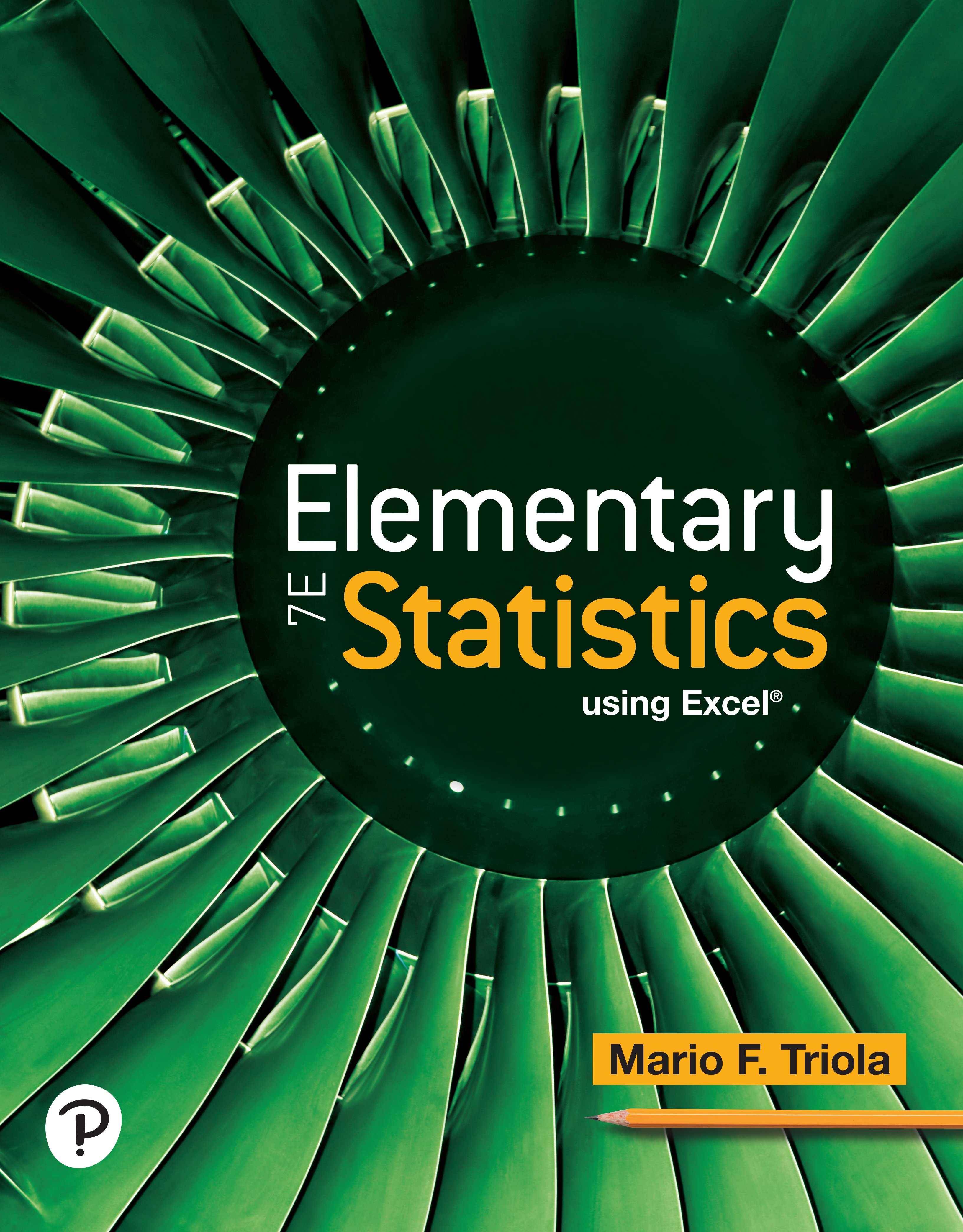


Elementary Statistics

7E

using Excel®

Mario F. Triola

Symbol Table

f	frequency with which a value occurs
Σ	capital sigma; summation
Σx	sum of the values
Σx^2	sum of the squares of the values
$(\Sigma x)^2$	square of the sum of all values
Σxy	sum of the products of each x value multiplied by the corresponding y value
n	number of values in a sample
N	number of values in a finite population; also used as the size of all samples combined
$n!$	n factorial
k	number of samples or populations or categories
$\bar{x}$	mean of the values in a sample
$\bar{R}$	mean of the sample ranges
μ	mu; mean of all values in a population
s	standard deviation of a set of sample values
σ	lowercase sigma; standard deviation of all values in a population
s^2	variance of a set of sample values
σ^2	variance of all values in a population
z	standard score
$z_{\alpha/2}$	critical value of z
t	t distribution
$t_{\alpha/2}$	critical value of t
df	number of degrees of freedom
F	F distribution
χ^2	chi-square distribution
χ^2_R	right-tailed critical value of chi-square
χ^2_L	left-tailed critical value of chi-square
p	probability of an event or the population proportion
q	probability or proportion equal to $1 - p$
$\hat{p}$	sample proportion
$\hat{q}$	sample proportion equal to $1 - \hat{p}$
$\bar{p}$	proportion obtained by pooling two samples
$\bar{q}$	proportion or probability equal to $1 - \bar{p}$
$P(A)$	probability of event A
$P(A \mid B)$	probability of event A, assuming event B has occurred
${}_nP_r$	number of permutations of n items selected r at a time
${}_nC_r$	number of combinations of n items selected r at a time
$\bar{A}$	complement of event A
H_0	null hypothesis
H_1	alternative hypothesis
α	alpha; probability of a type I error or the area of the critical region
β	beta; probability of a type II error
r	sample linear correlation coefficient
ρ	rho; population linear correlation coefficient
r^2	coefficient of determination
R^2	multiple coefficient of determination
r_s	Spearman's rank correlation coefficient
b_1	point estimate of the slope of the regression line
b_0	point estimate of the y-intercept of the regression line
$\hat{y}$	predicted value of y
d	difference between two matched values
$\bar{d}$	mean of the differences d found from matched sample data
s_d	standard deviation of the differences d found from matched sample data
s_e	standard error of estimate
T	rank sum; used in the Wilcoxon signed-ranks test

continued

Symbol Table

H	Kruskal-Wallis test statistic
R	sum of the ranks for a sample; used in the Wilcoxon rank-sum test
μ_R	expected mean rank; used in the Wilcoxon rank-sum test
σ_R	expected standard deviation of ranks; used in the Wilcoxon rank-sum test
G	number of runs in runs test for randomness
μ_G	expected mean number of runs; used in runs test for randomness
σ_G	expected standard deviation for the number of runs; used in runs test for randomness
$\mu_{\bar{x}}$	mean of the population of all possible sample means $\bar{x}$
$\sigma_{\bar{x}}$	standard deviation of the population of all possible sample means $\bar{x}$
E	margin of error of the estimate of a population parameter, or expected value
Q_1, Q_2, Q_3	quartiles
$D_1, D_2, \ldots, D_9$	deciles
$P_1, P_2, \ldots, P_{99}$	percentiles
x	data value

This page intentionally left blank

7th EDITION

ELEMENTARY STATISTICS USING EXCEL

This page intentionally left blank

7th EDITION

ELEMENTARY STATISTICS USING EXCEL

MARIO F. TRIOLA

Pearson

Content Development: Robert Carroll
Content Management: Suzanna Bainbridge, Amanda Brands
Content Production: Jean Choe, Peggy McMahon
Product Management: Karen Montgomery
Product Marketing: Alicia Wilson
Rights and Permissions: Tanvi Bhatia/Anjali Singh

Please contact https://support.pearson.com/getsupport/s/ with any queries on this content.
Cover Image by Bim/Getty Images

Library of Congress Cataloging-in-Publication Data

Names: Triola, Mario F., author.
Title: Elementary statistics using Excel / Mario F. Triola.
Description: 7th edition. | Hoboken, NJ : Pearson, [2022] | Includes bibliographical references and index. | Summary: "This book is designed to be a basic intro to Statistics with Excel included throughout as a key supplement"-- Provided by publisher.
Identifiers: LCCN 2020054082 | ISBN 9780136921721 (hardback)
Subjects: LCSH: Statistics. | Microsoft Excel (Computer file)
Classification: LCC QA276.45.M53 T75 2022 | DDC 519.50285/554--dc23
LC record available at https://lccn.loc.gov/2020054082

ScoutAutomatedPrintCode

Rental
ISBN-10: 0136921728
ISBN-13: 9780136921721

Pearson

To Ginny
Marc, Dushana, and Marisa
Scott, Anna, Siena, and Kaia

This page intentionally left blank

ABOUT THE AUTHOR

Mario F. Triola is a Professor Emeritus of Mathematics at Dutchess Community College, where he has taught statistics for over 30 years. Marty is the author of *Elementary Statistics,* 14th edition, *Essentials of Statistics,* 6th edition, *Elementary Statistics Using the TI-83/84 Plus Calculator,* 5th edition, and he is a co-author of *Biostatistics for the Biological and Health Sciences,* 2nd edition, *Statistical Reasoning for Everyday Life,* 5th edition, and *Business Statistics. Elementary Statistics* is currently available as an International Edition, and it has been translated into several foreign languages. Marty designed the original Statdisk statistical software, and he has written several manuals and workbooks for technology supporting statistics education. He has been a speaker at many conferences and colleges. Marty's consulting work includes the design of casino slot machines and fishing rods. He has worked with attorneys in determining probabilities in paternity lawsuits, analyzing data in medical malpractice lawsuits, identifying salary inequities based on gender, and analyzing disputed election results. He has also used statistical methods in analyzing medical school surveys, and in analyzing survey results for the New York City Transit Authority, and analyzing COVID-19 virus data for government officials. Marty has testified as an expert witness in the New York State Supreme Court. The Text and Academic Authors Association has awarded Marty a "Texty" for Excellence for his work on *Elementary Statistics*. As of this writing, Marty's *Elementary Statistics* book has been the #1 statistics book in the United States for the past 25 consecutive years.

Celebrating the past 25 years as the #1 statistics textbook author!

This page intentionally left blank

CONTENTS

PREFACE

The ancient Chinese philosopher Lao Tzu famously wrote: *A journey of a thousand miles must begin with a single step.* This textbook will lead you, step-by-step, on a journey through the important concepts of statistics and if you're reading this, you've already taken the first step! Thankfully, our journey will be much less physically taxing than a "journey of a thousand miles" and will only require use of your feet for determining skewness (see page 66).

We are now on the leading edge of a major revolution in technology, and the content of this text is key to that revolution. Artificial intelligence, machine learning, and deep learning are studied in data science, and the study of data science requires study of the discipline of statistics. Data science is now experiencing unprecedented growth. Projections indicate a 33% increased demand for statisticians in a few short years, and there is a projected shortage of workers with statistical skills. Also, as in past decades, statistics continues to be essential to a wide variety of disciplines, including medicine, polling, journalism, law, physical science, education, business, and economics. It is a gross understatement to suggest that it is now very wise to initiate a study of statistics.

Why Excel?

Microsoft Excel continues to be the dominant spreadsheet program used in business and industry. Motivated by a desire to better serve their students by better preparing them for their professional careers, many professors include Excel as the technology that is used throughout the statistics course. By using Excel, students acquire important professional skills along with concepts of statistics. This marriage of statistics concepts and spreadsheet applications is giving birth to a generation of students who can enter professional careers armed with knowledge and skills that were once desired, but are now required.

The marriage of statistics and Excel is not without its pitfalls. Excel's statistical capabilities are sometimes inadequate, so supplemental "add-ins" to Excel must be used. This book is designed to guide professors and students in choosing and using the best features of Excel and the appropriate add-in.

Using Excel

Elementary Statistics Using Excel, Seventh Edition, is designed to be an introduction to basic statistics. Instead of being a manual of computer instructions, this book places strong emphasis on understanding concepts of statistics, with Excel included throughout as the key supplement. Topics are presented with illustrative examples, identification of required assumptions, and underlying theory. Excel instructions are provided along with typical displays of results. In some cases, such as examples involving *formulas* and *graphs*, detailed instructions are presented so that Excel can be used effectively in all applications, instead of those relating only to statistics.

Features for Excel

This book has key features that are critical to an effective use of Excel as a statistical tool, including the following.

- **Introduction to Excel** Section 1-4 introduces important and basic features of using Excel.
- **Excel Instructions** Detailed Excel instructions are included throughout the book.

continued

- **Excel Displays** More than 300 displays of Excel functions and results are included throughout the book.
- **Excel Exercises** Exercises that are suitable for use with Excel are identified with a special icon.
- **Excel Answers** In some cases, exercise answers based on table values may be slightly different from answers based on the use of Excel. For those exercises, Appendix D generally gives both answers.
- **Excel Quick References** A concise list of Excel Quick References includes brief commands for executing most of the key statistical methods used in this book. The Excel Quick References are located at the end of the book.
- **Excel Projects** The wide variety of end-of-chapter features includes special Excel Projects.
- **XLSTAT Add-In** Excel lacks some important features, such as the ability to generate confidence intervals or to conduct hypothesis tests involving proportions. XLSTAT is a software add-in that enhances and expands the capability of Excel. XLSTAT can be installed to become part of the Excel program, so it is automatically available when Excel is launched. Instructions for installing XLSTAT are given in Section 1-4.
- **Analysis ToolPak Add-In** Section 1-4 includes instructions for installing Excel's Analysis ToolPak add-in. Instructions for using that add-in are included in appropriate sections throughout the book.
- **Downloadable Appendix B Data Sets** The data sets in Appendix B are available for download as Excel worksheets from www.TriolaStats.com. Some exercises use those data sets, so students can open Excel files instead of manually entering long lists of data.

  For those who choose to supplement Excel with other software, www.TriolaStats.com includes the Appendix B data sets in a variety of formats. In addition to Excel, the data sets are available in formats for Minitab and TI-83/84 Plus calculators.
- **Excel Instructional Videos** Instructional Excel videos and other technology resources are available at www.TriolaStats.com.
- **Statdisk** For those who choose to further supplement Excel, the separate program Statdisk has been designed specifically for the content of this book. Because Statdisk is easy to use and includes procedures for almost every major method in this book, it is extremely valuable as a tool for verifying results from Excel or obtaining results that cannot be obtained with Excel. Data can be easily transported between Excel and Statdisk by using Copy and Paste features. Statdisk is free to users of this book and can be accessed at www.Statdisk.com.

Goals of This Seventh Edition

- Foster personal growth of students through critical thinking, use of technology, collaborative work, and development of communication skills.
- Incorporate the latest and best methods used by professional statisticians.
- Include features that address all of the recommendations included in the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) as recommended by the American Statistical Association.
- Provide an abundance of new and interesting data sets, examples, and exercises, such as those involving biometric security, cybersecurity, drones, and Internet traffic.

- Present topics used in data science and many other applications, and include very large data sets that have become so important in our current culture.
- Enhance teaching and learning with the most extensive and best set of supplements and digital resources.

Audience / Prerequisites

Elementary Statistics Using Excel is written for students majoring in any subject. Algebra is used minimally. It is recommended that students have completed at least an elementary algebra course or that students should learn the relevant algebra components through an integrated or co-requisite course available through MyLab Statistics. In many cases, underlying theory is included, but this book does not require the mathematical rigor more appropriate for mathematics majors. Instead of being a "cookbook" devoid of any theory, this book includes the mathematics underlying important statistical methods, but the focus is on understanding and applying those methods along with interpreting results in a meaningful way.

Hallmark Features

Great care has been taken to ensure that each chapter of *Elementary Statistics Using Excel* will help students understand the concepts presented. The following features are designed to help meet that objective of conceptual understanding.

Real Data

Thousands of hours have been devoted to finding data that are real, meaningful, and interesting to students. 94% of the examples are based on real data, and 93% of the exercises are based on real data. Some exercises refer to the 46 data sets listed in Appendix B, and 20 of those data sets are new to this edition. Exercises requiring use of the Appendix B data sets are located toward the end of each exercise set and are marked with a special data set icon. These data sets are also available in MyLab Statistics, including data sets for StatCrunch.

Appendix B includes descriptions of the 46 data sets that can be downloaded from www.TriolaStats.com in formats for Excel™, Minitab™, JMP, SPSS, and TI-83/84™ Plus calculators. (Because TI-83/84 Plus calculators have limited memory, several larger data sets have been truncated for TI users, and answers have been annotated when appropriate.)

Readability

Great care, enthusiasm, and passion have been devoted to creating a book that is readable, understandable, interesting, and relevant. Students pursuing any major are sure to find applications related to their future work.

Website

This textbook is supported by www.pearsonhighered.com/triola and the author's website www.TriolaStats.com which are continually updated to provide the latest digital resources for the *Triola Statistics Series,* including:

- Statdisk: A free and robust browser-based statistical program designed specifically for this book. This is the only statistics textbook with dedicated and comprehensive statistics software.
- Downloadable Appendix B data sets in a variety of technology formats.
- Downloadable textbook supplements including Section 1-4 *Ethics in Statistics*, Section 6-6 *Normal as Approximation to Binomial*, *Glossary of Statistical Terms*, and *Formulas and Tables*.

- Interactive flow charts for key statistical procedures.
- Online instructional videos created specifically for the 7th Edition that provide step-by-step technology instructions.
- Contact link providing one-click access for instructors and students to contact the author, Marty Triola, with questions and comments.

Chapter Features

Chapter Opening Features

- Chapters begin with a ***Chapter Problem*** that uses real data and motivates the chapter material.
- ***Chapter Objectives*** provide a summary of key learning goals for each section in the chapter.

Exercises Many exercises require the *interpretation* of results. Great care has been taken to ensure their usefulness, relevance, and accuracy. Exercises are arranged in order of increasing difficulty and exercises are also divided into two groups: (1) *Basic Skills and Concepts* and (2) *Beyond the Basics*. *Beyond the Basics* exercises address more difficult concepts or require a stronger mathematical background. In a few cases, these exercises introduce a new concept.

End-of-Chapter Features

- ***Chapter Quick Quiz*** provides 10 review questions that require brief answers.
- ***Review Exercises*** offer practice on the chapter concepts and procedures.
- ***Cumulative Review Exercises*** reinforce earlier material.
- ***Excel Project*** provides an activity that can be used with Excel.
- ***Big (or Very Large) Data Projects*** encourage use of large data sets.
- ***From Data to Decision*** is a capstone problem that requires critical thinking and writing.
- ***Cooperative Group Activities*** encourage active learning in groups.

Other Features

Margin Essays There are 133 margin essays designed to highlight real-world topics and foster student interest. 36 of them are new to this edition. There are also many *Go Figure* items that briefly describe interesting numbers or statistics.

Flowcharts The text includes flowcharts that simplify and clarify more complex concepts and procedures. Animated versions of the text's flowcharts are available within MyLab Statistics.

Formulas and Tables This summary of key formulas, organized by chapter, gives students a quick reference for studying, or can be printed for use when taking tests (if allowed by the instructor). It also includes the most commonly used tables. This is available for download in MyLab Statistics, via pearson.com/math-stats-resources, or TriolaStats.com.

Excel Integration

As in the preceding edition, detailed Excel instructions and displays are provided throughout the book, and some exercises are based on displayed results from technology. The end-of-chapter features include an *Excel Project.*

In addition to Excel, the Statdisk statistical software package is designed specifically for this textbook and contains all Appendix B data sets. Statdisk is free to users of this book and it can be accessed at www.Statdisk.com.

Changes to This 7th Edition

New Features

New Content: This 7th edition includes an abundance of new exercises, new examples, and Chapter Problems, as summarized in the following table.

	Number	New to 7th Edition	Use Real Data
Exercises	1822	64% (1172)	93% (1703)
Examples	213	58% (124)	94% (201)
Chapter Problems	14	100% (14)	100% (14)

New Data Sets: This book includes a rich data set library in Appendix B so that professors and students have ready access to real and interesting data. Appendix B has been expanded from 32 data sets to 46 data sets. Twenty of those data sets are new, including *Internet Traffic*, *Queues*, *Car Data*, *Commute Times*, *Candies*, *Taxis*, and *Disney World Wait Times*.

Larger Data Sets: The largest data set in the previous edition had 600 cases. The data set library in this 7th edition includes data sets with 6068, 3982, 5755, 8959, and 1000 cases. In addition, there are *big* data sets with 465,506 cases and 31,784 cases. Working with such larger data sets is essential to students progressing into the age of big data and data science.

New Types of Exercises: To foster the development of critical thinking, the Cumulative Review Exercises near the end of Chapters 9, 10, and 11 consist of open-ended questions in which students are presented with a data set, and they are asked to pose a key question relevant to the data, identify a procedure for addressing that question, then analyze the data to form a conclusion.

New Margin Essays: This 7th edition of *Elementary Statistics Using Excel* includes 36 new margin essays.

Big (or Very Large) Data Projects: New to this edition, these projects are located near the end of each chapter and ask students to think critically while using large data sets.

New Chapter Problem Icon: Examples that relate to the Chapter Problem are now highlighted with this icon CP to show how different statistical concepts and procedures can be applied to the real-world issue highlighted in the chapter.

Organization Changes

New Technology: The previous edition of *Elementary Statistics Using Excel* introduced the resampling method of *bootstrapping* in Section 7-4. This 7th edition of *Elementary Statistics Using Excel* includes these methods of resampling using bootstrapping and randomization:

Bootstrap One Proportion
Bootstrap Two Proportions
Bootstrap One Mean
Bootstrap Two Means
Bootstrap Matched Pairs
--
Randomization One Proportion
Randomization Two Proportions
Randomization One Mean
Randomization Two Means
Randomization Matched Pairs
Randomization Correlation

New Section 4-5: *Simulations for Hypothesis Tests*

New Resampling Methods: Resampling methods are new to Sections 8-2, 8-3, 8-4, 8-5, 9-5, and 10-1.

New Section 8-5: *Resampling: Using Technology for Hypothesis Testing*

New Section 9-5: *Resampling: Using Technology for Inferences*

New Subsection 10-1, Part 3: *Randomization Test (for Correlation)*

New Chapter 15: *Holistic Statistics*

Removed Section: *The content of Section 6-6 (Normal as Approximation to Binomial) has been removed from the text and is now available for download (MyLab Statistics, pearson.com/math-stats-resources, or TriolaStats.com).*

Removed Section: *Ethics in Statistics has been moved from Chapter 15 to Section 1-4, and is available for download (MyLab Statistics, pearson.com/math-stats-resources, or TriolaStats.com).*

Technology Changes

New to Statdisk: The previous version of Statdisk for *Elementary Statistics Using Excel* included bootstrap resampling, but the new version of Statdisk for the 7th edition also includes all of the bootstrapping and randomization methods listed above under "*New Technology.*"

Statdisk Online: Statdisk is now a browser-based program that can be used on any device with a modern web browser, including laptops (Windows, macOS), Chromebooks, tablets and smartphones. Statdisk Online includes all of the statistical functions from earlier versions of Statdisk and is continually adding new functions and features.

Flexible Syllabus

This book's organization reflects the preferences of most statistics instructors, but there are two common variations:

- ***Early Coverage of Correlation and Regression:*** Some instructors prefer to cover the basics of correlation and regression early in the course. Section 2-4 includes basic concepts of scatterplots, correlation, and regression without the use of formulas and greater depth found in Sections 10-1 (*Correlation*) and 10-2 (*Regression*).
- ***Minimum Probability:*** Some instructors prefer extensive coverage of probability, while others prefer to include only basic concepts. Instructors preferring minimum coverage can include Section 4-1 while skipping the remaining sections of Chapter 4, as they are not essential for the chapters that follow. Many instructors prefer to cover the fundamentals of probability along with the basics of the addition rule and multiplication rule (Section 4-2).

GAISE This book reflects recommendations from the American Statistical Association and its *Guidelines for Assessment and Instruction in Statistics Education* (GAISE). Those guidelines suggest the following objectives and strategies.

1. ***Emphasize statistical literacy and develop statistical thinking:*** Each section exercise set begins with *Statistical Literacy and Critical Thinking* exercises. Many of the book's exercises are designed to encourage statistical thinking rather than the blind use of mechanical procedures.
2. ***Use real data:*** 94% of the examples and 93% of the exercises use real data.

3. ***Stress conceptual understanding rather than mere knowledge of procedures:*** Instead of seeking simple numerical answers, most exercises and examples involve conceptual understanding through questions that encourage practical interpretations of results. Also, each chapter includes a *From Data to Decision* project.
4. ***Foster active learning in the classroom:*** Each chapter ends with several *Cooperative Group Activities*.
5. ***Use technology for developing conceptual understanding and analyzing data:*** Computer software displays are included throughout the book. Special *Tech Center* subsections include instruction for using the software. Each chapter includes a *Technology Project*. When there are discrepancies between answers based on tables and answers based on technology, Appendix D provides *both* answers. The website www.TriolaStats.com includes free text-specific software (Statdisk), data sets formatted for several different technologies, and instructional videos for technologies. MyLab Statistics also includes support videos for different statistical software applications.
6. ***Use assessments to improve and evaluate student learning:*** Assessment tools include an abundance of section exercises, *Chapter Quick Quizzes*, *Chapter Review Exercises*, *Cumulative Review Exercises*, *Technology Projects*, *Big (or Very Large) Data Projects*, *From Data to Decision* projects, and *Cooperative Group Activities.*

Acknowledgments

I would like to thank the thousands of statistics professors and students who have contributed to the success of this book. I thank the reviewers for their suggestions for this seventh edition: Mary Kay Abbey, Vance Granville Community College; Kristin Cook, College of Western Idaho; Celia Cruz, Lehman College of CUNY; Don Davis, Lakeland Community College; Jean Ellefson, Alfred University; Matthew Harris, Ozarks Tech Community College; Stephen Krizan, Sait Polytechnic; Adam Littig, Los Angeles Valley College; Dr. Rick Silvey, University of Saint Mary – Leavenworth; Sasha Verkhovtseva, Anoka Ramsey Community College; William Wade, Seminole Community College. Special thanks to Laura Iossi of Broward College for her contributions to the Triola Statistics Series.

Other recent reviewers have included Raid W. Amin, University of West Florida; Robert Black, United States Air Force Academy; James Bryan, Merced College; Donald Burd, Monroe College; Keith Carroll, Benedictine University; Monte Cheney, Central Oregon Community College; Christopher Donnelly, Macomb Community College; Billy Edwards, University of Tennessee—Chattanooga; Marcos Enriquez, Moorpark College; Angela Everett, Chattanooga State Technical Community College; Joe Franko, Mount San Antonio College; Rob Fusco, Broward College; Sanford Geraci, Broward College; Eric Gorenstein, Bunker Hill Community College; Rhonda Hatcher, Texas Christian University; Laura Heath, Palm Beach State College; Richard Herbst, Montgomery County Community College; Richard Hertz; Diane Hollister, Reading Area Community College; Michael Huber, George Jahn, Palm Beach State College; Gary King, Ozarks Technical Community College; Kate Kozak, Coconino Community College; Dan Kumpf, Ventura College; Ladorian Latin, Franklin University; Mickey Levendusky, Pima County Community College; Mitch Levy, Broward College; Tristan Londre, Blue River Community College; Alma Lopez, South Plains College; Kim McHale, Heartland Community College; Carla Monticelli, Camden County Community College; Ken Mulzet, Florida State College at Jacksonville; Julia Norton, California State University Hayward; Michael Oriolo, Herkimer Community College; Jeanne Osborne, Middlesex Community College; Joseph Pick, Palm Beach

State College; Ali Saadat, University of California—Riverside; Radha Sankaran, Passaic County Community College; Steve Schwager, Cornell University; Pradipta Seal, Boston University; Kelly Smitch, Brevard College; Sandra Spain, Thomas Nelson Community College; Ellen G. Stutes, Louisiana State University, Eunice; Sharon Testone, Onondaga Community College; Chris Vertullo, Marist College; Dave Wallach, University of Findlay; Cheng Wang, Nova Southeastern University; Barbara Ward, Belmont University; Richard Weil, Brown College; Lisa Whitaker, Keiser University; Gail Wiltse, St. John River Community College; Claire Wladis, Borough of Manhattan Community College; Rick Woodmansee, Sacramento City College; Yong Zeng, University of Missouri at Kansas City; Jim Zimmer, Chattanooga State Technical Community College; Cathleen Zucco-Teveloff, Rowan University; Mark Z. Zuiker, Minnesota State University, Mankato.

This seventh edition of *Elementary Statistics Using Excel* is truly a team effort, and I consider myself fortunate to work with the dedication and commitment of the Pearson team. I thank Suzy Bainbridge, Amanda Brands, Deirdre Lynch, Peggy McMahon, Vicki Dreyfus, Jean Choe, Robert Carroll, Joe Vetere, and Rose Kernan of RPK Editorial Services.

I thank the following for their help in checking the accuracy of text and answers in this 7th edition: Paul Lorczak, and Dirk Tempelaar.

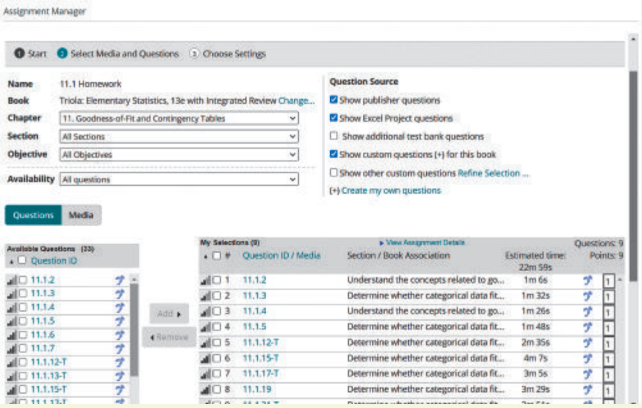
I extend special thanks to Marc Triola, M.D., New York University School of Medicine, for his outstanding work on creating the new Statdisk Online software. I thank Scott Triola for his very extensive help throughout the entire production process for this 7th edition.

M.F.T.
Madison, Connecticut
December 2020

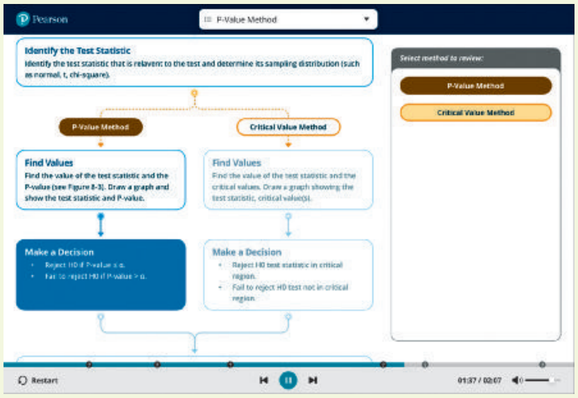
Resources for Success

MyLab Statistics is available to accompany Pearson's market-leading text options, including Elementary Statistics Using Excel, 7e by Mario F. Triola (access code required).

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



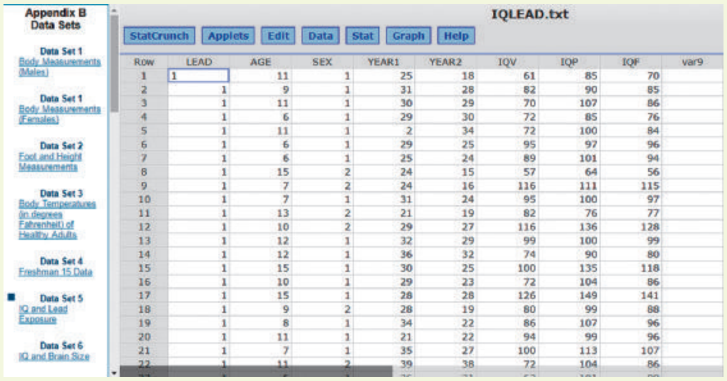
Expanded objective-based exercise coverage - Exercises in MyLab Statistics are designed to reinforce and support students' understanding of key statistical topics.



Enhanced video program to meet Introductory Statistics needs:

- **New! Animated Flow Charts -** Animated flow charts have been updated with a modern, interactive interface with assignable auto-graded assessment questions in MyLab Statistics.
- **New! Tech-Specific Video Tutorials -** These short, topical videos show how to use common statistical software to complete exercises.
- **Updated! Chapter Review Exercise Videos -** Watch the Chapter Review Exercises come to life with new review videos that help students understand key chapter concepts.

Real-World Data Examples - Help students understand how statistics applies to everyday life through the extensive current, real-world data examples and exercises provided throughout the text. MyLab Statistics allows students to easily launch data sets from their exercises to analyze real-world data.



pearson.com/mylab/statistics

Resources for Success

Supplements

Student Resources

Student's Solutions Manual, by James Lapp (Colorado Mesa University), provides detailed, worked-out solutions to all odd-numbered text exercises. Available for download in MyLab Statistics.

Student Workbook for the Triola Statistics Series, by Laura Iossi (Broward College) offers additional examples, concept exercises, and vocabulary exercises for each chapter. Available for download in MyLab Statistics. Can also be purchased separately.
ISBN: 0137363435 | 9780137363438

The following technology manuals include instructions, examples from the main text, and interpretations to complement those given in the text. They are all available for download in MyLab Statistics.

Excel Student Laboratory Manual and Workbook, (Download Only) by Laurel Chiappetta (University of Pittsburgh).

Instructor Resources

Annotated Instructor's Edition, by Mario F. Triola, contains answers to exercises in the margin, plus recommended assignments, and teaching suggestions.

ISBN-13: 9780136939702; ISBN-10: 0136939708

Instructor's Solutions Manual (Download Only), by James Lapp (Colorado Mesa University), contains solutions to all the exercises. These files are available to qualified instructors through Pearson Education's online catalog at www.pearsonhighered.com/irc or within MyLab Statistics.

Insider's Guide to Teaching with the Triola Statistics Series, (Download Only) by Mario F. Triola, contains sample syllabi and tips for incorporating projects, as well as lesson overviews, extra examples, minimum outcome objectives, and recommended assignments for each chapter.

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

PowerPoint® Lecture Slides: These files are available to instructors through Pearson's online catalog at www.pearsonhighered.com/irc or within MyLab Statistics.

Accessibility—Pearson works continuously to ensure our products are as accessible as possible to all students. Currently we work toward achieving WCAG 2.0 AA for our existing products (2.1 AA for future products) and Section 508 standards, as expressed in the Pearson Guidelines for Accessible Educational Web Media.

pearson.com/mylab/statistics

Get the Most Out of MyLab Statistics®

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

Integrated Review

Integrated Review in MyLab Statistics provides embedded and personalized review of prerequisite topics within relevant chapters.

- Students begin each chapter with a premade, assignable Skills Check assignment to check each student's understanding of prerequisite skills needed to be successful in that chapter.
- For any gaps in skill that are identified, a personalized review homework is populated. Students practice on the topics they need to focus on-no more, no less.
- A suite of resources are available to help students understand the objectives they missed on the Skills Check quiz, including worksheets and videos to help remediate.

Integrated Review in the MyLab is ideal for corequisite courses, where students are enrolled in a statistics course while receiving just-in-time remediation. But it can also be used simply to get underprepared students up to speed on prerequisite skills in order to be more successful in the Statistics content.

Used by nearly one million students a year, MyLab Statistics is the world's leading online program for teaching and learning statistics.

This page intentionally left blank

1

INTRODUCTION TO STATISTICS

CHAPTER PROBLEM

Is YouTube Becoming a More Important Learning Tool?

Surveys provide data that enable us to better understand the world in which we live and identify changes in the opinions, habits, and behaviors of others. Survey data guide public policy, influence business and educational practices, and affect many aspects of our daily lives. A recent Pearson survey, conducted by The Harris Poll, examined how technology has shaped students' learning habits and compared the responses from Gen-Z (ages 14–23) and millennials (ages 24–40). Among other topics, this survey asked respondents to identify their preferred learning tools, and YouTube was identified as one of the top tools by both Gen-Z and millennials. Figure 1-1 includes a graph that depicts the percentage of Gen-Z and millennials who identified YouTube as a preferred learning tool.

Critical Thinking Figure 1-1 on the next page makes it appear that Gen-Z is more than twice as likely to prefer YouTube as a learning tool compared to millennials. A quick glance might also give the impression that most millennials *do not* prefer YouTube as a learning tool. But wait! Look carefully at Figure 1-1 and see that the vertical axis has a scale that ranges from 52% to 60%. The graph in Figure 1-1 is *misleading* because it uses the scale of 52% to 60% instead of a scale that begins with 0%. As a result, the difference between the two bars is visually exaggerated in Figure 1-1. In Figure 1-2, the same data are shown in the graph, but we use a scale that begins with 0%. Figure 1-2 shows that the Gen-Z prefers YouTube as a learning tool only *slightly* more than millennials (actually 4% more to be exact). Figure 1-1 is misleading, whereas Figure 1-2 depicts the data fairly.

We might now consider how these survey data can be used to improve the learning experience for *Elementary Statistics*! Figure 1-2 shows that the majority of both Gen-Z and millennials prefer YouTube as a learning tool and this percentage has increased from one generation to the next. Knowing that YouTube and other videos are increasingly preferred learning tools, the author has created a YouTube channel with custom instructional videos to support this textbook (visit www.TriolaStats.com for the link). In addition, MyLab includes additional instructional videos and interactive content to support students.

The flaw shown in Figure 1-1 is among the most commonly used tactics to present misleading arguments, so it is especially important to recognize. Here are brief descriptions of common flaws:

Flaw 1: Misleading Graphs The bar chart in Figure 1-1 is very deceptive. By using a vertical scale that does not start at zero, the difference between the two percentages is grossly exaggerated. Deceptive graphs are discussed in more detail in Section 2-3.

Flaw 2: Bad Sampling Method Figure 1-1 and Figure 1-2 are based on data from the Pearson survey cited earlier. This study included 2587 respondents from a nationally representative sample, and the sampling method appears to be sound based on the description provided in the report. However, many other surveys obtain participants by using methods that are inappropriate and may lead to biased results, such as these:

- **Voluntary response sample:** Participants decide themselves whether to participate. *Example:* A survey question is posted on a website, and then Internet users decide whether to respond. With a voluntary response sample, it often happens that those with a strong interest in the topic are more likely to participate, so the results are very questionable.
- **Convenience sample:** Participants are selected because they are easy to reach and are readily available. *Example:* A student conducts a survey of fellow students relaxing in the cafeteria.

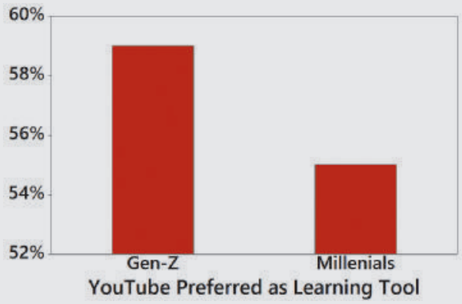


FIGURE 1-1 YouTube as a Preferred Learning Tool

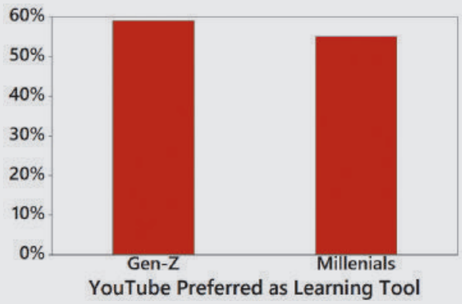


FIGURE 1-2 Same as Figure 1-1 but with scale beginning with 0%

When using sample data to learn something about a population, it is *extremely* important to obtain sample data that are representative of the population from which the data are drawn. As we proceed through this chapter and discuss types of data and sampling methods, we should focus on these key concepts:

- **Sample data must be collected in an appropriate way, such as through a process of *random* selection.**
- **If sample data are not collected in an appropriate way, the data may be so completely useless that no amount of statistical torturing can salvage them.**

It is all too easy to analyze sample data without thinking critically about how the data were collected. We could then develop conclusions that are fundamentally wrong and misleading. Instead, we should develop skills in statistical thinking and critical thinking so that we can distinguish between collections of sample data that are good and those that are seriously flawed.

CHAPTER OBJECTIVES

Here is the single most important concept presented in this chapter: When using methods of statistics with sample data to form conclusions about a population, it is absolutely essential to collect sample data in a way that is appropriate. Here are the chapter objectives:

1-1 Statistical and Critical Thinking

- Analyze sample data relative to context, source, and sampling method.
- Understand the difference between *statistical significance* and *practical significance*.
- Define and identify a *voluntary response sample* and know that statistical conclusions based on data from such a sample are generally not valid.

1-2 Types of Data

- Distinguish between a *parameter* and a *statistic.*
- Distinguish between *quantitative data* and *categorical* (or *qualitative* or *attribute*) *data*.
- Distinguish between *discrete* data and *continuous* data.
- Determine whether basic statistical calculations are appropriate for a particular data set.

1-3 Collecting Sample Data

- Define and identify a *simple random sample.*
- Understand the importance of sound sampling methods and the importance of good design of experiments.

1-4 Introduction to Excel

- Develop the ability to create an Excel worksheet, get help in Excel, undo an operation, and save an Excel file.
- Be able to open data sets from Appendix B.
- Learn the basics of using Excel for calculations, sorting data, and statistical functions.
- Be able to install Excel add-ins, including XLSTAT and the Analysis ToolPak.

1-5 Ethics in Statistics (available at www.TriolaStats.com)

- Analyze ethical issues in statistics, including those related to data collection, analysis, and reporting.

1-1 Statistical and Critical Thinking

Key Concept In this section we begin with a few very basic definitions, and then we consider an *overview* of the process involved in conducting a statistical study. This process consists of "prepare, analyze, and conclude." "Preparation" involves consideration of the *context*, the *source* of data, and *sampling method*. In future chapters we construct suitable graphs, explore the data, and execute computations required for the statistical method being used. In future chapters we also form conclusions by determining whether results have statistical significance and practical significance.

Importance of Accurate Census Results

The United States Constitution requires a census every ten years. Some factors affected by census results: Apportionment of congressional seats; distribution of billions of dollars of federal funds to states for transportation, schools, and hospitals; locations of sites for businesses and stores. Although accuracy of census results is extremely important, it is becoming more difficult to collect accurate census data due to the growing diversity of cultures and languages and increased distrust of the government. No amount of statistical analysis can salvage poor data, so it is critical that the census data is collected in an appropriate manner.

Statistical thinking involves critical thinking and the ability to make sense of results. Statistical thinking demands so much more than the ability to execute complicated calculations. Through numerous examples, exercises, and discussions, this text will help you develop the statistical thinking skills that are so important in today's world.

We begin with some very basic definitions.

DEFINITIONS

Data are collections of observations, such as measurements, genders, or survey responses. (A single data value is called a *datum*, a term rarely used. The term "data" is plural, so it is correct to say "data *are* . . ." not "data *is* . . .")

Statistics is the science of planning studies and experiments; obtaining data; and organizing, summarizing, presenting, analyzing, and interpreting those data and then drawing conclusions based on them.

A **population** is the complete collection of *all* measurements or data that are being considered.

A **census** is the collection of data from *every* member of the population.

A **sample** is a *subcollection* of members selected from a population.

Because populations are often very large, a common objective of the use of statistics is to obtain data from a sample and then use those data to form a conclusion about the population.

EXAMPLE 1 Watch What You Post Online

In a survey of 410 human resource professionals, 148 of them said that job candidates were disqualified because of information found on social media postings (based on data from *The Society for Human Resource Management*). In this case, the population and sample are as follows:

Population: All human resource professionals

Sample: The 410 human resource professionals who were surveyed

The objective is to use the sample as a basis for drawing a conclusion about the population of all human resource professionals, and methods of statistics are helpful in drawing such conclusions.

YOUR TURN. Do part (a) of Exercise 2 "Reported Versus Measured."

We now proceed to consider the process involved in a statistical study. See Figure 1-3 for a summary of this process and note that the focus is on critical thinking, not mathematical calculations. Thanks to wonderful developments in technology, we have powerful tools that effectively do the number crunching so that we can focus on understanding and interpreting results.

Prepare

Context Figure 1-3 suggests that we begin our preparation by considering the *context* of the data, so let's start with context by considering the data in Table 1-1. Table 1-1 includes shoe print lengths and heights of eight males. Forensic scientists measure shoe print lengths at burglary scenes and other crime scenes in order to estimate the height of the criminal. The format of Table 1-1 suggests the following goal: Determine whether there is a *relationship* between shoe print lengths

Go Figure

78%: The percentage of *female* veterinarian students who are women, according to *The Herald* in Glasgow, Scotland.

and heights of males. This goal suggests a reasonable hypothesis: Males with larger shoe print lengths tend to be taller. (We are using data for males only because 84% of burglaries are committed by males.)

TABLE 1-1 Shoe Print Lengths and Heights of Men

Shoe Print (cm)	27.6	29.7	29.7	31.0	31.3	31.4	31.8	34.5
Height (cm)	172.7	175.3	177.8	175.3	180.3	182.3	177.8	193.7

Source of the Data The second step in our preparation is to consider the source (as indicated in Figure 1-3). The data in Table 1-1 are from Data Set 9 "Foot and Height" in Appendix B, where the source is identified. The source certainly appears to be reputable.

Sampling Method Figure 1-3 suggests that we conclude our preparation by considering the sampling method. For the data in Table 1-1, individuals were randomly selected, so the sampling method appears to be sound.

Sampling methods and the use of random selection will be discussed in Section 1-3, but for now, we stress that a sound sampling method is absolutely essential for good results in a statistical study. It is generally a bad practice to use voluntary response (or self-selected) samples, even though their use is common.

Prepare

1. **Context**
   - What do the data represent?
   - What is the goal of study?
2. **Source of the Data**
   - Are the data from a source with a special interest so that there is pressure to obtain results that are favorable to the source?
3. **Sampling Method**
   - Were the data collected in a way that is unbiased, or were the data collected in a way that is biased (such as a procedure in which respondents volunteer to participate)?

↓

Analyze

1. **Graph the Data**
2. **Explore the Data**
   - Are there any outliers (numbers very far away from almost all of the other data)?
   - What important statistics summarize the data (such as the mean and standard deviation described in Chapter 3)?
   - How are the data distributed?
   - Are there missing data?
   - Did many selected subjects refuse to respond?
3. **Apply Statistical Methods**
   - Use technology to obtain results.

↓

Conclude

1. **Significance**
   - Do the results have statistical significance?
   - Do the results have practical significance?

FIGURE 1-3 Statistical and Critical Thinking

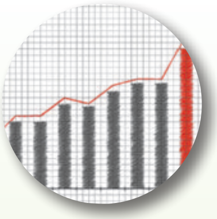
Survivorship Bias

In World War II, statistician Abraham Wald saved many lives with his work on the Applied Mathematics Panel. Military leaders asked the panel how they could improve the chances of aircraft bombers returning after missions. They wanted to add some armor for protection, and they recorded locations on the bombers where damaging holes were found. They reasoned that armor should be placed in locations with the most holes, but Wald said that strategy would be a big mistake. He said that armor should be placed where returning bombers were *not* damaged. His reasoning was this: The bombers that made it back with damage were *survivors*, so the damage they suffered could be survived. Locations on the aircraft that were not damaged were the most vulnerable, and aircraft suffering damage in those vulnerable areas were the ones that did not make it back. The military leaders would have made a big mistake with survivorship bias by studying the planes that survived instead of thinking about the planes that did not survive.

Go Figure

17%: The percentage of U.S. men between 20 and 40 years of age and taller than 7 feet who play basketball in the NBA.

Origin of "Statistics"



The word *statistics* is derived from the Latin word *status* (meaning "state"). Early uses of statistics involved compilations of data and graphs describing various aspects of a state or country. In 1662, John Graunt published statistical information about births and deaths. Graunt's work was followed by studies of mortality and disease rates, population sizes, incomes, and unemployment rates. Households, governments, and businesses rely heavily on statistical data for guidance. For example, unemployment rates, inflation rates, consumer indexes, and birth and death rates are carefully compiled on a regular basis, and the resulting data are used by business leaders to make decisions affecting future hiring, production levels, and expansion into new markets.

DEFINITION

A **voluntary response sample** (or **self-selected sample**) is one in which the respondents themselves decide whether to be included.

The following types of polls are common examples of voluntary response samples. By their very nature, all are seriously flawed because we should not make conclusions about a population on the basis of samples with a strong possibility of bias.

- Internet polls, in which people online decide whether to respond
- Mail-in polls, in which people decide whether to reply
- Telephone call-in polls, in which newspaper, radio, or television announcements ask that you voluntarily call a special number to register your opinion

See the following Example 2.

EXAMPLE 2 Voluntary Response Sample

The ABC television show *Nightline* asked viewers to call with their opinion about whether the United Nations headquarters should remain in the United States. Viewers then decided themselves whether to call with their opinions, and 67% of 186,000 respondents said that the United Nations should be moved out of the United States. In a separate and independent survey, 500 respondents were randomly selected and surveyed, and 38% of this group wanted the United Nations to move out of the United States. The two polls produced dramatically different results. Even though the *Nightline* poll involved 186,000 volunteer respondents, the much smaller poll of 500 randomly selected respondents is more likely to provide better results because of the far superior sampling method.

YOUR TURN. Do Exercise 1 "Computer Virus."

Analyze

Figure 1-3 indicates that after completing our preparation by considering the context, source, and sampling method, we begin to *analyze* the data.

Graph and Explore An analysis should begin with appropriate graphs and explorations of the data. Graphs are discussed in Chapter 2, and important statistics are discussed in Chapter 3.

Apply Statistical Methods Later chapters describe important statistical methods, but application of these methods is often made easy with technology (Excel, calculators, and/or statistical software packages). A good statistical analysis does not require strong computational skills. A good statistical analysis does require using common sense and paying careful attention to sound statistical methods.

Conclude

Figure 1-3 shows that the final step in our statistical process involves conclusions, and we should develop an ability to distinguish between *statistical significance* and *practical significance*.

Statistical Significance *Statistical significance* is achieved in a study when we get a result that is very unlikely to occur by chance. A common criterion has been this: We have statistical significance if the likelihood of an event occurring by chance is 5% or less.

- Getting 98 girls in 100 random births *is* statistically significant because such an extreme outcome is not likely to result from random chance.
- Getting 52 girls in 100 births *is not* statistically significant because that event could easily occur with random chance.

CAUTION An outcome can be statistically significant, and it may or may not be *important*. Don't associate statistical significance with importance.

Practical Significance It is possible that some treatment or finding is effective, but common sense might suggest that the treatment or finding does not make enough of a difference to justify its use or to be practical, as illustrated in Example 3.

EXAMPLE 3 Statistical Significance Versus Practical Significance

In a trial of weight loss programs, 21 subjects on the Atkins program lost an average (mean) of 2.1 kg (or 4.6 lb) after one year (based on data from "Comparison of the Atkins, Ornish, Weight Watchers, and Zone Diets for Weight Loss and Heart Disease Reduction," by Dansinger et al., *Journal of the American Medical Association*, Volume 93, Number 1). The results show that this loss is *statistically significant* and is not likely to occur by chance. However, many dieters believe that after following this diet for a year, a loss of only 2.1 kg is not worth the time, cost, and effort so that for these people, this diet does not have *practical significance*.

YOUR TURN. Do Exercise 13 "Diet and Exercise Program."

Example 3 includes a small sample of only 21 subjects, but with very large data sets (e.g., "big data"), statistically significant differences can often be found with very small differences. We should be careful to avoid the mistake of thinking that those small differences have practical significance.

Analyzing Data: Potential Pitfalls

Here are a few more items that could cause problems when analyzing data.

Misleading Conclusions When forming a conclusion based on a statistical analysis, we should make statements that are clear even to those who have no understanding of statistics and its terminology. We should carefully avoid making statements not justified by the statistical analysis. For example, later in this book we introduce the concept of a correlation, or association between two variables, such as shoe print lengths and heights of males. A statistical analysis might justify the statement that there is a correlation between shoe print length and height, but it would not justify a statement that an increase in the shoe print length *causes* an increase in height. Such a statement about causality can be justified by physical evidence, not by statistical analysis.

Correlation does not imply causation.

Sample Data *Reported* Instead of Measured When collecting data from people, it is better to take measurements yourself instead of asking subjects to *report* results. Ask people what they weigh and you are likely to get their *desired* weights, not their

Publication Bias

There is a "publication bias" in professional journals. It is the tendency to publish positive results (such as showing that some treatment is effective) much more often than negative results (such as showing that some treatment has no effect). In the article "Registering Clinical Trials" (*Journal of the American Medical Association*, Vol. 290, No. 4), authors Kay Dickersin and Drummond Rennie state that "the result of not knowing who has performed what (clinical trial) is loss and distortion of the evidence, waste and duplication of trials, inability of funding agencies to plan, and a chaotic system from which only certain sponsors might benefit, and is invariably against the interest of those who offered to participate in trials and of patients in general." They support a process in which *all* clinical trials are registered in one central system, so that future researchers have access to all previous studies, not just the studies that were published.

Statistician Jobs

In a recent year, *U.S. News and World Report* provided a list of the top 10 best jobs. Here are the first two jobs at the top of the list: (1) Software developer; (2) Statistician. It was noted that one reason for this high ranking is that the unemployment rate for statisticians is only 0.9 percent. That unemployment rate is lower than 1 person in 100. Not to mention how cool the contemporary discipline of statistics has become!

Go Figure

Five out of four people have some difficulty with statistics.

actual weights. People tend to round, usually down, sometimes *way* down. When asked, someone with a weight of 187 lb might respond that he or she weighs 160 lb. Accurate weights are collected by using a scale to *measure* weights, not by asking people what they weigh.

Loaded Questions If survey questions are not worded carefully, the results of a study can be misleading. Survey questions can be "loaded," or intentionally worded to elicit a desired response. Here are the actual rates of "yes" responses for the two different wordings of a question:

> 97% yes: "Should the President have the line item veto to eliminate waste?"
>
> 57% yes: "Should the President have the line item veto, or not?"

Order of Questions Sometimes survey questions are unintentionally loaded by such factors as the order of the items being considered. See the following two questions from a poll conducted in Germany, along with the very different response rates:

> "Would you say that traffic contributes more or less to air pollution than industry?" (45% blamed traffic; 27% blamed industry.)
>
> "Would you say that industry contributes more or less to air pollution than traffic?" (24% blamed traffic; 57% blamed industry.)

In addition to the order of items within a question, as illustrated above, the order of separate questions could also affect responses.

Nonresponse A *nonresponse* occurs when someone either refuses to respond to a survey question or is unavailable. When people are asked survey questions, some firmly refuse to answer. The refusal rate has been growing in recent years, partly because many persistent telemarketers try to sell goods or services by beginning with a sales pitch that initially sounds as though it is part of an opinion poll. (This "selling under the guise" of a poll is called *sugging*.) In *Lies, Damn Lies, and Statistics*, author Michael Wheeler makes this very important observation:

> **People who refuse to talk to pollsters are likely to be different from those who do not. Some may be fearful of strangers and others jealous of their privacy, but their refusal to talk demonstrates that their view of the world around them is markedly different from that of those people who will let poll-takers into their homes.**

Low Response Rates Related to the preceding item of nonresponses is the issue of low response rates. If a survey has a low response rate, the reliability of the results decreases. In addition to having a smaller sample size, there is an increased likelihood of having a bias among those who do respond. Some steps to help prevent a low response rate: (1) A survey should present an engaging argument for its importance; (2) a survey should not be very time consuming; (3) it is helpful to provide a reward for completing a survey, such as cash or a chance to win a prize. There are not definitive guidelines for acceptable response rates. A very good response rate is 80% or higher. Some suggest that response rates of at least 40% are acceptable. Pew Research Center reports that its typical telephone surveys have a response rate around 9%, but their surveys tend to be quite good. Sections 7-1, 7-2, and 7-3 include procedures for determining the sample size needed to estimate characteristics (proportion, mean, standard deviation) of a population, and those methods require sound sampling methods.

Percentages Some studies cite misleading or unclear percentages. Note that 100% of some quantity is *all* of it, but if there are references made to percentages that exceed 100%, such references are often not justified. If an advertiser claims that your utility costs can be reduced by 200%, that claim is misleading. Eliminating all utility costs would be a reduction of 100%, and a reduction of 200% doesn't make sense.

The following list identifies some key principles to apply when dealing with percentages. These principles all use the basic concept that % or "percent" really means "divided by 100." The first principle that follows is used often in this book.

Percentage of: To find a percentage of an amount, replace the % symbol with division by 100, and then interpret "of" to be multiplication. This example shows that 6% of 1200 is 72:

$$6\% \text{ of } 1200 \text{ responses} = \frac{6}{100} \times 1200 = 72$$

Decimal → Percentage: To convert from a decimal to a percentage, multiply by 100%. This example shows that 0.25 is equivalent to 25%:

$$0.25 \rightarrow 0.25 \times 100\% = 25\%$$

Fraction → Percentage: *To convert from a fraction to a percentage,* divide the denominator into the numerator to get an equivalent decimal number; then multiply by 100%. This example shows that the fraction $3/4$ is equivalent to 75%:

$$\frac{3}{4} = 0.75 \rightarrow 0.75 \times 100\% = 75\%$$

Percentage → Decimal: To convert from a percentage to a decimal number, replace the % symbol with division by 100. This example shows that 85% is equivalent to 0.85:

$$85\% = \frac{85}{100} = 0.85$$

1-1 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. Computer Virus In an AOL survey of Internet users, this question was posted online: "Have you ever been hit by a computer virus?" Among the 170,063 responses, 63% answered "yes." What term is used to describe this type of survey in which the people surveyed consist of those who chose to respond? What is wrong with this type of sampling method?

2. Reported Versus Measured In a survey of 1046 adults conducted by Bradley Corporation, subjects were asked how often they wash their hands when using a public restroom, and 70% of the respondents said "always."

a. Identify the sample and the population.

b. Why would better results be obtained by observing the hand washing instead of asking about it?

3. Statistical Significance Versus Practical Significance When testing a new treatment, what is the difference between statistical significance and practical significance? Can a treatment have statistical significance, but not practical significance?

4. Correlation One study showed that for a recent period of 10 years, there was a strong correlation (or association) between the per capita consumption of margarine and the divorce rate in Maine (based on data from National Vital Statistics reports and the U.S. Department of Agriculture). Does this imply that increasing margarine consumption is the cause of an increase in the divorce rate in Maine? Why or why not?

Consider the Source. ***In Exercises 5–8, determine whether the given source has the potential to create a bias in a statistical study.***

5. AAA The American Automobile Association (AAA) is a not-for-profit federation of motor clubs that provides automotive and travel services. AAA conducts a survey of its members about their use of public transportation versus private automobiles.

6. Body Data Data Set 1 "Body Data" in Appendix B includes pulse rates of subjects, and those pulse rates were recorded by examiners as part of a study conducted by the National Center for Health Statistics.

7. Brain Size A data set in Appendix B includes brain volumes from 10 pairs of monozygotic (identical) twins. The data were collected by researchers at Harvard University, Massachusetts General Hospital, Dartmouth College, and the University of California at Davis.

8. Chocolate An article in *Journal of Nutrition* (Vol. 130, No. 8) noted that chocolate is rich in flavonoids. The article notes "regular consumption of foods rich in flavonoids may reduce the risk of coronary heart disease." The study received funding from Mars, Inc., the candy company, and the Chocolate Manufacturers Association.

Sampling Method. ***In Exercises 9–12, determine whether the sampling method appears to be sound or is flawed.***

9. Nuclear Power Plants In a survey of 1368 subjects, the following question was posted on the *USA Today* website: "In your view, are nuclear plants safe?" The survey subjects were Internet users who chose to respond to the question posted on the electronic edition of *USA Today*.

10. Clinical Trials Researchers at Yale University conduct a wide variety of clinical trials by using subjects who volunteer after reading advertisements soliciting paid volunteers.

11. Sharing Passwords In a Password Boss survey of 2030 randomly selected adults, 39% said that they never share passwords with anyone.

12. Social Media Usage In a survey of social media usage, the Pew Research Center randomly selected 2002 adults in the United States.

Statistical Significance and Practical Significance. ***In Exercises 13–20, determine whether the results appear to have statistical significance, and also determine whether the results appear to have practical significance.***

13. Diet and Exercise Program In a study of the Ornish weight loss program, 40 subjects lost a mean of 3.3 lb after 12 months (based on data from "Comparison of the Atkins, Ornish, Weight Watchers, and Zone Diets for Weight Loss and Heart Disease Risk Reduction," by Dansinger et al., *Journal of the American Medical Association*, Vol. 293, No. 1). Methods of statistics can be used to show that if this diet had no effect, the likelihood of getting these results is roughly 3 chances in 1000.

14. Surgery versus Splints A study compared surgery and splinting for subjects suffering from carpal tunnel syndrome. It was found that among 73 patients treated with surgery, there was a 92% success rate. Among 83 patients treated with splints, there was a 72% success rate. Calculations using those results showed that if there really is no difference in success rates between surgery and splints, then there is about one chance in a thousand of getting success rates like the ones obtained in this study.

15. Mendel's Genetics Experiments One of Gregor Mendel's famous hybridization experiments with peas yielded 580 offspring with 152 of those peas (or 26%) having yellow pods. According to Mendel's theory, 25% of the offspring peas should have yellow pods.

16. IQ Scores Most people have IQ scores between 70 and 130. For $39.99, you can purchase a PC or Mac program from HighIQPro that is claimed to increase your IQ score by 10 to 20 points. The program claims to be "the only proven IQ increasing software in the brain training market," but the author of your text could find no substantial data supporting that claim, so let's suppose that these results were obtained: In a study of 12 subjects using the program, the average increase in IQ score is 3 IQ points. There is a 25% chance of getting such results if the program has no effect.

17. Election Fraud The County Clerk in Essex County, New Jersey, was responsible for randomly assigning the order in which candidates' names appeared on a recent election ballot. Among 41 different ballots, a Democrat was placed on the first line 40 times, and a Republican was placed on the first line once.

18. Football Overtime Games In "The Overtime Rule in the National Football League: Fair or Unfair?" by Gorgievski et al., *MathAMATYC Educator*, Vol. 2, No. 1, the authors report that among 414 football games won in overtime (prior to the overtime rule change in 2012), 235 were won by the team that won the coin toss at the beginning of overtime. If winning the coin toss does not provide an advantage, there is a 0.3% chance of getting such results.

19. Bias in Jury Selection In the case of *Casteneda v. Partida*, it was found that during a period of 11 years in Hidalgo County, Texas, 870 people were selected for grand jury duty, and 39% of them were Americans of Mexican ancestry. Among the people eligible for grand jury duty, 79.1% were Americans of Mexican ancestry.

20. Misleading Survey Responses In one presidential election, voting records showed that 61% of eligible voters actually did vote. In a survey of 1002 people, 70% *said* that they voted in that election (based on data from ICR Research Group). If survey respondents answered honestly and with accurate recall, there is about a 0.0000006% chance of getting such results.

In Exercises 21–24, refer to the sample of body temperatures (degrees Fahrenheit) in the table below. (The body temperatures are from Data Set 5 in Appendix B.)

	Subject				
	1	2	3	4	5
8 AM	97.0	98.5	97.6	97.7	98.7
12 AM	97.6	97.8	98.0	98.4	98.4

21. Context of the Data Refer to the table of body temperatures. Is there some meaningful way in which each body temperature recorded at 8 AM is matched with the 12 AM temperature?

22. Source The listed body temperatures were obtained from Dr. Steven Wasserman, Dr. Philip Mackowiak, and Dr. Myron Levine, who were researchers at the University of Maryland. Is the source of the data likely to be biased?

23. Conclusion Given the body temperatures in the table, what issue can be addressed by conducting a statistical analysis of the data?

24. Conclusion If we analyze the listed body temperatures with suitable methods of statistics, we conclude that when the differences are found between the 8 AM body temperatures and the 12 AM body temperatures, there is a 64% chance that the differences can be explained by random results obtained from populations that have the same 8 AM and 12 AM body temperatures. What should we conclude about the statistical significance of those differences?

In Exercises 25–28, refer to the data in the table below. The entries are for five different years, and they consist of weights (metric tons) of lemons imported from Mexico and U.S. car crash fatality rates per 100,000 population [based on data from "The Trouble with QSAR (or How I Learned to Stop Worrying and Embrace Fallacy)" by Stephen Johnson,* Journal of Chemical Information and Modeling, *Vol. 48, No. 1].

Lemon Imports	230	265	358	480	530
Crash Fatality Rate	15.9	15.7	15.4	15.3	14.9

25. Context Given that the data are matched and considering the units of the data, does it make sense to use the difference between each pair of values? Why or why not?

26. Analysis Given the context of the data in the table, what issue can be addressed by conducting a statistical analysis of the measurements?

27. Source of the Data Considering the source of the data, does that source appear to be biased in some way?

28. Conclusion If we were to use the sample data and conclude that there is a correlation or association between lemon imports and crash fatality rates, does it follow that lemon imports are the cause of fatal crashes?

What's Wrong? ***In Exercises 29–36, identify what is wrong.***

29. Potatoes In a poll sponsored by the Idaho Potato Commission, 1000 adults were asked to select their favorite vegetables, and the favorite choice was potatoes, which were selected by 26% of the respondents.

30. Healthy Water In a *USA Today* online poll, 951 Internet users chose to respond, and 57% of them said that they prefer drinking bottled water instead of tap water.

31. Motorcycles and Sour Cream In recent years, there has been a strong correlation between per capita consumption of sour cream and the numbers of motorcycle riders killed in noncollision accidents. Therefore, consumption of sour cream causes motorcycle fatalities.

32. Smokers The electronic cigarette maker V2 Cigs sponsored a poll showing that 55% of smokers surveyed say that they feel ostracized "sometimes," "often," or "always."

33. Cell Phones and Pirates In recent years, the numbers of cell phones and the numbers of pirates have both increased, so there is a correlation, or association, between those two variables. Therefore, pirates cause increases in cell phones.

34. Storks and Babies In the years following the end of World War II, it was found that there was a strong correlation, or association, between the number of human births and the stork population. It therefore follows that storks cause babies.

35. Future Optimism Survey The software package StatCrunch coordinated a "Future Optimism Survey" that included this question: Do you expect to be better off than your parents were over their lifetime? StatCrunch users could choose to complete the survey and obtain results. Among 1019 StatCrunch users, 587 said that they expected to be better off than their parents.

36. Diet Research Twelve nutritionists are each paid $100,000 to try a new celebrity diet on ten of their clients and then write a report summarizing the results. Based on the sample results, it is found that the diet is effective for 118 of the 120 people.

Percentages. ***In Exercises 37–44, answer the given questions, which are related to percentages.***

37. Workplace Attire In a survey conducted by Opinion Research Corporation, 1000 adults were asked to identify "what is inappropriate in the workplace." Of the 1000 subjects, 70% said that miniskirts were not appropriate in the workplace.

a. What is 70% of 1000?

b. Among the 1000 respondents, 550 said that shorts are unacceptable in the workplace. What percentage of respondents said that shorts are unacceptable in the workplace?

38. Checking Job Applicants In a study conducted by the Society for Human Resource Management, 347 human resource professionals were surveyed. Of those surveyed, 73% said that their companies conduct criminal background checks on all job applicants.

a. What is the exact value that is 73% of the 347 survey subjects?

b. Could the result from part (a) be the actual number of survey subjects who said that their companies conduct criminal background checks on all job applicants? Why or why not?

c. What is the actual number of survey subjects who said that their company conducts criminal background checks on all job applicants?

d. Assume that 112 of the survey subjects are females. What percentage of those surveyed are females?

39. Marriage Proposals In a survey conducted by TheKnot.com, 1165 engaged or married women were asked about the importance of a bended knee when making a marriage proposal. Among the 1165 respondents, 48% said that the bended knee was essential.

a. What is the exact value that is 48% of 1165 survey respondents?

b. Could the result from part (a) be the actual number of survey subjects who said that a bended knee is essential? Why or why not?

c. What is the actual number of survey respondents saying that the bended knee is essential?

d. Among the 1165 respondents, 93 said that a bended knee is corny and outdated. What percentage of respondents said that a bended knee is corny and outdated?

40. Texting While Driving *USA Today* reported results from an Arity survey in which 2018 drivers were asked if they text while driving.

a. Among the respondents, 42% said that they text while driving. What is the exact value that is 42% of the number of respondents?

b. Could the result from part (a) be the actual number of respondents who said that they text while driving? Why or why not?

c. What is the actual number of respondents who said that they text while driving?

d. What do the results suggest about highway safety?

41. Percentages in Advertising An ad for Big Skinny wallets included the statement that one of their wallets "reduces your filled wallet size by 50%–200%." What is wrong with this statement?

42. Percentages in Advertising Continental Airlines ran ads claiming that lost baggage is "an area where we've already improved 100% in the past six months." What is wrong with this statement?

43. Percentages in Advertising A *New York Times* editorial criticized a chart caption that described a dental rinse as one that "reduces plaque on teeth by over 300%." What is wrong with this statement?

44. Percentages in Advertising In an actual ad for the Club, a device used to discourage car thefts, it was stated that "The Club reduces your odds of car theft by 400%." What is wrong with this statement?

1-1 Beyond the Basics

45. Percentages in Negotiations When the author was negotiating a contract for the faculty and administration at a college, a dean presented the argument that if faculty receive a 4% raise and administrators receive a 4% raise, that's an 8% raise and it would never be approved. What's wrong with that argument?

46. What's Wrong with This Picture? The *Newport Chronicle* ran a survey by asking readers to call in their response to this question: "Do you support the development of atomic weapons that could kill millions of innocent people?" It was reported that 20 readers responded and that 87% said "no," while 13% said "yes." Identify four major flaws in this survey.

47. Falsifying Data A researcher at the Sloan-Kettering Cancer Research Center was once criticized for falsifying data. Among his data were figures obtained from 6 groups of mice, with 20 individual mice in each group. The following values were given for the percentage of successes in each group: 53%, 58%, 63%, 46%, 48%, 67%. What's wrong with those values?

1-2 Types of Data

Key Concept Because a major use of statistics is to collect and use sample data to make conclusions about populations, we should know and understand the meanings of the terms *statistic* and *parameter*, as defined below. In this section we describe a few different types of data. The type of data is one of the key factors that determine the statistical methods we use in our analysis.

In Part 1 of this section we describe the basics of different types of data, and then in Part 2 we consider "big data" and missing data.

PART 1 Basic Types of Data

Parameter / Statistic

DEFINITIONS

A **parameter** is a numerical measurement describing some characteristic of a *population*.

A **statistic** is a numerical measurement describing some characteristic of a *sample*.

HINT The alliteration in "population parameter" and "sample statistic" helps us remember the meanings of these terms.

If we have more than one statistic, we have "statistics." Another meaning of "statistics" was given in Section 1-1, where we defined *statistics* to be the science of planning studies and experiments; obtaining data; organizing, summarizing, presenting, analyzing, and interpreting those data; and then drawing conclusions based on them. We now have two different definitions of statistics, but we can determine which of these two definitions applies by considering the context in which the term *statistics* is used. The following example uses the first meaning of *statistics* as given on the previous page.

EXAMPLE 1 Parameter / Statistic

There are 250,342,875 adults in the United States. In a survey of 1659 randomly selected adults, 28% of them said that they own a credit card.

1. **Parameter:** The population size of 250,342,875 adults is a *parameter*, because it is the entire population of all adults in the United States. (If we somehow knew the percentage of all 250,342,875 adults who have a credit card, that percentage would also be a parameter.)
2. **Statistic:** The sample size of 1659 adults is a statistic, because it is based on a sample, not the entire population of all adults in the United States. The value of 28% is another statistic, because it is also based on the sample, not on the entire population.

YOUR TURN. Do Exercise 1 "Parameter and Statistic."

Quantitative / Categorical

Some data are numbers representing counts or measurements (such as heights of adults), whereas others are attributes (such as eye color of green or brown) that are not counts or measurements. The terms quantitative data and categorical data distinguish between these types.

DEFINITIONS

Quantitative (or **numerical) data** consist of *numbers* representing counts or measurements.

Categorical (or **qualitative** or **attribute) data** consist of names or labels (not numbers that represent counts or measurements).

CAUTION Categorical data are sometimes coded with numbers, with those numbers replacing names. Although such numbers might appear to be quantitative, they are actually categorical data. See the third part of Example 2 that follows.

Include Units of Measurement With quantitative data, it is important to use the appropriate units of measurement, such as dollars, hours, feet, or meters. We should carefully observe information given about the units of measurement, such as "all amounts are in *thousands of dollars*" or "all units are in *kilograms*." Ignoring such units of measurement can be very costly. The National Aeronautics and Space Administration (NASA) lost its $125 million Mars Climate Orbiter when the orbiter crashed

Units of Measurement

The television news show *60 Minutes* reported on the increased risk of cancer from using certain laminate flooring products made in China and sold by Lumber Liquidators. The Centers for Disease Control and Prevention (CDC) stated that the risk of cancer was actually three times higher than had previously been reported. The error occurred because in some of the first calculations, the CDC neglected to convert feet to meters. Lumber Liquidators discontinued sales of the problem laminate flooring.

Go Figure

7 billion: The world population that was exceeded in early 2012, which is 13 years after it passed 6 billion.

because the controlling software had acceleration data in *English* units, but they were incorrectly assumed to be in *metric* units.

Hopefully, the day will soon come when the United States adopts the metric system and joins almost all of the rest of the countries on planet Earth.

Validation Question

A question is sometimes used in a survey to confirm that a subject is attempting to seriously complete the survey questions instead of just mindlessly checking off answers. Here is an example:

This question is unlike the others. To confirm that you have read this question carefully, please select "Don't know" from the following list.

- *Definitely will*
- *Probably will*
- *Probably will not*
- *Definitely will not*
- *Don't know*

EXAMPLE 2 Quantitative / Categorical

1. **Quantitative Data:** The ages (in years) of subjects enrolled in a clinical trial
2. **Categorical Data as Labels:** The genders (male / female) of subjects enrolled in a clinical trial
3. **Categorical Data as Numbers:** The identification numbers 1, 2, 3, . . . , 25 are assigned randomly to the 25 subjects in a clinical trial. Those numbers are substitutes for names. They don't measure or count anything, so they are categorical data.

YOUR TURN. Do Exercise 2 "Quantitative/Categorical Data."

Discrete / Continuous

Quantitative data can be further described by distinguishing between *discrete* and *continuous* types.

DEFINITIONS

Discrete data result when the data values are quantitative and the number of values is finite, or "countable." (If there are infinitely many values, the collection of values is countable if it is possible to count them individually, such as the number of tosses of a coin before getting tails.)

Continuous (numerical) data result from infinitely many possible quantitative values, where the collection of values is not countable. (That is, it is impossible to count the individual items because at least some of them are on a continuous scale, such as the lengths of distances from 0 cm to 12 cm.)

CAUTION The concept of countable data plays a key role in the preceding definitions, but it is not a particularly easy concept to understand. Continuous data can be measured, but not counted. If you select a particular data value from continuous data, there is no "next" data value. See Example 3.

Continuous Data

Discrete Data

EXAMPLE 3 **Discrete / Continuous**

1. **Discrete Data of the Finite Type:** A statistics professor counts the number of students in attendance at each of her classes. The numbers are discrete because they are finite numbers resulting from a counting process.
2. **Discrete Data of the Infinite Type:** A statistics student plans to toss a fair coin until it turns up heads. It is theoretically possible to toss the coin forever without ever getting heads, but the number of tosses can be counted, even though the counting could go on forever. Because such numbers result from a counting procedure, the numbers are discrete.
3. **Continuous Data:** Burmese pythons are invading Florida. Researchers capture pythons and measure their lengths. So far, the largest python captured in Florida was 17 feet long. If the python lengths are between 0 feet and 17 feet, there are infinitely many values between 0 feet and 17 feet. Because it is impossible to count the number of different possible values on such a continuous scale, these lengths are continuous data.

YOUR TURN. Do Exercise 3 "Discrete/Continuous Data."

GRAMMAR: FEWER VERSUS LESS When describing smaller amounts, it is correct grammar to use "fewer" for discrete amounts and "less" for continuous amounts. It is correct to say that we drank *fewer* cans of cola and that, in the process, we drank *less* cola. The numbers of cans of cola are discrete data, whereas the volume amounts of cola are continuous data.

Levels of Measurement

Another common way of classifying data is to use four levels of measurement: nominal, ordinal, interval, and ratio, all defined below. (Also see Table 1-2 on page 19 for brief descriptions of the four levels of measurements.) When we are applying statistics to real problems, the level of measurement of the data helps us decide which procedure to use. There will be references to these levels of measurement in this book, but the important point here is based on common sense: *Don't do computations and don't use statistical methods that are not appropriate for the data.* For example, it would not make sense to compute an average (mean) of Social Security numbers, because those numbers are data that are used for identification, and they don't represent measurements or counts of anything.

DEFINITION

The **nominal level of measurement** is characterized by data that consist of names, labels, or categories only. The data cannot be arranged in some order (such as low to high).

EXAMPLE 4 **Nominal Level**

Here are examples of sample data at the nominal level of measurement.

1. **Yes / No / Undecided:** Survey responses of *yes*, *no*, and *undecided*
2. **Coded Survey Responses:** For an item on a survey, respondents are given a choice of possible answers, and they are coded as follows: "I agree" is coded as 1; "I disagree" is coded as 2; "I don't care" is coded as 3; "I refuse to answer" is coded as 4; "Go away and stop bothering me" is coded as 5. The numbers 1, 2, 3, 4, 5 don't measure or count anything.

YOUR TURN. Do Exercise 21 "College Students."

Measuring Disobedience

How are data collected about something that doesn't seem to be measurable, such as people's level of disobedience? Psychologist Stanley Milgram devised the following experiment: A researcher instructed a volunteer subject to operate a control board that gave increasingly painful "electrical shocks" to a third person. Actually, no real shocks were given, and the third person was an actor. The volunteer began with 15 volts and was instructed to increase the shocks by increments of 15 volts. The disobedience level was the point at which the subject refused to increase the voltage. Surprisingly, two-thirds of the subjects obeyed orders even when the actor screamed and faked a heart attack.

Big Data Study of Measles Vaccine and Autism

In 2019, there were measles outbreaks in U.S. geographic regions with large numbers of children who did not receive the MMR (measles, mumps, rubella) vaccine. Many parents opposed those vaccinations because they believed that they were associated with autism. Much of that belief was fueled by a 1998 "study" of 12 subjects showing an autism and MMR link that was reported in *Lancet*, but that article was later retracted. Based on a new ten-year study of 657,461 children, the *Annals of Internal Medicine* reported that the "MMR vaccination does not increase the risk for autism, does not trigger autism in susceptible children, and is not associated with clustering of autism cases after vaccination." An article in *The New York Times* reported about this study and emphasized the key point with this headline: "One More Time, With Big Data: Measles Vaccine Doesn't Cause Autism." In this case, the use of big data is being used to help overcome misunderstandings that result in unnecessary measles outbreaks.

Because nominal data lack any ordering or numerical significance, they should not be used for calculations. Numbers such as 1, 2, 3, and 4 are sometimes assigned to the different categories (especially when data are coded for computers), but these numbers have no real computational significance and any average (mean) calculated from them is meaningless and possibly misleading.

DEFINITION

Data are at the **ordinal level of measurement** if they can be arranged in some order, but differences (obtained by subtraction) between data values either cannot be determined or are meaningless.

EXAMPLE 5 Ordinal Level

Here is an example of sample data at the ordinal level of measurement.

Course Grades: A college professor assigns grades of A, B, C, D, or F. These grades can be arranged in order, but we can't determine differences between the grades. For example, we know that A is higher than B (so there is an ordering), but we cannot subtract B from A (so the difference cannot be found).

YOUR TURN. Do Exercise 23 "Movie Ratings."

Ordinal data provide information about relative comparisons but not the *magnitudes* of the differences. Ordinarily, ordinal data (such as course grades of A, B, C, D, F) should not be used for calculations such as the average (mean), but calculations are commonly used for some ordinal data, such as data from a survey question with a rating scale of 0 to 10. (A *Likert scale* is used to measure attitudes or opinions with a scale used for the level of agreement, usually with five to ten choices ranging from one extreme opinion to the opposite extreme.)

DEFINITION

Data are at the **interval level of measurement** if they can be arranged in order, and differences between data values can be found and are meaningful. *Data at the interval level do not have a natural zero starting point at which none of the quantity is present.*

EXAMPLE 6 Interval Level

These examples illustrate the interval level of measurement.

1. **Temperatures:** The lowest and highest temperatures recorded on earth are −129°F and 134°F. Those values are examples of data at the interval level of measurement. Those values are ordered, and we can determine that their difference is 263°F. However, there is no natural starting point. The value of 0°F is arbitrary and does not represent the total absence of heat (negative temperatures are common).
2. **Years:** The years 1492 and 1776 can be arranged in order, and the difference of 284 years can be found and is meaningful. However, time did not begin in the year 0, so the year 0 is arbitrary instead of being a natural zero starting point representing "no time." The years of 1492 and 1776 are therefore at the interval level of measurement.
3. **Shoe Sizes:** The shoe sizes of 10 and 5 can be arranged in order, and the difference is the same as the difference in shoe sizes of 8 and 13. However, size 0 is arbitrary.

YOUR TURN. Do Exercise 25 "Baseball."

DEFINITION

Data are at the **ratio level of measurement** if they can be arranged in order, differences can be found and are meaningful, and *there is a natural zero starting point* (where zero indicates that none of the quantity is present). For data at this level, differences and ratios are both meaningful.

EXAMPLE 7 Ratio Level

The following are examples of data at the ratio level of measurement. Note the presence of the natural zero value, and also note the use of meaningful ratios of "twice" and "three times."

1. **Heights of Students:** Heights of 180 cm and 90 cm for a high school student and a preschool student (0 cm represents no height, and 180 cm is *twice* as tall as 90 cm.)
2. **Class Times:** The times of 50 min and 100 min for a statistics class (0 min represents no class time, and 100 min is *twice* as long as 50 min.)

YOUR TURN. Do Exercise 27 "Areas of States."

TABLE 1-2 Levels of Measurement

Level of Measurement	Brief Description	Example
Ratio	There is a natural zero starting point and ratios make sense.	Heights, lengths, distances, volumes
Interval	Differences are meaningful, but there is no natural zero starting point and ratios are meaningless.	Body temperatures in degrees Fahrenheit or Celsius
Ordinal	Data can be arranged in order, but differences either can't be found or are meaningless.	Ranks of colleges in *U.S. News & World Report*
Nominal	Categories only. Data cannot be arranged in order.	Eye colors

HINT The distinction between the interval and ratio levels of measurement can be a bit tricky. Here are two tools to help with that distinction:

1. **Ratio Test** Focus on the term "ratio" and know that the term "twice" describes the ratio of one value to be double the other value. To distinguish between the interval and ratio levels of measurement, use a "ratio test" by asking this question: Does use of the term "twice" make sense? "Twice" makes sense for data at the ratio level of measurement, but it does not make sense for data at the interval level of measurement.
2. **True Zero** For ratios to make sense, there must be a value of "true zero," where the value of zero indicates that none of the quantity is present, and zero is not simply an arbitrary value on a scale. The temperature of 0°F is arbitrary and does not indicate that there is no heat, so temperatures on the Fahrenheit scale are at the interval level of measurement, not the ratio level.

Six Degrees of Separation

Social psychologists, historians, political scientists, and communications specialists are interested in "The Small World Problem": Given any two people in the world, how many intermediate links are necessary to connect the two original people? In the 1950s and 1960s, social psychologist Stanley Milgram conducted an experiment in which subjects tried to contact other target people by mailing an information folder to an acquaintance who they thought would be closer to the target. Among 160 such chains that were initiated, only 44 were completed, so the failure rate was 73%. Among the successes, the number of intermediate acquaintances varied from 2 to 10, with a median of 6 (hence "six degrees of separation"). The experiment has been criticized for its high failure rate and its disproportionate inclusion of subjects with above-average incomes. A more recent study conducted by Microsoft researcher Eric Horvitz and Stanford Assistant Professor Jure Leskovec involved 30 billion instant messages and 240 million people. This study found that for instant messages that used Microsoft, the mean length of a path between two individuals is 6.6, suggesting "seven degrees of separation." Work continues in this important and interesting field.

Big Data Instead of a Clinical Trial

Nicholas Tatonetti of Columbia University searched Food and Drug Administration databases for adverse reactions in patients that resulted from different pairings of drugs. He discovered that the Paxil (paroxetine) drug for depression and the pravastatin drug for high cholesterol interacted to create increases in glucose (blood sugar) levels. When taken separately by patients, neither drug raised glucose levels, but the increase in glucose levels occurred when the two drugs were taken together. This finding resulted from a general database search of interactions from many pairings of drugs, not from a clinical trial involving patients using Paxil and pravastatin.

EXAMPLE 8 Distinguishing Between the Ratio Level and Interval Level

For each of the following, determine whether the data are at the ratio level of measurement or the interval level of measurement:

a. Times (minutes) it takes students to complete a statistics test.

b. Body temperatures (Celsius) of statistics students.

SOLUTION

a. Apply the "ratio test" described in the preceding hint. If one student completes the test in 40 minutes and another student completes the test in 20 minutes, does it make sense to say that the first student used *twice* as much time? Yes! So the times are at the ratio level of measurement. We could also apply the "true zero" test. A time of 0 minutes does represent "no time," so the value of 0 is a true zero indicating that no time was used.

b. Apply the "ratio test" described in the preceding hint. If one student has a body temperature of 40°C and another student has a body temperature of 20°C, does it make sense to say that the first student is *twice* as hot as the second student? (Ignore subjective amounts of attractiveness and consider only science.) No! So the body temperatures are not at the ratio level of measurement. Because the difference between 40°C and 20°C is the same as the difference between 90°C and 70°C, the differences are meaningful, but because ratios do not make sense, the body temperatures are at the interval level of measurement. Also, the temperature of 0°C does not represent "no heat" so the value of 0 is not a true zero indicating that no heat is present.

YOUR TURN. Do Exercise 28 "Body Temperatures."

PART 2 Big Data and Missing Data: Too Much and Not Enough

When working with data, we might encounter some data sets that are ginormous, and we might also encounter some data sets with individual elements missing. Here in Part 2 we briefly discuss both cases.

Big Data

UPS delivers 20 million packages every day. UPS analyzes massive amounts of data in order to optimize routes and plan maintenance for its truck and aircraft fleets. Data analysis and optimization efforts to-date have enabled UPS to save 40 million gallons of fuel and shorten travel distances by 370 million miles. The need to analyze large data sets has led to the birth of *data science*. There is not universal agreement on the following definitions, and various other definitions can be easily found elsewhere.

DEFINITIONS

Big data refers to data sets so large and so complex that their analysis is beyond the capabilities of traditional software tools. Analysis of big data may require software simultaneously running in parallel on many different computers.

Data science involves applications of statistics, computer science, and software engineering, along with some other relevant fields (such as sociology or finance).

Examples of Data Set Magnitudes We can see from the definition of big data that there isn't a fixed number that serves as an exact boundary for determining whether a data set qualifies as being big data, but big data typically involves amounts of data such as the following:

- Terabytes (10^{12} or 1,000,000,000,000 bytes) of data
- Petabytes (10^{15} bytes) of data
- Exabytes (10^{18} bytes) of data
- Zettabytes (10^{21} bytes) of data
- Yottabytes (10^{24} bytes) of data

Examples of Applications of Big Data The following are a few other examples involving big data:

- Google provides live traffic maps by recording and analyzing GPS (global positioning system) data collected from the smartphones of people traveling in their vehicles.
- Netflix collects data on viewing records and uses the data to create original programming as well as identifying which movies to acquire.
- Attempts to forecast flu epidemics are made by analyzing Internet searches of flu symptoms.
- The Sloan Digital Sky Survey started in the year 2000, and it quickly collected more astronomy data than in the history of mankind up to 2000. It now has more than 140 terabytes of astronomy data.
- Walmart processes 2.5 petabytes (2,500,000,000,000,000 bytes) of data every hour. For online sales, Walmart developed the Polaris search engine that increased sales by 10% to 15%, worth billions of dollars.
- Amazon monitors and tracks about 6 million items shipped daily from its stores that are distributed across hundreds of fulfillment centers around the world.
- Uber feeds driver and customer data into algorithms that identify the most profitable driver/passenger matches.

Examples of Jobs According to Analytic Talent, there are 6000 companies hiring data scientists, and here are some job posting examples:

- Facebook: Data Scientist
- IBM: Data Scientist
- PayPal: Data Scientist
- The College Board: SAS Programmer/Data Scientist
- Netflix: Senior Data Engineer/Scientist

It was noted in the Preface that we are experiencing a new major revolution in technology that uses artificial intelligence, machine learning, and deep learning—topics studied in Data Science, which requires a study of statistics. Data Science and statistics are now experiencing unprecedented growth.

Statistics in Data Science The modern data scientist has a solid background in statistics and computer systems as well as expertise in fields that extend beyond statistics. The modern data scientist might be skilled with software of *R*, Python or Hadoop. The modern data scientist might also have a strong background in some other field

Statistics for Online Dating

The four founders of the online dating site OkCupid are mathematicians who use methods of statistics to analyze results from their website. The chief executive officer of OkCupid has been quoted as saying, "We're not psychologists. We're math guys" (from "Looking for a Date? A Site Suggests You Check the Data," by Jenna Wortham, *New York Times*). The OkCupid website is unique in its use of methods of statistics to match people more effectively.

By analyzing the photos and responses of 7000 users, analysts at OkCupid found that when creating a profile photo, men should not look directly at the camera, and they should not smile. For women, the appearance of being interesting produces much better results than the appearance of being sexy. They found that brevity is good for the first posted message; the ideal length of the first posted message is 40 words—about what a typical person can type in 1 minute.

such as psychology, biology, medicine, chemistry, or economics. Because of the wide range of disciplines required, a data science project might typically involve a team of collaborating individuals with expertise in different fields. An introductory statistics course is a great first step in becoming a data scientist.

Hawthorne and Experimenter Effects

The well-known placebo effect occurs when an untreated subject incorrectly believes that he or she is receiving a real treatment and reports an improvement in symptoms. The Hawthorne effect occurs when treated subjects somehow respond differently simply because they are part of an experiment. (This phenomenon was called the "Hawthorne effect" because it was first observed in a study of factory workers at Western Electric's Hawthorne plant.) An experimenter effect (sometimes called a Rosenthal effect) occurs when the researcher or experimenter unintentionally influences subjects through such factors as facial expression, tone of voice, or attitude.

Missing Data

When collecting sample data, it is quite common to find that some values are missing. Ignoring missing data can sometimes create misleading results. If you make the mistake of skipping over a few different sample values when you are manually typing them into a statistics software program, the missing values are not likely to have a serious effect on the results. However, if a survey includes many missing salary entries because those with very low incomes are reluctant to reveal their salaries, those missing low values will have the serious effect of making salaries appear higher than they really are.

For an example of missing data, see the following table. The body temperature for Subject 2 at 12 AM on Day 2 is missing. (The table below includes the first three rows of data from Data Set 5 "Body Temperatures" in Appendix B.)

Body Temperatures (in degrees Fahrenheit) of Healthy Adults

			Temperature Day 1		Temperature Day 2	
Subject	Sex	Smoke	8 AM	12 AM	8 AM	12 AM
1	M	Y	98.0	98.0	98.0	98.6
2	M	Y	97.0	97.6	97.4	----
3	M	Y	98.6	98.8	97.8	98.6

There are different categories of missing data, as described in the following definitions.

DEFINITION

A data value is **missing completely at random** if the likelihood of its being missing is independent of its value or any of the other values in the data set. That is, any data value is just as likely to be missing as any other data value.

(*NOTE:* More complete discussions of missing data will distinguish between *missing completely at random* and *missing at random*, which means that the likelihood of a value being missing is independent of its value after controlling for another variable. There is no need to know this distinction in this book.)

Example of Missing Data—Random: When using a keyboard to manually enter ages of survey respondents, the operator is distracted by a colleague singing "Daydream Believer" and makes the mistake of failing to enter the age of 37 years. This data value is missing completely at random.

DEFINITION

A data value is **missing not at random** if the missing value is related to the reason that it is missing.

Example of Missing Data—Not at Random A survey question asks each respondent to enter his or her annual income, but respondents with very low incomes skip this question because they find it embarrassing.

Biased Results? Based on the two definitions and examples from the previous page, it makes sense to conclude that if we ignore data *missing completely at random*, the remaining values are not likely to be biased and good results should be obtained. However, if we ignore data that are *missing not at random*, it is very possible that the remaining values are biased and results will be misleading.

Correcting for Missing Data There are different methods for dealing with missing data.

1. **Delete Cases:** One very common method for dealing with missing data is to delete all subjects having any missing values.
   - If the data are missing completely at random, the remaining values are not likely to be biased and good results can be obtained, but with a smaller sample size.
   - If the data are missing not at random, deleting subjects having any missing values can easily result in a bias among the remaining values, so results can be misleading.
2. **Impute Missing Values:** We "impute" missing data values when we substitute values for them. There are different methods of determining the replacement values, such as using the mean of the other values, or using a randomly selected value from other similar cases, or using a method based on regression analysis (which will make more sense after studying Chapter 10).

In this book we do not work much with missing data, but it is important to understand this:

> **When analyzing sample data with missing values, try to determine *why* they are missing, then decide whether it makes sense to treat the remaining values as being representative of the population. If it appears that there are missing values that are *missing not at random* (that is, their values are related to the reasons why they are missing), know that the remaining data may well be biased and any conclusions based on those remaining values may well be misleading.**

Declining Response Rate

The Pew Research Center is now using the Internet for most of its surveys conducted in the United States. One major factor precipitating that change is the low and declining response rate of telephone surveys. The response rate for telephone surveys was 36% in 1997, but it has now dropped to only 6%. A major cause of this declining response rate is the high and growing use of robocalls. Public opinion surveys conducted by telephone usually appear as an unknown source, so potential respondents are much more likely to reject such calls. However, Pew research has shown that low response rates do not cause inaccurate results. But low response rates for telephone surveys do result in higher survey costs.

1-2 Basic Skills and Concepts

Statistical Literacy and Critical Thinking

1. Parameter and Statistic In a Citrix Security survey of 1001 adults in the United States, it was found that 69% of those surveyed believe that having their personal information stolen is inevitable. Identify the population and sample. Is the value of 69% a statistic or a parameter?

2. Quantitative / Categorical Data Identify each of the following as quantitative data or categorical data.

a. The platelet counts in Data Set 1 "Body Data" in Appendix B

b. The cigarette brands in Data Set 16 "Cigarette Contents" in Appendix B

c. The colors of the M&M candies in Data Set 38 "Candies" in Appendix B

d. The weights of the M&M candies in Data Set 38 "Candies" in Appendix B

3. Discrete / Continuous Data Which of the following describe discrete data?

a. The exact heights of all NBA basketball players

b. The numbers of people surveyed in each of the Gallup polls preceding the next Presidential election

c. The exact times that randomly selected students spend on smart phones during the preceding week

4. E-Cigarette Survey In a survey of 36,000 adults, 3.7% said that they regularly use E-cigarettes (based on data from the National Center for Health Statistics).

a. Identify the sample and population.

b. Is the value of 3.7% a statistic or parameter?

c. What is the level of measurement of the value of 3.7%? (nominal, ordinal, interval, ratio)

d. Are the numbers of subjects in such surveys discrete or continuous?

In Exercises 5–12, identify whether the given value is a statistic or a parameter.

5. Lost Wallets A survey of a sample of consumers in the United States showed that among those who have found a wallet or purse, 89% either turned it in or located the owner (based on a Toluna Quicksurveys poll).

6. Drivers According to the Federal Highway Administration, there are 212 million licensed drivers in the United States.

7. Titanic Deaths The sinking of the Titanic on April 15, 1912, is one of the most infamous disasters in history. A population of 1503 passengers and crew died when the Titanic sank approximately 400 miles south of Newfoundland, Canada.

8. Birth Weight In a study of a sample of babies born at hospitals in New York State, it was found that the average (mean) weight at birth was 3152.0 grams.

9. Birth Genders In the same study cited in the preceding exercise, 51% of the babies were girls.

10. Smart Phones In a Pew Research Center poll, a sample of adults in the United States was obtained, and it was found that 72% of them own smart phones.

11. Super Bowl A study was conducted of all 70,081 people who attended Super Bowl LIII at Mercedes-Benz Stadium in Atlanta, Georgia.

12. Prisoners According to the U.S. Bureau of Justice, there are 2,227,318 prisoners in the United States.

In Exercises 13–20, determine whether the data are from a discrete or continuous data set.

13. Freshman 15 In a study of weight gains by college students in their freshman year, researchers record the amounts of weight gained by randomly selected students (as in Data Set 13 “Freshman 15” in Appendix B).

14. Fraud Detection While monitoring Internet traffic in order to detect fraudulent activity, a researcher records the interarrival times (sec) between incoming Internet queries.

15. House Attendance The Clerk of the U.S. House of Representatives records the number of representatives present at each session.

16. Students Your statistics professor records the number of students who pass each of her courses.

17. Amazon Sales In a study of service times at an Amazon fulfillment center, the times (minutes) it takes to process orders are recorded.

18. Texting Fatalities The Insurance Institute for Highway Safety collects data consisting of the numbers of motor vehicle fatalities caused by driving while texting.

19. Statistics Classes In each of her classes, a statistics professor records the number of students who earned a grade of A.

20. Criminal Forensics When studying the relationship between lengths of feet and heights so that footprint evidence at a crime scene can be used to estimate the height of the suspect, a researcher records the exact lengths of feet from a large sample of random subjects.

In Exercises 21–28, determine which of the four levels of measurement (nominal, ordinal, interval, ratio) best describes the given data.

21. College Students In order to better plan for the incoming freshman class, a college dean asks each newly admitted student to identify their likely major (physics, business, math, psychology, engineering, law, etc.).

22. Medical School Rankings *U.S. News & World Report* periodically provides its rankings of medical schools, and in a recent year the ranks for Harvard, Johns Hopkins, and New York University were 1, 2, and 3, respectively.

23. Movie Ratings In a college film studies course, students rate ten documentaries using a scale of 0 to 5 stars.

24. Criminology In a criminology study, the lengths of prison sentences are obtained for randomly selected subjects convicted of auto theft.

25. Baseball Baseball statistician Bill James records the years in which the baseball World Series is won by a team from the National League.

26. Art History In an art history course, students are asked to identify the painting styles (abstract, classical, expressionism, etc.) of several paintings.

27. Areas of States A data set consists of the areas (km^2) of each of the 50 United States.

28. Body Temperatures Body temperatures (in degrees Fahrenheit) listed in Data Set 5 "Body Temperatures" in Appendix B

In Exercises 29–32, identify the level of measurement of the data as nominal, ordinal, interval, or ratio. Also, explain what is wrong with the given calculation.

29. Super Bowl The first Super Bowl attended by the author was Super Bowl XLVIII. On the first play of the game, the Seattle defense scored on a safety. The defensive players wore jerseys numbered 31, 28, 41, 56, 25, 54, 69, 50, 91, 72, 29, and the average (mean) of those numbers is 49.6.

30. Social Security Numbers As part of a project in a statistics class, students report the last four digits of their Social Security numbers, and the average (mean) of those digits is computed to be 4.7.

31. Temperatures As this exercise is being written, it is 80°F in Paris, France, and it is 40°F in Anchorage, Alaska, so it is twice as warm in Paris as it is in Anchorage.

32. College Ranks As of this writing, *U.S. News & World Report* ranked the best global universities, including these results: Harvard (1), MIT (2), Stanford (3), University of California at Berkeley (4), and University of Oxford (5). The difference between Harvard and MIT is the same as the difference between Stanford and the University of California at Berkeley.

1-2 Beyond the Basics

33. Countable For each of the following, categorize the nature of the data using one of these three descriptions: (1) discrete because the number of possible values is finite; (2) discrete because the number of possible values is infinite but countable; (3) continuous because the number of possible values is infinite and not countable.

a. Exact lengths of the feet of members of the band the Monkees

b. Shoe sizes of members of the band the Monkees (such as 9, 9½, and so on)

c. The number of albums sold by the Monkees band

d. The numbers of monkeys sitting at keyboards before one of them randomly types the lyrics for the song "Daydream Believer."

34. Directions in Degrees Standard navigation systems used for aviation and boating are based on directions measured in degrees, with north represented by 0°. Relative to north, east is 90°, south is 180°, and west is 270°. What is the level of measurement of such directions measured in degrees?

1-3 Collecting Sample Data

Key Concept When analyzing sample data, it is essential to use an appropriate method for collecting those sample data. This section includes comments about various methods and sampling procedures. Of particular importance is the method of using a *simple random sample*. We will make frequent use of this sampling method throughout the remainder of this book.

As you read this section, remember this:

If sample data are not collected in an appropriate way, the data may be so utterly useless that no amount of statistical torturing can salvage them.

PART 1 Basics of Design of Experiments and Collecting Sample Data

The Gold Standard Randomness with placebo/treatment groups is sometimes called the "gold standard" because it is so effective.

DEFINITION

A **placebo** is a harmless and ineffective pill, medicine, or procedure sometimes used for psychological benefit or sometimes used by researchers for comparison to other treatments.

The following example describes how the gold standard was used in the largest health experiment ever conducted.

EXAMPLE 1 **The Salk Vaccine Experiment**

In 1954, an experiment was designed to test the effectiveness of the Salk vaccine in preventing polio, which had killed or paralyzed thousands of children. By random selection, 401,974 children were randomly assigned to two groups: (1) 200,745 children were given a *treatment* consisting of Salk vaccine injections; (2) 201,229 children were injected with a *placebo* that contained no drug. Children were assigned to the treatment or placebo group through a process of random selection, equivalent to flipping a coin. Among the children given the Salk vaccine, 33 later developed paralytic polio, and among the children given a placebo, 115 later developed paralytic polio.

YOUR TURN. Do Exercise 1 "Magnet Treatment of Pain."

Example 1 describes an *experiment* because subjects were given a treatment, but ethical, cost, time, and other considerations sometimes prohibit the use of an experiment. We would never want to conduct a driving/texting experiment in which we ask subjects to text while driving—some of them could die. It would be far better to observe past crash results to understand the effects of driving while texting. See the following definitions.

DEFINITIONS

In an **experiment**, we apply some *treatment* and then proceed to observe its effects on the individuals. (The individuals in experiments are called **experimental units**, and they are often called **subjects** when they are people.)

In an **observational study**, we observe and measure specific characteristics, but we don't attempt to *modify* the individuals being studied.

Experiments are often better than observational studies because well-planned experiments typically reduce the chance of having the results affected by some variable that is not part of a study. A *lurking variable* is one that affects the variables included in the study, but it is not included in the study.

EXAMPLE 2 **Ice Cream and Drownings**

Observational Study: Observe past data to incorrectly conclude that ice cream causes drownings (based on data showing that increases in ice cream sales are associated with increases in drownings). The mistake is to miss the lurking variable of temperature and the failure to see that as the temperature increases, ice cream sales increase and drownings increase because more people swim.

Experiment: Conduct an *experiment* with one group treated with ice cream while another group gets no ice cream. We would see that the rate of drowning victims is about the same in both groups, so ice cream consumption has no effect on drownings.

Here, the experiment is clearly better than the observational study.

YOUR TURN. Do Exercise 6 "Experiment or Observational Study."

Clinical Trials Versus Observational Studies

In a *New York Times* article about hormone therapy for women, reporter Denise Grady wrote about randomized clinical trials that involve subjects who were randomly assigned to a treatment group and another group not given the treatment. Such randomized clinical trials are often referred to as the "gold standard" for medical research. In contrast, observational studies can involve patients who decide themselves to undergo some treatment. Subjects who decide themselves to undergo treatments are often healthier than other subjects, so the treatment group might appear to be more successful simply because it involves healthier subjects, not necessarily because the treatment is effective. Researchers criticized observational studies of hormone therapy for women by saying that results might appear to make the treatment more effective than it really is.

The Mode Effect

There is a variety of different ways that polls can be conducted, and the way that a particular poll is conducted can strongly influence the responses. Survey questions on sensitive topics can be greatly influenced by the survey mode that was used. For example, Pew reports that when asked about their personal financial situation, 20% of those responding online reported that it was in poor shape, but 14% of those responding by phone gave that same response. When given multiple choices to a question, respondents tend to favor the first choice when they are reading it in an online poll, but they tend to favor the last choice when they hear the choices in a live phone interview. These are examples of the mode effect—the way that a poll is conducted has an effect on the responses that people give.

Design of Experiments

Good design of experiments includes *replication, blinding,* and *randomness*.

- **Replication** is the repetition of an experiment on more than one individual. Good use of replication requires sample sizes that are large enough so that we can see effects of treatments. In the Salk experiment in Example 1, the experiment used sufficiently large sample sizes, so the researchers could see that the Salk vaccine was effective.
- **Blinding** is used when the subject doesn't know whether he or she is receiving a treatment or a placebo. Blinding is a way to get around the **placebo effect,** which occurs when an untreated subject reports an improvement in symptoms. (The reported improvement in the placebo group may be real or imagined.) The Salk experiment in Example 1 was **double-blind,** which means that blinding occurred at two levels: (1) The children being injected didn't know whether they were getting the Salk vaccine or a placebo, and (2) the doctors who gave the injections and evaluated the results did not know either. Codes were used so that the researchers could objectively evaluate the effectiveness of the Salk vaccine.
- **Randomness** is used when individuals are assigned to different groups through a process of random selection, as in the Salk vaccine experiment in Example 1. The logic behind randomness is to use chance as a way to create two groups that are similar. The following definition refers to one common and effective way to collect sample data in a way that uses randomness.

DEFINITION

A **simple random sample** of *n* subjects is selected in such a way that every possible *sample of the same size n* has the same chance of being chosen. (A simple random sample is often called a random sample, but strictly speaking, a *random sample* has the weaker requirement that all members of the population have the same chance of being selected. That distinction is not so important in this text. See Exercise 37 "Simple Random Sample vs. Random Sample".)

Throughout, we will use various statistical procedures, and we often have a requirement that we have collected a *simple random sample*, as defined above.

Unlike careless or haphazard sampling, random sampling usually requires very careful planning and execution. Wayne Barber of Chemeketa Community College is quite correct when he tells his students that "randomness needs help."

Other Sampling Methods In addition to simple random sampling, here are some other sampling methods commonly used for surveys. Figure 1-4 illustrates these different sampling methods.

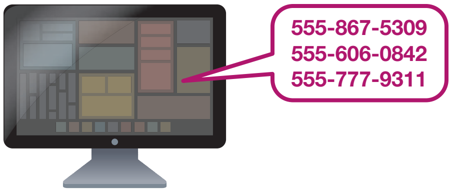
DEFINITIONS

With **systematic sampling**, we select some starting point and then select every *k*th (such as every 50th) element in the population.

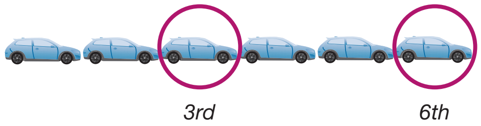
With **convenience sampling**, we simply use data that are very easy to get.

With **stratified sampling**, we subdivide the population into at least two different subgroups (or strata) so that subjects within the same subgroup share the same characteristics (such as gender). Then we draw a sample from each subgroup (or stratum).

With **cluster sampling**, we first divide the population area into sections (or clusters). Then we randomly select some of those clusters and choose *all* the members from those selected clusters.

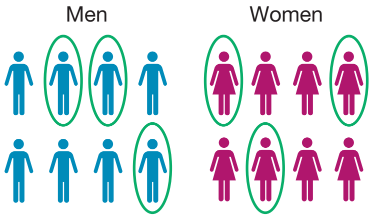


Simple Random Sample
A sample of *n* subjects is selected so that every sample of the same size *n* has the same chance of being selected.

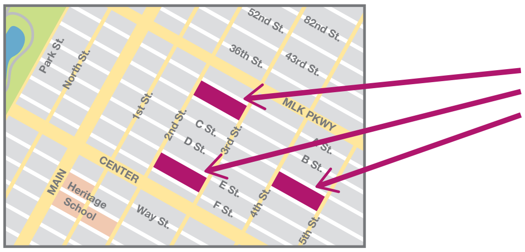


Systematic Sample
Select every *k*th subject.

Convenience Sample
Use data that are very easy to get.



Stratified Sample
Subdivide population into strata (groups) with the same characteristics, then randomly sample within those strata.



Cluster Sample
Partition the population in clusters (groups), then randomly select some clusters, then select all members of the selected clusters.

FIGURE 1-4 **Common Sampling Methods**

HINT Because it's difficult to remember the distinction between stratified sampling and cluster sampling, picture your entire class as one cluster among all classes at your college. Remember the alliteration of "cluster class" to recall that with cluster sampling, you choose *all* of the members of selected clusters. Associate "cluster" with "all." Then, stratified sampling is the other method of choosing samples from selected classes or subgroups.

Multistage Sampling Professional pollsters and government researchers often collect data by using some combination of the preceding sampling methods. In a multistage sample design, pollsters select a sample in different stages, and each stage might use different methods of sampling, as in the following example.

EXAMPLE 3 Multistage Sample Design

The U.S. government's unemployment statistics are based on surveys of households. It is impractical to personally survey each household in a simple random sample, because they would be scattered all over the country, making it nearly

continued

Wording in Surveys Can Affect Results

Pew Research Center pollsters conduct experiments to better understand how wording in questions can affect the responses of survey subjects. In a Pew Research Center survey of 1505 adults in the United States, half of the respondents were asked about finding "jobs" while the other half were asked about "*good* jobs." For the "jobs" group, 33% said that jobs were difficult to find. For the "*good* jobs" group, 45% said that good jobs were difficult to find.