <i>n</i>	Critical Value
5	0.880	13	0.932	21	0.952
6	0.888	14	0.935	22	0.954
7	0.898	15	0.939	23	0.956
8	0.906	16	0.941	24	0.957
9	0.912	17	0.944	25	0.959
10	0.918	18	0.946	30	0.960
11	0.923	19	0.949		
12	0.928	20	0.951		

Source: S. W. Looney and T. R. Gulledge, Jr. "Use of the Correlation Coefficient with Normal Probability Plots," *American Statistician* 39(Feb. 1985): 75–79.


Table V
Standard Normal Distribution

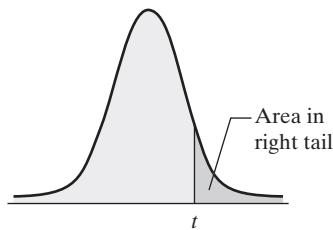
<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	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
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	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964	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

Confidence Interval Critical Values, $z_{\alpha/2}$

Level of Confidence	Critical Value, $z_{\alpha/2}$
0.90 or 90%	1.645
0.95 or 95%	1.96
0.98 or 98%	2.33
0.99 or 99%	2.575

Hypothesis Testing Critical Values

Level of Significance, α	Left-Tailed	Right-Tailed	Two-Tailed
0.10	-1.28	1.28	± 1.645
0.05	-1.645	1.645	± 1.96
0.01	-2.33	2.33	± 2.575

**Table VII**
***t*-Distribution
Area in Right Tail**

df	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.894	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
31	0.682	0.853	1.054	1.309	1.696	2.040	2.144	2.453	2.744	3.022	3.375	3.633
32	0.682	0.853	1.054	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.365	3.622
33	0.682	0.853	1.053	1.308	1.692	2.035	2.138	2.445	2.733	3.008	3.356	3.611
34	0.682	0.852	1.052	1.307	1.691	2.032	2.136	2.441	2.728	3.002	3.348	3.601
35	0.682	0.852	1.052	1.306	1.690	2.030	2.133	2.438	2.724	2.996	3.340	3.591
36	0.681	0.852	1.052	1.306	1.688	2.028	2.131	2.434	2.719	2.990	3.333	3.582
37	0.681	0.851	1.051	1.305	1.687	2.026	2.129	2.431	2.715	2.985	3.326	3.574
38	0.681	0.851	1.051	1.304	1.686	2.024	2.127	2.429	2.712	2.980	3.319	3.566
39	0.681	0.851	1.050	1.304	1.685	2.023	2.125	2.426	2.708	2.976	3.313	3.558
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
70	0.678	0.847	1.044	1.294	1.667	1.994	2.093	2.381	2.648	2.899	3.211	3.435
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
90	0.677	0.846	1.042	1.291	1.662	1.987	2.084	2.368	2.632	2.878	3.183	3.402
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.291

Table VIII

Degrees of Freedom	Chi-Square (χ^2) Distribution Area to the Right of Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	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
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

