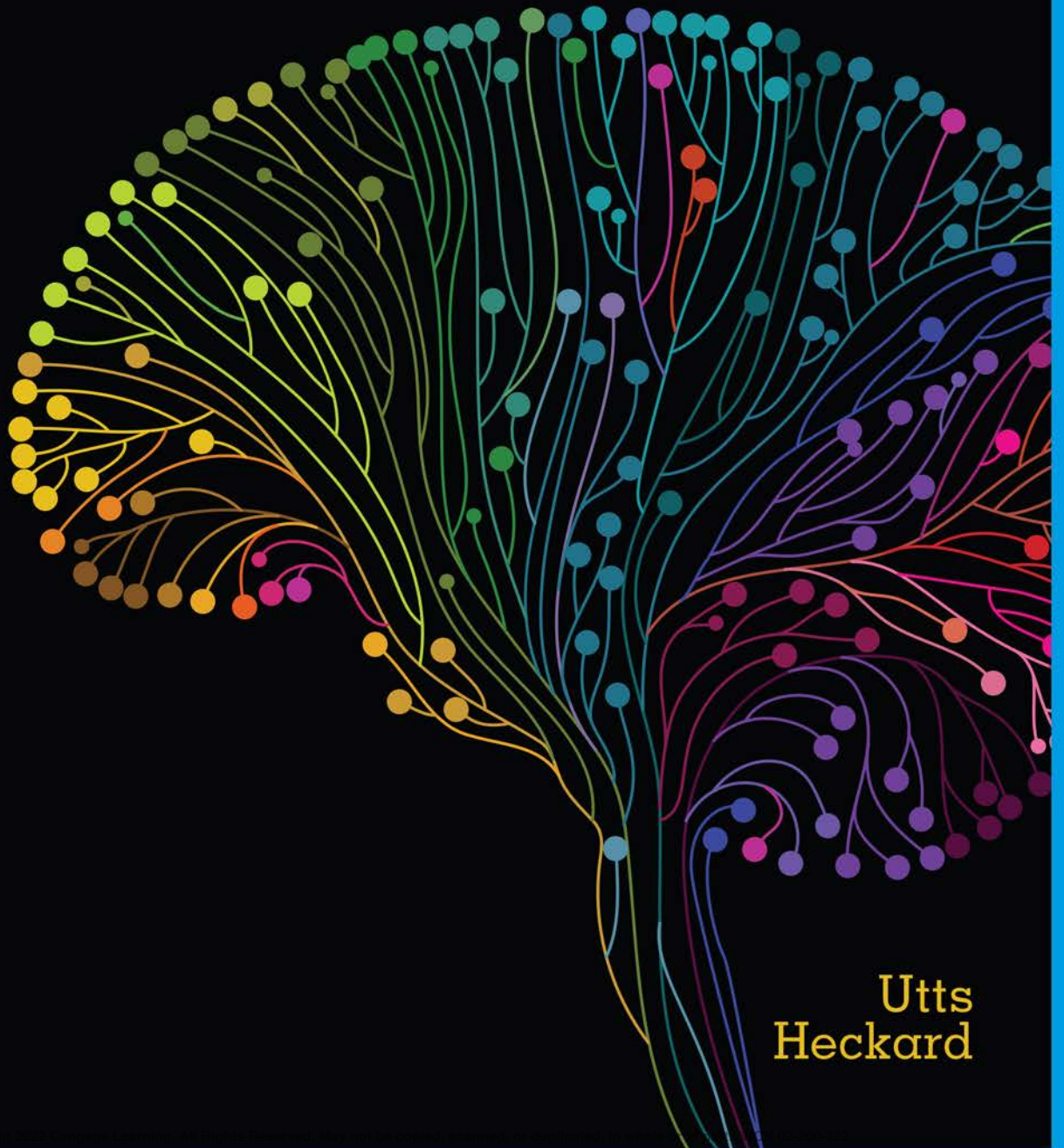


Mind on Statistics

6th Edition



Utts
Heckard

Mind on Statistics

Sixth Edition

Jessica M. Utts

University of California, Irvine

Robert F. Heckard

Pennsylvania State University



Australia • Brazil • Canada • Mexico • Singapore • United Kingdom • United States

Copyright 2022 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. WCN 02-200-322

Copyright 2022 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.

This is an electronic version of the print textbook. Due to electronic rights restrictions, some third party content may be suppressed. Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. The publisher reserves the right to remove content from this title at any time if subsequent rights restrictions require it. For valuable information on pricing, previous editions, changes to current editions, and alternate formats, please visit www.cengage.com/highered to search by ISBN#, author, title, or keyword for materials in your areas of interest.

Important Notice: Media content referenced within the product description or the product text may not be available in the eBook version.

Mind on Statistics, Sixth Edition
Jessica M. Utts and Robert F. Heckard

Product Director: Mark Santee

Product Manager: Andy Trus

Product Assistant: Kyra Kruger

Marketing Manager: Adam Kiszka

Associate Content Manager: Amanda Rose

Learning Designer: Elinor Gregory

IP Analyst: Ashley Maynard

IP Project Manager: Carly Belcher

Production Service: MPS Limited

Compositor: MPS Limited

Senior Designer: Angela Sheehan

Text Designer: Joe Devine, Red Hanger Design

Cover Designer: Angela Sheehan

Cover Image: VicW/Shutterstock.com

© 2022, 2015, 2012 Cengage Learning, Inc.

Unless otherwise noted, all content is © Cengage.

ALL RIGHTS RESERVED. No part of this work covered by the copyright herein may be reproduced or distributed in any form or by any means, except as permitted by U.S. copyright law, without the prior written permission of the copyright owner.

For product information and technology assistance, contact us at
Cengage Customer & Sales Support, 1-800-354-9706 or
support.cengage.com.

For permission to use material from this text or product,
submit all requests online at
www.cengage.com/permissions.

Library of Congress Control Number: 2020915913

Student Edition:

ISBN: 978-1-337-79360-5

Loose-leaf Edition:

ISBN: 978-1-337-79477-0

Cengage

200 Pier 4 Boulevard
Boston, MA 02210
USA

Cengage is a leading provider of customized learning solutions with employees residing in nearly 40 different countries and sales in more than 125 countries around the world. Find your local representative at **www.cengage.com.**

To learn more about Cengage platforms and services, register or access your online learning solution, or purchase materials for your course, visit **www.cengage.com.**

Printed in the United States of America
Print Number: 01 Print Year: 2020

*To Bill Harkness—energetic, generous, and innovative
educator, guide, and friend—who launched our careers
in statistics and continues to share his vision.*

And

*To our students, from whom we learned so much,
and who taught us how to be better teachers.*

Brief Contents

1	Statistics Success Stories and Cautionary Tales	3
2	Turning Data into Information	17
3	Relationships Between Quantitative Variables	71
4	Relationships Between Categorical Variables	117
5	Sampling: Surveys and How to Ask Questions	159
6	Gathering Useful Data for Examining Relationships	203
7	Probability	233
8	Random Variables	281
9	Understanding Sampling Distributions: Statistics as Random Variables	337
10	Estimating Proportions with Confidence	397
11	Estimating Means with Confidence	441
12	Testing Hypotheses About Proportions	497
13	Testing Hypotheses About Means	567
14	Inference About Simple Regression	623
15	More About Inference for Categorical Variables	661
16	Analysis of Variance	701
17	Turning Information into Wisdom	735

Contents

Preface xii

1 Statistics Success Stories and Cautionary Tales 3

- 1.1** What Is Statistics? 3
- 1.2** Eight Statistical Stories with Morals 4
- 1.3** The Common Elements in the Eight Stories 10
 - Key Terms 11
 - Exercises 12

2 Turning Data into Information 17

- 2.1** Raw Data 17
- 2.2** Types of Variables 19
- 2.3** Summarizing One or Two Categorical Variables 22
- 2.4** Exploring Features of Quantitative Data with Pictures 26
- 2.5** Numerical Summaries of Quantitative Variables 39
- 2.6** How to Handle Outliers 47
- 2.7** Bell-Shaped Distributions and Standard Deviations 48
 - Key Terms 54
 - In Summary Boxes 55
 - Simulations for Further Exploration 55
 - Exercises 56

3 Relationships Between Quantitative Variables 71

- 3.1** Looking for Patterns with Scatterplots 72
- 3.2** Describing Linear Patterns with a Regression Line 76
- 3.3** Measuring Strength and Direction with Correlation 85
- 3.4** Regression and Correlation Difficulties and Disasters 92
- 3.5** Correlation Does Not Prove Causation 97
 - Key Terms 101
 - In Summary Boxes 101
 - Simulations for Further Exploration 101
 - Exercises 102

4	Relationships Between Categorical Variables	117
4.1	Displaying Relationships Between Categorical Variables	117
4.2	Risk, Relative Risk, and Misleading Statistics About Risk	122
4.3	The Effect of a Third Variable and Simpson's Paradox	127
4.4	Assessing the Relationship in a 2×2 Table: Hypothesis Testing	128
4.5	Randomization Test for a 2×2 Table	139
	Key Terms	143
	In Summary Boxes	144
	Simulations for Further Exploration	144
	Exercises	144
5	Sampling: Surveys and How to Ask Questions	159
5.1	Collecting and Using Sample Data Wisely	159
5.2	Margin of Error, Confidence Intervals, and Sample Size	163
5.3	Choosing a Simple Random Sample	167
5.4	Additional Probability Sampling Methods	169
5.5	Difficulties and Disasters in Sampling	178
5.6	Pitfalls in Asking Survey Questions	183
	Key Terms	191
	In Summary Boxes	191
	Simulations for Further Exploration	191
	Exercises	191
6	Gathering Useful Data for Examining Relationships	203
6.1	Speaking the Language of Research Studies	203
6.2	Designing a Good Experiment	208
6.3	Designing a Good Observational Study	217
6.4	Difficulties and Disasters in Experiments and Observational Studies	219
	Key Terms	224
	In Summary Boxes	224
	Exercises	224
7	Probability	233
7.1	Random Circumstances	233
7.2	Interpretations of Probability	235
7.3	Probability Definitions and Relationships	240
7.4	Basic Rules for Finding Probabilities	245
7.5	Conditional Probabilities and Bayes' Rule	252
7.6	Using Simulation to Estimate Probabilities	259
7.7	Flawed Intuitive Judgments About Probability	261
	Key Terms	268
	In Summary Boxes	268

Simulations for Further Exploration	268
Exercises	269

8 Random Variables 281

8.1	What Is a Random Variable?	281
8.2	Discrete Random Variables	284
8.3	Expectations for Random Variables	289
8.4	Binomial Random Variables	294
8.5	Continuous Random Variables	299
8.6	Normal Random Variables	301
8.7	Approximating Binomial Distribution Probabilities	312
8.8	Linear Combinations and Linear Transformations of Random Variables	316
	Key Terms	323
	In Summary Boxes	323
	Simulations for Further Exploration	323
	Exercises	324

9 Understanding Sampling Distributions: Statistics as Random Variables 337

9.1	Parameters, Statistics, and Statistical Inference	337
9.2	From Curiosity to Questions About Parameters	340
9.3	SD Module 0: An Overview of Sampling Distributions	345
9.4	SD Module 1: Sampling Distribution for One Sample Proportion	348
9.5	SD Module 2: Sampling Distribution for the Difference in Two Sample Proportions	353
9.6	SD Module 3: Sampling Distribution for One Sample Mean	357
9.7	SD Module 4: Sampling Distribution for the Sample Mean of Paired Differences	361
9.8	SD Module 5: Sampling Distribution for the Difference in Two Sample Means	364
9.9	Preparing for Statistical Inference: Standardized Statistics	367
	Lesson 1: Standardized Statistics for Sampling Distributions	367
	Lesson 2: Standardized Statistics for Proportions	367
	Lesson 3: Standardized Statistics for Means	368
9.10	Generalizations Beyond the Big Five	371
	Key Terms	378
	In Summary Boxes	378
	Simulations for Further Exploration	379
	Exercises	379

10 Estimating Proportions with Confidence 397

10.1 CI Module 0: An Overview of Confidence Intervals 397

Lesson 1: The Basic Idea of a Confidence Interval 397

Lesson 2: Computing Confidence Intervals for the Big Five Parameters 400

10.2 CI Module 1: Confidence Intervals for Population Proportions 403

Lesson 1: Details of How to Compute a Confidence Interval for a Population Proportion 403

Lesson 2: Understanding the Formula 406

Lesson 3: Reconciling and Understanding Different Margin of Error Formulas 409

10.3 CI Module 2: Confidence Intervals for the Difference in Two Population Proportions 412

10.4 Using Simulation to Calculate Confidence Intervals: Bootstrapping 416

10.5 Using Confidence Intervals to Guide Decisions 422

Key Terms 426

In Summary Boxes 427

Simulations for Further Exploration 427

Exercises 427

11 Estimating Means with Confidence 441

11.1 Introduction to Confidence Intervals for Means 441

11.2 CI Module 3: Confidence Intervals for One Population Mean 449

Lesson 1: Finding a Confidence Interval for a Mean for Any Sample Size and Any Confidence Level 449

Lesson 2: Special Case: Approximate 95% Confidence Intervals for Large Samples 455

11.3 CI Module 4: Confidence Intervals for the Population Mean of Paired Differences 457

11.4 CI Module 5: Confidence Intervals for the Difference in Two Population Means (Independent Samples) 462

Lesson 1: The General (Unpooled) Case 462

Lesson 2: The Equal Variance Assumption and the Pooled Standard Error 468

11.5 Using Simulation to Calculate Confidence Intervals: Bootstrapping for Means and Other Parameters 472

11.6 Understanding Any Confidence Interval 479

Key Terms 482

In Summary Boxes 482

Simulations for Further Exploration 482

Exercises 483

12 Testing Hypotheses About Proportions 497

- 12.1 HT Module 0: An Overview of Hypothesis Testing 498
 - Lesson 1: Formulating Hypothesis Statements 498
 - Lesson 2: Test Statistic, p -Value, and Deciding Between the Hypotheses 501
 - Lesson 3: What Can Go Wrong: The Two Types of Errors and Their Probabilities 508
- 12.2 HT Module 1: Testing Hypotheses About a Population Proportion 512
- 12.3 HT Module 2: Testing Hypotheses About the Difference in Two Population Proportions 524
- 12.4 Using Resampling to Estimate the p -Value for Testing Hypotheses About Two Proportions 529
- 12.5 Sample Size, p -Values, and Power 533
- 12.6 Understanding and Addressing Criticisms of Significance Testing 538
 - Key Terms 548
 - In Summary Boxes 548
 - Simulations for Further Exploration 548
 - Exercises 549

13 Testing Hypotheses About Means 567

- 13.1 Introduction to Hypothesis Tests for Means 567
- 13.2 HT Module 3: Testing Hypotheses About One Population Mean 569
- 13.3 HT Module 4: Testing Hypotheses About the Population Mean of Paired Differences 575
- 13.4 HT Module 5: Testing Hypotheses About the Difference in Two Population Means (Independent Samples) 579
 - Lesson 1: The General (Unpooled) Case 579
 - Lesson 2: The Pooled Two-Sample t -Test 585
 - Lesson 3: Randomization Test for Comparing Two Means 588
- 13.5 The Relationship Between Significance Tests and Confidence Intervals 591
- 13.6 Choosing an Appropriate Inference Procedure 594
- 13.7 Effect Size 598
- 13.8 Evaluating Statistical Results in Research Reports 603
 - Key Terms 607
 - In Summary Boxes 607
 - Simulations for Further Exploration 608
 - Exercises 608

14 Inference About Simple Regression 623

- 14.1 Sample and Population Regression Models 624
- 14.2 Estimating the Standard Deviation for Regression 631

14.3	Inference About the Slope of a Linear Regression	635
14.4	Predicting y and Estimating Mean y at a Specific x	638
14.5	Checking Conditions for Using Regression Models for Inference	644
	Key Terms	650
	In Summary Boxes	650
	Simulations for Further Exploration	650
	Exercises	651

15 More About Inference for Categorical Variables 661

15.1	The Chi-Square Test for Two-Way Tables	661
15.2	Methods for Analyzing 2×2 Tables	673
15.3	Testing Hypotheses About One Categorical Variable: Goodness-of-Fit	680
	Key Terms	686
	In Summary Boxes	686
	Simulations for Further Exploration	686
	Exercises	687

16 Analysis of Variance 701

16.1	Comparing Means with an ANOVA F -Test	702
16.2	Details of One-Way Analysis of Variance	710
16.3	Other Methods for Comparing Populations	716
16.4	Two-Way Analysis of Variance	720
	Key Terms	724
	In Summary Boxes	724
	Simulations for Further Exploration	724
	Exercises	724

17 Turning Information into Wisdom 735

17.1	Beyond the Data	735
17.2	Transforming Uncertainty into Wisdom	738
17.3	Making Personal Decisions	738
17.4	Controlling Societal Risks	740
17.5	Understanding Our World	742
17.6	Getting to Know You	743
17.7	Words to the Wise	745
	In Summary Boxes	746
	Exercises	746

Appendix of Tables 751

References 759

Answers to Selected Odd-Numbered Exercises 765

Index 790

Instructors: The Supplemental Topics are available on the book companion website, <http://www.cengage.com/statistics/Utts6e>.

- SUPPLEMENTAL TOPIC* **1 Additional Discrete Random Variables**
- S1.1** Hypergeometric Distribution
 - S1.2** Poisson Distribution
 - S1.3** Multinomial Distribution
 - S1.4** Negative Binomial and Geometric Distributions
- Key Terms
Exercises
- SUPPLEMENTAL TOPIC* **2 Nonparametric Tests of Hypotheses**
- S2.1** The Sign Test
 - S2.2** The Two-Sample Rank-Sum Test
 - S2.3** The Wilcoxon Signed-Rank Test
 - S2.4** The Kruskal-Wallis Test
- Key Terms
Exercises
- SUPPLEMENTAL TOPIC* **3 Multiple Regression**
- S3.1** The Multiple Linear Regression Model
 - S3.2** Inference About Multiple Regression Models
 - S3.3** Checking Conditions for Multiple Linear Regression
- Key Terms
Exercises
- SUPPLEMENTAL TOPIC* **4 Two-Way Analysis of Variance**
- S4.1** Assumptions and Models for Two-Way ANOVA
 - S4.2** Testing for Main Effects and Interactions
- Key Terms
Exercises
- SUPPLEMENTAL TOPIC* **5 Ethics**
- S5.1** Ethical Treatment of Human and Animal Participants
 - S5.2** Assurance of Data Quality
 - S5.3** Appropriate Statistical Analyses
 - S5.4** Fair Reporting of Results
- Key Terms
Exercises

Preface

A Challenge

Before you continue, think about how you would answer the question in the first bullet, and read the statement in the second bullet. We will return to them a little later in this preface.

- What do you *really know* is true, and how do you know it?
- The diameter of the moon is about 2160 miles.

What Is Statistics, and Who Should Care?

Because people are curious about many things, chances are that your interests include topics to which statistics has made a useful contribution. As written in Chapter 17, “information developed through the use of statistics has enhanced our understanding of how life works, helped us learn about each other, allowed control over some societal issues, and helped individuals make informed decisions. There is almost no area of knowledge that has not been advanced by statistical studies.”

Statistical methods have contributed to our understanding of health, psychology, ecology, politics, music, lifestyle choices, business, commerce, and dozens of other topics. A quick look through this book, especially Chapters 1 and 17, should convince you of this. Watch for the influences of statistics in your daily life as you learn this material.

How Is This Book Different? Two Basic Premises of Learning

We wrote this book because we were tired of being told that what statisticians do is boring and difficult. We think statistics is useful and not difficult to learn, and yet the majority of college graduates we’ve met seemed to have had a negative experience taking a statistics class in college. We hope this book will help to overcome these misguided stereotypes.

Let’s return to the two bullets at the beginning of this preface. Without looking, do you remember the diameter of the moon? Unless you already had a pretty good idea or have an excellent memory for numbers, you probably don’t remember. One premise of this book is that **new material is much easier to learn and remember if it is related to something interesting or previously known**. The diameter of the moon is about the same as the air distance between Atlanta and Los Angeles, San Francisco and Chicago, London and Cairo, or Moscow and Madrid. Picture the moon sitting between any of those pairs of cities, and you are not likely to forget the size of the moon again. Throughout this book, new material is presented in the context of interesting and useful examples. The first and last chapters (1 and 17) are exclusively devoted to examples and case studies, which illustrate the wisdom that can be generated through statistical studies.

Now answer the question asked in the first bullet: What do you *really know* is true, and how do you know it? If you are like most people, you know because it's something you have experienced or verified for yourself. It is not likely to be something you were told or heard in a lecture. The second premise of this book is that **new material is easier to learn if you actively ask questions and answer them for yourself**. *Mind on Statistics* is designed to help you learn statistical ideas by actively thinking about them. Throughout most of the chapters there are queries titled *Thought Questions*. Thinking about these questions will help you to discover and verify important ideas for yourself. Most chapters have a section called "Simulations for Further Exploration" that will guide you through hands-on activities and present you with a "Challenge Question." Working through the simulations in those sections will help you actively engage with the material. We encourage you to think and question, rather than simply read and listen.

New to This Edition

Chapter-Specific Updates

- Coverage of hypothesis testing has been substantially revised to clarify misconceptions about p -values and statistical significance. The new coverage is consistent with current statistical practice and research journal guidelines.
- An expanded discussion of statistical decision making in research and in daily life has been added to Chapters 12 and 17.
- Coverage of randomization tests has been added to Chapters 4, 12, and 13, and coverage of resampling and bootstrap confidence intervals has been added to Chapters 10 and 11.
- A discussion of multivariable thinking and an introduction to logistic regression have been added to Chapter 14.

Other Content and Technology Updates

- Thirty-seven new examples and case studies were written for the new edition, most of which are derived from recent research and news reports with a focus on medical and societal research. Data in many existing examples, case studies, and exercises also have been updated to the latest information available.
- Over 125 new exercises have been written for this edition. Many other exercises have been updated either to reflect the revised coverage of hypothesis testing or to incorporate newer data.
- Wording has been revised with increased sensitivity to gender inclusiveness.
- Learning objectives have been included at the beginning of each chapter.
- Additional problems, Watch It example videos, and Master It step-by-step tutorials have been added into WebAssign.
- Over 20% of WebAssign questions have been integrated into SALT, our new Statistical Analysis and Learning Tool.
- Simulations have been rebuilt with an improved user experience.

Student Resources: Tools for Learning

There are a number of tools provided in this book and on the companion website to enhance your learning of statistics. These tools are designed to help you learn statistics in the context of real-world research, to focus on understanding key concepts rather than crunching numbers, and to take an active role in learning how to use statistics to enhance your personal and professional life. Explanations and examples of these resources are listed in the following pages.

Learning Objectives have been added, which specify the skills you should obtain by completing each chapter. Shown here are the Chapter 1 Learning Objectives.

Learning Objectives

Learning Objectives

After completing this chapter, you will be able to:

- Explain how data can be summarized to help guide decisions when faced with uncertainty.
- Distinguish between an observational study and a randomized experiment and the conclusions that can be made from each type.
- Recognize the difference between statistical significance and significance in everyday life when reading the results of a study.
- Determine whether a sample is representative of a larger population and whether the sample results can be generalized to that population.

Tools for Conceptual Understanding

Thought Questions appear throughout each chapter to encourage active thinking and questioning about statistical ideas. *Hints* are provided at the bottom of the page to help you develop this skill.

Thought Question 2.4

Redo the bar graph in Figure 2.4 using counts instead of percentages. The necessary data are given in Table 2.3. Would the comparison of frequency of myopia across the categories of lighting be as easy to make using the bar graph with counts? Generalize your conclusion to provide guidance about what should be done in similar situations.*

***HINT:** Which graph makes it easier to compare the percentage with myopia for the three groups? What could be learned from the graph of counts that isn't apparent from the graph of percentages?

UPDATED! Simulations for Further Exploration sections provide opportunities for in-class or independent hands-on exploration of key statistical concepts. The simulations that accompany this feature can be found on the book's companion website, as well as within WebAssign.

Simulations for Further Exploration

Gain a deeper understanding of the concepts covered in Chapter 7 by exploring two interactive simulations. The simulations and suggested activities are available at <http://www.cengage.com/statistics/Utts6e>.

Simulation 7.1: Illustrating Probability Rules with a Venn Diagram

This simulation uses a Venn Diagram to illustrate the probability rules in Section 7.4. It shows how probabilities related to events A and B change based on how much the events overlap.

Simulation 7.2: Using a Tree Diagram to Find Conditional Probabilities

This simulation uses a tree diagram to illustrate how to find the conditional probability $P(B|A)$ when you know $P(A|B)$ and $P(B)$, and what happens when those values change.

Supplemental Notes boxes provide additional technical discussion of key concepts.

Supplemental Note

A Philosophical Issue About Probability

There is some debate about how to represent probability when an outcome has been determined but is unknown, such as if you have flipped a coin but not looked at it. Technically, any particular outcome has either happened or not. If it has happened, its probability of happening is 1; if it hasn't, its probability of happening is 0. In statistics, an example of this type of situation is the construction of a 95% confidence interval, which was introduced in Chapter 5 and which we will study in detail in Chapters 10 and 11. Before the sample is chosen, a probability statement makes sense. The probability is .95 that a sample will be selected for which the computed 95% confidence interval covers the truth. After the sample has been chosen, "the die is cast." Either the computed confidence interval covers the truth or it doesn't, although we may never know which is the case. That's why we say that we have *95% confidence* that a computed interval is correct, rather than saying that *the probability* that it is correct is .95.

Investigating Real-Life Questions

UPDATED! Relevant **Examples** form the basis for discussion in each chapter and walk you through real-life uses of statistical concepts.

Example 4.1

Age and Main News Source Where do you get most of your information about current news events? This question was asked in the 2018 General Social Survey, a national survey of randomly selected Americans. Possible answers included television, Internet, and newspapers, as well as other possibilities such as radio, family, and friends.

Figure 4.2 graphs the row percentages using a bar chart, as described in Chapter 2. We can see clearly that the variables are related. The increasing percentage for television and the decreasing percentage for Internet are easy to see as you move from the youngest age group to the oldest, as is the increasing percentage for newspapers.

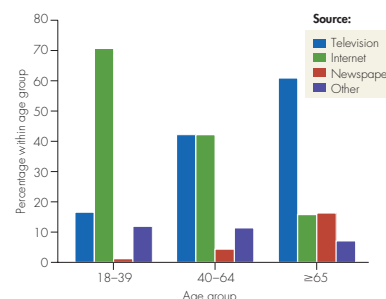


FIGURE 4.2 Age and main source of information about current news

UPDATED! Case Studies

apply statistical ideas to intriguing news stories. As the Case Studies are developed, they model the statistical reasoning process.

Case Study 12.1

Does Intensive Treatment to Lower Blood Pressure Reduce Dementia Risk?

A long-term randomized clinical trial was conducted to examine the effect of intensive treatment for high blood pressure on heart disease. A second goal of the study, discussed here, was to examine the treatment effect on cognitive impairment, also known as dementia. Participants with high blood pressure were randomly assigned to two groups. One group received standard blood pressure treatment aimed at reducing systolic blood pressure to 140. The other group received more intensive treatment with the goal of reducing systolic blood pressure to 120. The treatments will hereafter be called “standard” and “intensive.”

In 2019, the results of testing whether the two treatments had different outcomes on risk of dementia were published in the *Journal of the American Medical Association* (SPRINT MIND investigators, 2019). The *New York Times* reported the results using information from the research publication. Here is an excerpt from the *New York Times* story:

The primary outcome researchers measured was whether patients developed “probable dementia.” Fewer patients did so in

reporting these results we will use *relative risk* even though technically it isn't the precise statistic used. Here is what the researchers found after following the participants for a median time of about 5 years:

- In the sample, the relative risk of developing *dementia* for intensive treatment compared with standard treatment was 0.83, which translates into 7.2 versus 8.6 cases per 1000 person-years. A 95% confidence interval for the relative risk is 0.67 to 1.04. The *p*-value for testing whether the population relative risk was 1.0 is $p = .10$.
- In the sample, the relative risk of developing *mild cognitive impairment* for intensive treatment compared with standard treatment was 0.81, which translates into 14.6 versus 18.3 cases per 1000 person-years. A 95% confidence interval for the relative risk is 0.69 to 0.95. The *p*-value for testing whether the population relative risk was 1.0 is $p = .007$.

Getting Practice

Basic Exercises, comprising 25% of all exercises found in the text, focus on practice and review. These exercises, found under the header *Skillbuilder Exercises* and appearing at the beginning of each exercise section, complement the conceptual and data-analysis exercises. Basic exercises give you ample practice for these key concepts.

Exercises

 Denotes that the dataset is available on the companion website, <http://www.cengage.com/statistics/Utts6e>, but is not required to solve the exercise.

Bold exercises have answers in the back of the text.

Section 7.1

Skillbuilder Exercises

- 7.1** According to a U.S. Department of Transportation website (<https://www.bts.gov/newsroom/air-travel-consumer-report-september-2019-and-3rd-quarter-2019-numbers>), 78.1% of domestic flights flown by the top 17 U.S. airlines in the first nine months of 2019 arrived on time. Represent this in terms of a random circumstance and an associated probability.
- 7.2** This exercise is Thought Question 7.1. In Case Study 7.1, the names of 50 students in Alicia's class were placed in a bag and mixed well. Random Event 3 was defined as drawing a name from the bag and seeing whether Alicia's name was drawn. State the two possible outcomes of this random event and specify their probabilities.
- 7.3** Jayla is a member of a class with 20 students that meets daily. Each day for a week (Monday to Friday), a student in Jayla's class is randomly selected to explain how to solve a homework problem. Once a student has been selected, they are not selected again that week. If Jayla was not one of the four

- 7.7** Give an example of a random circumstance in which:

- The outcome is not determined until we observe it.
- The outcome is already determined, but our knowledge of it is uncertain.

Section 7.2

Skillbuilder Exercises

- 7.8** Is each of the following values a legitimate probability value? Explain any “no” answers.
- .50
 - .00
 - 1.00
 - 1.25
 - .25
- 7.9** Suppose that you live in a city that has 125,000 households and a polling organization randomly selects 1000 of them to contact for a survey. What is the probability that your household will be selected?
- 7.10** A car dealer has noticed that 1 out of 25 new-car buyers returns the car for warranty work within the first month.
- Write a sentence expressing this fact as a proportion.
 - Write a sentence expressing this fact as a percent.
 - Write a sentence expressing this fact as a probability.

Relevant conceptual and data analysis **Exercises** have been added and updated throughout the text. All exercises are found at the end of each chapter, with exercise sets written for each section and chapter. You will find more than 1700 exercises, providing a rich resource for testing your understanding.


Answers to Most Odd-Numbered Exercises,

indicated by bold numbers in the Exercise sections, have final answers or partial solutions found in the back of the text for checking your answers and guiding your thinking on similar exercises.

- 11.48 A researcher was interested in knowing whether the mean weight of the second baby is higher, lower, or about the same as the mean weight of the first baby for women who have at least two children. She selected a representative sample of 40 women who had at least two children and asked them for the weights of the oldest two at birth. She found that the mean of the differences (*first* – *second*) was 5 ounces, and the standard deviation of the differences was 7 ounces.
- Why was the researcher's study design better than taking independent samples of mothers and asking the first sample about the weight of their first baby and the second sample about the weight of their second baby?
 - Define the parameter of interest. Use appropriate notation.

 Dataset available but not required **Bold** exercises answered in the back

Lesson 1 Skillbuilder Exercises

- 11.51 Suppose that you were given a 95% confidence interval for the difference in two population means. What could you conclude about the population means if
- The confidence interval did not cover zero?
 - The confidence interval *did* cover zero?
- 11.52  Each of 63 students in a statistics class used her or his nondominant hand to print as many letters of the alphabet, in order, as they could in 15 seconds. The following output for this exercise gives results for a 95% confidence interval for the difference in population means for females and males. The “unpooled” procedure was used (Data source: **letters** dataset on the companion website).

Answers to Selected Odd-Numbered Exercises

The following are partial or complete answers to the exercises numbered in **bold** in the text.

Chapter 1

- a. 150 mph. b. 55 mph. c. 95 mph. d. 1/2. e. 51.
- a. .00043. b. .00043. c. Rate is based on past data; risk uses past data to predict an individual's likelihood of developing cervical cancer.
- a. All teens in the United States at the time the poll was taken. b. All teens in the United States who had dated at the time the poll was taken.
- a. All adults in the United States at the time the poll was taken. b. $\frac{1}{\sqrt{1048}} = .031$ or 3.1%. c. 30.9% to 37.1%.
- a. 400.
- a. Self-selected or volunteer sample. b. No; readers with strong opinions will respond.

Chapter 2

- a. 4. b. State in the United States. c. $n = 50$.
- a. Whole population. b. Sample.
- a. Population parameter. b. Sample statistic. c. Sample statistic.
- Gender and self-reported fastest ever driven speed. b. Students in a statistics class. c. Answer depends on whether interest is in this class only or in a larger group represented by this class.
- Population summary if we restrict interest to fiscal year 1998. Sample summary if 1998 value is used to represent errors in other years.
- a. Categorical. b. Quantitative. c. Quantitative. d. Categorical.

Technology for Developing Concepts and Analyzing Data

SALT (Statistical Analysis and Learning Tool) is a data analysis tool for introductory level statistics courses that helps students gain improved conceptual understanding of statistics through visualization and analysis of datasets.

SALT can be used both on its own by visiting statistics.cengage.com, or as a tool to answer SALT-enabled questions in WebAssign.

Minitab, Excel, TI-84, and SPSS Tips in the text offer key details on the use of technology.

Minitab Tip

Computing a Chi-Square Test for a Two-Way Table

- If the raw data are stored in columns of the worksheet, use **Stat > Tables > Chi-Square Test for Association**. Specify a categorical variable in the “Rows” box and a second categorical variable in the “Columns” box. Then click the **Okay** button.
- If the data are already summarized into counts, enter the table of counts (excluding totals) into columns of the worksheet, and then use **Stat > Tables > Chi-Square Test for Association**. From the pull-down menu, select “Summarized data in a two-way table.” In the dialog box, specify the two columns that contain the counts.

Excel Tip

The p -value can also be computed by using Microsoft Excel. The function $\text{CHIDIST}(x, df)$ provides the p -value, where x is the value of the chi-square statistic and df is a number called “degrees of freedom,” which will be explained later in this book. The formula for df is $(\# \text{ of rows} - 1)(\# \text{ of columns} - 1)$. For instance, corresponding to the information in Example 4.13, $df = (2 - 1)(2 - 1) = 1$, and the p -value is $\text{CHIDIST}(7.659, 1) = .005649$, or about .006 as given by Minitab.

Companion Website for Students

To access additional course materials and companion resources, please visit <http://www.cengage.com/statistics/Utts6e> or www.cengage.com. At the Cengage.com home page, search for the ISBN of your title (from the back cover of your book) using the search box at the top of the page. This will take you to the product page where the following free companion resources can be found:

- Conceptual simulations to accompany almost all chapters, with instructions and exercises
- Activities manual with engaging activities to accompany every chapter
- Step-by-Step technology guides for TI-84 Plus calculators, Microsoft® Excel®, Minitab®, SPSS®, R, and JMP
- Downloadable datasets (in ASCII as well as the native file formats for each software and calculator model covered by the Step-by-Step technology guides)
- Additional guidance for bootstrapping and randomization tests
- Examples of how some common surveys ask and utilize questions about gender and sex

Tools for Review

Key Terms at the end of each chapter, organized by section, can be used as a “quick-finder” and as a review tool.

Key Terms

Section 3.1

scatterplot, 71, 72, 76
explanatory variable, 72
response variable, 72
dependent variable, 72
y variable, 72
x variable, 72
positive association, 72, 73, 76
linear relationship, 72, 73, 76, 77
negative association, 73, 76
nonlinear relationship, 74
curvilinear relationship, 74, 76
outlier, 75

Section 3.2

regression analysis, 76
regression equation, 71, 76, 77, 78

prediction, 77
regression line, 77, 78, 83, 84
simple linear regression, 77
slope of a straight line, 77, 80, 84
intercept of a straight line, 77, 80
y-intercept, 77, 80
predicted y (\hat{y}), 78
estimated y , 78
predicted value, 78, 84
deterministic relationship, 79
statistical relationship, 79
prediction error, 82, 84
residual, 82, 84
least squares, 83
least squares line, 83
least squares regression, 83, 84
sum of squared errors (SSE), 83

Section 3.3

correlation, 71, 85
Pearson product moment correlation, 85
correlation coefficient, 85, 91
squared correlation (r^2), 88
proportion of variation explained by x , 88
sum of squares total (SSTO), 90
sum of squares due to regression (SSR), 90

Section 3.4

extrapolation, 92
interpolation, 92
influential observations, 92
multivariable thinking, 94

Section 3.5

causation versus correlation, 97–98

UPDATED! In Summary boxes serve as a useful study tool, appearing at appropriate points to enhance key concepts and calculations. More In Summary boxes have been added for this edition.

In Summary

Statistics on Risk, Relative Risk, Odds, and Odds Ratios

Let's summarize the various ways in which risk, relative risk, odds, and odds ratios are constructed from a two-way contingency table. These measures are usually employed when there is a definable explanatory and response variable and when there is a baseline condition, so we present them using those distinctions. In situations in which those distinctions cannot be made, simply be clear about which condition is in the numerator and which is in the denominator.

Explanatory Variable	Response Variable		
	Response 1	Response 2	Total
Category of Interest	A_1	A_2	T_A
Baseline Category	B_1	B_2	T_B

- Risk (of response 1) for category of interest = A_1/T_A
- Odds (of response 1 to response 2) for category of interest = A_1 to A_2
- Relative risk = $\frac{A_1/T_A}{B_1/T_B}$
- Odds ratio = $\frac{A_1/A_2}{B_1/B_2}$

In Summary Boxes

Basic Data Concepts, 19
Types of Variables and Roles for Variables, 22
Bar Graphs for Categorical Variables, 26
Using Visual Displays to Identify Interesting Features of Quantitative Data, 38

Numerical Summaries of Quantitative Variables, 46
Possible Reasons for Outliers and Reasonable Actions, 48
Bell-Shaped Distributions and Standard Deviation, 54

Tools for Active Learning

To access additional course materials and companion resources, please visit www.cengage.com. At the Cengage.com home page, search for the ISBN of your title (from the back cover of your book) using the search box at the top of the page. This will take you to the product page where free companion resources can be found.

The **Student Solutions Manual** (ISBN 9781337794619), prepared by Jessica M. Utts and Robert F. Heckard, provides worked-out solutions to most of the odd-numbered problems in the text.

The online **Activities Manual**, written by Jessica M. Utts and Robert F. Heckard, includes a variety of activities for students to explore individually or in teams. These activities guide students through key features of the text, help them understand statistical concepts, provide hands-on data collection and interpretation team-work, include exercises with tips incorporated for solution strategies, and provide bonus dataset activities.

Flexible Options for Covering Inference

Chapters 9 to 13, which contain the core material on sampling distributions and statistical inference, are organized in a modular, flexible format. There are six modules for each of the topics: sampling distributions, confidence intervals, and hypothesis testing. The first module presents an introduction and the remaining five modules each deal with a specific parameter, such as one mean, one proportion, or the difference in two means. Chapter 9 covers sampling distributions, Chapters 10 and 11 cover confidence intervals, and Chapters 12 and 13 cover hypothesis testing.

Organization of Chapters 9 to 13

Parameter	Chapter 9: Sampling Distributions (SD)	Chapter 10: Confidence Intervals (CI)	Chapter 11: Confidence Intervals (CI)	Chapter 12: Hypothesis Tests (HT)	Chapter 13: Hypothesis Tests (HT)
0 Introductory	SD Module 0 Overview of sampling distributions	CI Module 0 Overview of confidence intervals		HT Module 0 Overview of hypothesis testing	
1 Population proportion (p)	SD Module 1 SD for one sample proportion	CI Module 1 CI for one population proportion		HT Module 1 HT for one population proportion	
2 Difference in two population proportions ($p_1 - p_2$)	SD Module 2 SD for difference in two sample proportions	CI Module 2 CI for difference in two population proportions		HT Module 2 HT for difference in two population proportions	
3 Population mean (μ)	SD Module 3 SD for one sample mean		CI Module 3 CI for one population mean		HT Module 3 HT for one population mean
4 Population mean of paired differences (μ_d)	SD Module 4 SD for sample mean of paired differences		CI Module 4 CI for population mean of paired differences		HT Module 4 HT for population mean of paired differences
5 Difference in two population means ($\mu_1 - \mu_2$)	SD Module 5 SD for difference in two sample means		CI Module 5 CI for difference in two population means		HT Module 5 HT for difference in two population means

This structure emphasizes the similarity among the inference procedures for the five parameters discussed. It allows instructors to illustrate that each procedure covered is a specific instance of the same process. We recognize that instructors have different preferences for the order in which to cover inference topics. For instance, some prefer to first cover all topics about proportions and then cover all topics about means. Others prefer to first cover everything about confidence intervals and then cover everything about hypothesis testing. **With the modular format, instructors can cover these topics in the order they prefer.**

To aid in the navigation through these modular chapters, the book contains **color-coded, labeled tabs that correspond to the introductory and parameter modules.** The table above, also found in Chapter 9, lays out the color-coding system as well as the flexibility of these core chapters. The table also is a useful course planning tool.

To add to the flexibility of topic coverage, Supplemental Topics 1 to 5, on discrete random variables, nonparametric tests, multiple regression, two-way ANOVA, and ethics, are available for use on the book companion website.

Instructor Resources

Tools for Assessment



WebAssign for Utts/Heckard's *Mind on Statistics*, Sixth Edition, is a flexible and fully customizable online instructional solution that puts powerful tools in the hands of instructors, enabling them to deploy assignments, instantly assess individual student and class performance, and help students master the course concepts. With WebAssign's powerful digital platform and *Mind on Statistics* specific content, instructors can tailor their course with a wide range of assignment settings, add their own questions and content, and access student and course analytics and communication tools.

Cengage Learning Testing Powered by Cognero is a flexible, online system that allows you to:

- Author, edit, and manage test bank content from multiple Cengage Learning solutions
- Create multiple test versions in an instant
- Deliver tests from your learning management system (LMS), your classroom, or wherever you want

Companion Website for Instructors

The companion website at <http://www.cengage.com/Utts6e> contains a variety of resources for instructors.

- Microsoft® PowerPoint® lecture slides for all chapters
- Figures from the book
- Complete solutions manual
- Test banks
- Course outlines and syllabi
- Suggestions for class projects for Chapters 2 to 16
- Suggested discussions for the Thought Questions located throughout the text
- Supplemental Topics: Chapters S.1 to S.5 and Supplemental Topic solutions
- List of applications and methods
- Index of exercises by subject matter

A Note to Instructors

The entire *Mind on Statistics* learning package has been informed by the recommendations put forth by the ASA/MAA Joint Curriculum Committee, the original (2005) GAISE (Guidelines for Assessment and Instruction in Statistics Education) College Report, for which Jessica Utts was one of the authors, and the updated 2016 GAISE College Report. Each of the pedagogical features and ancillaries listed in the sections entitled “Student Resources: Tools for Learning” and “Instructor Resources” has been categorized by suggested use to provide you with options for designing a course that best fits the needs of your students.

In this edition, the material on significance testing and p -values has been substantially revised. The revisions were informed by the American Statistical Association’s statement on p -values and related changes in publication guidelines by many scientific journals. (See <https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf>)

Acknowledgments

We thank William Harkness, Professor Emeritus of Statistics at Penn State University, for his support and feedback throughout our careers and during the writing of this book, and for his remarkable dedication to undergraduate statistics education. Preliminary editions of *Mind on Statistics*, the basis for this text, were used at Penn State; the University of California, Davis; and Texas A&M University, and we thank the many students who provided comments and suggestions on those and on subsequent editions. Thanks to Deb Niemeier, University of Maryland, for suggesting that we add a supplemental chapter on Ethics (available on the companion website). We are indebted to Neal Rogness, Grand Valley State University, for help with the SPSS Tips, and Larry Schroeder and Darrell Clevidence, Carl Sandburg College, for help with the TI-84 Tips. At Penn State, Dave Hunter, Steve Arnold, and Tom Hettmansperger provided many helpful insights. At the University of California, Davis, Rodney Wong provided insights as well as material for some exercises and the test bank. We extend special thanks to George Pasles for providing hundreds of valuable suggestions for improving the previous edition of the book, and to Deborah Delanoy for providing valuable feedback on the revised discussion of significance testing and p -values in this edition.

For providing datasets used in the book and available at the companion website, we thank Susan Jelsing, as well as William Harkness and Laura Simon from Penn State University.

The following reviewers offered valuable suggestions for this and previous editions:

Erica Bernstein, *University of Hawaii at Hilo*
 Patricia M. Buchanan, *Penn State University*
 Elizabeth Clarkson, *Wichita State University*
 Ian Clough, *University of Cincinnati–Clermont College*
 Patti B. Collings, *Brigham Young University*
 James Curl, *Modesto Junior College*
 Boris Djokic, *Keiser University*
 Wade Ellis, *West Valley College*
 Patricia Erickson, *Taylor University*
 Linda Ernst, *Mt. Hood Community College*
 Anda Gadidov, *Kennesaw State University*
 Joan Garfield, *University of Minnesota*
 Jonathan Graham, *University of Montana*
 Jay Gregg, *Colorado State University*
 Brenda Gunderson, *University of Michigan*
 Donnie Hallstone, *Green River Community College*
 Hasan Hamdan, *James Madison University*
 Glenn Hansen, *University of Oklahoma*

Donald Harden, *Georgia State University*
 Sarai Hedges, *University of Cincinnati*
 Rosemary Hirschfelder, *University of South*
 Sue Holt, *Cabrillo Community College*
 Mortaza Jamshidian, *California State University-Fullerton*
 Mark Johnson, *University of Central Florida*
 Tom Johnson, *North Carolina University*
 Yevgeniya Kleiman, *University of Michigan*
 Danny Lau, *Gainesville State College*
 Andre Mack, *Austin Community College*
 Jean-Marie Magnier, *Springfield Technical Community College*
 Suman Majumdar, *University of Connecticut*
 D'Arcy Mays, *Virginia Commonwealth*
 Megan Meece, *University of Florida*
 Jack Osborn Morse Jr., *University of Georgia*
 Emily Murphree, *Miami University*
 Mary Murphy, *Texas A&M University*
 Helen Noble, *San Diego State University*
 Thomas Nygren, *Ohio State University*
 Wanda O'Connor, *Austin Community College*
 Jamis Perrett, *Texas A&M University*
 Thomas J. Pfaff, *Ithaca College*
 Nancy Pfenning, *University of Pittsburgh*
 Joel Pitt, *Georgian Court University*
 Jennifer Lewis Priestley, *Kennesaw State University*
 John Racquet, *State University of New York at Albany*
 Lawrence Ries, *University of Missouri*
 David Robinson, *St. Cloud State University*
 Neal Rogness, *Grand Valley State University*
 Kelly Sakkinen, *Lansing Community College*
 Heather Sasinouska, *Clemson University*
 Stephen Soltys, *Elizabethtown College*
 Kirk Steinhurst, *University of Idaho*
 Engin Sungur, *University of Minnesota, Morris*
 Robert Talbert, *Franklin College*
 Mary Ann Teel, *University of North Texas*
 Gwen Terwilliger, *University of Toledo*
 Ruth Trygstad, *Salt Lake Community College*
 Sasha Verkhovtseva, *Anoka-Ramsey Community College*
 Eric Westlund, *Luther College*
 Jim Wiseman, *Agnes Scott College*
 Robert Alan Wolf, *University of San Francisco*

Our sincere appreciation and gratitude also goes to Amanda Rose, Elinor Gregory, Kyra Kruger, Andy Trus, and the staff at Cengage Learning, as well as Lori Hazzard of MPS Limited, who oversaw the development and production of the sixth edition. We also wish to thank Carolyn Crockett Lewis and Danielle Derbenti, without whom this book could not have been written, and Martha Emry, who kept us on track throughout the editing and production of the first three editions of the book. Finally, for their support, patience, and numerous prepared dinners, we thank our families and friends, especially Candace Heckard, Molly Heckard, Wes Johnson, Claudia Utts-Smith, and Dennis Smith.

Jessica M. Utts
Robert F. Heckard

1



KURHAN/SHUTTERSTOCK.COM

Will her breakfast cereal influence the baby's sex?

See Case Study 1.8 (p. 9)

Statistics Success Stories and Cautionary Tales

Learning Objectives

After completing this chapter, you will be able to:

- Explain how data can be summarized to help guide decisions when faced with uncertainty.
- Distinguish between an observational study and a randomized experiment and the conclusions that can be made from each type.
- Recognize the difference between statistical significance and significance in everyday life when reading the results of a study.
- Determine whether a sample is representative of a larger population and whether the sample results can be generalized to that population.

Let's face it. You're a busy person. Why should you spend your time learning about statistics? In this chapter, we give eight examples of situations in which statistics provide either enlightenment or misinformation. After reading these examples, we hope you will agree that learning about statistics may be interesting and useful.

Each of the stories in this chapter illustrates one or more concepts that will be developed in this book. These concepts are given as "the moral of the story" after a case is presented. Definitions of some terms used in the story also are provided following each case. By the time you have read all of these stories, you already will have an overview of what statistics is all about.

Cautionary note: Many of the examples in this book present results of medical studies or studies about lifestyle choices. The examples are chosen to illustrate statistical concepts, and should not be interpreted as recommendations regarding medical or lifestyle choices.

1.1 What Is Statistics?

When you hear the word *statistics* you probably think of lifeless or gruesome numbers, such as the population of your state or the number of violent crimes committed in your city last year. The word *statistics*, however, actually is used to mean two different things. The better-known definition is that statistics are numbers measured for some purpose. A more complete definition, and the one that forms the substance of this book, is the following:

Definition

Statistics is a collection of procedures and principles for gathering data and analyzing information to help people make decisions when faced with uncertainty.

The eight stories in this chapter are meant to bring life to this definition. After reading them, if you think the subject of statistics is lifeless or gruesome, check your pulse!

1.2 Eight Statistical Stories with Morals

The best way to gain an understanding of some of the ideas and methods used in statistical studies is to see them in action. As you read each story presented in this section, think about how the situation was used to extract information from data. The methods and ideas used differ for each of these eight stories. Together they will give you an excellent overview of why it is useful to study statistics. To help you understand some basic statistical principles, each case study is accompanied by a “moral of the story” and by some definitions. All of the ideas and definitions will be discussed in greater detail in subsequent chapters.

A Note About Gender: Statistical studies have historically used an assumption of a gender binary. Many examples in this book use data collected in surveys by the authors in their classes. Students were asked to self-identify as male/female or leave the response blank, so the terms “male” and “female” are used as nouns when discussing results of those surveys, as in Case Study 1.1. As you come across studies and data sets throughout this text and elsewhere that compare only “males and females” or “men and women”, we encourage you to consider how this practice might ignore non-binary or gender non-conforming experience and reality.

Case Study 1.1

Who Are Those Speedy Drivers?

A survey taken in a large statistics class at Penn State University contained the question “What’s the fastest you have ever driven a car? ____ mph.” The *data* provided by the 87 self-identified males and 102 self-identified females who responded are listed here.

Males: 110 109 90 140 105 150 120 110 110 90 115 95 145 140
110 105 85 95 100 115 124 95 100 125 140 85 120 115 105 125
102 85 120 110 120 115 94 125 80 85 140 120 92 130 125 110
90 110 110 95 95 110 105 80 100 110 130 105 105 120 90 100
105 100 120 100 100 80 100 120 105 60 125 120 100 115 95
110 101 80 112 120 110 115 125 55 90

Females: 80 75 83 80 100 100 90 75 95 85 90 85 90 90 120 85
100 120 75 85 80 70 85 110 85 75 105 95 75 70 90 70 82 85 100
90 75 90 110 80 80 110 110 95 75 130 95 110 110 80 90 105 90
110 75 100 90 110 85 90 80 80 85 50 80 100 80 80 80 95 100
90 100 95 80 80 50 88 90 90 85 70 90 30 85 85 87 85 90 85 75
90 102 80 100 95 110 80 95 90 80 90

From these numbers, can you tell which gender tends to have driven faster and by how much? Notice how difficult it is to make sense of the *data* when you are simply presented with a list. Even if the numbers had been presented in numerical order, it would be difficult to compare the two groups.

Your first lesson in using statistics is how to formulate a simple summary of a long list of numbers. The **dotplot** shown in Figure 1.1 helps us see the pattern in the data. In the plot, each dot represents the response of an individual student. We can see that the male students tend to claim a higher “fastest ever driven” speed than do the female students.

The graph shows us a lot, and calculating some **summary statistics** will provide additional insight. There are a variety of ways to do so, but for this example, we examine a **five-number summary** of the data for males and females. The five numbers are the lowest value; the cut-off points for one-fourth,

one-half, and three-fourths of the ordered data; and the highest value. The three middle values of the summary (the cutoff points for one-fourth, one-half, and three-fourths of the ordered data) are called the *lower quartile*, *median*, and *upper quartile*, respectively. Five-number summaries can be represented as shown in the table underneath Figure 1.1.

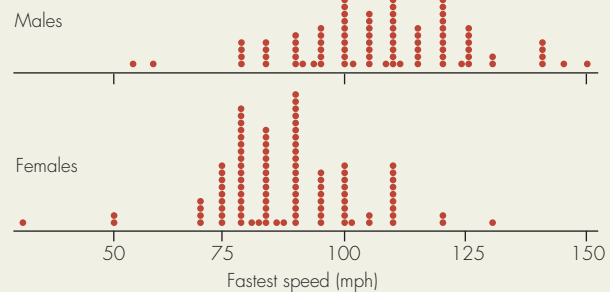


FIGURE 1.1 Responses to “What’s the fastest you’ve ever driven?”

	Males (87 Students)		Females (102 Students)	
Median	110		89	
Quartiles	95	120	80	95
Extremes	55	150	30	130

Some interesting facts become immediately obvious from these summaries. By looking at the medians, you see that half of the male students have driven 110 miles per hour or more, whereas the halfway point for female students is only 89 miles per hour. Interestingly, 95 miles per hour is the lower quartile for male students, but is the upper quartile for female students. This tells us that three-fourths of the males have driven 95 miles per hour or

more, but only one-fourth of the females have done so. These facts were not at all obvious from the original lists of numbers.

Moral of the Story: Simple summaries of data can tell an interesting story and are easier to digest than long lists.

Definitions: **Data** is a plural word referring to numbers or nonnumerical labels (such as male/female) collected from a

set of entities (people, cities, and so on). The **median** of a numerical list of data is the value in the middle when the numbers are put in order. For an even number of entities, the median is the average of the middle two values. The **lower quartile** and **upper quartile** are (roughly) the medians of the lower and upper halves of the ordered data.

Case Study 1.2

Safety in the Skies?

If you fly often, you may have been relieved to see the *New York Times* headline on October 1, 2007, proclaiming “Fatal airline crashes drop 65%” (Wald, 2007). And you may have been dismayed if you had seen an earlier headline in *USA Today* that read, “Planes get closer in midair as traffic control errors rise” (Levin, 1999). The details were even more disturbing: “Errors by air traffic controllers climbed from 746 in fiscal 1997 to 878 in fiscal 1998, an 18% increase.”

So, are the risks of a fatal airline crash or an air traffic control error something that should be a major concern for airline passengers? Don’t cancel your next vacation yet. A look at the statistics indicates that the news is actually pretty good! The low risk becomes obvious when we are told the *base rate* or *baseline risk* for these problems. According to the *New York Times* article, “the drop in the accident rate [from 1997 to 2007] will be about 65%, to one fatal accident in about 4.5 million departures, from 1 in nearly 2 million in 1997.” And according to the 1999 *USA Today* story, “The errors per million flights handled by controllers climbed from 4.8 to 5.5.” So the *rate* of fatal accidents decreased from about 1 in 2 million departures in 1997 to 1 in 4.5 million departures in 2007. And, the supposedly ominous rise in air traffic controller errors in 1998 only led to a very low rate of 5.5 errors per million flights.

Fortunately, the rates of these occurrences were provided in both stories. This is not always the case in news

reports of changes in rates or risk. For instance, an article may say that the risk of a certain type of cancer is doubled if you eat a certain unhealthy food. But what good is that information unless you know the actual risk? Doubling your chance of getting cancer from 1 in a million to 2 in a million is trivial, but doubling your chance from 1 in 50 to 2 in 50 is not.

Moral of the Story: When you read about the change in the rate or risk of occurrence of something, make sure you also find out the base rate or baseline risk.

Definitions: The **rate** at which something occurs is simply the number of times it occurs per number of opportunities for it to occur. In fiscal year 1998, the rate of air traffic controller errors was 5.5 per million flights. The **risk** of a bad outcome in the future can be estimated using the past rate for that outcome, if it is assumed the future will be like the past. Based on recent data, the risk of a fatal accident in 2019 was just 1 in 5.58 million flights, which is $1/5,580,000$ or about .00000018. (<https://www.reuters.com/article/us-airlines-safety/major-commercial-plane-crash-deaths-worldwide-fell-by-more-than-50-in-2019-group-idUSKBN1Z0242>). The **base rate** or **baseline risk** is the rate or risk at a beginning time period or under specific conditions. For instance, the base rate of fatal airline crashes from which the 65% decrease for 2007 was calculated was about 1 crash per 2 million flights for fiscal year 1997.

Case Study 1.3

Did Anyone Ask Whom You’ve Been Dating?

In the late 1990s interracial dating was a sensitive topic. So it was newsworthy to learn that “According to a new *USA Today*/Gallup Poll of teenagers across the country, 57% of teens who go out on dates say they’ve been out with someone of another race or ethnic group” (Peterson, 1997). That was over half of the dating teenagers, so it was natural for the headline in the *Sacramento Bee* to read, “Interracial dates common among today’s teenagers.” The article contained other information as well, such as “In most cases, parents aren’t a major obstacle. Sixty-four percent of teens say their parents don’t mind that they date interracially, or wouldn’t mind if they did.”

There were millions of teenagers in the United States whose experiences appeared to be being reflected in this

story. How could the polltakers manage to ask so many teenagers these questions? The answer is that they didn’t. The article states that “the results of the new poll of 602 teens, conducted Oct. 13–20, reflect the ubiquity of interracial dating today.” They asked only 602 teens? Could such a small sample possibly tell us anything about the millions of teenagers in the United States? The answer is “yes” if those teens constituted a *random sample* from the *population* of interest.

The featured statistic of the article is that “57 percent of teens who go out on dates say they’ve been out with someone of another race or ethnic group.” Only 496 of the 602 teens in the poll said that they date, so the 57% value is actually a percentage based on 496 responses. In other words, the

Continues

Continued

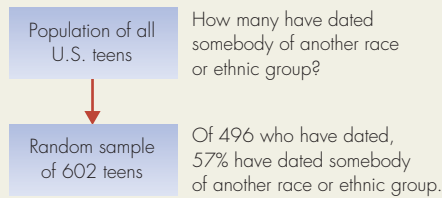


FIGURE 1.2 Population and sample for the survey

pollsters were using information from only 496 teenagers to estimate something about all teenagers who date. Figure 1.2 illustrates this situation.

How accurate could this *sample survey* possibly be? The answer may surprise you. The results of this *poll* are accurate to within a *margin of error* of about 4.5%. As surprising as it may seem, the true percentage of all dating teens in the United States at that time who had dated interracially is reasonably likely to be within 4.5% of the reported percentage that's based only on the 496 teens asked! We'll be conservative and round the 4.5% margin of error up to 5%. At the time the poll was taken, the percentage of all dating teenagers in the United States that would say they had dated someone of another race or ethnic group was likely to be in the range $57\% \pm 5\%$, or between 52% and 62%. (The symbol \pm is read "plus and minus" and means that the value on the right should be added to and subtracted from the value on the left to create an interval.)

Polls and *sample surveys* are frequently used to assess public opinion and to estimate population characteristics such as the percent of teens who have dated interracially or the proportion of voters who plan to vote for a certain candidate.

Many sophisticated methods have been developed that allow pollsters to gain the information they need from a very small number of individuals. The trick is to know how to select those individuals. In Chapter 5, we examine a number of other strategies that are used to ensure that sample surveys provide reliable information about populations.

Moral of the Story: *A representative sample of only a few thousand, or perhaps even a few hundred, can give reasonably accurate information about a population of many millions.*

Definitions: A **population** is a collection of all individuals about which information is desired. The "individuals" are usually people, but could also be schools, cities, pet dogs, agricultural fields, and so on. A **random sample** is a subset of the population selected so that every individual has a specified probability of being part of the sample. (Often, but not always, it is specified that every individual has the same chance of being selected for the sample.) In a **poll** or **sample survey**, the investigators gather opinions or other information from each individual included in the sample. The **margin of error** for a properly conducted survey is a number that is added to and subtracted from the sample information to produce an interval that is 95% certain to contain the true value for the population. In the simplest types of sample surveys, the margin of error is approximately equal to 1 divided by the square root of the number of individuals in the sample.

Hence, a sample of 496 teenagers who have dated produces a margin of error of about $1/\sqrt{496} = .045$, or about 4.5%. In some polls the margin of error is called the **margin of sampling error** to distinguish it from other sources of errors and biases that can distort the results. The next Case Study illustrates a common source of bias that can occur in surveys, discussed more fully in Chapter 5.

Case Study 1.4

Who Are Those Angry Women?

A well-conducted survey can be very informative, but a poorly conducted one can be a complete disaster. As an extreme example, Moore (1997, p. 11) reports that Shere Hite sent questionnaires to 100,000 women asking about love, sex, and relationships for her book *Women and Love* (1987). Only 4.5% of the women responded, and Hite used those responses to write her book. As Moore notes, "The women who responded were fed up with men and eager to fight them. For example, 91% of those who were divorced said that they had initiated the divorce. The anger of women toward men became the theme of the book." Do you think that women who were angry with men would be likely to answer questions about love relationships in the same way as the general population of women?

The Hite sample exemplifies one of the most common problems with surveys: The sample data may not represent the population. Extensive *nonparticipation* (*nonresponse*) from a random sample, or the use of a *self-selected* (i.e., a *volunteer*) sample, will probably produce biased results. Those who voluntarily respond

to surveys tend to care about the issue and therefore have stronger and different opinions than those who do not respond.

Moral of the Story: *An unrepresentative sample, even a large one, tells you almost nothing about the population.*

Definitions: **Nonparticipation bias** (also called **nonresponse bias**) can occur when many people who are selected for the sample either do not respond at all or do not respond to some of the key survey questions. This may occur even when an appropriate random sample is selected and contacted. The survey is then based on a nonrepresentative sample, usually those who feel strongly about the issues. Some surveys don't even attempt to contact a random sample but instead ask anyone who wishes to respond to do so. Magazines, television stations, and websites routinely conduct this kind of poll, and those who respond are called a **self-selected sample** or a **volunteer sample**. In most cases, this kind of sample tells you nothing about the larger population at all; it tells you only about those who responded.

Case Study 1.5 Does Prayer Lower Blood Pressure?

News headlines are notorious for making one of the most common mistakes in the interpretation of statistical studies: jumping to unwarranted conclusions. A headline in *USA Today* read, “Prayer can lower blood pressure” (Davis, 1998). The story that followed continued the possible fallacy it began by stating, “Attending religious services lowers blood pressure more than tuning into religious TV or radio, a new study says.” The words “attending religious services lowers blood pressure” imply a direct cause-and-effect relationship. This is a strong statement, but it is not justified by the research project described in the article.

The article was based on an *observational study* conducted by the U.S. National Institutes of Health, which collected data on 2391 people aged 65 or older for 6 years (Figure 1.3). The article described one of the study’s principal findings: “People who attended a religious service once a week and prayed or

studied the Bible once a day were 40% less likely to have high blood pressure than those who don’t go to church every week and prayed and studied the Bible less” (Davis, 1998). So the researchers did observe a relationship, but it’s a mistake to think that this justifies the conclusion that prayer actually *causes* lower blood pressure.

When groups are compared in an observational study, the groups usually differ in many important ways that may contribute to the observed relationship. In this example, people who attended church and prayed regularly may have been less likely than the others to smoke or to drink alcohol. These could affect the results because smoking and alcohol use are both believed to affect blood pressure. The regular church attendees may have had a better social network, a factor that could lead to reduced stress, which in turn could reduce blood pressure. People who were generally somewhat ill may not have been as willing or able to go out to church. We’re sure you can think of other possibilities for *confounding variables* that may have contributed to the observed relationship between prayer and lower blood pressure.

Moral of the Story: *Cause-and-effect conclusions cannot generally be made on the basis of an observational study.*

Definitions: An **observational study** is one in which participants are observed and measured but not asked to do anything differently. Comparisons based on observational studies are comparisons of naturally occurring groups. A **variable** is a characteristic that differs from one individual to the next. It may be numerical, such as blood pressure, or it may be categorical, such as whether or not someone attends church regularly. A **confounding variable** is a variable that is not the main concern of the study but may be partially responsible for the observed results.

Source: International Journal of Psychiatry in Medicine by Koenig, H. G., L. K. George, J. C. Hays, and D. B. Larson. [See p. 761 for complete credit.]

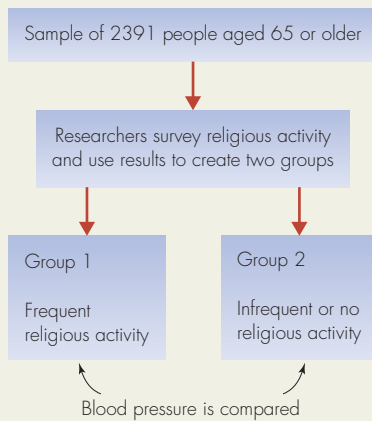


FIGURE 1.3 An observational study in Case Study 1.5. Researchers survey religious activity and compare blood pressure of frequent and not-frequent activity group

Case Study 1.6 Does Aspirin Reduce Heart Attack Rates?

In 1988, the Steering Committee of the Physicians’ Health Study Research Group released the results of a 5-year *randomized experiment* conducted using 22,071 male physicians between the ages of 40 and 84. The purpose of the experiment was to determine whether or not taking aspirin reduces the risk of a heart attack. The physicians had been *randomly assigned* to one of the two *treatment* groups. One group took an ordinary aspirin tablet every other day, while the other group took a *placebo*. None of the physicians knew whether he was taking the actual aspirin or the placebo. Figure 1.4 illustrates the design of this experiment.

The results, shown in Table 1.1, support the conclusion that taking aspirin does indeed help to reduce the risk of

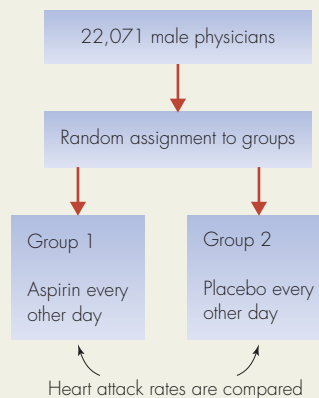


FIGURE 1.4 Randomized experiment for Case Study 1.6. Physicians were assigned to regularly take either aspirin or a placebo

Continues

Continued

TABLE 1.1 The Effect of Aspirin on Heart Attacks

Treatment	Heart Attacks	Doctors in Group	Attacks per 1000 Doctors
Aspirin	104	11,037	9.42
Placebo	189	11,034	17.13

having a heart attack. The rate of heart attacks in the group taking aspirin was only about half the rate of heart attacks in the placebo group. In the aspirin group, there were 9.42 heart attacks per 1000 participating doctors, while in the placebo group, there were 17.13 heart attacks per 1000 participants.

Because the men in this experiment were randomly assigned to the two conditions, other important risk factors such as age, amount of exercise, and dietary habits should have been similar for the two groups. The only important difference between the two groups should have been whether they took aspirin or a placebo. This makes it possible to conclude that taking aspirin actually *caused* the lower rate of heart attacks for that group. In a later chapter, you will learn how to determine that the difference seen in this sample is *statistically significant*. That term is used to indicate that the observed sample difference probably reflects a true difference within the population. A cautionary note, however, is that the designation of *statistical significance* does not provide any information about the size of the difference in the groups. It's quite possible that a difference can be labeled as statistically significant,

but have little practical importance, as will be illustrated in Case Study 1.7.

To what population does the conclusion of this study apply? The participants were all male physicians, so the conclusion that aspirin reduces the risk of a heart attack may not hold for the general population of men. No women were included, so the conclusion may not apply to women at all. More recent evidence, however, has provided additional support for the benefit of aspirin in broader populations.

Moral of the Story: *Unlike with observational studies, cause-and-effect conclusions can generally be made on the basis of randomized experiments.*

Definitions: A **randomized experiment** is a study in which treatments are randomly assigned to participants. A **treatment** is a specific regimen or procedure assigned to participants by the experimenter. A **random assignment** is one in which each participant has a specified probability of being assigned to each treatment. A **placebo** is a pill or treatment designed to look just like the active treatment but with no active ingredients. A **statistically significant** relationship or difference is one that is unlikely to have occurred in the sample if there was no relationship or difference in the population.

Source: The Steering Committee of the Physicians' Health Study Research Group (1988). [See p. 763 for complete credit.]

Case Study 1.7

Social Media Use and Depression in Teens

In July 2019, CNN posted a story on its website with the headline "Increasing social media use tied to rise in teens depressive symptoms, study says" (Howard, July 15, 2019). The story was based on a University of Montreal press release with the title "Increases in social media use and television viewing associated with increases in teen depression." (https://www.eurekalert.org/pub_releases/2019-07/uom-iis071119.php) The original study was published in the journal *JAMA Pediatrics* (Boers et al., 2019).

The study used data collected from 2012 to 2018 on almost 4000 students in 31 high schools in the Montreal area of Canada. The same students were contacted annually as they moved from 7th grade to 11th grade. Each year they were asked about their screen-time use and their symptoms of depression. Specifically, the students were asked to estimate how much time on average they spent per day on four kinds of screen time: playing video games, social networking, watching television and movies at home, and computer use not covered by the other categories. Depression was measured by asking students to indicate how much they experienced each of seven symptoms of depression, such as feeling lonely and feeling no interest in things, on a scale from 0 (not at all) to 4 (very much). The total depression score, the sum of scores for the seven

symptoms, could thus range from 0 to 28, with a higher score indicating a higher level of depression.

As you probably can imagine, students were not randomly assigned to differing amounts of screen-time use, and thus the results are based on an observational study, not a randomized experiment. Notice that the two headlines quoted above carefully avoid stating that increased social media use *causes* increased depression, as is appropriate for an observational study. However, a reader who does not understand the distinction between observational studies and randomized experiments may be led to conclude a causal relationship from statements such as the second sentence of the press release. It reads "Changes in adolescent social media use and television use predict increases in symptoms of depression." Be careful that you don't interpret such statements to mean that there is a cause-and-effect relationship.

Although complicated statistical methods were used to analyze the data, the main result that led to the media reports was stated simply in the original journal article. It said "A significant between-person association indicated that a 1-hour increase in social media use was associated with a 0.64 unit (on a scale from 0 to 28) increase in the severity of depression symptoms over 4 years."

Wait, what? A 1-hour increase in average daily social media use was associated with a *significant* increase in severity of depression symptoms? But that increase was only 0.64 point on a scale from 0 to 28? That level of increase doesn't even equate to a one-point change in one of the seven depression symptoms! How could that be "significant"?

The answer is that the use of the word *significant* in the quoted results refers to *statistical significance* and not to *practical* significance or importance. As noted in Case Study 1.6, the definition of statistical significance is that the observed increase in depression score was unlikely to be seen in the sample of 4000 students if there was no association at all in the population of students similar to those in this study. The label of "statistical significance" by itself provides no information about the size of the difference or whether it represents an important difference.

Unfortunately, neither the press release nor the CNN story reported the size of the increase in depression score. The press release summarized the finding by stating that "The study, published July 15 in *JAMA Pediatrics*, revealed that a higher than average frequency of social media and television viewing over four years predicts more severe symptoms of depression over that same time frame" (Gazaille, 2019). And the CNN story simply stated that "For every additional hour young people spend on social media or watching television, the severity of depressive symptoms they experience goes up" (Howard,

2019). The size of the increase is learned only by consulting the original journal article in *JAMA Pediatrics*.

Assuming that there really is an association between social media use and depression, what might explain it? You probably can think of confounding variables that might lead to higher social media use and also be associated with higher levels of depression. One example is participation in extracurricular activities such as sports or club activities. Time spent on those activities may result in less time being available for social media, and might also be associated with lower levels of depression. It's also possible that there is a causal relationship in the other direction, with higher levels of depression leading to the desire to spend more time on social media.

Moral of the Story: A statistically significant finding does not necessarily have practical significance or practical importance. When a study reports a statistically significant finding, find out the magnitude of relationship or difference. When the term "significant" is used in a news report, determine whether it is used in the statistical sense, or in the everyday usage of the term equating it with "important." A secondary moral to this story is that the implied direction of the cause and effect may be wrong. In this case, it could be that people who were more depressed were more drawn to social media.

Definitions: A statistically significant relationship or difference has **practical significance** or **practical importance** if it is large enough to matter in the real world.

Case Study 1.8

Did Your Mother's Breakfast Determine Your Sex?

You've probably heard that "you are what you eat," but did it ever occur to you that you might be who you are because of what your mother ate? A study published in 2008 by the British Royal Society seemed to find just that. The researchers reported that mothers who ate breakfast cereal prior to conception were more likely to have boys than mothers who did not (Mathews et al., 2008). But 9 months later, just enough time for the potential increased cereal sales to have produced a plethora of little baby boys, another study was published that dashed cold milk on the original claim (Young et al., 2009).

The dispute centered on something statisticians call *multiple testing*, which can lead to erroneous findings of statistical significance. The authors of the original study had asked 740 women about 133 different foods they might have eaten just before getting pregnant. They found that 59% of the women who consumed breakfast cereal daily gave birth to a boy, compared to only 43% of the women who rarely or never ate cereal (<http://www.cbsnews.com/stories/2008/04/22/health/webmd/main4036102.shtml>). The result was highly statistically significant, but almost none of the other foods tested showed a statistically significant difference in the ratio of male to female births.

As previously discussed, statistical significance is how statisticians assess whether a difference found in a sample, in this case of 740 women, provides evidence that the difference is likely to represent more than just chance. But sometimes what looks like a statistically significant difference is actually a *false positive*—a difference that looks like it wasn't due to chance when it really was. The more differences that are tested, the more likely it is that one of them will be a false positive. The criticism by Young et al. was based on this idea. When 133 food items that in fact do not affect the sex of a baby are all tested, it is likely that at least one of them will show up as a false positive, showing a big enough difference in the proportion of male to female births to be statistically significant when in fact the difference is due to chance.

The authors of the original study defended their work (Mathews et al., 2009). They noted that they only tested the individual food items after an initial test based on total pre-conception calorie consumption showed a difference in male and female births. They found that 56% of the mothers in the top third of calorie consumption had boys, compared with only 45% of the mothers in the bottom third of calorie consumption. That was one of only two initial tests they did; the other had to do with vitamin intake. With only two tests, it is

Continues

Continued

unlikely that either of them would be a false positive. Unfortunately the media found the cereal connection to be the most interesting result in the study, and that's what received overwhelming publicity. The best way to resolve the debate, as in most areas of science, is to ask the same questions in a new study and see if the results are consistent.

Moral of the Story: *When you read about a study that found a relationship or difference, try to find out how many different things were tested. The more tests that are performed, the more likely it is that a statistically significant difference is a false*

positive that can be explained by chance. You should be especially wary if dozens of things are tested and only one or two of them are statistically significant.

Definitions: **Multiple testing** or **multiple comparisons** in statistics refers to the fact that researchers often test many different hypotheses in the same study. This practice may result in statistically significant findings by mistake, called **false positive** results. Sometimes this practice is called **data snooping** because researchers snoop around in their data until they find something interesting to report.

Thought Question 1.1

According to the American Psychological Association, "Gender is a social construct and a social identity" while "sex refers to biological sex assignment." (Source: <https://apastyle.apa.org/style-grammar-guidelines/bias-free-language/gender>.) Many statistical studies do not make this distinction, yet it is important to understand which meaning applies when interpreting the results of studies. Consider Case Studies 1.1, 1.6 and 1.8. For each study, explain whether gender or sex was involved in collecting the data and interpreting the results, and if it would have been possible for the other one (gender or sex) to have been used instead. If it would have been possible, would it have changed the interpretation of the results?*

1.3 The Common Elements in the Eight Stories

The eight stories were meant to bring life to our definition of statistics. Let's consider that definition again:

STATISTICS is a collection of procedures and principles for gathering data and analyzing information to help people make decisions when faced with uncertainty.

Think back over the stories. In each of them, *data are used to make a judgment about a situation*. This common theme is what statistics is all about. The stories should also help you realize that you can be misled by the use of data. Learning to recognize how that happens is one of the themes of this book.

The Discovery of Knowledge

Each story illustrates part of the process of discovery of new knowledge, for which statistical methods can be very useful. The basic steps in this process are as follows:

1. *Asking the right question(s)*
2. *Collecting useful data*, which includes deciding how much is needed
3. *Summarizing and analyzing data*, with the goal of answering the questions
4. *Making decisions and generalizations* based on the observed data
5. *Turning the data and subsequent decisions into new knowledge*

We'll explore these five steps throughout the book, concluding with a chapter on "Turning Information into Wisdom." We're confident that your active participation in this exploration will benefit you in your everyday life and in your professional career.

***Hint:** In each case, is the biological trait of sex measured, or does the respondent choose a self-identified gender?

In a practical sense, almost all decisions in life are based on knowledge obtained by gathering and using data. Sometimes the data are quantitative, as when an instructor must decide what grades to give based on a collection of homework and exam scores. Sometimes the information is more qualitative and the process of using it to make a decision is informal, such as when you decide what you are going to wear to a party. In either case, the principles in this book will help you to understand how to be a better decision maker.

Thought Question 1.2

Think about a decision that you recently had to make. What “data” did you use to help you make the decision? Did you have as much information as you would have liked? How would you use the principles in this chapter to help you gain more useful information?*

In Summary

Some Important Statistical Principles

The “moral of the story” items for the case studies presented in this chapter give a good overview of many of the important ideas covered in this book. Here is a summary:

- Simple summaries of data can tell an interesting story and are easier to digest than long lists.
- When you read about the change in the rate or risk of occurrence of something, make sure you also find out the base rate or baseline risk.
- A representative sample of only a few thousand, or perhaps even a few hundred, can give reasonably accurate information about a population of many millions.
- An unrepresentative sample, even a large one, tells you almost nothing about the population.
- Cause-and-effect conclusions cannot generally be made on the basis of an observational study.
- Unlike with observational studies, cause-and-effect conclusions can generally be made on the basis of randomized experiments.
- A statistically significant finding does not necessarily have practical significance or importance. When a study reports a statistically significant finding, find out the magnitude of the relationship or difference.
- When you read about a study that found a relationship or difference, try to find out how many different things were tested. The more tests that are done, the more likely it is that a statistically significant difference is a false positive that can be explained by chance.

***HINT:** As an example, how did you decide to live where you are living? What additional data, if any, would have been helpful?

Key Terms

Every term in this chapter is discussed more extensively in later chapters, so don’t worry if you don’t understand all of the terminology that has been introduced here. The following list indicates the page number(s) where the important terms in this chapter are introduced and defined.

Section 1.1
statistics, 3

Case Study 1.1
dotplot, 4

summary statistics, 4
five-number summary, 4
data, 4
median, 4

lower quartile, 4
upper quartile, 4

Case Study 1.2
rate, 5

risk, 5
base rate, 5
baseline risk, 5

Case Study 1.3

population, 5, 6
random sample, 5, 6
poll, 5, 6
sample survey, 5, 6
margin of error, 5, 6
(margin of) sampling error, 6

Case Study 1.4

nonparticipation bias, 6
nonresponse bias, 6
self-selected sample, 6
volunteer sample, 6

Case Study 1.5

observational study, 7
variable, 7
confounding variable, 7

Case Studies 1.6 and 1.7

randomized experiment, 7, 8

treatment, 7, 8
random assignment, 7, 8
placebo, 7, 8
statistically significant, 8, 9
practical significance, 8, 9
practical importance, 8, 9

Case Study 1.8

multiple testing, 9, 10
multiple comparisons, 10
false positive, 9, 10
data snooping, 10

Exercises

Bold exercises have answers in the back of the text.

Note: Many of these exercises will be repeated in later chapters in which the relevant material is covered in more detail.

Skillbuilder Exercises

- 1.1** Refer to the data and five-number summaries given in Case Study 1.1. Give a numerical value for each of the following.
- The fastest speed driven by anyone in the class.
 - The slowest of the “fastest speeds” driven by a male.
 - The speed for which one-fourth of the females had driven at that speed or faster.
 - The proportion of females who had driven 89 mph or faster.
 - The number of females who had driven 89 mph or faster.
- 1.2** A five-number summary for the heights in inches of the self-identified female students who participated in the survey in Case Study 1.1 is as shown:

	Female Students' Heights (inches)	
Median	65	
Quartiles	63.5	67.5
Extremes	59	71

- What is the median height for these students?
 - What is the range of heights—that is, the difference in heights between the shortest and the tallest of these students?
 - What is the interval of heights containing the shortest one-fourth of these students?
 - What is the interval of heights containing the middle one-half of these students?
- 1.3** In recent years, Vietnamese Americans have had the highest rate of cervical cancer in the country. Suppose that among 200,000 Vietnamese American women, 86 developed cervical cancer in the past year. (Source: <https://rtips.cancer.gov/rtips/programDetails.do?programId=1427816>)
- Calculate the rate of cervical cancer for these women.
 - What is the estimated risk of developing cervical cancer for Vietnamese American women in the next year?

- Explain the conceptual difference between the rate and the risk, in the context of this example.

- 1.4** The risk of getting lung cancer at some point in one's life for men who have never smoked is about 13 in 1000. The risk for men who smoke is just over 13 times the risk for nonsmokers. (Source: Villeneuve and Lau, 1994)
- What is the base rate for lung cancer in men over a lifetime?
 - What is the approximate lifetime risk of getting lung cancer for men who smoke?
- 1.5** Refer to Case Study 1.3, in which teens were asked about their dating behavior.
- What population is represented by the random sample of 602 teens?
 - What population is represented by the 496 teens in the sample who had dated?
- 1.6** Using Case Study 1.6 as an example, explain the difference between a population and a sample.
- 1.7** A CBS News poll taken in December 2009 asked a random sample of 1048 adults in the United States, “In general, do you think the education most children are getting today in public schools is better, is about the same, or is worse than the education you received?” About 34% said “Better,” 24% said “About the same,” and 38% said “Worse.” (The remaining 4% were unsure.)
- What is the population for this survey?
 - What is the approximate margin of error for this survey?
 - Provide an interval that is 95% certain to cover the true percentage of U.S. adults in December 2009 who would have answered “Better” to this question if asked.
- 1.8** A telephone survey of 2000 Canadians conducted March 20–30, 2001, found that “Overall, about half of Canadians in the poll say the right number of immigrants are coming into the country and that immigration has a positive effect on Canadian communities. Only 16 percent view it as a negative impact while one third said it had no impact at all” (*The Ottawa Citizen*, August 17, 2001, p. A6).
- What is the population for this survey?
 - How many people were in the sample used for this survey?

Bold exercises answered in the back

- c. What is the approximate margin of error for this survey?
- d. Provide an interval of numbers that is 95% certain to cover the true percentage of Canadians who view immigration as having a negative impact.
- 1.9** In Case Study 1.3, the margin of error for the sample of 496 teenagers was about 4.5%. How many teenagers should be in the sample to produce an approximate margin of error of .05 or 5%?
- 1.10** About how many people would need to be in a random sample from a large population to produce an approximate margin of error of .10 or 10%?
- 1.11** A popular Sunday newspaper magazine often includes a yes-or-no survey question such as “Do you think there is too much violence on television?” or “Do you think parents should use physical discipline?” Readers are asked to send their answers to the magazine, and the results are reported in a subsequent issue.
- What is this type of sample called?
 - Do you think the results of these polls represent the opinions of all readers of the magazine? Explain.
- 1.12** A proposed study design is to leave 100 questionnaires by the checkout line in a student cafeteria. The questionnaire can be picked up by any student and returned to the cashier. Explain why this volunteer sample is a poor study design.
- 1.13** For each of the examples given here, decide whether the study was an observational study or a randomized experiment.
- A group of students enrolled in an introductory statistics course were randomly assigned to take either an online course or a traditional lecture course. The two methods were compared by giving the same final examination in both courses.
 - A group of smokers and a group of nonsmokers who visited a particular clinic were asked to come in for a physical exam every 5 years for the rest of their lives to monitor and compare their health status.
 - CEOs of major corporations were compared with other employees of the corporations to see if the CEOs were more likely to have been the first child born in their families than were the other employees.
- 1.14** For each of the studies described, explain whether the study was an observational study or a randomized experiment.
- A group of 100 students was randomly divided, with 50 assigned to receive vitamin C and the remaining 50 to receive a placebo, to determine whether or not vitamin C helps to prevent colds.
 - A random sample of patients who received a hip transplant operation at Stanford University Hospital during 2000 to 2010 were followed for 10 years after their operation to determine the success (or failure) of the transplant.
 - Volunteers with high blood pressure were randomly divided into two groups. One group was taught to practice meditation and the other group was given a low-fat diet. After 8 weeks, reduction in blood pressure was compared for the two groups.
- 1.15** Read Case Study 1.5. Give an example of a confounding variable that might explain why elderly people who attended religious services might have lower blood pressure than those who did not. Do not use one of the variables already mentioned in the Case Study.
- 1.16** Suppose that an observational study showed that students who got at least 7 hours of sleep performed better on exams than students who got less than 7 hours of sleep. Which of the following are possible confounding variables, and which are not? Explain why in each case.
- Number of courses the student took that term.
 - Weight of the student.
 - Number of hours the student spent partying in a typical week.
- 1.17** A randomized experiment was done in which overweight men were randomly assigned to either exercise or go on a diet for a year. At the end of the study there was a statistically significant difference in average weight loss for the two groups. What additional information would you need in order to determine if the difference in average weight loss had *practical* importance?
- 1.18** Explain the distinction between statistical significance and practical significance. Can the result of a study be statistically significant but not practically significant? Explain your answer.
- 1.19** A (hypothetical) study of what people do in their spare time found that people born under the astrological sign of Aries were significantly more likely to be regular swimmers than people born under other signs. What additional information would you want to know to help you determine if this result is a false positive?
- 1.20** Explain what is meant by a “false positive” in the context of conclusions in statistical studies.

Chapter Exercises

- 1.21** Refer to Case Study 1.6, in which the relationship between aspirin and heart attack rates was examined. Using the results of this experiment, what do you think is the base rate of heart attacks for men like the ones in this study? Explain.
- 1.22** Students in a statistics class at Penn State were asked, “About how many minutes do you typically exercise in a week?” Responses from the self-identified *women* in the class were
- 60, 240, 0, 360, 450, 200, 100, 70, 240, 0, 60, 360, 180, 300, 0, 270
- Responses from the self-identified *men* in the class were
- 180, 300, 60, 480, 0, 90, 300, 14, 600, 360, 120, 0, 240
- Compare the women to the men using a dotplot. What does your plot show you about the difference between the men and the women?
 - For each gender, determine the median response.
 - Do you think there’s a “significant” difference between the weekly amount that men and women exercise? Explain.
- 1.23** Refer to Exercise 1.22.
- Create a five-number summary for the men’s responses. Show how you found your answer.

- b. Use your five-number summary to describe in words the exercise behavior of this group of men.
- 1.24 Refer to Exercise 1.22.
- Create a five-number summary for the women's responses. Show how you found your answer.
 - Use your five-number summary to describe in words the exercise behavior of this group of women.
- 1.25 An article in the magazine *Science* (Service, 1994) discussed a study comparing the health of 6000 vegetarians and a similar number of their friends and relatives who were not vegetarians. The vegetarians had a 28% lower death rate from heart attacks and a 39% lower death rate from cancer, even after the researchers accounted for differences in smoking, weight, and social class. In other words, the reported percentages were the remaining differences after adjusting for differences in death rates due to those factors.
- Is this an observational study or a randomized experiment? Explain.
 - On the basis of this information, can we conclude that a vegetarian diet causes lower death rates from heart attacks and cancer? Explain.
 - Give an example of a potential confounding variable and explain what it means to say that it is a confounding variable.
- 1.26 Refer to Exercise 1.25, comparing vegetarians and nonvegetarians for two causes of death. Were base rates given for the two causes of death? If so, what were they? If not, explain what a base rate would be for this study.
- 1.27 An article in the *Sacramento Bee* (March 8, 1984, p. A1) reported on a study finding that "men who drank 500 ounces or more of beer a month (about 16 ounces a day) were three times more likely to develop cancer of the rectum than non-drinkers." In other words, the rate of cancer in the beer-drinking group was three times that of the non-beer drinkers in this study. What important numerical information is missing from this report?
- 1.28 Dr. Richard Hurt and his colleagues (Hurt et al., 1994) randomly assigned volunteers who wanted to quit smoking to wear either a nicotine patch or a placebo patch to determine whether wearing a nicotine patch improves the chance of quitting. After 8 weeks of use, 46% of those wearing the nicotine patch had quit smoking, but only 20% of those wearing the placebo patch had quit.
- Was this a randomized experiment or an observational study?
 - The difference in the percentage of participants who quit (20% versus 46%) was statistically significant. What conclusion can be made on the basis of this study?
 - Why was it advisable to assign some of the participants to wear a placebo patch?
- 1.29 Refer to the study in Exercise 1.28, in which there was a statistically significant difference in the percentage of smokers who quit using a nicotine patch and a placebo patch. Now read the two cautions in the "moral of the story" for Case Study 1.7. Discuss each of them in the context of this study.
- 1.30 Refer to the study in Exercises 1.28 and 1.29, comparing the percentage of smokers who quit using a nicotine patch and a placebo patch. Refer to the definition of statistics given on page 3, and explain how it applies to this study.
- 1.31 Case Study 1.6 reported that the use of aspirin reduces the risk of heart attack and that the relationship was found to be "statistically significant." Does either of the cautions in the "moral of the story" for Case Study 1.7 apply to this result? Explain.
- 1.32 A random sample of 1001 University of California faculty members taken in December 1995 was asked, "Do you favor or oppose using race, religion, sex, color, ethnicity, or national origin as a criterion for admission to the University of California?" (Roper Center, 1996). Fifty-two percent responded "favor."
- What is the population for this survey?
 - What is the approximate margin of error for the survey?
 - Based on the results of the survey, could it be concluded that a majority (over 50%) of *all* University of California faculty members favor using these criteria? Explain.
- 1.33 A Pew Research Center poll conducted in 2009 asked people if they had ever seen a ghost. Of the 2003 respondents, 360 said "yes." (<https://www.pewresearch.org/fact-tank/2015/10/30/18-of-americans-say-theyve-seen-a-ghost/>)
- What is the approximate margin of error that accompanies this result?
 - What is the interval that is 95% certain to contain the actual proportion of people in 2009 who would have said that they had seen a ghost?
- 1.34 Refer to Exercise 1.33. What is the risk of someone in this population having seen a ghost?
- 1.35 Refer to Exercise 1.33. The Pew Research Center selected a random sample of adults in the United States for this poll. Suppose listeners to a late-night radio talk show were asked to call and report whether or not they had ever seen a ghost.
- What is this type of sample called?
 - Do you think the proportion reporting that they had seen a ghost for the radio poll would be higher or lower than the proportion for the Pew Research Center poll? Explain.
- 1.36 The CNN website sometimes has a small box called "Quick vote" that contains a question about an interesting topic in the news that day. For example, one question in February 2010 asked "Should the U.S. military let gays and lesbians serve openly?" Visitors to the website are invited to click their response and to view the results. When the results are displayed they contain the message "This is not a scientific poll."
- What type of sample is obtained in this Quick vote?
 - What do you think is meant by the message that "This is not a scientific poll?"
- 1.37 Explain what is meant by "data snooping."
- 1.38 A headline in a major newspaper read, "Breast-fed youth found to do better in school."
- Do you think this statement was based on an observational study or a randomized experiment? Explain.

- b. Given your answer in part (a), which of these two alternative headlines do you think would be preferable: "Breast-feeding leads to better school performance" or "Link found between breast-feeding and school performance"? Explain.
- 1.39** In this chapter, you learned that cause and effect can be concluded from randomized experiments but generally not from observational studies. Why don't researchers simply conduct all studies as randomized experiments rather than observational studies?
- 1.40** Why was the study described in Case Study 1.5 conducted as an observational study instead of an experiment?
- 1.41** Give an example of a question you would like to have answered, such as "Does eating chocolate help to prevent depression?" Then explain how a randomized experiment or an observational study could be done to study this question.
- 1.42** Suppose you were to read the following news story: "Researchers compared a new drug to a placebo for treating high blood pressure, and it seemed to work. But the researchers were concerned because they found that significantly more people got headaches when taking the new drug than when taking the placebo. Headaches were the only problem out of the 20 possible side effects the researchers tested."
- Do you think the research used an observational study or a randomized experiment? Explain.
 - Do you think the researchers are justified in thinking the new drug would cause more headaches in the population than the placebo would? Explain.
- 1.43** Refer to Case Study 1.5. Explain what mistakes were made in the implementation of steps 4 and 5 of "The Discovery of Knowledge" when *USA Today* reported the results of this study.
- 1.44** Refer to Case Study 1.6. Go through the five steps listed under "The Discovery of Knowledge" in Section 1.3, and show how each step was addressed in this study.
- 1.45** According to the American Psychological Association, "Gender is a social construct and a social identity" while "sex refers to biological sex assignment." (Source: <https://apastyle.apa.org/style-grammar-guidelines/bias-free-language/gender>.) In each of the following proposed studies, would categorizing adults by sex or by gender be more appropriate? Explain your reasoning.
- A study of the most common forms of cancer for adult males and separately, the most common forms of cancer for adult females.
 - A study of the relationship between man/woman non-binary and opinion on the death penalty.
 - A national survey asking at what age the respondent was first married, for respondents who are married.
 - A study to compare the average body temperature for male twelve-graders and female twelfth-graders.

2



ANON_TAE/SHUTTERSTOCK.COM

Who would you expect to see behind the wheel of this speeding car?

See Example 2.13 (p. 43)

Turning Data into Information

Learning Objectives

After completing this chapter, you will be able to:

- Distinguish between categorical variables and quantitative variables.
- Create visual and numerical summaries for one or two categorical variables.
- Create visual displays for one quantitative variable.
- Describe a dataset with one quantitative variable using numerical summaries.
- Explain how to identify and manage outliers.
- Describe how the mean, standard deviation, z-scores, and the Empirical Rule are used to display the possible values in a bell-shaped distribution.

In Case Study 1.1, we analyzed the responses that 189 college students gave to the question “What’s the fastest you’ve ever driven a car?” The “moral of the story” for that case study was that *simple summaries of data can tell an interesting story and are easier to digest than long lists*. In this chapter, you will learn how to create simple summaries and pictures from various kinds of raw data.

2.1 Raw Data

Raw data is a term used for numbers and category labels that have been collected but have not yet been processed in any way. For example, here is a list of questions asked in a large statistics class and the “raw data” given by one of the students:

- | | |
|---|---------------------|
| 1. What is your sex (m = male, f = female)? | Raw data: m |
| 2. How many hours did you sleep last night? | Raw data: 5 hours |
| 3. Randomly pick a letter—S or Q. | Raw data: S |
| 4. What is your height in inches? | Raw data: 67 inches |
| 5. Randomly pick a number between 1 and 10. | Raw data: 3 |
| 6. What’s the fastest you’ve ever driven a car (mph)? | Raw data: 110 mph |
| 7. What is your right handspan in centimeters? | Raw data: 21.5 cm |
| 8. What is your left handspan in centimeters? | Raw data: 21.5 cm |

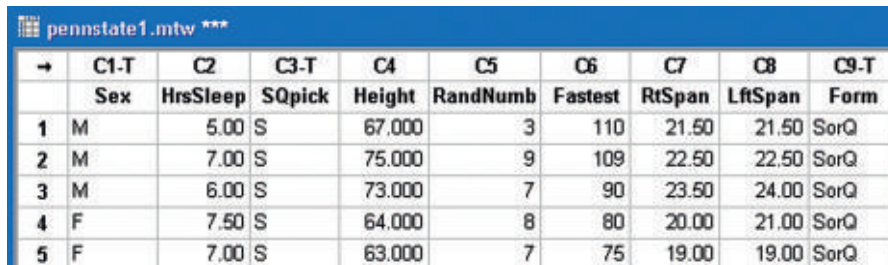
For questions 7 and 8, a centimeter ruler was provided on the survey form, and handspan was defined as the distance covered on the ruler by a stretched hand from the tip of the thumb to the tip of the small finger. For question 3, about one-half of the students saw the choice of letters in reverse order, so their question was “Randomly pick a letter — Q or S.” This was done to learn whether or not students might be more likely to pick the first choice offered, regardless of whether it was the S or the Q. If you do Exercise 2.35 at the end of this chapter, you will learn the result. You may be wondering why question 5 was asked. Your curiosity will be satisfied as you keep reading this chapter.

Datasets, Observations, and Variables

A **variable** is a characteristic that can differ from one individual to the next. Students in the statistics class provided raw data for eight variables: sex, hours of sleep, choice of a letter, height, choice of a number, fastest speed ever driven, right handspan, and left handspan. The instructor imposed a ninth variable: the order of listing S and Q in question 3.

An **observational unit** is a single individual entity, a person for instance, in a study. More simply, an individual entity may be called an **observation**. The word *observation* might also be used to describe the value of a single measurement, such as height = 67 inches. The **sample size** for a study is the total number of observational units. The letter n is used to represent the sample size; one hundred and ninety students participated in the class survey, so the sample size is $n = 190$. (One student did not report a “fastest speed”, so $n = 189$ for Case Study 1.1.)

A **dataset** is the complete set of raw data, for all observational units and variables, in a survey or experiment. When statistical software or a spreadsheet is used to summarize the raw data, the dataset typically is organized so that each row gives the data for one observational unit and each column gives the raw data for a particular variable. For the statistics class survey, Figure 2.1 shows the first five rows of a dataset created for the Minitab statistical software program. These five rows give the raw data for five of the 190 students in the dataset. Note that each column label indicates which variable is in the column. (T after a column number indicates text.) The final column indicates the order of presenting the letters in question 3.



	C1-T	C2	C3-T	C4	C5	C6	C7	C8	C9-T
	Sex	HrsSleep	SQpick	Height	RandNumb	Fastest	RtSpan	LftSpan	Form
1	M	5.00	S	67.000	3	110	21.50	21.50	SorQ
2	M	7.00	S	75.000	9	109	22.50	22.50	SorQ
3	M	6.00	S	73.000	7	90	23.50	24.00	SorQ
4	F	7.50	S	64.000	8	80	20.00	21.00	SorQ
5	F	7.00	S	63.000	7	75	19.00	19.00	SorQ

FIGURE 2.1 Minitab worksheet with dataset

Data from Samples and Populations

Researchers often use sample data to make inferences about the larger population represented by the data. Occasionally, in a **census**, data are collected from all members of a population.

- **Sample data** are collected when measurements are taken from a subset of a population.
- **Population data** are collected when all individuals in a population are measured.

Sometimes the reason for collecting the data creates this distinction. For instance, data collected from all students in a statistics class are sample data when we use them to represent a larger collection of students, but are population data if we only care about describing the students in that class.

It is generally important to determine whether raw data are sample data or population data. However, most of the descriptive methods for summarizing data explained in this chapter are the same for both sample and population data. Therefore, in this chapter we will only distinguish between sample and population data when the notation differs for the two situations. We will begin emphasizing the distinction between samples and populations in Chapter 9.

Parameters and Statistics

The generic names used for summary measures from sample and population data also differ. A summary measure computed from sample data is called a **statistic**, while a summary measure using data for an entire population is called a **parameter**. This distinction is often overlooked when we are interested only in numerical summaries for either a sample or a population. In that case, the summary numbers are simply called **descriptive statistics** for either a sample or a population.



2.1 Exercises are on page 56.

In Summary

Basic Data Concepts

- An **observational unit**, or **observation**, is an individual entity in a study. An individual measurement is also called an *observation*.
- A **variable** is a characteristic that may differ (vary) among individuals.
- **Sample data** are collected from a subset of a larger population.
- **Population data** are collected when all individuals in a population are measured.
- A **statistic** is a summary measure of sample data.
- A **parameter** is a summary measure of population data.

Thought Question 2.1

There were almost 200 students who answered the survey questions shown on page 17. Formulate four interesting questions that you would like to answer using the data from these students. What kind of summary information would help you answer your questions?*

2.2 Types of Variables

We learned in the previous section that a **variable** is a characteristic that may differ from one individual to the next. A variable may be a *categorical* characteristic, such as a person's blood type, or a *numerical* characteristic, such as hours of sleep last night. To determine what type of summary might provide meaningful information, you first have to recognize which type of variable you want to summarize.

For a **categorical variable**, the raw data consist of group or category names that don't necessarily have any logical ordering. Each individual falls into one and only one category. For a categorical variable, the most fundamental summaries are how many individuals and what percent of the total fall into each category.

The term **ordinal variable** may be used to describe the data when a categorical variable has ordered categories. For example, suppose that you are asked to rate your driving skills compared to the skills of other drivers, using the codes 1 = better, 2 = the same, and 3 = worse. The response is an ordinal variable because the response categories are ordered perceptions of driving skills.

Following are a few examples of categorical variables and their possible categories. The final variable in the list, the rating of a teacher, is ordinal because the response categories convey an ordering.

Categorical Variable	Possible Categories
Dominant hand	Left-handed, Right-handed, Ambidextrous
Regular church attendance	Yes, No
Opinion about marijuana legalization	In favor, Opposed, Not sure
Eye color	Brown, Blue, Green, Hazel, Other
Teacher Rating	Scale of 1 to 7, 1 = Poor, 7 = Excellent

For a **quantitative variable**, the raw data are either numerical measurements or counts collected from each individual. All individuals can be meaningfully ordered

***HINT:** An example is, "What was the average amount of sleep for these students?" Case Study 1.1 could be utilized to generate another example.

according to these values, and averaging and other arithmetic operations make sense for these data. A few examples of quantitative variables follow:

Quantitative Variable	Possible Responses
Height	Measured height in inches
Weight	Measured weight in pounds
Amount of sleep last night	Self-reported sleep in hours
Classes missed last week	Count of missed classes
Number of siblings	Count of brothers and sisters

Not all numbers fit the definition of a quantitative variable. For instance, Social Security numbers or student identification numbers may carry some information (such as region of the country where the Social Security number was obtained), but it is not generally meaningful to put them into numerical order or to determine the average Social Security number.

Supplemental Note

Summarizing Ordinal Variables

The way we summarize an ordinal variable can depend on the purpose. As an example, consider the self-rating of driver skill with 1 = better than, 2 = the same as, and 3 = worse than other drivers. We can summarize the responses in the way we usually do for a categorical variable, which is to find the number and percentage of the sample who responded in each category. It could also be informative to treat the variable as quantitative data and find the average of the responses to see whether or not it is close to 2, which it should be if all respondents give an honest appraisal of their abilities. On the other hand, it would not make sense to talk about “average household income” using the numerical codes attached to broad income categories like the ones shown on this page.

Measurement variable and **numerical variable** are synonyms for a quantitative variable. The term **continuous variable** can also be used for quantitative data when every value within some interval is a possible response. For example, height is a continuous quantitative variable because any height within a particular range is possible. The limitations of measuring tapes, however, don’t allow us to measure heights accurately enough to find that a person’s actual height is 66.5382617 inches. Even if we could measure that accurately, we would usually prefer to round such a height to 66.5 inches. The distinction between quantitative variables that are continuous and those that are not will be expanded in Chapter 8 when we study probability distributions.

A variable type can depend on how something is measured. For instance, household income is a numerical value with two digits after the decimal place, and if it is recorded this way, it is a quantitative variable. Researchers often collect household income data using ordered categories, however, such as 1 = less than \$25,000, 2 = \$25,000 to \$49,999, 3 = \$50,000 to \$74,999, and so on. With categories like these, household income becomes a categorical variable or, more specifically, an ordinal variable. In some situations, household income could be categorized very broadly, as when it is used to determine whether or not someone qualifies for a loan. In that case, income may be either “high enough” to qualify for the loan or “not high enough.”

Raw data for quantitative and categorical variables are summarized differently. It makes sense, for example, to calculate the average number of hours of sleep last night for the members of a group, but it doesn’t make sense to calculate the average blood type (A+, B+, etc.) for the group. For blood type data, it makes more sense to determine the number and proportion of the group who are in each blood type category. Usually ordinal variables are summarized using the same methods used for categorical variables, although occasionally they are summarized as quantitative variables.

Thought Question 2.2

Review the data collected in the statistics class, listed in Section 2.1, and identify a type for each variable. The only one that is ambiguous is question 5. That question asks for a numerical response, but as we will see later in this chapter, it is more interesting to summarize the responses as if they are categorical.*

Asking the Right Questions

As with most situations in life, the information you get when you summarize a dataset depends on how careful you are about asking for what you want. Here are some examples of the types of questions that are most commonly of interest for different kinds of variables and combinations of variables.

***HINT:** For each variable, consider whether the raw data are meaningful quantities or category names.

One Categorical Variable

Example: What percentage of college students favors the legalization of marijuana, and what percentage of college students opposes legalization of marijuana?

Opinion about the legalization of marijuana is a categorical variable with three possible response categories (favor, oppose, or not sure). For one categorical variable, it is useful to ask what percentage of individuals falls into each category.

Two Categorical Variables

Example: In Case Study 1.6, the researchers asked if the likelihood of a male physician having a heart attack depends on whether he has been taking aspirin or taking a placebo.

The two categorical variables here are whether or not a physician had a heart attack and whether the physician took aspirin or a placebo. For two categorical variables, we ask if there is a relationship between the two variables. Does the chance of falling into a particular category for one variable depend on which category an individual is in for the other variable?

One Quantitative Variable

Example: What is the average body temperature for adults, and how much variability is there in body temperature measurements?

Body temperature is a quantitative variable. To summarize one quantitative variable, we typically ask about summary measures such as the average or the range of values.

One Categorical and One Quantitative Variable

Example: On average, is “fastest ever driven speed” the same for female and male drivers?

We are considering how a quantitative variable (fastest speed ever driven) is related to a categorical variable (gender). A general question about this type of situation is whether the quantitative measurements are similar across the categories or whether they differ. This question could be approached by examining whether or not the average measurement (such as average fastest speed ever driven) is different for the two categories (male drivers and female drivers). We might also ask whether or not the range of measurements is different across the categories.

Two Quantitative Variables

Example: Does average body temperature change as people age?

Age and body temperature, the two variables in this example, are both quantitative variables. A question we ask about two quantitative variables is whether they are related so that when measurements are high (low) on one variable the measurements for the other variable also tend to be high (low).

Explanatory and Response Variables

Three of the questions just listed were about the relationship between two variables. In these instances, we usually can identify one variable as the **explanatory variable** and the other variable as the **response variable**. The value of the *explanatory variable* might partially explain the value of the *response variable* for an individual. For example, in the relationship between smoking and lung cancer, whether or not an individual smokes is the explanatory variable, and whether or not they develop lung cancer is the response variable. If we note that people with higher education levels generally have higher incomes, education level is the explanatory variable and income is the response variable.

The identification of one variable as “explanatory” and the other as “response” does not imply that there is a *causal* relationship. It simply implies that knowledge of the value of the explanatory variable may help provide knowledge about the value of the response variable for an individual. Occasionally, we simply want to know if two variables are related, but there is no clear explanatory or response variable. An example is handspan and foot length. We expect people with bigger hands to have bigger feet, but we would not consider one of the variables to be explaining the other.



2.2 Exercises are on pages 56–57.

In Summary

Types of Variables and Roles for Variables

- A **categorical variable** is a variable for which the raw data are group or category names that don't necessarily have a logical ordering. Examples include eye color and country of residence.
- An **ordinal variable** is a categorical variable for which the categories have a logical ordering or ranking. Examples include highest educational degree earned and T-shirt size (S, M, L, XL).
- A **quantitative variable** is a variable for which the raw data are numerical measurements or counts collected from each individual. Examples include height and number of siblings.
- In a relationship between two variables, regardless of type, an **explanatory variable** is one that might partially explain the value of a **response variable** for an individual.

2.3 Summarizing One or Two Categorical Variables

Numerical Summaries

To summarize a categorical variable, first count how many individuals fall into each possible category. Percentages usually are more informative than counts, so the second step is to calculate the percentage in each category. These two easy steps can also be used to summarize a combination of two categorical variables.

Example 2.1

Seatbelt Use by Twelfth-Graders One question asked in a 2003 nationwide survey of American high school students was, "How often do you wear a seatbelt when driving a car?" The biennial survey, organized by the U.S. Centers for Disease Control and Prevention, is conducted as part of a federal program called the Youth Risk Behavior Surveillance System. Possible answers for the seatbelt question were Always, Most times, Sometimes, Rarely, and Never. Respondents could also say that they didn't drive.

Table 2.1 summarizes responses given by twelfth-grade students who said that they drive. The total sample size for the table is $n = 3042$ students. Note that a majority, $1686/3042 = .554$, or 55.4%, said that they always wear a seatbelt when driving, while just $115/3042 = .038$, or 3.8%, said that they never wear a seatbelt. To find the percentage who either rarely or never wears a seatbelt, we sum the percentages in the Rarely and Never categories. This is $8.2\% + 3.8\% = 12\%$.

The survey also asked "What is your sex?" and the response options were "female" or "male." One stereotype about males and females is that males are more likely to engage in risky behaviors than females are. Are female drivers more likely to say that they always wear a seatbelt? Are male drivers more likely to say they rarely or never wear a seatbelt? Table 2.2 summarizes seatbelt use for male and female twelve-grade students in the sample. Percentages are given within each sex. Among female drivers, 915 out of 1467 = 62.4% said that they always wear a seatbelt compared to 771 out of 1575 = 49.0% of the

TABLE 2.1 Seatbelt Use by Twelfth-Graders When Driving

Response	Count	Percent
Always	1686	55.4%
Most times	578	19.0%
Sometimes	414	13.6%
Rarely	249	8.2%
Never	115	3.8%
Total	3042	100%

Source: Centers for Disease Control and Prevention, <http://www.cdc.gov/HealthyYouth/yrbs/index.htm>.

TABLE 2.2 Sex and Seatbelt Use by Twelfth-Graders When Driving

	Always	Most Times	Sometimes	Rarely	Never	Total
Female	915 (62.4%)	276 (18.8%)	167 (11.4%)	84 (5.7%)	25 (1.7%)	1467 (100%)
Male	771 (49.0%)	302 (19.2%)	247 (15.7%)	165 (10.5%)	90 (5.7%)	1575 (100%)

Source: <http://www.cdc.gov/HealthyYouth/yrbs/index.htm>.

male drivers. Male respondents were more likely than female respondents to rarely or never use seatbelts. Adding the percentages for Rarely and Never gives $10.5\% + 5.7\% = 16.2\%$ for the males and $5.7\% + 1.7\% = 7.4\%$ for the females.

Do these sample data provide enough information for us to infer that sex and seatbelt use are related variables in the larger population of all U.S. twelfth-grade drivers? We will learn how to answer this type of question in Chapters 4 and 15.

Frequency and Relative Frequency

In general, the **distribution** of a variable describes how often the possible responses occur.

- A **frequency distribution** for a categorical variable is a listing of all categories along with their frequencies (counts).
- A **relative frequency distribution** is a listing of all categories along with their relative frequencies (given as proportions or percentages).

It is commonplace to give the frequency and relative frequency distributions together, as was done in Table 2.1.

Example 2.2

Lighting the Way to Nearsightedness A survey of 479 children found that those who had slept with a nightlight or in a fully lit room before the age of 2 had a higher incidence of nearsightedness (myopia) later in childhood (*Sacramento Bee*, May 13, 1999, pp. A1, A18). The raw data for each child consisted of two categorical variables, each with three categories. Table 2.3 gives the categories and the number of children falling into each combination of them. The table also gives percentages (relative frequencies) falling into each eyesight category, where percentages are computed within each nighttime lighting category. For example, among the 172 children who slept in darkness, about 90% ($155/172 = .90$) had no myopia.

TABLE 2.3 Nighttime Lighting in Infancy and Eyesight

Slept with:	No Myopia	Myopia	High Myopia	Total
Darkness	155 (90%)	15 (9%)	2 (1%)	172 (100%)
Nightlight	153 (66%)	72 (31%)	7 (3%)	232 (100%)
Full Light	34 (45%)	36 (48%)	5 (7%)	75 (100%)
Total	342 (71%)	123 (26%)	14 (3%)	479 (100%)

Source: From Nature 1999, Vol. 399, pp. 113–114.

The pattern in Table 2.3 is striking. As the amount of sleeptime light increases, the incidence of myopia also increases. However, this study does not prove that sleeping with light actually *caused* myopia in more children. There are other possible explanations. For example, myopia has a genetic component, so those children whose parents have myopia are more likely to suffer from it themselves. Maybe nearsighted parents are more likely to provide light while their children are sleeping.

Thought Question 2.3

Can you think of possible explanations for the observed relationship between use of nightlights and myopia, other than direct cause and effect? What additional information might help to provide an explanation?*

***HINT:** Reread Example 2.2, in which one possible explanation is mentioned. What data would we need to investigate the possible explanation mentioned there?

Explanatory and Response Variables for Categorical Variables

In many summaries of two categorical variables, we can identify one variable as an explanatory variable and the other as a response variable (**outcome variable**). For instance, in Example 2.1, sex (male, female) was an explanatory variable and how often a student wears a seatbelt when driving was the response variable. In Example 2.2, the amount of sleeptime lighting was an explanatory variable and the degree of myopia was the response variable.

In both Tables 2.2 and 2.3, the explanatory variable categories defined the rows and the response variable categories defined the columns. Tables often are formed this way, although not always. When they are, row percentages are more informative than column percentages. In Tables 2.2 and 2.3, percentages were given across rows. For instance, Table 2.3 shows that 90% of children who slept in darkness did not have myopia but only 45% of those who had slept in full light did not have myopia.

No matter how the table is constructed, determine whether one variable is an explanatory variable and the other is a response variable. Within each explanatory variable category we are interested in the percentage falling into each response variable category.

Minitab Tip

Numerically Describing One or Two Categorical Variables

- To determine how many and what percentage fall into the categories of a single categorical variable, use **Stat > Tables > Tally Individual Variables**. In the dialog box, specify a column containing the raw data for a categorical variable. Click on any desired options for counts and percentages under “Display.”
- To create a two-way table for two categorical variables, use **Stat > Tables > Crosstabulation and Chi-Square**. Specify a categorical variable in the “For rows” box and another categorical variable in the “For columns” box. Select any desired percentages (row, column, and/or total) under “Display.”

SPSS Tip

Numerically Describing One or Two Categorical Variables

- To create a frequency table for one categorical variable, use **Analyze > Descriptive Statistics > Frequencies**.
- To create a two-way table for two categorical variables, use **Analyze > Descriptive Statistics > Crosstabs**. Use the **Cells** button to request row and/or column percentages.

Visual Summaries for Categorical Variables

Two simple visual summaries are used for categorical data:

- **Pie charts** are useful for summarizing a single categorical variable if there are not too many categories.
- **Bar graphs** are useful for summarizing one or two categorical variables. They are particularly useful for making comparisons when there are two categorical variables.

Both of these simple graphical displays are easy to construct and interpret, as the following examples demonstrate.

Example 2.3

Humans Are Not Good Randomizers Question 5 in the class survey described in Section 2.1 asked students to “Randomly pick a number between 1 and 10.” The pie chart shown in Figure 2.2 illustrates that the responses are not even close to being evenly distributed across the numbers. Note that almost 30% of the students chose 7, while only just over 1% chose the number 1.

Figure 2.3 illustrates the same results with a bar graph. This bar graph shows the frequencies of responses on the vertical axis and the possible response categories on the horizontal axis. The display makes it obvious that the number of students who chose 7 was more than double that of the next most popular choice. We also see that very few students chose either 1 or 10.

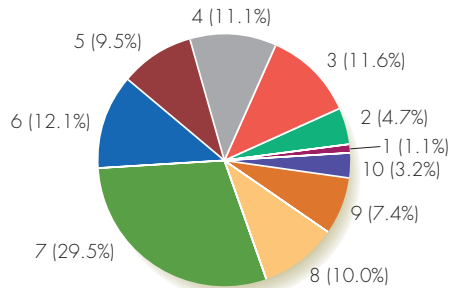


FIGURE 2.2 Pie chart of numbers picked

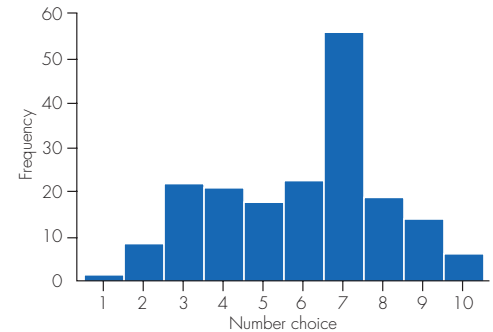


FIGURE 2.3 Bar graph of numbers picked

Example 2.4

Revisiting Example 2.2: Nightlight and Nearsightedness Figure 2.4 illustrates the data presented in Example 2.2 with a bar chart showing, for each lighting group, the percentage that ultimately had each level of myopia. This bar chart differs from the one in Figure 2.3 in two respects. First, it is used to present data for two categorical variables instead of just one. Second, the vertical axis represents percentages instead of counts, with the percentages for myopia status computed separately within each lighting category. Within each sleep-time lighting category, the percentages add to 100%.

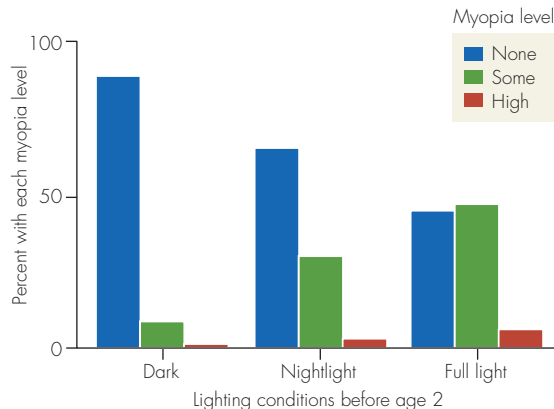


FIGURE 2.4 Bar chart for myopia and nighttime lighting in infancy

Thought Question 2.4

Redo the bar graph in Figure 2.4 using counts instead of percentages. The necessary data are given in Table 2.3. Would the comparison of frequency of myopia across the categories of lighting be as easy to make using the bar graph with counts? Generalize your conclusion to provide guidance about what should be done in similar situations.*

***HINT:** Which graph makes it easier to compare the percentage with myopia for the three groups? What could be learned from the graph of counts that isn't apparent from the graph of percentages?

In Summary**Bar Graphs for Categorical Variables**

In a bar graph for *one categorical variable*, you can choose one of the following to display as the height of a bar for each category, indicated by labeling the vertical axis:

- Frequency or count
- Relative frequency = number in category/total number
- Percentage = relative frequency \times 100%

In a bar graph for *two categorical variables*, if an explanatory and response variable can be identified, it is most common to:

- Draw a separate group of bars for each category of the explanatory variable.
- Within each group of bars, draw one bar for each category of the response variable.
- Label the vertical axis with percentages and make the heights of the bars for the response categories sum to 100% within each explanatory category group. It can sometimes be useful to make the heights of the bars equal the counts in the category groups instead of percentages.

Minitab Tip**Graphically Describing One or Two Categorical Variables**

- To draw a *bar graph*, use **Graph > Bar Chart**. In the resulting display, select **Simple** to graph one variable or select **Cluster** to graph the relationship between two variables. Then, in the “Categorical variables” box, specify the column(s) containing the raw data for the variable(s). To graph percentages rather than counts, use the Bar Chart Options button.
- To draw a *pie chart*, use **Graph > Pie Chart**. Use the **Multiple Graphs** button to create separate pie charts for subgroups within the dataset.



2.3 Exercises are on pages 58–59.

2.4 Exploring Features of Quantitative Data with Pictures

Looking at a long, disorganized list of data values is about as informative as looking at a scrambled set of letters. To begin finding the information in quantitative data, we have to organize it using visual displays and numerical summaries. In this section, we focus on interpreting the main features of quantitative variables. More specific details will be given in the following sections.

Table 2.4 displays the raw data for the right handspan measurements (in centimeters) made in the student survey described in Section 2.1. The measurements are listed separately for male and female students but are not organized in any other way. Imagine that you know a female student whose stretched right handspan is 20.5 cm. Can you see how she compares to the female students in Table 2.4? That will probably be hard because the list of data values is disorganized.

In Case Study 1.1, we graphed the “fastest speed ever driven” responses with a simple **dotplot**. We also summarized the data using a **five-number summary**, which consists of the median, the quartiles (roughly, the medians of the lower and upper halves of the data), and the extremes (low, high). Let’s use those methods to organize the handspan data in Table 2.4.

TABLE 2.4 Stretched Right Handspans (centimeters) of 190 College Students**Males (87 students):**

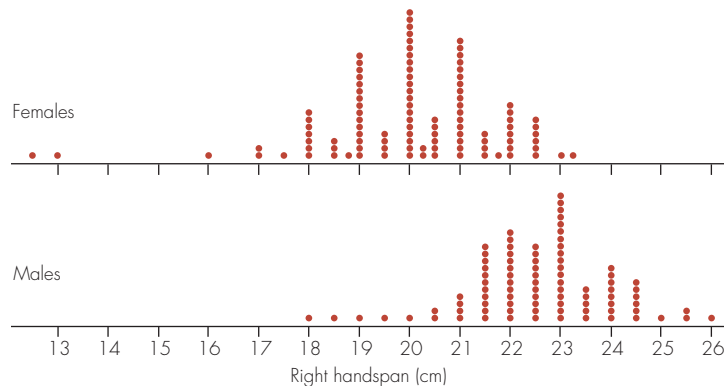
21.5, 22.5, 23.5, 23, 24.5, 23, 26, 23, 21.5, 21.5, 24.5, 23.5, 22, 23.5, 22, 22, 24.5, 23, 22.5, 19.5, 22.5, 22, 23, 22.5, 20.5, 21.5, 23, 22.5, 21.5, 25, 24, 21.5, 21.5, 18, 20, 22, 24, 22, 23, 22, 23, 22.5, 25.5, 24, 23.5, 21, 25.5, 23, 22.5, 24, 21.5, 22, 22.5, 23, 18.5, 21, 24, 23.5, 24.5, 23, 22, 23, 23, 24, 24.5, 20.5, 24, 22, 23, 21, 22.5, 21.5, 24.5, 22, 22, 21, 23, 22.5, 24, 22.5, 23, 23, 23, 21.5, 19, 21.5

Females (103 students):

20, 19, 20.5, 20.5, 20.25, 20, 18, 20.5, 22, 20, 21.5, 17, 16, 22, 22, 20, 20, 20, 20, 21.7, 22, 20, 21, 21, 19, 21, 20.25, 21, 22, 18, 20, 21, 19, 22.5, 21, 20, 19, 21, 20.5, 21, 22, 20, 20, 18, 21, 22.5, 22.5, 19, 19, 19, 22.5, 20, 13, 20, 22.5, 19.5, 18.5, 19, 17.5, 18, 21, 19.5, 20, 19, 21.5, 18, 19, 19.5, 20, 22.5, 21, 18, 22, 18.5, 19, 22, 17, 12.5, 18, 20.5, 19, 20, 21, 19, 19, 21, 18.5, 19, 21.5, 21.5, 23, 23.25, 20, 18.8, 21, 21, 20, 20.5, 20, 19.5, 21, 21, 20

Example 2.5

Right Handspans In Figure 2.5, each dot represents the handspan of an individual student, with the value of the measurement shown along the horizontal axis. From this dotplot, we learn that a majority of the female students had handspans between 19 and 21 cm and a good number of the male students had handspans between 21.5 and 23 cm. We also see that there were two female students with unusually small handspans compared to those of the other female students.

**FIGURE 2.5** Stretched right handspans (in centimeters) of college students

Here are five-number summaries for the handspan measurements given in Table 2.4 and graphed in Figure 2.5:

	Males (87 Students)		Females (103 Students)	
Median	22.5		20.0	
Quartiles	21.5	23.5	19.0	21.0
Extremes	18.0	26.0	12.5	23.25

Remember that the five-number summary approximately divides the dataset into quarters. For example, about 25% of the female handspan measurements are between 12.5 and 19.0 cm, about 25% are between 19.0 and 20.0 cm, about 25% are between 20.0 and 21.0 cm, and about 25% are between 21.0 and 23.25 cm. The five-number summary, along with the dotplot, gives us a good idea of where our imagined female student with the 20.5-cm handspan fits into the distribution of handspans for female students. She is in the third quarter of the data, slightly above the median (the middle value).

Summary Features of Quantitative Variables

The **distribution** of a quantitative variable is the overall pattern of how often the possible values occur. For most quantitative variables, three summary characteristics of the overall distribution of the data tend to be of the most interest. These are the **location** (center, average), the **spread** (variability), and the **shape** of the data. We also will be interested in whether or not there are any **outliers**—individual values that are unusual compared to the bulk of the other values—in the data.

Location (Center, Average)

The first concept for summarizing a quantitative variable is the idea of the “center” of the distribution of values, also called the *location* of the data. The **median**, approximately the middle value in the data, is one estimate of location. The **mean**, which is the usual arithmetic average, is another. Details about how to compute these are given in Section 2.5.

Spread (Variability)

The **variability** among the individual measurements is an important feature of any dataset. How spread out are the values? Are all values about the same? Are most of them together but with a few that are unusually high or low?

In a five-number summary, we can assess the amount of spread (variability) in the data by looking at the difference between the two extremes (called the *range*) and the difference between the two quartiles (called the *interquartile range*). Later in this chapter, you will learn about the *standard deviation*, another important measure of variability.

An assessment of variability is particularly important in interpreting data. For instance, to know whether or not the amount of rainfall during a year at a location is unusual, we have to know about the natural variation in annual rainfall amounts. To determine whether a 1-year-old child might be growing abnormally, we need to know about the natural variation in the heights of 1-year-old children.

Shape

A third feature to consider is the shape of how the values of a quantitative variable are distributed. Using appropriate visual displays, we can address questions about *shape* such as the following: Are most of the values clumped in the middle, with values tailing off at each end (like the handspan measurements shown in Figure 2.5)? Are most of the values clumped together on one end (either high or low), with the remaining few values stretching relatively far toward the other end? We will discuss *shape* more completely later in this section, on pages 34–35.


Outliers

We will also want to consider whether or not any individual values are outliers. There is no precise definition for an outlier, but in general, an **outlier** is a data point that is not consistent with the bulk of the data. For a single variable, an outlier is a value that is unusually high or low. When two variables are considered, an outlier is an unusual combination of values. For instance, in Example 2.5 about handspans, a female student with a handspan of 24.5 cm would be an outlier because this handspan is well past the largest of the measurements made by the 103 female students. A male student with a handspan of 24.5 cm, however, is not an outlier because this measurement is consistent with the data for male students.

The extreme values, low and high, in a dataset do not automatically qualify as outliers. To qualify as an outlier a data value must be unusually low or high compared to the rest of the data. We will describe a method for identifying outliers in Section 2.5.

Example 2.6

Annual Compensation for Highest Paid CEOs in the United States Figure 2.6 is a dotplot of the paid compensation (in millions of \$) for the 50 highest-paid CEOs in 2008 for companies on *Fortune* magazine’s list of Top 500 companies in the United States. Somewhat vague indications of *location* and *spread* are shown on the figure.

 The data are given in the **ceodata08** dataset on the companion website, <http://www.cengage.com/statistics/Utts6e>.

The median compensation for these 50 CEOs was about \$35.6 million, and that's approximately where "location" is indicated on Figure 2.6. Overall, the data spread from \$24.3 million to \$557 million, although the value at \$557 million looks to be an outlier, a data value inconsistent with the bulk of the data. By the way, this astounding amount was paid to Lawrence J. Ellison, CEO of Oracle. The *shape* of the dataset is that most values are clumped on the lower end of the scale with the remaining values stretching relatively far toward the high end (called a skewed shape).

Source: http://www.forbes.com/lists/2009/12/best-boss-09_CEO-Compensation_CompTotDisp.html

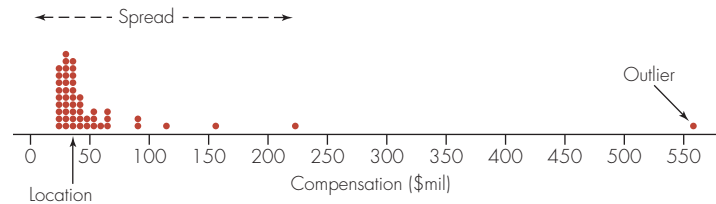


FIGURE 2.6 Dotplot of CEO compensation in 2008

The next example illustrates that sometimes the extreme points are the most interesting features of a dataset, even if they might not be outliers.

Example 2.7

Ages of Death of U.S. First Ladies Much has been written about ages of U.S. presidents when elected and at death, but what about their spouses? Do they tend to live short lives or long lives? Table 2.5 lists the approximate ages at death for first ladies of the United States, or their year of birth if they were not yet deceased as of early 2020. It is not

TABLE 2.5 The First Ladies of the United States of America

Name	Born–Died	Age at Death
Martha Dandridge Custis Washington	1731–1802	71
Abigail Smith Adams	1744–1818	74
Martha Wayles Skelton Jefferson	1748–1782	34
Dolley Payne Todd Madison	1768–1849	81
Elizabeth Kortright Monroe	1768–1830	62
Louisa Catherine Johnson Adams	1775–1852	77
Rachel Donelson Jackson	1767–1828	61
Hannah Hoes Van Buren	1783–1819	36
Anna Tuthill Symmes Harrison	1775–1864	89
Letitia Christian Tyler	1790–1842	52
Julia Gardiner Tyler	1820–1889	69
Sarah Childress Polk	1803–1891	88
Margaret Mackall Smith Taylor	1788–1852	64
Abigail Powers Fillmore	1798–1853	55
Jane Means Appleton Pierce	1806–1863	57
Harriet Lane	1830–1903	73
Mary Todd Lincoln	1818–1882	64
Eliza McCordle Johnson	1810–1876	66
Julia Dent Grant	1826–1902	76
Lucy Ware Webb Hayes	1831–1889	58
Lucretia Rudolph Garfield	1832–1918	86

Continues

TABLE 2.5 *Continued*

Name	Born–Died	Age at Death
Ellen Lewis Herndon Arthur	1837–1880	43
Frances Folsom Cleveland	1864–1947	83
Caroline Lavinia Scott Harrison	1832–1892	60
Ida Saxton McKinley	1847–1907	60
Edith Kermit Carow Roosevelt	1861–1948	87
Helen Herron Taft	1861–1943	82
Ellen Louise Axson Wilson	1860–1914	54
Edith Bolling Galt Wilson	1872–1961	89
Florence Kling Harding	1860–1924	64
Grace Anna Goodhue Coolidge	1879–1957	78
Lou Henry Hoover	1874–1944	70
Anna Eleanor Roosevelt Roosevelt	1884–1962	78
Elizabeth Virginia Wallace Truman	1885–1982	97
Mamie Geneva Doud Eisenhower	1896–1979	83
Jacqueline Lee Bouvier Kennedy Onassis	1929–1994	65
Claudia Taylor Johnson	1912–2007	95
Patricia Ryan Nixon	1912–1993	81
Elizabeth Bloomer Ford	1918–2011	93
Rosalynn Smith Carter	1927–	
Nancy Davis Reagan	1921–2016	94
Barbara Pierce Bush	1925–2018	92
Hillary Rodham Clinton	1947–	
Laura Welch Bush	1946–	
Michelle Robinson Obama	1964–	
Melania Knavs Trump	1970–	

Source: http://en.wikipedia.org/wiki/List_of_First_Ladies_of_the_United_States.

completely accurate to label all of these women “first ladies” if the strict definition is “the wife of a president while in office.” For example, Harriet Lane served socially as “first lady” to President James Buchanan, but he was unmarried and she was his niece. A few of the women listed died before their husband’s term in office. Nonetheless, we will use the data as provided by the White House and summarize the ages at death for these women. Following is a five-number summary for these ages:

First Ladies’ Ages at Death		
Median	73	
Quartiles	60.5	83
Extremes	34	97

If you are at all interested in history, this summary will make you curious about the extreme points. Who died at 34? Who lived to be 97? The extremes are more interesting features of this dataset than is the summary of ages in the middle, which tend to match what we would expect for ages at death. From Table 2.5, you can see that Thomas Jefferson’s wife, Martha, died in 1782 at age 34, almost 20 years before he entered office. He reportedly was devastated, and he never remarried, although historians believe that he may have had other children in his relationship with Sally Hemings. At the other

extreme, Bess Truman died in 1982 at age 97; her husband, Harry, preceded her in death by 10 years, but he too lived a long life—he died at age 88.

Should we attach the label “outlier” to either of the most extreme points in the list of ages at death for the first ladies? To study this issue, we have to examine all of the data to see whether or not the two extremes clearly stand apart from the other values. If you look over Table 2.5, you may be able to form an opinion about whether Martha Jefferson and Bess Truman should be called outliers by comparing them to the other first ladies. Making sense of a list of numbers, however, is difficult. The most effective way to look for outliers is to graph the data, which we will learn more about in the remainder of this section.

Pictures of Quantitative Data

Three similar types of pictures are used to represent quantitative variables, all of which are valuable for assessing location, spread, shape, and outliers. **Histograms** are similar to bar graphs and can be used for any number of data values, although they are not particularly informative when the sample size is small. **Stem-and-leaf plots** and **dotplots** present all individual values, so for very large datasets, they are more cumbersome than histograms.

Figures 2.7 to 2.9 illustrate a histogram, a stem-and-leaf plot, and a dotplot, respectively, for the female students’ right handspans displayed in Table 2.4. Figure 2.9 is merely a portion of the dotplot shown previously in Figure 2.5 on page 27. Examine the three figures. Note that if the stem-and-leaf plot were turned on its side, all three pictures would look similar. Each picture shows the *distribution* of the data—the pattern of how often the various measurements occurred.

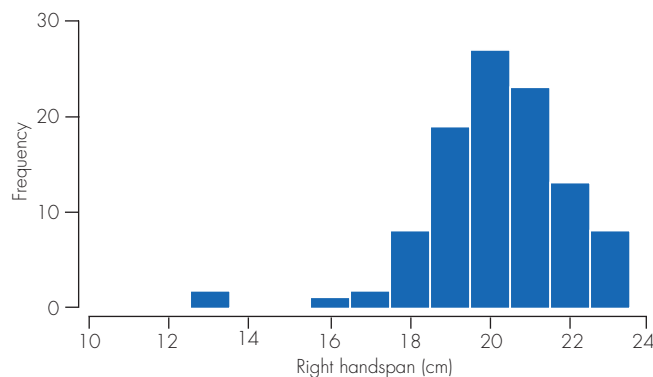


FIGURE 2.7 Histogram of females’ right handspans

```

12 | 5
13 | 0
14 |
15 |
16 | 0
17 | 005
18 | 00000005558
19 | 00000000000000005555
20 | 000000000000000000022555555
21 | 000000000000000000055557
22 | 00000000555555
23 | 02

```

Example: $|12|5 = 12.5$

FIGURE 2.8 Stem-and-leaf plot of females’ right handspans

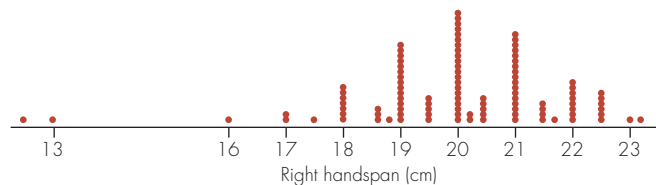


FIGURE 2.9 Dotplot of females’ right handspans

A fourth kind of picture, called a **boxplot** or **box-and-whisker plot**, displays the information given in a five-number summary. It is especially useful for comparing two or more groups and for identifying outliers. We will examine boxplots at the end of this section.

Interpreting Histograms, Stem-and-Leaf Plots, and Dotplots

Each of these pictures is useful for assessing the location, spread, and shape of a distribution, and each is also useful for detecting outliers. For the data presented in Figures 2.7 to 2.9, note that the values are *centered* at about 20 cm, which we learned in Example 2.5 is indeed the median value. There are two possible *outlier* values that are low in comparison to the bulk of the data. These are identifiable in the stem-and-leaf plot as 12.5 and 13.0 cm, but are evident in the other two pictures as well. Except for those values, the handspans have a *range* of about 7 cm, extending from about 16 to 23 cm. They tend to be clumped around 20 and taper off toward 16 and 23.

There are many computer programs that can be used to create these pictures. Figures 2.7, 2.8, and 2.9, for instance, are slight modifications of pictures created using Minitab. We will go through the steps for creating each type of picture by hand, but keep in mind that statistical software such as Minitab automates most of the process.

Creating a Histogram

A histogram is a bar chart of a quantitative variable that shows how many values are in various intervals of the data. The steps in creating a histogram are as follows:

- Step 1:** Decide how many *equally spaced* intervals to use for the horizontal axis. The experience of many researchers is that somewhere between about 6 and 15 intervals is a good number for displaying the shape and spread of the dataset, although occasionally more may be needed to accommodate outliers. Use intervals that make the range of each interval convenient.
- Step 2:** Decide whether to use *frequencies* or *relative frequencies* on the vertical axis. A frequency is the actual number of observations in an interval. A relative frequency is either the proportion or the percent in an interval.
- Step 3:** Draw the appropriate number of equally spaced intervals on the horizontal axis; be sure to cover the entire data range. Determine the frequency or relative frequency of data values in each interval and, for each interval, draw a bar with the corresponding height. If a value is on a boundary, count it in the interval that begins with that value.

Example 2.8

Revisiting Example 2.7: Histograms for Ages of Death of U.S. First Ladies

Figures 2.10 and 2.11 show two different histograms for the ages of death for the first ladies of the United States. The raw data were given in Table 2.5 on pages 29–30. In each histogram, the horizontal axis gives age at death and the vertical axis gives the frequency of how many first ladies died within the age interval represented by any particular bar.

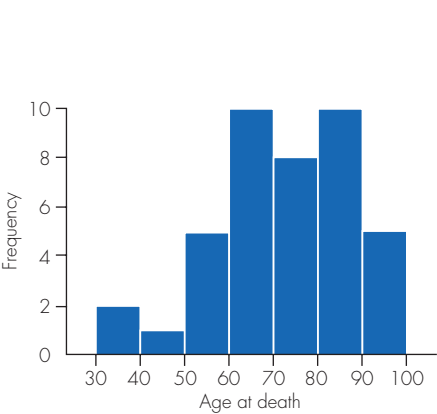


FIGURE 2.10 Histogram of ages of death of U.S. first ladies using seven 10-year intervals

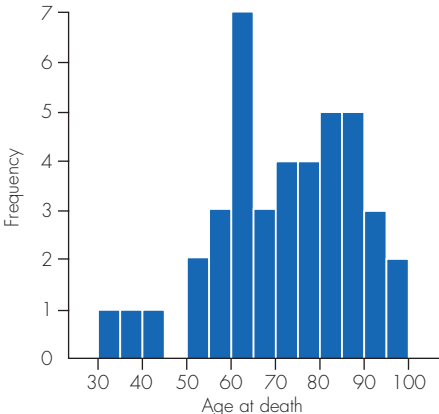


FIGURE 2.11 Histogram of ages of death of U.S. first ladies using fourteen 5-year intervals