UNDERSTANDING BASIC STATISTICS

FREQUENTLY USED FORMULAS

n = sample size N = population size f = frequency

Chapter 2

Class width $= \dfrac{\text{High} - \text{Low}}{\text{Number of classes}}$ (increase to next integer)

Class midpoint $= \dfrac{\text{Upper limit} + \text{Lower limit}}{2}$

Lower boundary = Lower boundary of previous class + Class width

Chapter 3

Sample mean $\bar{x} = \dfrac{\Sigma x}{n}$

Population mean $\mu = \dfrac{\Sigma x}{N}$

Weighted average $= \dfrac{\Sigma xw}{\Sigma w}$

Range = Largest data value − Smallest data value

Sample standard deviation $s = \sqrt{\dfrac{\Sigma(x - \bar{x})^2}{n - 1}}$

Computation formula $s = \sqrt{\dfrac{\Sigma x^2 - (\Sigma)^2/n}{n - 1}}$

Population standard deviation $\sigma = \sqrt{\dfrac{\Sigma(x - \mu)^2}{N}}$

Sample variance s^2

Population variance σ^2

Sample coefficient of variation $CV = \dfrac{s}{\bar{x}} \cdot 100\%$

Sample mean for grouped data $\bar{x} = \dfrac{\Sigma xf}{n}$

Sample standard deviation for grouped data

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2 f}{n - 1}} = \sqrt{\frac{\Sigma x^2 f - (\Sigma xf)^2/n}{n - 1}}$$

Chapter 4

Regression and Correlation

Pearson product moment correlation coefficient

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

Least-squares line $\hat{y} = a + bx$

where $b = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$

$a = \bar{y} - b\bar{x}$

$b = r(s_y/s_x)$

Coefficient of determination $= r^2$

Chapter 5

Probability of the complement of event A

$P(A^c) = 1 - P(A)$

Multiplication rule for independent events

$P(A \text{ and } B) = P(A) \cdot P(B)$

General multiplication rules

$P(A \text{ and } B) = P(A) \cdot P(B|A)$

$P(A \text{ and } B) = P(B) \cdot P(A|B)$

Addition rule for mutually exclusive events

$P(A \text{ or } B) = P(A) + P(B)$

General addition rule

$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Permutation rule $P_{n,r} = \dfrac{n!}{(n - r)!}$

Combination rule $C_{n,r} = \dfrac{n!}{r!(n - r)!}$

Chapter 6

Mean of a discrete probability distribution $\mu = \Sigma xP(x)$

Standard deviation of a discrete probability distribution

$\sigma = \sqrt{\Sigma(x - \mu)^2 P(x)}$

For Binomial Distributions

r = number of successes; p = probability of success; $q = 1 - p$

Binomial probability distribution $P(r) = C_{n,r}p^r q^{n-r}$

Mean $\mu = np$

Standard deviation $\sigma = \sqrt{npq}$

Chapter 7

Raw score $x = z\sigma + \mu$ Standard score $z = \dfrac{x - \mu}{\sigma}$

Mean of $\bar{x}$ distribution $\mu_{\bar{x}} = \mu$

Standard deviation of $\bar{x}$ distribution $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Standard score for $\bar{x}$; $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

Mean of $\hat{p}$ distribution $\mu_{\hat{p}} = p$

Standard deviation for $\hat{p}$; $\sigma_{\hat{p}} = \sqrt{\dfrac{pq}{n}}$; $q = 1 - p$

Chapter 8

Confidence Interval

for μ

$\bar{x} - E < \mu < \bar{x} + E$

where $E = z_c \dfrac{\sigma}{\sqrt{n}}$ when σ is known

$$E = t_c \frac{s}{\sqrt{n}} \text{ when } \sigma \text{ is unknown}$$

with $d.f. = n - 1$

for p given $np > 5$ and $n(1 - p) > 5$

$$\hat{p} - E < p < \hat{p} + E$$

$$\text{where } E = z_c \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

$$\hat{p} = \frac{r}{n}$$

Sample Size for Estimating

$$\text{means } n = \left(\frac{z_c \sigma}{E}\right)^2$$

proportions

$$n = p(1 - p)\left(\frac{z_c}{E}\right)^2 \text{ with preliminary estimate for } p$$

$$n = \frac{1}{4}\left(\frac{z_c}{E}\right)^2 \text{ without preliminary estimate for } p$$

Chapter 9

Sample Test Statistics for Tests of Hypotheses

for μ (σ known) $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

for μ (σ unknown) $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$; $d.f. = n - 1$

for p ($np > 5$ and $nq > 5$) $z = \dfrac{\hat{p} - p}{\sqrt{pq/n}}$

where $q = 1 - p$; $\hat{p} = r/n$

Chapter 10

Sample Test Statistics for Tests of Hypotheses

for paired differences d $t = \dfrac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$; $d.f. = n - 1$

for difference of means, σ_1 and σ_2 known

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

for difference of means, σ_1 or σ_2 unknown

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

$d.f.$ = smaller of $n_1 - 1$ and $n_2 - 1$

(*Note:* Software uses Satterthwaite's approximation for degrees of freedom $d.f.$)

for difference of proportions

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\bar{p}\,\bar{q}}{n_1} + \dfrac{\bar{p}\,\bar{q}}{n_2}}}$$

$$\text{where } \bar{p} = \frac{r_1 + r_2}{n_1 + n_2} \text{ and } \bar{q} = 1 - \bar{p}$$

$$\hat{p}_1 = r_1/n_1;\ \hat{p}_2 = r_2/n_2$$

Confidence Interval

for $\mu_1 - \mu_2$ (independent samples)

$$(\bar{x}_1 - \bar{x}_2) - E < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + E$$

$$\text{where } E = z_c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \text{ when } \sigma_1 \text{ and } \sigma_2 \text{ are known}$$

$$E = t_c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \text{ when } \sigma_1 \text{ or } \sigma_2 \text{ is unknown}$$

with $d.f.$ = smaller of $n_1 - 1$ and $n_2 - 1$

(*Note:* Software uses Satterthwaite's approximation for degrees of freedom $d.f.$)

for difference of proportions $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) - E < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + E$$

$$\text{where } E = z_c \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\hat{p}_1 = r_1/n_1;\ \hat{p}_2 = r_2/n_2$$

$$\hat{q}_1 = 1 - \hat{p}_1;\ \hat{q}_2 = 1 - \hat{p}_2$$

Chapter 11

$$\chi^2 = \sum \frac{(O - E)^2}{E} \text{ where } O = \text{observed frequency}$$

For tests of independence and tests of homogeneity

$$E = \frac{\text{(Row total)(Column total)}}{\text{Sample size}}$$

For goodness of fit test E = (given percent)(sample size)

Tests of independence $d.f. = (R - 1)(C - 1)$

Test of homogeneity $d.f. = (R - 1)(C - 1)$

Goodness of fit $d.f.$ = (number of categories) − 1

Sample test statistic for σ^2

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2} \text{ with } d.f. = n - 1$$

Linear Regression

Sample test statistic for r

$$t = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}} \text{ with } d.f. = n - 2$$

Standard error of estimate $S_e = \sqrt{\dfrac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n - 2}}$

Confidence interval for y: $\hat{y} - E < y < \hat{y} + E$

$$\text{where } E = t_c S_e \sqrt{1 + \frac{1}{n} + \frac{n(x - \bar{x})^2}{n\Sigma x^2 - (\Sigma x)^2}}$$

with $d.f. = n - 2$

Sample test statistic for slope b

$$t = \frac{b}{S_e}\sqrt{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2} \text{ with } d.f. = n - 2$$

Understanding Basic Statistics

EIGHTH EDITION

Understanding Basic Statistics

Charles Henry Brase

REGIS UNIVERSITY

Corrinne Pellillo Brase

ARAPAHOE COMMUNITY COLLEGE

CENGAGE

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

This book is dedicated to the memory of
a great teacher, mathematician, and friend

Burton W. Jones
Professor Emeritus, University of Colorado

CENGAGE

Understanding Basic Statistics,
Eighth Edition

Brase/Brase

Product Director: Mark Santee

Product Manager: Cassie Van Der Laan

Associate Content Developer: Gabriela Carrascal

Content Developer: Spencer Arritt

Marketing Manager: Mike Saver

Senior Content Project Manager: Andrea Wagner

Manufacturing Planner: Doug Bertke

IP Analyst: Reba Frederics

IP Project Manager: Sarah Shainwald

Production Service: MPS Limited

Compositor: MPS Limited

Art Director: Vernon Boes

Text Designer: Diane Beasley

Cover Designer: Cheryl Carrington

Cover Image: blickwinkel/Alamy Stock Photo

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

For permission to use material from this text or product,
submit all requests online at **www.cengage.com/permissions.**
Further permissions questions can be emailed to
permissionrequest@cengage.com.

Library of Congress Control Number: 2017954452

Student Edition:
ISBN: 978-133-755807-5

Loose-leaf Edition:
ISBN: 978-133-755819-8

Cengage
20 Channel Center Street
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**.

Cengage products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage platforms and services, visit **www.cengage.com**.

To register or access your online learning solution or purchase materials for your course, visit **www.cengagebrain.com**.

Printed in the United States of America
Print Number: 01 Print Year: 2017

Contents

11 Additional Topics Using Inference 536

Appendix: Tables A1

Critical Thinking

Students need to develop critical thinking skills in order to understand and evaluate the limitations of statistical methods. *Understanding Basic Statistics* makes students aware of method appropriateness, assumptions, biases, and justifiable conclusions.

CRITICAL THINKING

UNUSUAL VALUES

Chebyshev's theorem tells us that no matter what the data distribution looks like, at least 75% of the data will fall within 2 standard deviations of the mean. As we will see in Chapter 7, when the distribution is mound-shaped and symmetric, about 95% of the data are within 2 standard deviations of the mean. Data values beyond 2 standard deviations from the mean are less common than those closer to the mean.

In fact, one indicator that a data value might be an outlier is that it is more than 2.5 standard deviations from the mean (Source: *Statistics,* by G. Upton and I. Cook, Oxford University Press).

UNUSUAL VALUES

For a binomial distribution, it is unusual for the number of successes r to be higher than $\mu + 2.5\sigma$ or lower than $\mu - 2.5\sigma$.

We can use this indicator to determine whether a specified number of successes out of n trials in a binomial experiment are unusual.

For instance, consider a binomial experiment with 20 trials for which probability of success on a single trial is $p = 0.70$. The expected number of successes is $\mu = 14$, with a standard deviation of $\sigma \approx 2$. A number of successes above 19 or below 9 would be considered unusual. However, such numbers of successes are possible.

◀ Critical Thinking

Critical thinking is an important skill for students to develop in order to avoid reaching misleading conclusions. The Critical Thinking feature provides additional clarification on specific concepts as a safeguard against incorrect evaluation of information.

Interpretation ▶

Increasingly, calculators and computers are used to generate the numeric results of a statistical process. However, the student still needs to correctly interpret those results in the context of a particular application. The Interpretation feature calls attention to this important step. Interpretation is stressed in examples, in guided exercises, and in the problem sets.

SOLUTION: Since we want to know the number of standard deviations from the mean, we want to convert 6.9 to standard z units.

$$z = \frac{x - \mu}{\sigma} = \frac{6.9 - 8}{0.5} = -2.20$$

Interpretation The amount of cheese on the selected pizza is only 2.20 standard deviations below the mean. The fact that z is negative indicates that the amount of cheese is 2.20 standard deviations *below* the mean. The parlor will not lose its franchise based on this sample.

6. **Interpretation** A campus performance series features plays, music groups, dance troops, and stand-up comedy. The committee responsible for selecting the performance groups include three students chosen at random from a pool of volunteers. This year the 30 volunteers came from a variety of majors. However, the three students for the committee were all music majors. Does this fact indicate there was bias in the selection process and that the selection process was not random? Explain.

7. **Critical Thinking** Greg took a random sample of size 100 from the population of current season ticket holders to State College men's basketball games. Then he took a random sample of size 100 from the population of current season ticket holders to State College women's basketball games.
   (a) What sampling technique (stratified, systematic, cluster, multistage, convenience, random) did Greg use to sample from the population of current season ticket holders to all State College basketball games played by either men or women?
   (b) Is it appropriate to pool the samples and claim to have a random sample of size 200 from the population of current season ticket holders to all State College home basketball games played by either men or women? Explain.

◀ Critical Thinking and Interpretation Exercises

In every section and chapter problem set, Critical Thinking problems provide students with the opportunity to test their understanding of the application of statistical methods and their interpretation of their results. Interpretation problems ask students to apply statistical results to the particular application.

Statistical Literacy

No language, including statistics, can be spoken without learning the vocabulary. *Understanding Basic Statistics* introduces statistical terms with deliberate care.

What Does the Level of Measurement Tell Us?

The level of measurement tells us which arithmetic processes are appropriate for the data. This is important because different statistical processes require various kinds of arithmetic. In some instances all we need to do is count the number of data that meet specified criteria. In such cases nominal (and higher) data levels are all appropriate. In other cases we need to order the data, so nominal data would not be suitable. Many other statistical processes require division, so data need to be at the ratio level. Just keep the nature of the data in mind before beginning statistical computations.

◀ What Does (concept, method, statistical result) Tell Us?

This feature gives a brief summary of the information we obtain from the named concept, method, or statistical result.

Important Features of a (concept, method, or result) ▶

In statistics we use many different types of graphs, samples, data, and analytical methods. The features of each such tool help us select the most appropriate ones to use and help us interpret the information we receive from applications of the tools.

Important Features of a Simple Random Sample

For a simple random sample

- Every sample of specified size n from the population has an equal chance of being selected.
- No researcher bias occurs in the items selected for the sample.
- A random sample may not always reflect the diversity of the population. For instance, from a population of 10 cats and 10 dogs, a random sample of size 6 could consist of all cats.

SECTION 7.1 PROBLEMS

1. **Statistical Literacy** Which, if any, of the curves in Figure 7-7 look(s) like a normal curve? If a curve is not a normal curve, tell why.
2. **Statistical Literacy** Look at the normal curve in Figure 7-8, and find μ, $\mu + \sigma$, and σ.

FIGURE 7-7

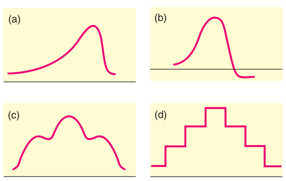
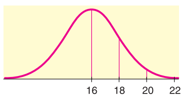


FIGURE 7-8



8. **Terminology** Consider the following statement: If you answered 90% of the questions on a test correctly, then your score is in the 90th percentile. Is the statement true or false? Explain.
9. **Statistical Literacy**
   (a) What measures of variation indicate spread about the mean?
   (b) Which graphic display shows the median and the data spread about the median?

◀ Statistical Literacy and NEW Terminology Problems

In every section and chapter problem set, Statistical Literacy problems test student understanding of terminology, statistical methods, and the appropriate conditions for use of the different processes. NEW terminology problems in each chapter problem set provide multiple choice or true/false questions related to terminology.

Definition Boxes ▶

Whenever important terms are introduced in text, blue definition boxes appear within the discussions. These boxes make it easy to reference or review terms as they are used further.

Box-and-Whisker Plots

The quartiles together with the low and high data values give us a very useful *five-number summary* of the data and their spread.

FIVE-NUMBER SUMMARY

Lowest value, Q_1, median, Q_3, highest value

We will use these five numbers to create a graphic sketch of the data called a *box-and-whisker plot*. Box-and-whisker plots provide another useful technique from exploratory data analysis (EDA) for describing data.

IMPORTANT WORDS & SYMBOLS

SECTION 5.1

Probability of an event A, $P(A)$ 194
Intuition 194
Relative frequency 194
Equally likely outcomes 194
Law of large numbers 196
Statistical experiment 196
Event 196
Simple event 196
Sample space 194
Complement of event A^c 198

SECTION 5.2

Independent events 206
Dependent events 206
Probability of event *A and B* 206
Event $A \mid B$ 206
Conditional probability 206
$P(A \mid B)$ 206
Multiplication rules of probability (for independent and dependent events) 206
More than two independent events 211
Probability of event *A or B* 211
Event *A and B* 211
Event *A or B* 211
Mutually exclusive events 213
Addition rules (for mutually exclusive and general events) 213
More than two mutually exclusive events 215
Contingency table 215
Basic probability rules 218

SECTION 5.3

Multiplication rule of counting 225
Tree diagram 225
Factorial notation 229
Permutations rule 229
Combinations rule 231

◀ Important Words & Symbols

The Important Words & Symbols within the Chapter Review feature at the end of each chapter summarizes the terms introduced in the Definition Boxes for student review at a glance. Page numbers for first occurrence of term are given for easy reference.

e following topics in class or review the topics on your own. Then omplete essay in which you summarize the main points. Please nd graphs as appropriate.
mean to say that we are going to use a sample to draw an inference lation? Why is a random sample so important for this process? If random sample of students in the cafeteria, why couldn't we just tudents who order Diet Pepsi with their lunch? Comment on the A random sample is like a miniature population, whereas samples andom are likely to be biased." Why would the students who order ith lunch not be a random sample of students in the cafeteria?
words, explain the differences among the following sampling techniques: simple random sample, stratified sample, systematic sample, cluster sample, multistage sample, and convenience sample. Describe situations in which each type might be useful.

5. **Basic Computation: Central Limit Theorem** Suppose x has a distribution with a mean of 8 and a standard deviation of 16. Random samples of size $n = 64$ are drawn.
(a) Describe the $\bar{x}$ distribution and compute the mean and standard deviation of the distribution.
(b) Find the z value corresponding to $\bar{x} = 9$.
(c) Find $P(\bar{x} > 9)$.
(d) **Interpretation** Would it be unusual for a random sample of size 64 from the x distribution to have a sample mean greater than 9? Explain.

◀ Basic Computation Problems

These problems focus student attention on relevant formulas, requirements, and computational procedures. After practicing these skills, students are more confident as they approach real-world applications.

Expand Your Knowledge Problems ▶

Expand Your Knowledge problems present optional enrichment topics that go beyond the material introduced in a section. Vocabulary and concepts needed to solve the problems are included at point-of-use, expanding students' statistical literacy.

30. **Expand Your Knowledge: Geometric Mean** When data consist of percentages, ratios, compounded growth rates, or other rates of change, the *geometric mean* is a useful measure of central tendency. For n data values,

$$\text{Geometric mean} = \sqrt[n]{\text{product of the } n \text{ data values}}, \quad \text{assuming all data values are positive}$$

To find the *average growth factor* over 5 years of an investment in a mutual fund with growth rates of 10% the first year, 12% the second year, 14.8% the third year, 3.8% the fourth year, and 6% the fifth year, take the geometric mean of 1.10, 1.12, 1.148, 1.038, and 1.16. Find the average growth factor of this investment.

Direction and Purpose

Real knowledge is delivered through direction, not just facts. *Understanding Basic Statistics* ensures the student knows what is being covered and why at every step along the way to statistical literacy.

Chapter Preview ▶ Questions

Preview Questions at the beginning of each chapter give the student a taste of what types of questions can be answered with an understanding of the knowledge to come.

Normal Curves and Sampling Distributions

PREVIEW QUESTIONS

PART I

What are some characteristics of a normal distribution? What does the empirical rule tell you about data spread around the mean? How can this information be used in quality control? (SECTION 7.1)

Can you compare apples and oranges, or maybe elephants and butterflies? In most cases, the answer is no—unless you first standardize your measurements. What are a standard normal distribution and a standard *z* score? (SECTION 7.2)

How do you convert any normal distribution to a standard normal distribution? How do you find probabilities of "standardized events"? (SECTION 7.3)

Pressmaster/Shutterstock.com

FOCUS PROBLEM

Benford's Law: The Importance of Being Number 1

Benford's Law states that in a wide variety of circumstances, numbers have "1" as their first nonzero digit disproportionately often. Benford's Law applies to such diverse topics as the drainage areas of rivers; properties of chemicals; populations of towns; figures in newspapers, magazines, and government reports; and the half-lives of radioactive atoms!

Specifically, such diverse measurements begin with "1" about 30% of the time, with "2" about 18% of time, and with "3" about 12.5% of the time. Larger digits occur less often. For example, less than 5% of the numbers in circumstances such as these begin with the digit 9. This is in dramatic contrast to a random sampling situation, in which each of the digits 1 through 9 has an equal chance of appearing.

BRENDAN SMIALOWSKI/AFP/Getty Images

8. **Focus Problem: Benford's Law** Again, suppose you are the auditor for a very large corporation. The revenue file contains millions of numbers in a large computer data bank (see Problem 7). You draw a random sample of $n = 228$ numbers from this file and $r = 92$ have a first nonzero digit of 1. Let p represent the population proportion of all numbers in the computer file that have a leading digit of 1.
   i. Test the claim that p is more than 0.301. Use $\alpha = 0.01$.
   ii. If p is in fact larger than 0.301, it would seem there are too many numbers in the file with leading 1s. Could this indicate that the books have been "cooked" by artificially lowering numbers in the file? Comment from the point of view of the Internal Revenue Service. Comment from the perspective of the Federal Bureau of Investigation as it looks for "profit skimming" by unscrupulous employees.
   iii. Comment on the following statement: "If we reject the null hypothesis at level of significance α, we have not *proved* H_0 to be false. We can say that the probability is α that we made a mistake in rejecting H_0." Based on the outcome of the test, would you recommend further investigation before accusing the company of fraud?

▲ Chapter Focus Problems

The Preview Questions in each chapter are followed by a Focus Problem, which serves as a more specific example of what questions the student will soon be able to answer. The Focus Problems are set within appropriate applications and are incorporated into the end-of-section exercises, giving students the opportunity to test their understanding.

Focus Points ▶

Each section opens with bulleted Focus Points describing the primary learning objectives of the section.

SECTION 3.1

Measures of Central Tendency: Mode, Median, and Mean

FOCUS POINTS

- **Compute mean, median, and mode from raw data.**
- **Interpret what mean, median, and mode tell you.**
- **Explain how mean, median, and mode can be affected by extreme data values.**
- **What is a trimmed mean? How do you compute it?**
- **Compute a weighted average.**

This section can be covered quickly. Good discussion topics include *The Story of Old Faithful* in Data Highlights, Problem 1; Linking Concepts, Problem 1; and the trade winds of Hawaii (Using Technology).

The average price of an ounce of gold is $1350. The Zippy car averages 39 miles per gallon on the highway. A survey showed the average shoe size for women is size 9.

Average

In each of the preceding statements, *one* number is used to describe the entire sample or population. Such a number is called an *average*. There are many ways to compute averages, but we will study only three of the major ones.

The easiest average to compute is the *mode*.

Mode

The **mode** of a data set is the value that occurs most frequently. *Note:* If a data set has no single value that occurs more frequently than any other, then that data set has no mode.

EXAMPLE 1 *Mode*

Count the letters in each word of this sentence and give the mode. The numbers of letters in the words of the sentence are

5 3 7 2 4 4 2 4 8 3 4 3 4

LOOKING FORWARD

In later chapters we will use information based on a sample and sample statistics to estimate population parameters (Chapter 8) or make decisions about the value of population parameters (Chapters 9 and 10).

◀ Looking Forward

This feature shows students where the presented material will be used later. It helps motivate students to pay a little extra attention to key topics.

CHAPTER REVIEW

SUMMARY

In this chapter, you've seen that statistics is the study of how to collect, organize, analyze, and interpret numerical information from populations or samples. This chapter discussed some of the features of data and ways to collect data. In particular, the chapter discussed

- Individuals or subjects of a study and the variables associated with those individuals
- Data classification as qualitative or quantitative, and levels of measurement of data
- Sample and population data. Summary measurements from sample data are called statistics, and those from populations are called parameters.
- Sampling strategies, including simple random, stratified, systematic, multistage, and convenience. Inferential techniques presented in this text are based on simple random samples.
- Methods of obtaining data: Use of a census, simulation, observational studies, experiments, and surveys
- Concerns: Undercoverage of a population, nonresponse, bias in data from surveys and other factors, effects of confounding or lurking variables on other variables, generalization of study results beyond the population of the study, and study sponsorship

▲ Chapter Summaries

The Summary within each Chapter Review feature now also appears in bulleted form, so students can see what they need to know at a glance.

Ethics and Real-World Skills

Statistics is not done in a vacuum. *Understanding Basic Statistics* gives students valuable skills for the real world with ethical considerations, technology instruction, genuine applications, actual data, and group projects.

Discussion of Ethics in Statistics ▶

There are many essential ethical considerations in the practice of sampling and research based on statistics. The American Statistical Association has published ethical guidelines that are discussed in this textbook.

ETHICAL PRACTICE IN STATISTICS REQUIRES

- Transparent and openly stated assumptions regarding data collection and statistical methods used
- Advancement of useful knowledge while avoiding harm to others, especially subjects under study

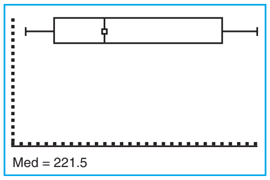
◀ REVISED! Tech Notes

Tech Notes appearing throughout the text give students helpful hints on using TI-84Plus and TI-*n*spire (with TI-84Plus keypad) and TI-83Plus calculators, Microsoft Excel 2013, Minitab, and Minitab Express to solve a problem. They include display screens to help students visualize and better understand the solution.

> Tech Notes *Box-and-Whisker Plot*

Both Minitab and the TI-84Plus/TI-83Plus/TI-*n*spire calculators support box-and-whisker plots. On the TI-84Plus/TI-83Plus/TI-*n*spire, the quartiles Q_1 and Q_3 are calculated as we calculate them in this text. In Minitab and Excel 2013, they are calculated using a slightly different process.

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus Keypad) Press **STATPLOT ➤On.** Highlight box plot. Use **Trace** and the arrow keys to display the values of the five-number summary. The display shows the plot for calories in ice cream bars.



Excel 2013 Does not produce box-and-whisker plot. However, each value of the five-number summary can be found. On the **Home** ribbon, click the **Insert Function** f_x. In the dialogue box, select **Statistical** as the category and scroll to **Quartile**. In the dialogue box, enter the data location and then enter the number of the value you want. For instance, enter 1 in the quartile box for the first quartile.

Minitab Press **Graph ➤ Boxplot.** In the dialogue box, set Data View to IQRange Box.

MinitabExpress Press **Graph ➤ Boxplot ➤ simple.**

REVISED! Using Technology ▶

Further technology instruction is available at the end of each chapter in the Using Technology section. Problems are presented with real-world data from a variety of disciplines that can be solved by using TI-84Plus and TI-*n*spire (with TI-84Plus keypad) and TI-83Plus calculators, Microsoft Excel 2013, Minitab, and Minitab Express.

> USING TECHNOLOGY

Binomial Distributions

Although tables of binomial probabilities can be found in most libraries, such tables are often inadequate. Either the value of p (the probability of success on a trial) you are looking for is not in the table, or the value of n (the number of trials) you are looking for is too large for the table. In Chapter 7, we will study the normal approximation to the binomial. This approximation is a great help in many practical applications. Even so, we sometimes use the formula for the binomial probability distribution on a computer or graphing calculator to compute the probability we want.

Applications

The following percentages were obtained over many years of observation by the U.S. Weather Bureau. All data listed are for the month of December.

Location	Long-Term Mean % of Clear Days in Dec.
Juneau, Alaska	18%
Seattle, Washington	24%
Hilo, Hawaii	36%

You may find that using cumulative probabilities and appropriate subtraction of probabilities, rather than addition of probabilities, will make finding the solutions to Applications 3 to 7 easier.

3. Estimate the probability that Juneau will have at most 7 clear days in December.
4. Estimate the probability that Seattle will have from 5 to 10 (including 5 and 10) clear days in December.
5. Estimate the probability that Hilo will have at least 12 clear days in December.
6. Estimate the probability that Phoenix will have 20 or more clear days in December.
7. Estimate the probability that Las Vegas will have from 20 to 25 (including 20 and 25) clear days in December.

Technology Hints

TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84 Plus keypad), Excel 2013, Minitab/MinitabExpress

The Tech Note in Section 6.2 gives specific instructions for binomial distribution functions on the TI-84Plus/

EXAMPLE 11 Central Limit Theorem

A certain strain of bacteria occurs in all raw milk. Let x be the bacteria count per milliliter of milk. The health department has found that if the milk is not contaminated, then x has a distribution that is more or less mound-shaped and symmetric. The mean of the x distribution is $\mu = 2500$, and the standard deviation is $\sigma = 300$. In a large commercial dairy, the health inspector takes 42 random samples of the milk produced each day. At the end of the day, the bacteria count in each of the 42 samples is averaged to obtain the sample mean bacteria count $\bar{x}$.

(a) Assuming the milk is not contaminated, what is the distribution of $\bar{x}$?

SOLUTION: The sample size is $n = 42$. Since this value exceeds 30, the central limit theorem applies, and we know that $\bar{x}$ will be approximately normal, with mean and standard deviation

◀ UPDATED! Applications

Real-world applications are used from the beginning to introduce each statistical process. Rather than just crunching numbers, students come to appreciate the value of statistics through relevant examples.

Most exercises in each section are applications problems. ▶

8. **Ranching: Cattle** You are the foreman of the Bar-S cattle ranch in Colorado. A neighboring ranch has calves for sale, and you are going to buy some to add to the Bar-S herd. How much should a healthy calf weigh? Let x be the age of the calf (in weeks), and let y be the weight of the calf (in kilograms). The following information is based on data taken from *The Merck Veterinary Manual* (a reference used by many ranchers).

x	1	3	10	16	26	36
y	42	50	75	100	150	200

Complete parts (a) through (e), given $\Sigma x = 92$, $\Sigma y = 617$, $\Sigma x^2 = 2338$, $\Sigma y^2 = 82{,}389$, $\Sigma xy = 13{,}642$, and $r \approx 0.998$.

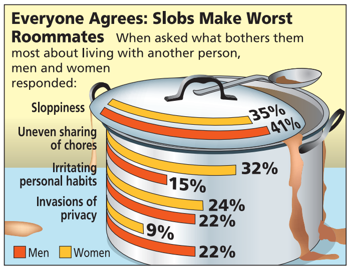
(f) The calves you want to buy are 12 weeks old. What does the least-squares line predict for a healthy weight?

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

1.Examine Figure 2-20, "Everyone Agrees: Slobs Make Worst Roommates." This is a clustered bar graph because two percentages are given for each response category: responses from men and responses from women. Comment about how the artistic rendition has slightly changed the format of a bar graph. Do the bars seem to have lengths that accurately reflect the relative percentages of the responses? In your own opinion, does the artistic rendition enhance or confuse the information? Explain. Which characteristic of "worst roommates" does the graphic seem to illustrate? Can this graph be considered a Pareto chart for men? for women? Why or why not? From the information given in the figure, do you think the survey just listed the four given annoying characteristics? Do you think a respondent could choose more than one characteristic? Explain

FIGURE 2-20



Source: Advantage Business Research for Mattel *Compatibility*

◀ Data Highlights: Group Projects

Using Group Projects, students gain experience working with others by discussing a topic, analyzing data, and collaborating to formulate their response to the questions posed in the exercise.

Making the Jump

Get to the "Aha!" moment faster. *Understanding Basic Statistics* provides the push students need to get there through guidance and example.

PROCEDURE

How to test μ when σ is known

Requirements

Let x be a random variable appropriate to your application. Obtain a simple random sample (of size n) of x values from which you compute the sample mean $\bar{x}$. The value of σ is already known (perhaps from a previous study). If you can assume that x has a normal distribution, then any sample size n will work. If you cannot assume this, then use a sample size $n \geq 30$.

Procedure

1. In the context of the application, state the *null and alternate hypotheses* and set the *level of significance* α.
2. Use the known σ, the sample size n, the value of $\bar{x}$ from the sample, and μ from the null hypothesis to compute the standardized *sample test statistic*.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

3. Use the standard normal d
two-tailed, to find the *P*-v
4. *Conclude* the test. If *P*-val
then do not reject H_0.
5. *Interpret your conclusion*

◀ Procedures and Requirements

Procedure display boxes summarize simple step-by-step strategies for carrying out statistical procedures and methods as they are introduced. Requirements for using the procedures are also stated. Students can refer back to these boxes as they practice using the procedures.

GUIDED EXERCISE 10 *Probability Regarding $\bar{x}$*

In mountain country, major highways sometimes use tunnels instead of long, winding roads over high passes. However, too many vehicles in a tunnel at the same time can cause a hazardous situation. Traffic engineers are studying a long tunnel in Colorado. If x represents the time for a vehicle to go through the tunnel, it is known that the x distribution has mean $\mu = 12.1$ minutes and standard deviation $\sigma = 3.8$ minutes under ordinary traffic conditions. From a histogram of x values, it was found that the x distribution is mound-shaped with some symmetry about the mean.

Engineers have calculated that, *on average*, vehicles should spend from 11 to 13 minutes in the tunnel. If the time is less than 11 minutes, traffic is moving too fast for safe travel in the tunnel. If the time is more than 13 minutes, there is a problem of bad air quality (too much carbon monoxide and other pollutants).

Under ordinary conditions, there are about 50 vehicles in the tunnel at one time. What is the probability that the mean time for 50 vehicles in the tunnel will be from 11 to 13 minutes?

We will answer this question in steps.

(a) Let $\bar{x}$ represent the sample mean based on samples of size 50. Describe the $\bar{x}$ distribution.

➡ From the central limit theorem, we expect the $\bar{x}$ distribution to be approximately normal, with mean and standard deviation

$$\mu_{\bar{x}} = \mu = 12.1 \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.8}{\sqrt{50}} \approx 0.54$$

(b) Find $P(11 < \bar{x} < 13)$.

Jupiterimages/Stockbyte/Getty Images

➡ We convert the interval

$11 < \bar{x} < 13$

to a standard z interval and use the standard normal probability table to find our answer. Since

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx \frac{\bar{x} - 12.1}{0.54}$$

$\bar{x} = 11$ converts to $z \approx \frac{11 - 12.1}{0.54} = -2.04$

and $\bar{x} = 13$ converts to $z \approx \frac{13 - 12.1}{0.54} = 1.67$

Therefore,

$$\begin{aligned} P(11 < \bar{x} < 13) &= P(-2.04 < z < 1.67) \\ &= 0.9525 - 0.0207 \\ &= 0.9318 \end{aligned}$$

(c) **Interpret** your answer to part (b).

➡ It seems that about 93% of the time, there should be no safety hazard for average traffic flow.

Guided Exercises ▶

Students gain experience with new procedures and methods through Guided Exercises. Beside each problem in a Guided Exercise, a completely worked-out solution appears for immediate reinforcement.

Preface

Welcome to the exciting world of statistics! We have written this text to make statistics accessible to everyone, including those with a limited mathematics background. Statistics affects all aspects of our lives. Whether we are testing new medical devices or determining what will entertain us, applications of statistics are so numerous that, in a sense, we are limited only by our own imagination in discovering new uses for statistics.

Overview

The eighth edition of *Understanding Basic Statistics* continues to emphasize *concepts of statistics*. Statistical methods are carefully presented with a focus on understanding both the *suitability of the method* and the *meaning of the result.* Statistical methods and measurements are developed in the context of applications.

Critical thinking and interpretation are essential in understanding and evaluating information. Statistical literacy is fundamental for applying and comprehending statistical results. In this edition we have expanded and highlighted the treatment of statistical literacy, critical thinking, and interpretation.

We have retained and expanded features that made the first seven editions of the text very readable. Definition boxes highlight important terms. Procedure displays summarize steps for analyzing data. Examples, exercises, and problems touch on applications appropriate to a broad range of interests.

The eighth edition continues to have extensive online support. Instructional videos are available on DVD. The companion website at http://www.cengage.com/statistics/brase8e contains more than 100 data sets (in JMP, Microsoft Excel, Minitab, SPSS, and TI-84Plus/TI-83Plus/TI-*n*spire with TI-84Plus keypad ASCII file formats) and technology guides.

Major Changes in the Eighth Edition

With each new edition, the authors reevaluate the scope, appropriateness, and effectiveness of the text's presentation and reflect on extensive user feedback. Revisions have been made throughout the text to clarify explanations of important concepts and to update problems.

Current Issues in Statistics Addressed

Brief discussions of ethics in statistics (Section 1.3) and the appropirate use of *P*-values (Section 9.3) as recommended by the American Statistical Association are included. Also included is a brief presentation about data science and big data (Using Technology, Chapter 2).

New Terminology Problems

These problems are included in the chapter problems. They are primarily true/false or multiple choice, highlighting main features of terms or concepts.

Additional Flexibility Provided by Dividing Longer Chapters into Parts

The chapter on Normal Curves and Sampling Distributions (Chapter 7) has been divided into two parts that give appropriate places to pause, review, and summarize content. Each part has a brief introduction, a brief summary, and a list of chapter problems applicable to the part. The same features were aded to the last chapter, which includes applications of the chi-square distribution in Part I and inferences for linear regression in Part II.

Revised Examples and New Section Problems

Examples, guided exercises, and problem sets have been updated and revised as necessary.

Updates in Technology Including Minitab Express

Instructions for Excel 2013, Minitab 17, and new Minitab Express are included in the *Tech Notes* and *Using Technology.*

JMP Statistical Software

All new copies of the text now come with JMP statistical software at no additional cost. We realize the vital importance of having a data analysis tool and understand the desirability for students to acquire familiarity with a commercial software package whose use might continue outside of this course. The text can still be used with many other popular software packages as well. See Instructor Resources on page xxv for a full description of the software.

WebAssign from Cengage

Understanding Basic Statistics, Eighth Edition is a fully customizable online solution for STEM disciplines that empowers you to help your students learn, not just do homework. Insightful tools save you time and highlight exactly where your students are struggling. Decide when and what type of help students can access while working on assignments—and incentivize independent work so help features aren't abused. Meanwhile, your students get an engaging experience, instant feedback and better outcomes. A total win-win!

To try a sample assignment, learn about LMS integration or connect with our digital course support visit www.webassign.com/cengage

Continuing Content

Critical Thinking, Interpretation, and Statistical Literacy

The eighth edition of this text continues and expands the emphasis on critical thinking, interpretation, and statistical literacy. Calculators and computers are very good at providing numerical results of statistical processes. However, numbers from a computer or calculator display are meaningless unless the user knows how to interpret the results and if the statistical process is appropriate. This text helps students determine whether or not a statistical method or process is appropriate. It helps students understand what a statistic measures. It helps students interpret the results of a confidence interval, a hypothesis test, or a linear regression model.

Introduction of Hypothesis Testing Using *P*-Values

In keeping with the use of computer technology and standard practice in research, hypothesis testing is introduced using *P*-values. The critical region method is still supported but not given primary emphasis.

Use of Student's *t* Distribution in Confidence Intervals and Testing of Means

If the normal distribution is used in confidence intervals and testing of means, then the *population standard deviation must be known.* If the population standard deviation is not known, then under conditions described in the text, the Student's t distribution is used. This is the most commonly used procedure in statistical research. It is also used in statistical software packages such as JMP, Microsoft Excel 2013, Minitab, Minitab Express, SPSS, and TI-84Plus/TI-83Plus/TI-*n*spire calculators.

Confidence Intervals and Hypothesis Tests of Difference of Means

If the normal distribution is used, then both population standard deviations must be known. When this is not the case, the Student's t distribution incorporates an approximation for t, with a commonly used conservative choice for the degrees of freedom. Satterthwaite's approximation for the degrees of freedom as used in computer software is also discussed. The pooled standard deviation is presented for appropriate applications ($\sigma_1 \approx \sigma_2$).

Features in the Eighth Edition

Chapter and Section Lead-ins

- *Preview Questions* at the beginning of each chapter are keyed to the sections.
- *Focus Problems* at the beginning of each chapter demonstrate types of questions students can answer once they master the concepts and skills presented in the chapter.
- *Focus Points* at the beginning of each section describe the primary learning objectives of the section.

Carefully Developed Pedagogy

- *Examples* show students how to select and use appropriate procedures.
- *Guided Exercises* within the sections give students an opportunity to work with a new concept. Completely worked-out solutions appear beside each exercise to give immediate reinforcement.
- *Definition boxes* highlight important definitions throughout the text.
- *Procedure displays* summarize key strategies for carrying out statistical procedures and methods. Conditions required for using the procedure are also stated.
- *What Does (a concept method or result) Tell Us?* summarizes information we obtain from the named concepts and statistical processes and gives insight for additional application.
- *Important Features of a (concept, method, or result)* summarizes the features of the listed item.
- NEW! Recommendations from the Statistical Association of America regarding the importance of ethics in statistics and appropriate use of *P*-values are discussed.

- *Looking Forward* features give a brief preview of how a current topic is used later.
- *Labels* for each example or guided exercise highlight the technique, concept, or process illustrated by the example or guided exercise. In addition, labels for section and chapter problems describe the field of application and show the wide variety of subjects in which statistics is used.
- *Section and chapter problems* require the student to use all the new concepts mastered in the section or chapter. Problem sets include a variety of real-world applications with data or settings from identifiable sources. Key steps and solutions to odd-numbered problems appear at the end of the book.
- *Basic Computation problems* ask students to practice using formulas and statistical methods on very small data sets. Such practice helps students understand what a statistic measures.
- *Statistical Literacy problems* ask students to focus on correct terminology and processes of appropriate statistical methods. Such problems occur in every section and chapter problem set.
- *Interpretation problems* ask students to explain the meaning of the statistical results in the context of the application.
- *Critical Thinking problems* ask students to analyze and comment on various issues that arise in the application of statistical methods and in the interpretation of results. These problems occur in every section and chapter problem set.
- *Expand Your Knowledge problems* present enrichment topics such as dotplots, grouped data, estimation of standard deviation from a range of data values, residual plots, relationship between confidence intervals and two-tailed hypothesis tests, and more.
- NEW! *Terminology Problems* at the end of each chapter are true/false or multiple choice questions regarding important features or terms or concepts.
- *Cumulative review problem sets* occur after every third chapter and include key topics from previous chapters. Answers to *all* cumulative review problems are given at the end of the book.
- *Data Highlights and Linking Concepts* provide group projects and writing projects.
- *Viewpoints* are brief essays presenting diverse situations in which statistics is used.
- *Design and photos* are appealing and enhance readability.

Expand Your Knowledge Problems and Quick Overview Topics With Additional Applications

Expand Your Knowledge problems do just that! These are optional but contain very useful information taken from the vast literature of statistics. These topics are not included in the main text but are easily learned using material from the section or previous sections. Although these topics are optional, the authors feel that they add depth and enrich a student's learning experience. Each topic was chosen for its relatively straightforward presentation and useful applications. All such problems and their applications are flagged with a sun logo.

Expand Your Knowledge problems in the eighth edition provide additional topics found in Advanced Placement courses. These topics include logarithmic transformation problems featuring exponential growth and power law applications, linear functions of random variables, and confidence intervals for the slope of the least-squares line. Also new are problems involving stratified sampling and the best estimate for the population mean μ.

Some of the other topics in *Expand Your Knowledge problems* or quick overviews include graphs such as dotplots, donut graphs, and variations on stem-and-leaf plots; outliers in stem-and-leaf plots; harmonic and geometric means; moving averages; calculating odds in favor and odds against; continuous uniform distribution, quick estimate of the standard deviation using the Empirical rule; plus four confidence intervals for proportions; Satterthwaite's approximation for degrees of freedom in confidence intervals and hypothesis tests; relationship between confidence intervals and two-tailed hypothesis testing; pooled two-sample procedures for confidence intervals and hypothesis tests; resampling (also known as bootstrap); simulations of confidence intervals and hypothesis tests using different samples of the same size; and serial correlation (also called autocorrelation).

For location of these optional topics in the text, please see the index.

Technology Within the Text

- *Tech Notes* within sections provide brief point-of-use instructions for the TI-84Plus, TI-83Plus, and TI-*n*spire (with 84Plus keypad) calculators, Microsoft Excel 2013, Minitab and Minitab Express.
- *Using Technology* sections show the use of SPSS as well as the TI-84Plus, TI-83Plus, and TI-*n*spire (with TI-84Plus keypad) calculators, Microsoft Excel 2013, Minitab 17 and Minitab Express.

Most Recent Operating System for the TI-84Plus/TI-83Plus Calculators

The latest operating system (v2.55MP) for the TI-84Plus/TI-83Plus calculators is also discussed, with new functions such as the inverse t distribution and the chi-square goodness-of-fit test described. One convenient feature of the operating system is that it provides on-screen prompts for inputs required for many probability and statistical functions. This operating system is already on new TI-84Plus/TI-83Plus calculators and is available for download to older calculators at the Texas Instruments website.

Alternate Routes Through the Text

Understanding Basic Statistics, eighth edition, is designed to be flexible. It offers the professor a choice of teaching possibilities. In most one-semester courses, it is not practical to cover all the material in depth. However, depending on the emphasis of the course, the professor may choose to cover various topics. For help in topic selection, refer to the Table of Prerequisite Material on page 1.

- *Linear regression.* Chapter 4, Correlation and Regression, may be delayed until after Chapter 9. The descriptive topics of linear regression may then be followed immediately by the inferential topics of linear regression presented in Chapter 11.
- *Probability.* For courses requiring minimal probability, Section 5.1 (What Is Probability?) and the first part of Section 5.2 (Some Probability Rules—Compound Events) will be sufficient.

Acknowledgments

It is our pleasure to acknowledge the prepublication reviewers of this text. All of their insights and comments have been very valuable to us. Reviewers of this text include

Bill Burgin, Gaston College
Robert Faux, Pennsylvania State University
Sharon Giles, Grossmont College
Tammy Hoyle, Gardner-Webb University
Dale Joyner, Caldwell Community College & Technical Institute
Kewal Krishan, Hudson County Community College
Theodore Lai, Hudson County Community College
Marilyn Milovich, Missouri Valley College
Stacy Reagan, Caldwell Community College
Cynthia Roemer, Union County College
Risharra Stulc, Colorado Mesa University
Jason Willis, Gardner-Webb

We also would like to acknowledge reviewers of previous editions:

Reza Abbasian, Texas Lutheran University
Paul Ache, Kutztown University
Kathleen Almy, Rock Valley College
Polly Amstutz, University of Nebraska at Kearney
Delores Anderson, Truett-McConnell College
Robert J. Astalos, Feather River College
Lynda L. Ballou, Kansas State University
Kimberly Benien, Wharton County Junior College
Mary Benson, Pensacola Junior College
Larry Bernett, Benedictine University
Kiran Bhutani, The Catholic University of America
Abraham Biggs, Broward Commuity College
Melissa Bingham, University of Wisconsin-La Crosse
Kristy E. Bland, Valdosta State University
John Bray, Broward Community College
Bill Burgin, Gaston College
Toni Carroll, Siena Heights University
Chand Chauhan, Indiana University-Purdue University-Fort Wayne
Pinyuen Chen, Syracuse University
Emmanuel des-Bordes, James A. Rhodes State College
Jennifer M. Dollar, Grand Rapids Community College
Larry E. Dunham, Wor-Wic Community College
Andrew Ellett, Indiana University
Ruby Evans, Keiser University
Mary Fine, Moberly Area Community College
Rebecca Fouguette, Santa Rosa Junior College
Rene Garcia, Miami-Dade Community College
Larry Green, Lake Tahoe Community College
Shari Harris, John Wood Community College
Janice Hector, DeAnza College
Tammy Hoyle, Gardner-Webb University
Jane Keller, Metropolitan Community College
Raja Khoury, Collin County Community College
Diane Koenig, Rock Valley College
Charles G. Laws, Cleveland State Community College
Michael R. Lloyd, Henderson State University
Beth Long, Pellissippi State Technical and Community College
Lewis Lum, University of Portland
Darcy P. Mays, Virginia Commonwealth University
Charles C. Okeke, College of Southern Nevada, Las Vegas
Peg Pankowski, Community College of Allegheny County
Ram Polepeddi, Westwood College, Denver North Campus
Azar Raiszadeh, Chattanooga State Technical Community College
David Read, Paradise Valley Community College
Traei Reed, St. Johns River Community College
Michael L. Russo, Suffolk County Community College
Janel Schultz, Saint Mary's University of Minnesota
Sankara Sethuraman, Augusta State University
Stephen Soltys, West Chester University of Pennsylvania
Ron Spicer, Colorado Technical University
Winson Taam, Oakland University
Jennifer L. Taggart, Rockford College
William Truman, University of North Carolina at Pembroke
Diane Van Deusen, Napa Valley College

Daniel Weiner, Boston University
Bill White, University of South Carolina Upstate
Jim Wienckowski, State University of New York at Buffalo
Stephen M. Wilkerson, Susquehanna University
Hongkai Zhang, East Central University
Shunpu Zhang, University of Alaska, Fairbanks
Cathy Zucco-Teveloff, Trinity College

We would especially like to thank Roger Lipsett for his careful accuracy review of this text. We are especially appreciative of the excellent work by the editorial and production professionals at Cengage Learning. In particular we thank Catherine Van Der Laan, Spencer Arritt, Gabriela Carrascal, and Andrea Wagner.

Without their creative insight and attention to detail, a project of this quality and magnitude would not be possible. Finally, we acknowledge the cooperation of Minitab, Inc., SPSS, Texas Instruments, and Microsoft.

Charles Henry Brase
Corrinne Pellillo Brase

Additional Resources—Get More from Your Textbook!

JMP is a statistics software for Windows and Macintosh computers from SAS, the market leader in analytics software and services for industry. JMP Student Edition is a streamlined, easy-to-use version that provides all the statistical analysis and graphics covered in this textbook. Once data are imported, students will find that most procedures require just two or three mouse clicks. JMP can import data from a variety of formats, including Excel and other statistical packages, and you can easily copy and paste graphs and output into documents.

JMP also provides an interface to explore data visually and interactively, which will help your students develop a healthy relationship with their data, work more efficiently with data, and tackle difficult statistical problems more easily. Because its output provides both statistics and graphs together, the student will better see and understand the application of concepts covered in this book as well. JMP Student Edition also contains some unique platforms for student projects, such as mapping and scripting. JMP functions in the same way on both Windows and Macintosh platforms and instructions contained with this book apply to both platforms.

Access to this software is available with new copies of the book. Students can purchase JMP standalone via CengageBrain.com or www.jmp.com/getse.

WebAssign from Cengage

Understanding Basic Statistics, Eighth Edition is a fully customizable online solution for STEM disciplines that empowers you to help your students learn, not just do homework. Insightful tools save you time and highlight exactly where your students are struggling. Decide when and what type of help students can access while working on assignments—and incentivize independent work so help features aren't abused. Meanwhile, your students get an engaging experience, instant feedback and better outcomes. A total win-win!

To try a sample assignment, learn about LMS integration or connect with our digital course support visit www.webassign.com/cengage

Instructor Resources

Annotated Instructor's Edition (AIE) Answers to all exercises, teaching comments, and pedagogical suggestions appear in the margin or at the end of the text in the case of large graphs.

Complete Solutions Manual This manual contains solutions to all exercises from the text, including Chapter Review Exercises, and Cumulative Review Exercises. This manual can be found on the Instructors Companion Site.

Cengage Learning Testing Powered by Cognero A flexible, online system that 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 LMS, your classroom or wherever you want.

Companion Website The companion website at http://www.cengage.com/brase contains a variety of resources:

- Microsoft® PowerPoint® lecture slides
- More than 100 data sets in a variety of formats, including
  - JMP
  - Microsoft Excel 2013
  - Minitab
  - SPSS
  - TI-84Plus/TI-83Plus/TI-nspire with 84plus keypad ASCII file formats
- Technology guides for the following programs
  - JMP
  - TI-84Plus, TI-83Plus, and TI-*n*spire graphing calculators
  - Minitab software (version 14)
  - Microsoft Excel (2010/2007)
  - SPSS Statistics software

Student Resources

Student Solutions Manual Provides solutions to the odd-numbered section and chapter exercises and to all the Cumulative Review exercises in the student textbook.

Minitab® and IBM SPSS These statistical software packages manipulate and interpret data to produce textual, graphical, and tabular results. Minitab® and/or SPSS may be packaged with the textbook. Student versions are available.

The companion website at http://www.cengage.com/statistics/brase8e contains useful assets for students:

- **Technology Guides** Separate guides exist with information and examples for various technology tools. Guides are available for JMP; the TI-84Plus, TI-83Plus, and TI-*n*spire graphing calculators; Minitab software (version 14); Microsoft Excel (2010/2007); and SPSS Statistics software.
- **Interactive Teaching and Learning Tools** include glossary flashcards, online data sets (in JMP, Microsoft Excel 2013, Minitab, SPSS, and Tl-84Plus/TI-83Plus/TI-*n*spire with TI-84Plus keypad ASCII file formats), statistical tables and formulas, and more.

CENGAGEbrain.com

CengageBrain.com Provides the freedom to purchase online homework and other materials à la carte—exactly what you need, when you need it.

For more information, visit **http://www.cengage.com/statistics/brase8e** or contact your local Cengage Learning sales representative.

Table of Prerequisite Material

Chapter		Prerequisite Sections
1	Getting Started	None
2	Organizing Data	1.1, 1.2
3	Averages and Variation	1.1, 1.2, 2.1
4	Correlation and Regression	1.1, 1.2, 3.1, 3.2
5	Elementary Probability Theory	1.1, 1.2, 2.1
6	The Binomial Probability Distribution and Related Topics	1.1, 1.2, 2.1, 3.1, 3.2, 5.1, 5.2 5.3 useful but not essential
7	Normal Curves and Sampling	
	Distributions (omit 7.6)	1.1, 1.2, 2.1, 3.1, 3.2, 5.1, 5.2, 6.1
	(include 7.6)	also 6.2, 6.3
8	Estimation	
	(omit 8.3)	1.1, 1.2, 2.1, 3.1, 3.2, 5.1, 5.2, 6.1, 7.1, 7.2, 7.3, 7.4, 7.5
	(include 8.3)	also 6.2, 6.3, 7.6
9	Hypothesis Testing	
	(omit 9.3)	1.1, 1.2, 2.1, 3.1, 3.2, 5.1, 5.2, 6.1, 7.1, 7.2, 7.3, 7.4, 7.5
	(include 9.3)	also 6.2, 6.3, 7.6
10	Inferences about Differences	
	(omit 10.3)	1.1, 1.2, 2.1, 3.1, 3.2, 5.1, 5.2, 6.1, 7.1, 7.2, 7.3, 7.4, 7.5, 8.1, 8.2, 9.1, 9.2
	(include 10.3)	also 6.2, 6.3, 7.6, 9.3
11	Additional Topics Using Inference	
	Part I Chi-Square Applications: 11.1, 11.2, 11.3	1.1, 1.2, 2.1, 3.1, 3.2, 5.1, 5.2, 6.1, 7.1, 7.2, 7.3, 7.4, 7.5, 9.1
	(Part II Inferences for Linear Regression: 11.4)	Chapter 4, 8.1, 8.2, 9.1, 9.2 also

1

Paul Spinelli/Major League Baseball/Getty Images

Bettmann/Historical/Corbis

Chance favors the prepared mind.

—Louis Pasteur

Statistical techniques are tools of thought . . . not substitutes for thought.

—Abraham Kaplan

Louis Pasteur (1822–1895) is the founder of modern bacteriology. At age 57, Pasteur was studying cholera. He accidentally left some bacillus culture unattended in his laboratory during the summer. In the fall, he injected laboratory animals with this bacilli. To his surprise, the animals did not die—in fact, they thrived and were resistant to cholera.

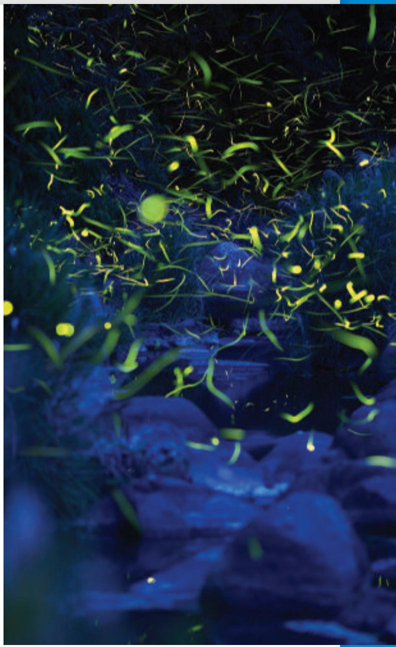
When the final results were examined, it is said that Pasteur remained silent for a minute and then exclaimed, as if he had seen a vision, "Don't you see; they have been vaccinated!" Pasteur's work ultimately saved many human lives.

Most of the important decisions in life involve incomplete information. Such decisions often involve so many complicated factors that a complete analysis is not practical or even possible. We are often forced into the position of making a guess based on limited information.

As the first quote reminds us, our chances of success are greatly improved if we have a "prepared mind." The statistical methods you will learn in this book will help you achieve a prepared mind for the study of many different fields. The second quote reminds us that statistics is an important tool, but it is not a replacement for an in-depth knowledge of the field to which it is being applied.

The authors of this book want you to understand and enjoy statistics. The reading material will *tell you* about the subject. The examples will *show you* how it works. To understand, however, you must *get involved*. Guided exercises, calculator and computer applications, section and chapter problems, and writing exercises are all designed to get you involved in the subject. As you grow in your understanding of statistics, we believe you will enjoy learning a subject that has a world full of interesting applications.

Getting Started



iStock/Getty Images Plus/north-tail/Getty Images

PREVIEW QUESTIONS

Why is statistics important? (SECTION 1.1)
What is the nature of data? (SECTION 1.1)
How can you draw a random sample? (SECTION 1.2)
What are other sampling techniques? (SECTION 1.2)
How can you design ways to collect data? (SECTION 1.3)
Why are ethical considerations important? (SECTION 1.3)

FOCUS PROBLEMS

Where Have All the Fireflies Gone?

A feature article in *The Wall Street Journal* discusses the disappearance of fireflies. In the article, Professor Sara Lewis of Tufts University and other scholars express concern about the decline in the worldwide population of fireflies.

There are a number of possible explanations for the decline, including habitat reduction of woodlands, wetlands, and open fields; pesticides; and pollution. Artificial nighttime lighting might interfere with the Morse–code–like mating ritual of the fireflies. Some chemical companies pay a bounty for fireflies because the insects contain two rare chemicals used in medical research and electronic detection systems used in spacecraft.

What does any of this have to do with statistics?

The truth, at this time, is that no one really knows (a) how much the world firefly population has declined or (b) how to explain the decline. The population of all fireflies is simply too large to study in its entirety.

In any study of fireflies, we must rely on incomplete information from samples. Furthermore, from these samples we must draw realistic conclusions that have statistical integrity. This is the kind of work that makes use of statistical methods to determine ways to collect, analyze, and investigate data.

Suppose you are conducting a study to compare firefly populations exposed to normal daylight/darkness conditions with firefly populations exposed to continuous light (24 hours a day). You set up two firefly colonies in a laboratory environment. The two colonies are identical except that one colony is exposed to normal daylight/darkness conditions and the other is exposed to continuous light. Each colony is populated with the same number of mature fireflies. After 72 hours, you count the number of living fireflies in each colony.

After completing this chapter, you will be able to answer the following questions:

(a) Is this an experiment or an observation study? Explain.
(b) Is there a control group? Is there a treatment group?

Courtesy of Corrinne and Charles Brase

Adapted from Ohio State University Firefly Files logo

For online student resources, visit the Brase/Brase, *Understanding Basic Statistics*, 8th edition web site at http://www.cengage.com/statistics/brase.

(c) What is the variable in this study?
(d) What is the level of measurement (nominal, interval, ordinal, or ratio) of the variable?

(See Problem 11 of the Chapter 1 Review Problems.)

SECTION 1.1 What Is Statistics?

FOCUS POINTS

- **Identify variables in a statistical study.**
- **Distinguish between quantitative and qualitative variables.**
- **Identify populations and samples.**
- **Distinguish between parameters and statistics.**
- **Determine the level of measurement.**
- **Compare descriptive and inferential statistics.**

Introduction

Decision making is an important aspect of our lives. We make decisions based on the information we have, our attitudes, and our values. Statistical methods help us examine information. Moreover, statistics can be used for making decisions when we are faced with uncertainties. For instance, if we wish to estimate the proportion of people who will have a severe reaction to a flu shot without giving the shot to everyone who wants it, statistics provides appropriate methods. Statistical methods enable us to look at information from a small collection of people or items and make inferences about a larger collection of people or items.

Procedures for analyzing data, together with rules of inference, are central topics in the study of statistics.

Statistics

Statistics is the study of how to collect, organize, analyze, and interpret numerical information from data.

The subject of statistics is multifaceted. The following definition of statistics is found in the *International Encyclopedia of Statistical Science*, edited by Miodrag Lovric. Professor David Hand of Imperial College London—the president of the Royal Statistical Society—presents the definition in his article "Statistics: An Overview."

Statistics is both the science of uncertainty and the technology of extracting information from data.

The statistical procedures you will learn in this book should supplement your built-in system of inference—that is, the results from statistical procedures and good sense should dovetail. Of course, statistical methods themselves have no power to work miracles. These methods can help us make some decisions, but not all conceivable decisions. Remember, even a properly applied statistical procedure is no more accurate than the data, or facts, on which it is based. Finally, statistical results should be interpreted by one who understands not only the methods, but also the subject matter to which they have been applied.

The general prerequisite for statistical decision making is the gathering of data. First, we need to identify the individuals or objects to be included in the study and the characteristics or features of the individuals that are of interest.

Individuals
Variable

Individuals are the people or objects included in the study.
A **variable** is a characteristic of the individual to be measured or observed.

For instance, if we want to do a study about the people who have climbed Mt. Everest, then the individuals in the study are all people who have actually made it to the summit. One variable might be the height of such individuals. Other variables might be age, weight, gender, nationality, income, and so on. Regardless of the variables we use, we would not include measurements or observations from people who have not climbed the mountain.

The variables in a study may be *quantitative* or *qualitative* in nature.

Quantitative variable
Qualitative variable

A **quantitative variable** has a value or numerical measurement for which operations such as addition or averaging make sense. A **qualitative variable** describes an individual by placing the individual into a category or group, such as male or female.

For the Mt. Everest climbers, variables such as height, weight, age, or income are *quantitative* variables. *Qualitative variables* involve nonnumerical observations such as gender or nationality. Sometimes qualitative variables are referred to as *categorical variables.*

Categorical variable

Another important issue regarding data is their source. Do the data comprise information from *all* individuals of interest, or from just *some* of the individuals?

Population data
Sample data

In **population data,** the data are from *every* individual of interest.
In **sample data,** the data are from *only some* of the individuals of interest.

It is important to know whether the data are population data or sample data. Data from a specific population are fixed and complete. Data from a sample may vary from sample to sample and are *not* complete.

Population parameter

Sample statistic

A **population parameter** is a numerical measure that describes an aspect of a population.
A **sample statistic** is a numerical measure that describes an aspect of a sample.

For instance, if we have data from *all* the individuals who have climbed Mt. Everest, then we have population data. The proportion of males in the *population* of all climbers who have conquered Mt. Everest is an example of a *parameter.*

On the other hand, if our data come from just some of the climbers, we have sample data. The proportion of male climbers in the *sample* is an example of a *statistic.* Note that different samples may have different values for the proportion of male climbers. One of the important features of sample statistics is that they can vary from sample to sample, whereas population parameters are fixed for a given population.

LOOKING FORWARD

In later chapters we will use information based on a sample and sample statistics to estimate population parameters (Chapter 8) or make decisions about the value of population parameters (Chapters 9 and 10).

EXAMPLE 1 *Using Basic Terminology*

The Hawaii Department of Tropical Agriculture is conducting a study of ready-to-harvest pineapples in an experimental field.

(a) The pineapples are the *objects* (individuals) of the study. If the researchers are interested in the individual weights of pineapples in the field, then the *variable* consists of weights. At this point, it is important to specify units of measurement

and degrees of accuracy of measurement. The weights could be measured to the nearest ounce or gram. Weight is a *quantitative* variable because it is a numerical measure. If weights of *all* the ready-to-harvest pineapples in the field are included in the data, then we have a *population*. The average weight of all ready-to-harvest pineapples in the field is a *parameter.*

(b) Suppose the researchers also want data on taste. A panel of tasters rates the pineapples according to the categories "poor," "acceptable," and "good." Only some of the pineapples are included in the taste test. In this case, the *variable* is taste. This is a *qualitative* or *categorical* variable. Because only some of the pineapples in the field are included in the study, we have a *sample*. The proportion of pineapples in the sample with a taste rating of "good" is a *statistic*.

Throughout this text, you will encounter *guided exercises* embedded in the reading material. These exercises are included to give you an opportunity to work immediately with new ideas. The questions guide you through appropriate analysis. Cover the answers on the right side (an index card will fit this purpose). After you have thought about or written down *your own response,* check the answers. If there are several parts to an exercise, check each part before you continue. You should be able to answer most of these exercise questions, but don't skip them—they are important.

GUIDED EXERCISE 1 *Using Basic Terminology*

How important is music education in school (K–12)? *The Harris Poll* did an online survey of 2286 adults (aged 18 and older) within the United States. Among the many questions, the survey asked if the respondents agreed or disagreed with the statement "Learning and habits from music education equip people to be better team players in their careers." In the most recent survey, 71% of the study participants agreed with the statement.

(a) Identify the individuals of the study and the variable.

→ The individuals are the 2286 adults who participated in the online survey. The variable is the response agree or disagree with the statement that music education equips people to be better team players in their careers.

(b) Do the data comprise a sample? If so, what is the underlying population?

→ The data comprise a sample of the population of responses from all adults in the United States.

(c) Is the variable qualitative or quantitative?

→ Qualitative—the categories are the two possible responses: agrees or disagrees with the statement that music education equips people to be better team players in their careers.

(d) Identify a quantitative variable that might be of interest.

→ Age or income might be of interest.

(e) Is the proportion of respondents in the sample who agree with the statement regarding music education and effect on careers a statistic or a parameter?

→ Statistic—the proportion is computed from sample data.

Levels of Measurement: Nominal, Ordinal, Interval, Ratio

We have categorized data as either qualitative or quantitative. Another way to classify data is according to one of the four *levels of measurement*. These levels indicate the type of arithmetic that is appropriate for the data, such as ordering, taking differences, or taking ratios.

LEVELS OF MEASUREMENT

Levels of Measurement

Nominal level

The **nominal level of measurement** applies to data that consist of names, labels, or categories. There are no implied criteria by which the data can be ordered from smallest to largest.

Ordinal level

The **ordinal level of measurement** applies to data that can be arranged in order. However, differences between data values either cannot be determined or are meaningless.

Interval level

The **interval level of measurement** applies to data that can be arranged in order. In addition, differences between data values are meaningful.

Ratio level

The **ratio level of measurement** applies to data that can be arranged in order. In addition, both differences between data values and ratios of data values are meaningful. Data at the ratio level have a true zero.

EXAMPLE 2 *Levels of Measurement*

Identify the type of data.

(a) Taos, Acoma, Zuni, and Cochiti are the names of four Native American pueblos from the population of names of all Native American pueblos in Arizona and New Mexico.

Michelle Dulieu/Shutterstock.com

SOLUTION: These data are at the *nominal* level. Notice that these data values are simply names. By looking at the name alone, we cannot determine if one name is "greater than or less than" another. Any ordering of the names would be numerically meaningless.

(b) In a high school graduating class of 319 students, Jim ranked 25th, June ranked 19th, Walter ranked 10th, and Julia ranked 4th, where 1 is the highest rank.

SOLUTION: These data are at the *ordinal* level. Ordering the data clearly makes sense. Walter ranked higher than June. Jim had the lowest rank, and Julia the highest. However, numerical differences in ranks do not have meaning. The difference between June's and Jim's ranks is 6, and this is the same difference that exists between Walter's and Julia's ranks. However, this difference doesn't really mean anything significant. For instance, if you looked at grade point average, Walter and Julia may have had a large gap between their grade point averages, whereas June and Jim may have had closer grade point averages. In any ranking system, it is only the relative standing that matters. Computed differences between ranks are meaningless.

iStockphoto.com/Korban Schwabi

(c) Body temperatures (in degrees Celsius) of trout in the Yellowstone River.

SOLUTION: These data are at the interval level. We can certainly order the data, and we can compute meaningful differences. However, for Celsius-scale temperatures, there is not an inherent starting point. The value 0°C may seem to be a starting point, but this value does not indicate the state of "no heat." Furthermore, it is not correct to say that 20°C is twice as hot as 10°C.

(d) Length of trout swimming in the Yellowstone River.

SOLUTION: These data are at the *ratio* level. An 18-inch trout is three times as long as a 6-inch trout. Observe that we can divide 6 into 18 to determine a meaningful *ratio* of trout lengths.

In summary, there are four levels of measurement. The nominal level is considered the lowest, and in ascending order we have the ordinal, interval, and ratio levels. In general, calculations based on a particular level of measurement may not be appropriate for a lower level.

PROCEDURE

How to determine the level of measurement

The levels of measurement, listed from lowest to highest, are nominal, ordinal, interval, and ratio. To determine the level of measurement of data, state the *highest level* that can be justified for the entire collection of data. Consider which calculations are suitable for the data.

Level of Measurement	Suitable Calculation
Nominal	We can put the data into categories.
Ordinal	We can order the data from smallest to largest or "worst" to "best." Each data value can be *compared* with another data value.
Interval	We can order the data and also take the differences between data values. At this level, it makes sense to compare the differences between data values. For instance, we can say that one data value is 5 more than or 12 less than another data value.
Ratio	We can order the data, take differences, and also find the ratio between data values. For instance, it makes sense to say that one data value is twice as large as another.

What Does the Level of Measurement Tell Us?

The level of measurement tells us which arithmetic processes are appropriate for the data. This is important because different statistical processes require various kinds of arithmetic. In some instances all we need to do is count the number of data that meet specified criteria. In such cases nominal (and higher) data levels are all appropriate. In other cases we need to order the data, so nominal data would not be suitable. Many other statistical processes require division, so data need to be at the ratio level. Just keep the nature of the data in mind before beginning statistical computations.

GUIDED EXERCISE 2 *Levels of Measurement*

The following describe different data associated with a state senator. For each data entry, indicate the corresponding *level of measurement.*

(a) The senator's name is Sam Wilson. → Nominal level

(b) The senator is 58 years old. → Ratio level. Notice that age has a meaningful zero. It makes sense to give age ratios. For instance, Sam is twice as old as someone who is 29.

Continued

Guided Exercise 2 *continued*

(c) The years in which the senator was elected to the Senate are 2000, 2006, and 2012.

→ Interval level. Dates can be ordered, and the difference between dates has meaning. For instance, 2006 is 6 years later than 2000. However, ratios do not make sense. The year 2000 is not twice as large as the year 1000. In addition, the year 0 does not mean "no time."

(d) The senator's total taxable income last year was $878,314.

→ Ratio level. It makes sense to say that the senator's income is 10 times that of someone earning $87,831.

(e) The senator surveyed his constituents regarding his proposed water protection bill. The choices for response were strongly support, support, neutral, against, or strongly against.

→ Ordinal level. The choices can be ordered, but there is no meaningful numerical difference between two choices.

(f) The senator's marital status is "married."

→ Nominal level

(g) A leading news magazine claims that the senator is ranked seventh for his voting record on bills regarding public education.

→ Ordinal level. Ranks can be ordered, but differences between ranks may vary in meaning.

CRITICAL THINKING

"Data! Data! Data!" he cried impatiently. "I can't make bricks without clay." Sherlock Holmes said these words in *The Adventure of the Copper Beeches* by Sir Arthur Conan Doyle.

Reliable statistical conclusions require reliable data. This section has provided some of the vocabulary used in discussing data. As you read a statistical study or conduct one, pay attention to the nature of the data and the ways they were collected.

When you select a variable to measure, be sure to specify the process and requirements for measurement. For example, if the variable is the weight of ready-to-harvest pineapples, specify the unit of weight, the accuracy of measurement, and maybe even the particular scale to be used. If some weights are in ounces and others in grams, the data are fairly useless.

Another concern is whether or not your measurement instrument truly measures the variable. Just asking people if they know the geographic location of the island nation of Fiji may not provide accurate results. The answers may reflect the fact that the respondents want you to think they are knowledgeable. Asking people to locate Fiji on a map may give more reliable results.

The level of measurement is also an issue. You can put numbers into a calculator or computer and do all kinds of arithmetic. However, you need to judge whether the operations are meaningful. For ordinal data such as restaurant rankings, you can't conclude that a 4-star restaurant is "twice as good" as a 2-star restaurant, even though the number 4 is twice 2.

Continued

Critical Thinking *continued*

Are the data from a sample, or do they comprise the entire population? Sample data can vary from one sample to another! This means that if you are studying the same statistic from two different samples of the same size, the data values may be different. In fact, the ways in which sample statistics vary among different samples of the same size will be the focus of our study from Section 7.4 on.

INTERPRETATION **When you work with sample data, carefully consider the population from which they are drawn. Observations and analysis of the sample are applicable to only the population from which the sample is drawn.**

Descriptive statistics

Inferential statistics

Descriptive statistics involves methods of organizing, picturing, and summarizing information from samples or populations.
Inferential statistics involves methods of using information from a sample to draw conclusions regarding the population.

LOOKING FORWARD

The purpose of collecting and analyzing data is to obtain information. Statistical methods provide us tools to obtain information from data. These methods break into two branches.

We will look at methods of descriptive statistics in Chapters 2, 3, and 4. These methods may be applied to data from samples or populations.

Sometimes we do not have access to an entire population. At other times, the difficulties or expense of working with the entire population is prohibitive. In such cases, we will use inferential statistics together with probability. These are the topics of Chapters 5 through 11.

VIEWPOINT *The First Measured Century*

The 20th century saw measurements of aspects of American life that had never been systematically studied before. Social conditions involving crime, sex, food, fun, religion, and work were numerically investigated. The measurements and survey responses taken over the entire century reveal unsuspected statistical trends. *The First Measured Century* is a book by Caplow, Hicks, and Wattenberg. It is also a PBS documentary available on video. For more information, visit the PBS web site and find the link to *The First Measured Century* documentary.

SECTION 1.1 PROBLEMS

1. **Statistical Literacy** In a statistical study, what is the difference between an individual and a variable?

2. **Statistical Literacy** Are data at the nominal level of measurement quantitative or qualitative?

3. **Statistical Literacy** What is the difference between a parameter and a statistic?

4. **Statistical Literacy** For a set population, does a parameter ever change? If there are three different samples of the same size from a set population, is it possible to get three different values for the same statistic?

5. **Critical Thinking** Numbers are often assigned to data that are categorical in nature.
   (a) Consider these number assignments for category items describing electronic ways of expressing personal opinions:

   1 = Twitter; 2 = e-mail; 3 = text message; 4 = Facebook; 5 = blog

   Are these numerical assignments at the ordinal data level or higher? Explain.
   (b) Consider these number assignments for category items describing usefulness of customer service:

   1 = not helpful; 2 = somewhat helpful; 3 = very helpful;
   4 = extremely helpful

   Are these numerical assignments at the ordinal data level? Explain. What about at the interval level or higher? Explain.

6. **Interpretation** Lucy conducted a survey asking some of her friends to specify their favorite type of TV entertainment from the following list of choices:

   sitcom; reality; documentary; drama; cartoon; other

   Do Lucy's observations apply to *all* adults? Explain. From the description of the survey group, can we draw any conclusions regarding age of participants, gender of participants, or education level of participants?

7. **Marketing: Fast Food** A national survey asked 1261 U.S. adult fast-food customers which meal (breakfast, lunch, dinner, snack) they ordered.
   (a) Identify the variable.
   (b) Is the variable quantitative or qualitative?
   (c) What is the implied population?

8. **Advertising: Auto Mileage** What is the average miles per gallon (mpg) for all new hybrid small cars? Using *Consumer Reports*, a random sample of such vehicles gave an average of 35.7 mpg.
   (a) Identify the variable.
   (b) Is the variable quantitative or qualitative?
   (c) What is the implied population?

9. **Ecology: Wetlands** Government agencies carefully monitor water quality and its effect on wetlands (Reference: *Environmental Protection Agency Wetland Report* EPA 832-R-93-005). Of particular concern is the concentration of nitrogen in water draining from fertilized lands. Too much nitrogen can kill fish and wildlife. Twenty-eight samples of water were taken at random from a lake. The nitrogen concentration (milligrams of nitrogen per liter of water) was determined for each sample.
   (a) Identify the variable.
   (b) Is the variable quantitative or qualitative?
   (c) What is the implied population?

10. **Archaeology: Ireland** The archaeological site of Tara is more than 4000 years old. Tradition states that Tara was the seat of the high kings of Ireland. Because of its archaeological importance, Tara has received extensive study (Reference: *Tara: An Archaeological Survey* by Conor Newman, Royal Irish Academy, Dublin). Suppose an archaeologist wants to estimate the density of ferromagnetic artifacts in the Tara region. For this purpose, a random sample

of 55 plots, each of size 100 square meters, is used. The number of ferromagnetic artifacts for each plot is determined.
(a) Identify the variable.
(b) Is the variable quantitative or qualitative?
(c) What is the implied population?

11. **Student Life: Levels of Measurement** Categorize these measurements associated with student life according to level: nominal, ordinal, interval, or ratio.
(a) Length of time to complete an exam
(b) Time of first class
(c) Major field of study
(d) Course evaluation scale: poor, acceptable, good
(e) Score on last exam (based on 100 possible points)
(f) Age of student

12. **Business: Levels of Measurement** Categorize these measurements associated with a robotics company according to level: nominal, ordinal, interval, or ratio.
(a) Salesperson's performance: below average, average, above average
(b) Price of company's stock
(c) Names of new products
(d) Temperature (°F) in CEO's private office
(e) Gross income for each of the past 5 years
(f) Color of product packaging

13. **Fishing: Levels of Measurement** Categorize these measurements associated with fishing according to level: nominal, ordinal, interval, or ratio.
(a) Species of fish caught: perch, bass, pike, trout
(b) Cost of rod and reel
(c) Time of return home
(d) Guidebook rating of fishing area: poor, fair, good
(e) Number of fish caught
(f) Temperature of water

14. **Education: Teacher Evaluation** If you were going to apply *statistical methods* to analyze teacher evaluations, which question form, A or B, would be better?

Form A: In your own words, tell how this teacher compares with other teachers you have had.

Form B: Use the following scale to rank your teacher as compared with other teachers you have had.

1	2	3	4	5
worst	**below average**	**average**	**above average**	**best**

15. **Critical Thinking** You are interested in the weights of backpacks students carry to class and decide to conduct a study using the backpacks carried by 30 students.
(a) Give some instructions for weighing the backpacks. Include unit of measure, accuracy of measure, and type of scale.
(b) Do you think each student asked will allow you to weigh his or her backpack?
(c) Do you think telling students ahead of time that you are going to weigh their backpacks will make a difference in the weights?

SECTION 1.2

Random Samples

Focus Points

- **Explain the importance of random samples.**
- **Construct a simple random sample using random numbers.**
- **Simulate a random process.**
- **Describe stratified sampling, cluster sampling, systematic sampling, multistage sampling, and convenience sampling.**

Simple Random Samples

Eat lamb—20,000 coyotes can't be wrong!

This slogan is sometimes found on bumper stickers in the western United States. The slogan indicates the trouble that ranchers have experienced in protecting their flocks from predators. Based on their experience with this sample of the coyote population, the ranchers concluded that *all* coyotes are dangerous to their flocks and should be eliminated! The ranchers used a special poison bait to get rid of the coyotes. Not only was this poison distributed on ranch land, but with government cooperation, it also was distributed widely on public lands.

The ranchers found that the results of the widespread poisoning were not very beneficial. The sheep-eating coyotes continued to thrive while the general population of coyotes and other predators declined. What was the problem? The sheep-eating coyotes that the ranchers had observed were not a representative sample of all coyotes. Modern methods of predator control, however, target the sheep-eating coyotes. To a certain extent, the new methods have come about through a closer examination of the sampling techniques used.

In this section, we will examine several widely used sampling techniques. One of the most important sampling techniques is a *simple random sample.*

Simple random sample

A **simple random sample** of n measurements from a population is a subset of the population selected in such a manner that every sample of size n from the population has an equal chance of being selected.

In a simple random sample, not only does every sample of the specified size have an equal chance of being selected, but every individual of the population also has an equal chance of being selected. However, the fact that each individual has an equal chance of being selected does not necessarily imply a simple random sample. Remember, for a simple random sample, every sample of the given size must also have an equal chance of being selected.

Important Features of a Simple Random Sample

For a simple random sample

- Every sample of specified size n from the population has an equal chance of being selected.
- No researcher bias occurs in the items selected for the sample.
- A random sample may not always reflect the diversity of the population. For instance, from a population of 10 cats and 10 dogs, a random sample of size 6 could consist of all cats.

GUIDED EXERCISE 3 *Simple Random Sample*

Is open space around metropolitan areas important? Players of the Colorado Lottery might think so, since some of the proceeds of the game go to fund open space and outdoor recreational space. To play the game, you pay $1 and choose any six different numbers from the group of numbers 1 through 42. If your group of six numbers matches the winning group of six numbers selected by simple random sampling, then you are a winner of a grand prize of at least $1.5 million.

(a) Is the number 25 as likely to be selected in the winning group of six numbers as the number 5?

➡ Yes. Because the winning numbers constitute a simple random sample, each number from 1 through 42 has an equal chance of being selected.

(b) Could all the winning numbers be even?

➡ Yes, since six even numbers is one of the possible groups of six numbers.

(c) Your friend always plays the numbers

1 2 3 4 5 6

Could she ever win?

➡ Yes. In a simple random sample, the listed group of six numbers is *as likely* as any of the 5,245,786 possible groups of six numbers to be selected as the winner. (See Section 5.43 to learn how to compute the number of possible groups of six numbers that can be selected from 42 numbers.)

How do we get random samples? Suppose you need to know if the emission systems of the latest shipment of Toyotas satisfy pollution-control standards. You want to pick a random sample of 30 cars from this shipment of 500 cars and test them. One way to pick a random sample is to number the cars 1 through 500. Write these numbers on cards, mix up the cards, and then draw 30 numbers. The sample will consist of the cars with the chosen numbers. If you mix the cards sufficiently, this procedure produces a random sample.

Random-number table

An easier way to select the numbers is to use a *random-number table*. You can make one yourself by writing the digits 0 through 9 on separate cards and mixing up these cards in a hat. Then draw a card, record the digit, return the card, and mix up the cards again. Draw another card, record the digit, and so on. Table 1 in the Appendix is a ready-made random-number table (adapted from Rand Corporation, *A Million Random Digits with 100,000 Normal Deviates*). Let's see how to pick our random sample of 30 Toyotas by using this random-number table.

EXAMPLE 3 *Random-Number Table*

Use a random-number table to pick a random sample of 30 cars from a population of 500 cars.


Vibrant Image Studio/Shutterstock.com

SOLUTION: Again, we assign each car a different number between 1 and 500, inclusive. Then we use the random-number table to choose the sample. Table 1 in the Appendix has 50 rows and 10 blocks of five digits each; it can be thought of as a solid mass of digits that has been broken up into rows and blocks for user convenience.

You read the digits by beginning anywhere in the table. We dropped a pin on the table, and the head of the pin landed in row 15, block 5. We'll begin there and list all the digits in that row. If we need more digits, we'll move on to row 16, and so on. The digits we begin with are

99281 59640 15221 96079 09961 05371

Since the highest number assigned to a car is 500, and this number has three digits, we regroup our digits into blocks of 3:

992	815	964	015	221	960	790	996	105	371

To construct our random sample, we use the first 30 car numbers we encounter in the random-number table when we start at row 15, block 5. We skip the first three groups—992, 815, and 964—because these numbers are all too large. The next group of three digits is 015, which corresponds to 15. Car number 15 is the first car included in our sample, and the next is car number 221. We skip the next three groups and then include car numbers 105 and 371. To get the rest of the cars in the sample, we continue to the next line and use the random-number table in the same fashion. If we encounter a number we've used before, we skip it.

COMMENT When we use the term *(simple) random sample,* we have very specific criteria in mind for selecting the sample. One proper method for selecting a simple random sample is to use a computer- or calculator-based random-number generator or a table of random numbers as we have done in the example. The term *random* should not be confused with *haphazard*!

PROCEDURE

How to draw a random sample

1. Number all members of the population sequentially.
2. Use a table, calculator, or computer to select random numbers from the numbers assigned to the population members.
3. Create the sample by using population members with numbers corresponding to those randomly selected.

LOOKING FORWARD

Simple random samples are key components in methods of inferential statistics that we will study in Chapters 8–11. In fact, in order to draw conclusions about a population, the methods we will study *require* that we have simple random samples from the populations of interest.

Another important use of random-number tables is in *simulation.* We use the word *simulation* to refer to the process of providing numerical imitations of "real" phenomena. Simulation methods have been productive in studying a diverse array of subjects such as nuclear reactors, cloud formation, cardiology (and medical science in general), highway design, production control, shipbuilding, airplane design, war games, economics, and electronics. A complete list would probably include something from every aspect of modern life. In Guided Exercise 4 we'll perform a brief simulation.

Simulation

A **simulation** is a numerical facsimile or representation of a real-world phenomenon.

GUIDED EXERCISE 4 *Simulation*

Use a random-number table to simulate the outcomes of tossing a balanced (that is, fair) penny 10 times.

(a) How many outcomes are possible when you toss a coin once? → Two—heads or tails

Continued

Guided Exercise 4 *continued*

(a) There are several ways to assign numbers to the two outcomes. Because we assume a fair coin, we can assign an even digit to the outcome "heads" and an odd digit to the outcome "tails." Then, starting at block 3 of row 2 of Table 1 in the Appendix, list the first 10 single digits.	7 1 5 4 9 4 4 8 4 3
(b) What are the outcomes associated with the 10 digits?	T T T H T H H H H T
(c) If you start in a different block and row of Table 1 in the Appendix, will you get the same sequence of outcomes?	It is possible, but not very likely. (In Section 5.3 you will learn how to determine that there are 1024 possible sequences of outcomes for 10 tosses of a coin.)

> Tech Notes

Sampling with replacement

Most statistical software packages, spreadsheet programs, and statistical calculators generate random numbers. In general, these devices sample with replacement. *Sampling with replacement* means that although a number is selected for the sample, it is *not removed* from the population. Therefore, the same number may be selected for the sample more than once. If you need to sample without replacement, generate more items than you need for the sample. Then sort the sample and remove duplicate values. Specific procedures for generating random samples using the TI-84Plus/TI-83Plus/TI-*n*spire (with TI-84Plus keypad) calculator, Excel 2013, Minitab, Minitab Express, and SPSS are shown in Using Technology at the end of this chapter. More details are given in the separate *Technology Guides* for each of these technologies.

Other Sampling Techniques

Although we will assume throughout this text that (simple) random samples are used, other methods of sampling are also widely used. Appropriate statistical techniques exist for these sampling methods, but they are beyond the scope of this text.

Stratified sampling

One of these sampling methods is called *stratified sampling*. Groups or classes inside a population that share a common characteristic are called *strata* (plural of *stratum*). For example, in the population of all undergraduate college students, some strata might be freshmen, sophomores, juniors, or seniors. Other strata might be men or women, in-state students or out-of-state students, and so on. In the method of stratified sampling, the population is divided into at least two distinct strata. Then a (simple) random sample of a certain size is drawn from each stratum, and the information obtained is carefully adjusted or weighted in all resulting calculations.

The groups or strata are often sampled in proportion to their actual percentages of occurrence in the overall population. However, other (more sophisticated) ways to determine the optimal sample size in each stratum may give the best results. In general, statistical analysis and tests based on data obtained from stratified samples are somewhat different from techniques discussed in an introductory course in statistics. Such methods for stratified sampling will not be discussed in this text.

Systematic sampling

Another popular method of sampling is called *systematic sampling*. In this method, it is assumed that the elements of the population are arranged in some natural sequential order. Then we select a (random) starting point and select every *k*th element for our sample. For example, people lining up to buy rock concert tickets are "in order." To generate a systematic sample of these people (and ask questions regarding topics such as age, smoking habits, income level, etc.), we could include every fifth person in line. The "starting" person is selected at random from the first five.

The advantage of a systematic sample is that it is easy to get. However, there are dangers in using systematic sampling. When the population is repetitive or cyclic in nature, systematic sampling should not be used. For example, consider a fabric mill that produces dress material. Suppose the loom that produces the material makes a mistake every 17th yard, but we check only every 16th yard with an automated electronic scanner. In this case, a random starting point may or may not result in detection of fabric flaws before a large amount of fabric is produced.

Cluster sampling

Cluster sampling is a method used extensively by government agencies and certain private research organizations. In cluster sampling, we begin by dividing the demographic area into sections. Then we randomly select sections or clusters. Every member of the cluster is included in the sample. For example, in conducting a survey of schoolchildren in a large city, we could first randomly select five schools and then include all the children from each selected school.

Multistage sampling

Often a population is very large or geographically spread out. In such cases, samples are constructed through a *multistage sample design* of several stages, with the final stage consisting of clusters. For instance, the government Current Population Survey interviews about 60,000 households across the United States each month by means of a multistage sample design.

For the Current Population Survey, the first stage consists of selecting samples of large geographic areas that do not cross state lines. These areas are further broken down into smaller blocks, which are stratified according to ethnic and other factors. Stratified samples of the blocks are then taken. Finally, housing units in each chosen block are broken into clusters of nearby housing units. A random sample of these clusters of housing units is selected, and each household in the final cluster is interviewed.

Convenience sampling

Convenience sampling simply uses results or data that are conveniently and readily obtained. In some cases, this may be all that is available, and in many cases, it is better than no information at all. However, convenience sampling does run the risk of being severely biased. For instance, consider a newsperson who wishes to get the "opinions of the people" about a proposed seat tax to be imposed on tickets to all sporting events. The revenues from the seat tax will then be used to support the local symphony. The newsperson stands in front of a concert hall and surveys the first five people exiting after a symphony performance who will cooperate. This method of choosing a sample will produce some opinions, and perhaps some human interest stories, but it certainly has bias. It is hoped that the city council will not use these opinions as the sole basis for a decision about the proposed tax. It is good advice to be very cautious indeed when the data come from the method of convenience sampling.

SAMPLING TECHNIQUES

Random sampling: Use a simple random sample from the entire population.

Stratified sampling: Divide the entire population into distinct subgroups called strata. The strata are based on a specific characteristic such as age, income, education level, and so on. All members of a stratum share the specific characteristic. Draw random samples from each stratum.

Systematic sampling: Number all members of the population sequentially. Then, from a starting point selected at random, include every *k*th member of the population in the sample.

Cluster sampling: Divide the entire population into pre-existing segments or clusters. The clusters are often geographic. Make a random selection of clusters. Include every member of each selected cluster in the sample.

Multistage sampling: Use a variety of sampling methods to create successively smaller groups at each stage. The final sample consists of clusters.

Convenience sampling: Create a sample by using data from population members that are readily available.

CRITICAL THINKING

Sampling frame

We call the list of individuals from which a sample is actually selected the *sampling frame*. Ideally, the sampling frame is the entire population. However, from a practical perspective, not all members of a population may be accessible. For instance, using a telephone directory as the sample frame for residential telephone contacts would not include unlisted numbers.

Undercoverage

When the sample frame does not match the population, we have what is called *undercoverage.* In demographic studies, undercoverage could result if the homeless, fugitives from the law, and so forth are not included in the study.

A **sampling frame** is a list of individuals from which a sample is actually selected.

Undercoverage results from omitting population members from the sample frame.

In general, even when the sampling frame and the population match, a sample is not a perfect representation of a population. Therefore, information drawn from a sample may not exactly match corresponding information from the population. To the extent that sample information does not match the corresponding population information, we have an error, called a *sampling error.*

Sampling error

A **sampling error** is the difference between measurements from a sample and corresponding measurements from the respective population. It is caused by the fact that the sample does not perfectly represent the population.

Nonsampling error

A **nonsampling error** is the result of poor sample design, sloppy data collection, faulty measuring instruments, bias in questionnaires, and so on.

Sampling errors do not represent mistakes! They are simply the consequences of using samples instead of populations. However, be alert to nonsampling errors, which may sometimes occur inadvertently.

VIEWPOINT *Extraterrestrial Life?*

Do you believe intelligent life exists on other planets? Using methods of random sampling, a Fox News opinion poll found that about 54% of all U.S. men do believe in intelligent life on other planets, whereas only 47% of U.S. women believe there is such life. How could you conduct a random survey of students on your campus regarding belief in extraterrestrial life?

SECTION 1.2 PROBLEMS

1. **Statistical Literacy** Explain the difference between a stratified sample and a cluster sample.
2. **Statistical Literacy** Explain the difference between a simple random sample and a systematic sample.

3. **Statistical Literacy** Marcie conducted a study of the cost of breakfast cereal. She recorded the costs of several boxes of cereal. However, she neglected to take into account the number of servings in each box. Someone told her not to worry because she just had some sampling error. Comment on that advice.

4. **Statistical Literacy** A random sample of students who use the college recreation center were asked if they approved increasing student fees for all students in order to add a climbing wall to the recreation center. Describe the sample frame. Does the sample frame include all students enrolled in the college? Explain.

5. **Interpretation** In a random sample of 50 students from a large university, all the students were between 18 and 20 years old. Can we conclude that the entire population of students at the university is between 18 and 20 years old? Explain.

6. **Interpretation** A campus performance series features plays, music groups, dance troops, and stand-up comedy. The committee responsible for selecting the performance groups includes three students chosen at random from a pool of volunteers. This year, the 30 volunteers came from a variety of majors. However, the three students for the committee were all music majors. Does this fact indicate there was bias in the selection process and that the selection process was not random? Explain.

7. **Critical Thinking** Greg took a random sample of size 100 from the population of current season ticket holders to State College men's basketball games. Then he took a random sample of size 100 from the population of current season ticket holders to State College women's basketball games.
(a) What sampling technique (stratified, systematic, cluster, multistage, convenience, random) did Greg use to sample from the population of current season ticket holders to all State College basketball games played by either men or women?
(b) Is it appropriate to pool the samples and claim to have a random sample of size 200 from the population of current season ticket holders to all State College home basketball games played by either men or women? Explain.

8. **Critical Thinking** Consider the students in your statistics class as the population and suppose they are seated in four rows of 10 students each. To select a sample, you toss a coin. If it comes up heads, you use the 20 students sitting in the first two rows as your sample. If it comes up tails, you use the 20 students sitting in the last two rows as your sample.
(a) Does every student have an equal chance of being selected for the sample? Explain.
(b) Is it possible to include students sitting in row 3 with students sitting in row 2 in your sample? Is your sample a simple random sample? Explain.
(c) Describe a process you could use to get a simple random sample of size 20 from a class of size 40.

9. **Critical Thinking** Suppose you are assigned the number 1, and the other students in your statistics class call out consecutive numbers until each person in the class has his or her own number. Explain how you could get a random sample of four students from your statistics class.
(a) Explain why the first four students walking into the classroom would not necessarily form a random sample.
(b) Explain why four students coming in late would not necessarily form a random sample.
(c) Explain why four students sitting in the back row would not necessarily form a random sample.
(d) Explain why the four tallest students would not necessarily form a random sample.

10. **Critical Thinking** In each of the following situations, the sampling frame does not match the population, resulting in undercoverage. Give examples of population members that might have been omitted.
(a) The population consists of all 250 students in your large statistics class. You plan to obtain a simple random sample of 30 students by using the sampling frame of students present next Monday.
(b) The population consists of all 15-year-olds living in the attendance district of a local high school. You plan to obtain a simple random sample of 200 such residents by using the student roster of the high school as the sampling frame.

11. **Sampling: Random** Use a random-number table to generate a list of 10 random numbers between 1 and 99. Explain your work.

12. **Sampling: Random** Use a random-number table to generate a list of eight random numbers from 1 to 976. Explain your work.

13. **Sampling: Random** Use a random-number table to generate a list of six random numbers from 1 to 8615. Explain your work.

14. **Simulation: Coin Toss** Use a random-number table to simulate the outcomes of tossing a quarter 25 times. Assume that the quarter is balanced (i.e., fair).

15. **Computer Simulation: Roll of a Die** A die is a cube with dots on each face. The faces have 1, 2, 3, 4, 5, or 6 dots. The table below is a computer simulation (from the software package Minitab) of the results of rolling a fair die 20 times.

DATA DISPLAY

ROW	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	5	2	2	2	5	3	2	3	1	4
2	3	2	4	5	4	5	3	5	3	4

(a) Assume that each number in the table corresponds to the number of dots on the upward face of the die. Is it appropriate that the same number appears more than once? Why? What is the outcome of the fourth roll?
(b) If we simulate more rolls of the die, do you expect to get the same sequence of outcomes? Why or why not?

16. **Simulation: Birthday Problem** Suppose there are 30 people at a party. Do you think any two share the same birthday? Let's use the random-number table to simulate the birthdays of the 30 people at the party. Ignoring leap year, let's assume that the year has 365 days. Number the days, with 1 representing January 1, 2 representing January 2, and so forth, with 365 representing December 31. Draw a random sample of 30 days (with replacement). These days represent the birthdays of the people at the party. Were any two of the birthdays the same? Compare your results with those obtained by other students in the class. Would you expect the results to be the same or different?

17. **Education: Test Construction** Professor Gill is designing a multiple-choice test. There are to be 10 questions. Each question is to have five choices for answers. The choices are to be designated by the letters *a, b, c, d,* and *e*. Professor Gill wishes to use a random-number table to determine which letter choice should correspond to the correct answer for a question. Using the number correspondence 1 for *a,* 2 for *b,* 3 for *c,* 4 for *d,* and 5 for *e,* use a random-number table to determine the letter choice for the correct answer for each of the 10 questions.

18. **Education: Test Construction** Professor Gill uses true–false questions. She wishes to place 20 such questions on the next test. To decide whether to place a true statement or a false statement in each of the 20 questions, she uses a random-number table. She selects 20 digits from the table. An even digit tells her to use a true statement. An odd digit tells her to use a false statement. Use

a random-number table to pick a sequence of 20 digits, and describe the corresponding sequence of 20 true–false questions. What would the test key for your sequence look like?

19. **Sampling Methods: Benefits Package** An important part of employee compensation is a benefits package, which might include health insurance, life insurance, child care, vacation days, retirement plan, parental leave, bonuses, etc. Suppose you want to conduct a survey of benefits packages available in private businesses in Hawaii. You want a sample size of 100. Some sampling techniques are described below. Categorize each technique as *simple random sample, stratified sample, systematic sample, cluster sample,* or *convenience sample.*
(a) Assign each business in the Island Business Directory a number, and then use a random-number table to select the businesses to be included in the sample.
(b) Use postal ZIP Codes to divide the state into regions. Pick a random sample of 10 ZIP Code areas, and then include all the businesses in each selected ZIP Code area.
(c) Send a team of five research assistants to Bishop Street in downtown Honolulu. Let each assistant select a block or building and interview an employee from each business found. Each researcher can have the rest of the day off after getting responses from 20 different businesses.
(d) Use the Island Business Directory. Number all the businesses. Select a starting place at random, and then use every 50th business listed until you have 100 businesses.
(e) Group the businesses according to type: medical, shipping, retail, manufacturing, financial, construction, restaurant, hotel, tourism, other. Then select a random sample of 10 businesses from each business type.

20. **Sampling Methods: Health Care** Modern Managed Hospitals (MMH) is a national for-profit chain of hospitals. Management wants to survey patients discharged this past year to obtain patient satisfaction profiles. It wishes to use a sample of such patients. Several sampling techniques are described below. Categorize each technique as *simple random sample, stratified sample, systematic sample, cluster sample,* or *convenience sample.*
(a) Obtain a list of patients discharged from all MMH facilities. Divide the patients according to length of hospital stay (2 days or less, 3–7 days, 8–14 days, more than 14 days). Draw simple random samples from each group.
(b) Obtain lists of patients discharged from all MMH facilities. Number these patients, and then use a random-number table to obtain the sample.
(c) Randomly select some MMH facilities from each of five geographic regions, and then include all the patients on the discharge lists of the selected hospitals.
(d) At the beginning of the year, instruct each MMH facility to survey every 500th patient discharged.
(e) Instruct each MMH facility to survey 10 discharged patients this week and send in the results.

Ariel Skelley/Blend Images/Getty Images

SECTION 1.3 Introduction to Experimental Design

FOCUS POINTS

- **Discuss what it means to take a census.**
- **Describe simulations, observational studies, and experiments.**
- **Identify control groups, placebo effects, completely randomized experiments, and randomized block experiments.**
- **Discuss potential pitfalls that might make your data unreliable.**

Planning a Statistical Study

Planning a statistical study and gathering data are essential components of obtaining reliable information. Depending on the nature of the statistical study, a great deal of expertise and resources may be required during the planning stage. In this section, we look at some of the basics of planning a statistical study.

PROCEDURE

Basic guidelines for planning a statistical study

1. First, identify the individuals or objects of interest.
2. Specify the variables as well as the protocols for taking measurements or making observations.
3. Determine if you will use an entire population or a representative sample. If using a sample, decide on a viable sampling method.
4. In your data collection plan, address issues of ethics, subject confidentiality, and privacy. If you are collecting data at a business, store, college, or other institution, be sure to be courteous and to obtain permission as necessary.
5. Collect the data.
6. Use appropriate descriptive statistics methods (Chapters 2–4) and make decisions using appropriate inferential statistics methods (Chapters 8–11).
7. Finally, note any concerns you might have about your data collection methods and list any recommendations for future studies.

Census

One issue to consider is whether to use the entire population in a study or a representative sample. If we use data from the entire population, we have a *census*.

In a **census,** measurements or observations from the *entire* population are used.

When the population is small and easily accessible, a census is very useful because it gives complete information about the population. However, obtaining a census can be both expensive and difficult. Every 10 years, the U.S. Department of Commerce Census Bureau is required to conduct a census of the United States. However, contacting some members of the population—such as the homeless—is almost impossible. Sometimes members of the population will not respond. In such cases, statistical estimates for the missing responses are often supplied.

Overcounting, that is, counting the same person more than once, is also a problem the Census Bureau is addressing. In fact, in 2000, slightly more people were counted twice than the estimated number of people missed. For instance, a college student living on campus might be counted on a parent's census form as well as on his or her own census form.

Sample

If we use data from only part of the population of interest, we have a *sample.*

In a **sample,** measurements or observations from *part* of the population are used.

In the previous section, we examined several sampling strategies: simple random, stratified, cluster, systematic, multistage, and convenience. In this text, we will study methods of inferential statistics based on simple random samples.

Simulation

As discussed in Section 1.2, s*imulation* is a numerical facsimile of real-world phenomena. Sometimes simulation is called a "dry lab" approach, in the sense that it is a mathematical imitation of a real situation. Advantages of simulation are that numerical and statistical simulations can fit real-world problems extremely well. The researcher can also explore procedures through simulation that might be very dangerous in real life.

Experiments and Observation

When gathering data for a statistical study, we want to distinguish between observational studies and experiments.

Observational study

In an **observational study,** observations and measurements of individuals are conducted in a way that doesn't change the response or the variable being measured.

Experiment

In an **experiment,** a *treatment* is deliberately imposed on the individuals in order to observe a possible change in the response or variable being measured.

EXAMPLE 4 *Experiment*

In 1778, Captain James Cook landed in what we now call the Hawaiian Islands. He gave the islanders a present of several goats, and over the years these animals multiplied into wild herds totaling several thousand. They eat almost anything, including the famous silver sword plant, which was once unique to Hawaii. At one time, the silver sword grew abundantly on the island of Maui (in Haleakala, a national park on that island, the silver sword can still be found), but each year there seemed to be fewer and fewer plants. Biologists suspected that the goats were partially responsible for the decline in the number of plants and conducted a statistical study that verified their theory.

(a) To test the theory, park biologists set up stations in remote areas of Haleakala. At each station two plots of land similar in soil conditions, climate, and plant count were selected. One plot was fenced to keep out the goats, while the other was not. At regular intervals a plant count was made in each plot. This study involved an *experiment* because a *treatment* (the fence) was imposed on one plot.

(b) The experiment involved two plots at each station. The plot that was not fenced represented the *control* plot. This was the plot on which a treatment was specifically not imposed, although the plot was similar to the fenced plot in every other way.

Tim Davis/Science Source

Silver sword plant, Haleakala National Park

Placebo effect

Statistical experiments are commonly used to determine the effect of a treatment. However, the design of the experiment needs to *control* for other possible causes of the effect. For instance, in medical experiments, the *placebo effect* is the improvement or change that is the result of patients just believing in the treatment, whether or not the treatment itself is effective.