Elementary

14E
Statistics

Mario F. Triola

This page intentionally left blank

ELEMENTARY STATISTICS

This page intentionally left blank

14th EDITION

ELEMENTARY STATISTICS

MARIO F. TRIOLA

P Pearson

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 *Essentials of Statistics,* 6th edition, *Elementary Statistics Using Excel™,* 7th 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 57).

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.

Goals of This Fourteenth 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 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* 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 14th 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.
- *Technology Project* provides an activity that can be used with a variety of technologies.
- *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.

Technology Integration

As in the preceding edition, there are many displays of screens from technology throughout the book, and some exercises are based on displayed results from technology. Where appropriate, sections end with a **Tech Center** subsection that includes detailed instructions for Statdisk, Minitab®, Excel®, StatCrunch, *R* (new to this edition), or a TI-83/84 Plus® calculator. (Throughout this text, "TI-83/84 Plus" is used to identify a TI-83 Plus or TI-84 Plus calculator). The Tech Centers also include references to new technology-specific instructional videos. The end-of-chapter features include a *Technology Project.*

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 14th Edition

New Features

New Content: This 14th edition includes an abundance of new exercises, new examples, and Chapter Problems, as summarized in the following table.

	Number	New to 14th 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 14th 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 14th edition of *Elementary Statistics* 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* introduced the resampling method of *bootstrapping* in Section 7-4. This 14th edition of *Elementary Statistics* 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* included bootstrap resampling, but the new version of Statdisk for the 14th 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.

New Technology: Where it is appropriate, the end-of-section *Tech Centers* include *R* as an additional technology. (The technologies of Statdisk, Excel, StatCrunch, Minitab, and TI-83/84 Plus calculators continue to be included in the Tech Centers.)

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 instruc-
   tional 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

Resources for Success

MyLab Statistics is available to accompany Pearson's market-leading text options, including Elementary Statistics, 14e 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.
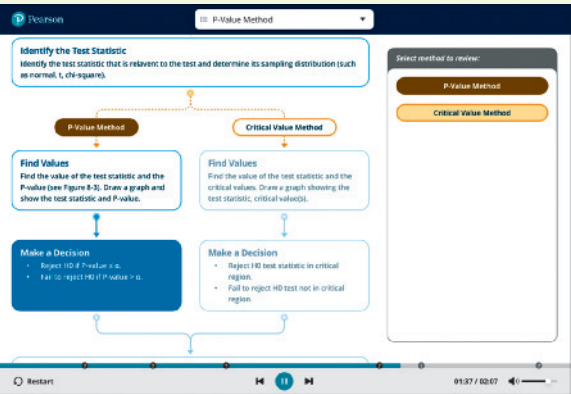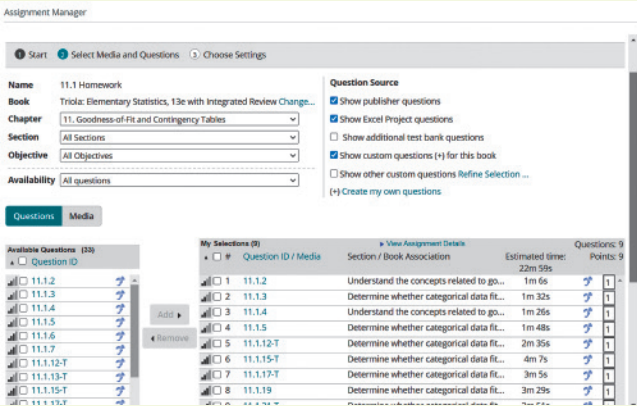
Enhanced video program to meet Introductory Statistics needs:

- **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).

Graphing Calculator Manual for the TI-83 Plus, TI-84 Plus, TI-84 Plus C and TI-84 Plus CE, (Download Only) by Kathleen McLaughlin (University of Connecticut) & Dorothy Wakefield (University of Connecticut Health Center).

Statdisk Student Laboratory Manual and Workbook (Download Only), by Mario F. Triola. Available at www.TriolaStats.com or within MyLab Statistics.

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: 9780136803065; ISBN-10: 0136803067)

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.

pearson.com/mylab/statistics

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
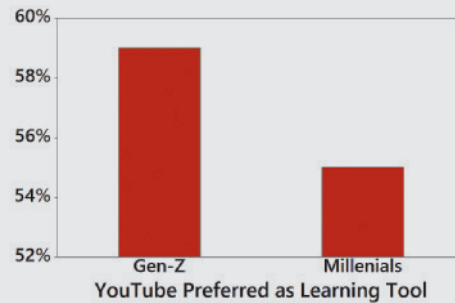
FIGURE 1-1 YouTube as a Preferred Learning Tool
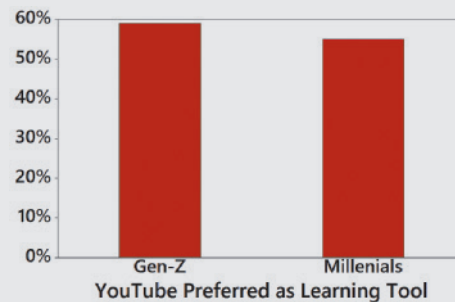
FIGURE 1-2 Same as Figure 1-1 but with scale beginning with 0%

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.

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 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.

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

---

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.

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.

Origin of "Statistics"
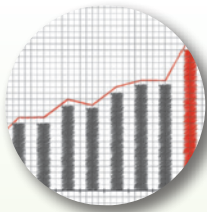
The word *statistics* is derived from the Latin word *status* (meaning "state"). Early uses of statistics involved compilations of data and graphs describing various aspects of a state or country. In 1662, John Graunt published statistical information about births and deaths. Graunt's work was followed 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 (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.

<p style="text-align:center">Correlation does not imply causation.</p>

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.

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?