



openstax™

# Stat- istics

HIGH SCHOOL



# Statistics

SENIOR CONTRIBUTING AUTHORS

**BARBARA ILLOWSKY, DE ANZA COLLEGE**

**SUSAN DEAN, DE ANZA COLLEGE**



**OpenStax**

Rice University  
6100 Main Street MS-375  
Houston, Texas 77005

To learn more about OpenStax, visit <https://openstax.org>.  
Individual print copies and bulk orders can be purchased through our website.

©2020 Texas Education Agency (TEA). This instructional material was initially created through a Texas Education Agency (TEA) initiative to provide high-quality open-source instructional materials to districts free of charge. Changes were made to the original material, including updates to art, structure, and other content updates. The original material is available at: <https://www.texasgateway.org/book/tea-statistics>. Textbook content is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>) (CC BY 4.0).

For questions regarding this licensing, please contact [support@openstax.org](mailto:support@openstax.org).

**Trademarks**

The OpenStax name, OpenStax logo, OpenStax book covers, OpenStax CNX name, OpenStax CNX logo, OpenStax Tutor name, OpenStax Tutor logo, Connexions name, Connexions logo, Rice University name, and Rice University logo are not subject to the license and may not be reproduced without the prior and express written consent of Rice University.

**HARDCOVER BOOK ISBN-13**  
**B&W PAPERBACK BOOK ISBN-13**  
**DIGITAL VERSION ISBN-13**  
**ORIGINAL PUBLICATION YEAR**  
10 9 8 7 6 5 4 3 2 1

**978-1-975076-53-5**  
**978-1-975076-52-8**  
**978-1-951693-22-0**  
**2020**

## OPENSTAX

OpenStax provides free, peer-reviewed, openly licensed textbooks for introductory college and Advanced Placement® courses and low-cost, personalized courseware that helps students learn. A nonprofit ed tech initiative based at Rice University, we're committed to helping students access the tools they need to complete their courses and meet their educational goals.

## RICE UNIVERSITY

OpenStax, OpenStax CNX, and OpenStax Tutor are initiatives of Rice University. As a leading research university with a distinctive commitment to undergraduate education, Rice University aspires to path-breaking research, unsurpassed teaching, and contributions to the betterment of our world. It seeks to fulfill this mission by cultivating a diverse community of learning and discovery that produces leaders across the spectrum of human endeavor.



---

## PHILANTHROPIC SUPPORT

OpenStax is grateful for our generous philanthropic partners, who support our vision to improve educational opportunities for all learners.

Laura and John Arnold Foundation

Arthur and Carlyse Ciocca Charitable Foundation

Ann and John Doerr

Bill & Melinda Gates Foundation

Girard Foundation

Google Inc.

The William and Flora Hewlett Foundation

Rusty and John Jagers

The Calvin K. Kazanjian Economics Foundation

Charles Koch Foundation

Leon Lowenstein Foundation, Inc.

The Maxfield Foundation

Burt and Deedee McMurtry

Michelson 20MM Foundation

National Science Foundation

The Open Society Foundations

Jumee Yhu and David E. Park III

Brian D. Patterson USA-International Foundation

The Bill and Stephanie Sick Fund

Robin and Sandy Stuart Foundation

The Stuart Family Foundation

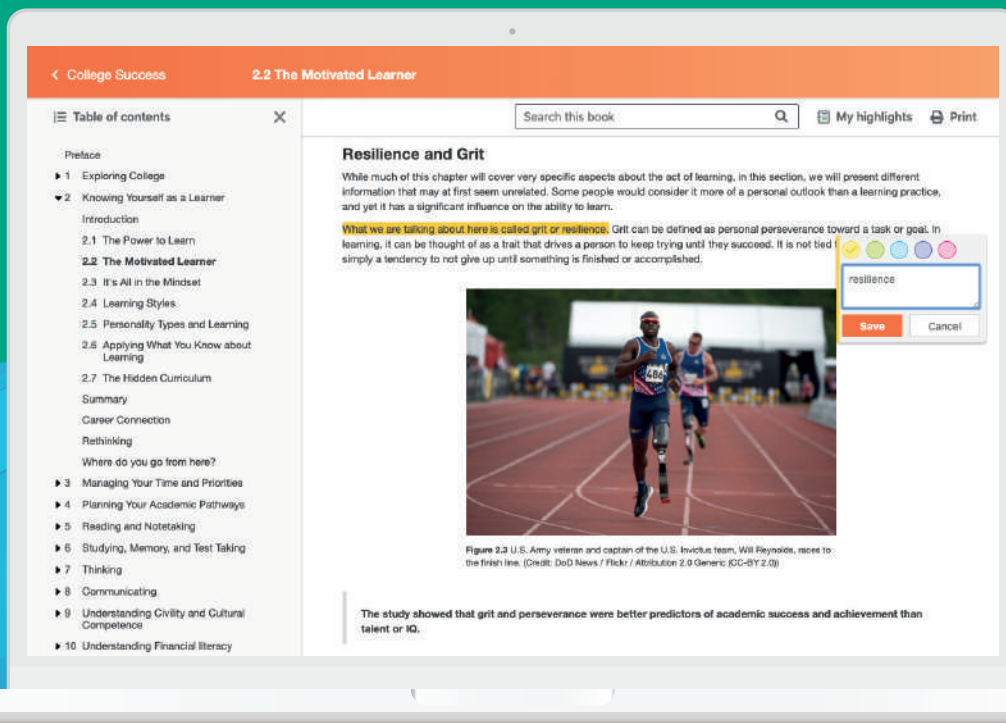
Tammy and Guillermo Treviño

# Study where you want, what you want, **when you want.**

When you access College Success in our web view, you can use our new online **highlighting and note-taking** features to create your own study guides.

Our books are free and flexible, forever.

Get started at [openstax.org/details/books/statistics](https://openstax.org/details/books/statistics)



## Table of Contents

<b>Preface</b>	<b>1</b>
<b>Chapter 1: Sampling and Data</b>	<b>5</b>
1.1 Definitions of Statistics, Probability, and Key Terms	5
1.2 Data, Sampling, and Variation in Data and Sampling	10
1.3 Frequency, Frequency Tables, and Levels of Measurement	28
1.4 Experimental Design and Ethics	37
1.5 Data Collection Experiment	41
1.6 Sampling Experiment	43
<b>Chapter 2: Descriptive Statistics</b>	<b>71</b>
2.1 Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs	72
2.2 Histograms, Frequency Polygons, and Time Series Graphs	82
2.3 Measures of the Location of the Data	93
2.4 Box Plots	102
2.5 Measures of the Center of the Data	106
2.6 Skewness and the Mean, Median, and Mode	112
2.7 Measures of the Spread of the Data	116
2.8 Descriptive Statistics	128
<b>Chapter 3: Probability Topics</b>	<b>181</b>
3.1 Terminology	182
3.2 Independent and Mutually Exclusive Events	188
3.3 Two Basic Rules of Probability	198
3.4 Contingency Tables	203
3.5 Tree and Venn Diagrams	210
3.6 Probability Topics	221
<b>Chapter 4: Discrete Random Variables</b>	<b>253</b>
4.1 Probability Distribution Function (PDF) for a Discrete Random Variable	255
4.2 Mean or Expected Value and Standard Deviation	257
4.3 Binomial Distribution (Optional)	264
4.4 Geometric Distribution (Optional)	271
4.5 Hypergeometric Distribution (Optional)	276
4.6 Poisson Distribution (Optional)	279
4.7 Discrete Distribution (Playing Card Experiment)	284
4.8 Discrete Distribution (Lucky Dice Experiment)	289
<b>Chapter 5: Continuous Random Variables</b>	<b>325</b>
5.1 Continuous Probability Functions	327
5.2 The Uniform Distribution	330
5.3 The Exponential Distribution (Optional)	340
5.4 Continuous Distribution	352
<b>Chapter 6: The Normal Distribution</b>	<b>379</b>
6.1 The Standard Normal Distribution	380
6.2 Using the Normal Distribution	385
6.3 Normal Distribution—Lap Times	394
6.4 Normal Distribution—Pinkie Length	396
<b>Chapter 7: The Central Limit Theorem</b>	<b>413</b>
7.1 The Central Limit Theorem for Sample Means (Averages)	414
7.2 The Central Limit Theorem for Sums (Optional)	419
7.3 Using the Central Limit Theorem	422
7.4 Central Limit Theorem (Pocket Change)	431
7.5 Central Limit Theorem (Cookie Recipes)	434
<b>Chapter 8: Confidence Intervals</b>	<b>459</b>
8.1 A Single Population Mean Using the Normal Distribution	461
8.2 A Single Population Mean Using the Student's t-Distribution	472
8.3 A Population Proportion	477
8.4 Confidence Interval (Home Costs)	483
8.5 Confidence Interval (Place of Birth)	485
8.6 Confidence Interval (Women's Heights)	487
<b>Chapter 9: Hypothesis Testing with One Sample</b>	<b>523</b>

9.1 Null and Alternative Hypotheses . . . . .	524
9.2 Outcomes and the Type I and Type II Errors . . . . .	526
9.3 Distribution Needed for Hypothesis Testing . . . . .	529
9.4 Rare Events, the Sample, and the Decision and Conclusion . . . . .	529
9.5 Additional Information and Full Hypothesis Test Examples . . . . .	532
9.6 Hypothesis Testing of a Single Mean and Single Proportion . . . . .	548
<b>Chapter 10: Hypothesis Testing with Two Samples . . . . .</b>	<b>583</b>
10.1 Two Population Means with Unknown Standard Deviations . . . . .	584
10.2 Two Population Means with Known Standard Deviations . . . . .	592
10.3 Comparing Two Independent Population Proportions . . . . .	595
10.4 Matched or Paired Samples (Optional) . . . . .	600
10.5 Hypothesis Testing for Two Means and Two Proportions . . . . .	606
<b>Chapter 11: The Chi-Square Distribution . . . . .</b>	<b>637</b>
11.1 Facts About the Chi-Square Distribution . . . . .	638
11.2 Goodness-of-Fit Test . . . . .	639
11.3 Test of Independence . . . . .	649
11.4 Test for Homogeneity . . . . .	654
11.5 Comparison of the Chi-Square Tests . . . . .	657
11.6 Test of a Single Variance . . . . .	657
11.7 Lab 1: Chi-Square Goodness-of-Fit . . . . .	660
11.8 Lab 2: Chi-Square Test of Independence . . . . .	664
<b>Chapter 12: Linear Regression and Correlation . . . . .</b>	<b>695</b>
12.1 Linear Equations . . . . .	696
12.2 The Regression Equation . . . . .	698
12.3 Testing the Significance of the Correlation Coefficient (Optional) . . . . .	707
12.4 Prediction (Optional) . . . . .	712
12.5 Outliers . . . . .	713
12.6 Regression (Distance from School) (Optional) . . . . .	721
12.7 Regression (Textbook Cost) (Optional) . . . . .	723
12.8 Regression (Fuel Efficiency) (Optional) . . . . .	725
<b>Chapter 13: F Distribution and One-way Anova . . . . .</b>	<b>761</b>
13.1 One-Way ANOVA . . . . .	762
13.2 The F Distribution and the F Ratio . . . . .	763
13.3 Facts About the F Distribution . . . . .	767
13.4 Test of Two Variances . . . . .	775
13.5 Lab: One-Way ANOVA . . . . .	778
<b>Appendix A: Appendix A Review Exercises (Ch 3–13) . . . . .</b>	<b>805</b>
<b>Appendix B: Appendix B Practice Tests (1–4) and Final Exams . . . . .</b>	<b>833</b>
<b>Appendix C: Data Sets . . . . .</b>	<b>889</b>
<b>Appendix D: Group and Partner Projects . . . . .</b>	<b>893</b>
<b>Appendix E: Solution Sheets . . . . .</b>	<b>899</b>
<b>Appendix F: Mathematical Phrases, Symbols, and Formulas . . . . .</b>	<b>903</b>
<b>Appendix G: Notes for the TI-83, 83+, 84, 84+ Calculators . . . . .</b>	<b>909</b>
<b>Appendix H: Tables . . . . .</b>	<b>921</b>
<b>Index . . . . .</b>	<b>923</b>



# PREFACE

Welcome to *Statistics*, an OpenStax resource. This textbook was written to increase teacher and student access to high-quality learning materials, maintaining the highest standards of academic rigor at little to no cost.

## About OpenStax

OpenStax is a nonprofit based at Rice University, and it's our mission to improve student access to education. Our first openly licensed college textbook was published in 2012, and our library has since scaled to over 35 books used by hundreds of thousands of students for college and AP<sup>®</sup> courses. OpenStax Tutor and Rover, our low-cost personalized learning tools, are being used in college and high school courses throughout the country. Through our partnerships with philanthropic foundations and our alliance with other educational resource organizations, OpenStax is breaking down the most common barriers to learning and empowering students and instructors to succeed.

## About OpenStax Resources Customization

*Statistics* is licensed under a Creative Commons Attribution 4.0 International (CC BY) license, which means that you can distribute, remix, and build upon the content, as long as you provide attribution to OpenStax and its content contributors.

Because our books are openly licensed, you are free to use the entire book or pick and choose the sections that are most relevant to the needs of your students. Feel free to remix the content by assigning your students certain chapters and sections in your syllabus, in the order that you prefer. You can even provide a direct link in your syllabus or student assignment system to the sections in the web view of your book.

Instructors also have the option of creating a customized version of their OpenStax book. The custom version can be made available to students in low-cost print or digital form through their campus bookstore. Visit the Instructor Resources section of your book page on [openstax.org](https://openstax.org) for more information.

## Art Attribution in *Statistics*

In *Statistics*, most art contains attribution to its title, creator or rights holder, host platform, and license within the caption. For art that is openly licensed, anyone may reuse the art as long as they provide the same attribution to its original source. Some art has been provided through permissions and should only be used with the attribution or limitations provided in the credit.

## Errata

All OpenStax textbooks undergo a rigorous review process. However, like any professional-grade textbook, errors sometimes occur. The good part is, since our books are web-based, we can make updates periodically. If you have a correction to suggest, submit it through our errata reporting tool. We will review your suggestion and make necessary changes.

## Format

You can access this textbook for free in web view or PDF through [openstax.org](https://openstax.org), and for a low cost in print.

## About *Statistics*

This instructional material was initially created through a Texas Education Agency (TEA) initiative to provide high-quality open-source instructional materials to districts free of charge. Funds were allocated by the 84th Texas Legislature (2015) for the creation of state-developed, open-source instructional materials with the request that advanced secondary courses supporting the study of science, technology, engineering, and mathematics should be prioritized.

*Statistics* covers the scope and sequence requirements of a typical one-year statistics course. The text provides comprehensive coverage of statistical concepts, including quantitative examples, collaborative activities, and practical applications. *Statistics* was designed to meet and exceed the requirements of the relevant **Texas Essential Knowledge and Skills (TEKS)** (<http://ritter.tea.state.tx.us/rules/tac/chapter111/ch111c.html#111.47>), while allowing significant flexibility for instructors.

Qualified and experienced Texas faculty were involved throughout the development process, and the textbooks were reviewed extensively to ensure effectiveness and usability in each course. Reviewers considered each resource's clarity, accuracy, student support, assessment rigor and appropriateness, alignment to TEKS, and overall quality. Their invaluable suggestions provided the basis for continually improved material and helped to certify that the books are ready for use.

The writers and reviewers also considered common course issues, effective teaching strategies, and student engagement to provide instructors and students with useful, supportive content and drive effective learning experiences.

## Coverage and Scope

*Statistics* presents the appropriate statistical concepts and skills in a logical and engaging progression that should be familiar to faculty.

Chapter 1: Sampling and Data  
Chapter 2: Descriptive Statistics  
Chapter 3: Probability Topics  
Chapter 4: Discrete Random Variables  
Chapter 5: Continuous Random Variables  
Chapter 6: The Normal Distribution  
Chapter 7: The Central Limit Theorem  
Chapter 8: Confidence Intervals  
Chapter 9: Hypothesis Testing with One Sample  
Chapter 10: Hypothesis Testing with Two Samples  
Chapter 11: The Chi-Square Distribution  
Chapter 12: Linear Regression and Correlation  
Chapter 13: F Distribution and One-Way ANOVA

## Flexibility

Like any OpenStax content, this textbook can be modified as needed for use by the instructor depending on the needs of the students in the course. Each set of materials created by OpenStax is organized into units and chapters and can be used like a traditional textbook as the entire syllabus for each course. The materials can also be accessed in smaller chunks for more focused use with a single student or an entire class. Instructors are welcome to download and assign the PDF version of the textbook through a learning management system or can use their LMS to link students to specific chapters and sections of the book relevant to the concept being studied. The entire textbook will be available during the fall of 2020 in an editable Google document, and until then instructors are welcome to copy and paste content from the textbook to modify as needed prior to instruction.

## Student-Centered Focus

*Statistics* was developed with detailed and practical guidance from experienced high school teachers and curriculum experts. Their contributions helped create a resource that provides easy-to-follow explanations with ample opportunities for enrichment and practice. In addition to clear and grade-level appropriate main text coverage, the following features are meant to enhance student understanding of statistics concepts:

- *Examples* are placed strategically throughout the text to show students the step-by-step process of interpreting and solving statistical problems. To keep the text relevant for students, the examples are drawn from a broad spectrum of practical topics, including examples from academic life and learning, health and medicine, retail and business, and sports and entertainment.
- *Try It* practice problems immediately follow many examples and give students the opportunity to practice as they read the text. Like the Examples, the Try It problems are usually based on practical and familiar topics.
- *Collaborative Exercises* provide an in-class scenario for students to work together and learn from each other as they explore course concepts.
- *Calculator Guidance* shows students step-by-step instructions for input using the TI-83, 83+, 84, and 84+ calculators and helps them consider how to use these tools in their studies. The Technology Icon indicates where the use of a TI calculator or computer software is recommended.
- *Practice, Homework, and Bringing It Together* problems give the students problems at various degrees of difficulty while including real-world scenarios to engage students.

## Statistics Labs

These innovative activities were developed by Barbara Illowsky and Susan Dean (both of De Anza College) and allow students to design, implement, and interpret statistical analyses. They are drawn from actual experiments and data-gathering processes and offer a unique hands-on and collaborative experience. Statistics Labs appear at the end of each chapter and begin with student learning outcomes, general estimates for time on task, and global implementation notes. Students are then provided with step-by-step guidance, including sample data tables and calculation prompts. This detailed assistance will help the students successfully apply statistics concepts and lay the groundwork for future collaborative or individual work.

## Additional Resources

### Student and Instructor Resources

We've compiled additional resources for both students and instructors, including Getting Started Guides, PowerPoint slides, and an instructor answer guide. Instructor resources require a verified instructor account, which you can apply for when you log in or create your account on OpenStax.org. Take advantage of these resources to supplement your OpenStax book.

### Partner Resources

OpenStax Partners are our allies in the mission to make high-quality learning materials affordable and accessible to students and instructors everywhere. Their tools integrate seamlessly with our OpenStax titles at a low cost. To access the partner resources for your text, visit your book page on OpenStax.org.

## About the Authors

### Senior Contributing Authors

**Barbara Illowsky, De Anza College**

**Susan Dean, De Anza College**

### Contributing Authors

Daniel Birmajer, Nazareth College

Bryan Blount, Kentucky Wesleyan College

Sheri Boyd, Rollins College

Matthew Einsohn, Prescott College

James Helmreich, Marist College

Lynette Kenyon, Collin County Community College

Sheldon Lee, Viterbo University

Jeff Taub, Maine Maritime Academy

### Reviewers of Prior Editions

Laurel Chiappetta, University of Pittsburgh

Lenore Desilets, De Anza College

Lisa Markus, De Anza College

Mary Teegarden, San Diego Mesa College

Carol Olmstead, De Anza College

Carol Weideman, St. Petersburg College

Charles Ashbacher, Upper Iowa University, Cedar Rapids

Charles Klein, De Anza College

Cheryl Wartman, University of Prince Edward Island

David French, Tidewater Community College

Dennis Walsh, Middle Tennessee State University

Diane Mathios, De Anza College

John Thomas, College of Lake County

Jing Chang, College of Saint Mary

Sara Lenhart, Christopher Newport University

Sarah Boslaugh, Kennesaw State University

Abdulhamid Sukar, Cameron University

Abraham Biggs, Broward Community College

Adam Pennell, Greensboro College

Alexander Kolovos

Ann Flanigan, Kapiolani Community College

Robert McDevitt, Germanna Community College

Roberta Bloom, De Anza College

Rupinder Sekhon, De Anza College

Sudipta Roy, Kankakee Community College

Cindy Moss, Skyline College

Ernest Bonat, Portland Community College

Kathy Plum, De Anza College

Andrew Wiesner, Pennsylvania State University

Jonathan Oaks, Macomb Community College  
 Michael Greenwich, College of Southern Nevada  
 Miriam Masullo, SUNY Purchase  
 Mo Geraghty, De Anza College  
 Larry Green, Lake Tahoe Community College  
 Nydia Nelson, St. Petersburg College  
 Philip J. Verrecchia, York College of Pennsylvania  
 Robert Henderson, Stephen F. Austin State University  
 Benjamin Ngwudike, Jackson State University  
 Mel Jacobsen, Snow College  
 Birgit Aquilonius, West Valley College  
 Jim Lucas, De Anza College  
 David Bosworth, Hutchinson Community College  
 Frank Snow, De Anza College  
 George Bratton, University of Central Arkansas  
 Inna Grushko, De Anza College  
 Janice Hector, De Anza College  
 Javier Rueda, De Anza College  
 Lisa Rosenberg, Elon University  
 Mark Mills, Central College  
 Mary Jo Kane, De Anza College  
 Travis Short, St. Petersburg College  
 Valier Hauber, De Anza College  
 Vladimir Logvenenko, De Anza College  
 Wendy Lightheart, Lane Community College  
 Yvonne Sandoval, Pima Community College

### Editorial Review Board

**Linda Gann** (6-12 Mathematics Coordinator, Boerne ISD) taught mathematics and statistics for over twenty-five years at Northside ISD, and currently serves as the 6-12 Mathematics Coordinator in Boerne ISD. She was awarded the Presidential Award for Excellence in Teaching, the Radio Shack National Teacher Award, the HEB Teaching Excellence Award (State Finalist), and the AP Siemens Award. For many years, Linda worked for the College Board as a consultant for AP Calculus AB, BC, and Statistics, and as a reader for AP Statistics. She has also served as the co-chair for the College and Career Readiness Standards for Mathematics for all three writing phases. Her educational background consists of a B.S. in Mathematics from Illinois State University and an M.S. in Mathematics from the University of Texas, San Antonio. Additionally, she is nearing completion of her Ph.D. in Interdisciplinary of Learning and Teaching from UTSA. She presently serves as president of the Alamo District Council of Teachers of Mathematics and scholarship chair for the Priest Holmes Foundation.

**Wendy Martinez** (Cedar Park High School) has been a teacher since 1994. She currently teaches PreAP Geometry and on-level Statistics at Cedar Park High School in Leander ISD. She has taught at Rouse High School, Lake Travis High School, and Pflugerville Middle School.

**Alexander Teich** (Rice University Graduate Student, Master's Degree in Applied Mathematics) has teaching experience back to 2004 and has taught math classes at Spring Woods High School, in Cambridge, Massachusetts, and Philadelphia, Pennsylvania. He formerly sponsored the Spring Wood Chess Club, and has a wealth of varied practical experience outside the classroom.

**Amanda Yowell** (Pleasant Grove High School) earned a Bachelors of Science in Business Administration and Finance from the University of Arkansas and worked in Financial Management. She teaches Mathematics classes at Pleasant Grove High School in Texarkana, TX. In her free time, she enjoys spending time with her husband and their three children.

# 1 | SAMPLING AND DATA



**Figure 1.1** We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (David Sim)

## Introduction

### Chapter Objectives

By the end of this chapter, the student should be able to do the following:

- Recognize and differentiate between key terms
- Apply various types of sampling methods to data collection
- Create and interpret frequency tables

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or *fact*. Statistical methods can help you make the *best educated guess*.

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what *good* data can be distinguished from *bad*.

## 1.1 | Definitions of Statistics, Probability, and Key Terms

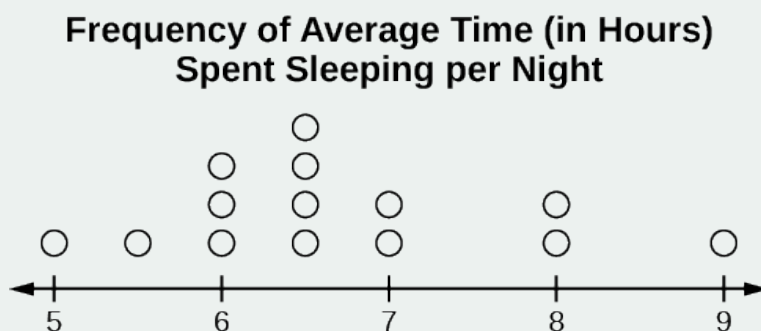
The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives.

## Collaborative Exercise

In your classroom, try this exercise. Have class members write down the average time—in hours, to the nearest half-hour—they sleep per night. Your instructor will record the data. Then create a simple graph, called a dot plot, of the data. A dot plot consists of a number line and dots, or points, positioned above the number line. For example, consider the following data:

5, 5.5, 6, 6, 6.5, 6.5, 6.5, 6.5, 7, 7, 8, 8, 9.

The dot plot for this data would be as follows:



**Figure 1.2**

Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not?

Where do your data appear to cluster? How might you interpret the clustering?

The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers, for example, finding an average. After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from *good* data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data, or inference, is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

## Statistical Models

Statistics, like all other branches of mathematics, uses **mathematical models** to describe phenomena that occur in the real world. Some mathematical models are deterministic. These models can be used when one value is precisely determined from another value. Examples of deterministic models are the quadratic equations that describe the acceleration of a car from rest or the differential equations that describe the transfer of heat from a stove to a pot. These models are quite accurate and can be used to answer questions and make predictions with a high degree of precision. Space agencies, for example, use deterministic models to predict the exact amount of thrust that a rocket needs to break away from Earth's gravity and achieve orbit.

However, life is not always precise. While scientists can predict to the minute the time that the sun will rise, they cannot say precisely where a hurricane will make landfall. **Statistical models** can be used to predict life's more uncertain situations. These special forms of mathematical models or functions are based on the idea that one value affects another value. Some statistical models are mathematical functions that are more precise—one set of values can predict or determine another set of values. Or some statistical models are mathematical functions in which a set of values do not precisely determine other



values. Statistical models are very useful because they can describe the probability or likelihood of an event occurring and provide alternative outcomes if the event does not occur. For example, weather forecasts are examples of statistical models. Meteorologists cannot predict tomorrow's weather with certainty. However, they often use statistical models to tell you how likely it is to rain at any given time, and you can prepare yourself based on this probability.

## Probability

**Probability** is a mathematical tool used to study randomness. It deals with the chance of an event occurring. For example, if you toss a **fair** coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is  $\frac{1}{2}$  or .5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl **Pearson** who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction  $\frac{996}{2,000}$  is equal to .498 which is very close to .5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments.

## Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion, or subset, of the larger population and study that portion—the sample—to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16-ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class as a sample of the population of *all* math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. Since we do not have the data for all math classes, that statistic is our best estimate of the average for the entire population of math classes. If we happen to have data for *all* math classes, we can find the population parameter. A **parameter** is a numerical characteristic of the whole population that can be estimated by a statistic. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. In order to have an accurate sample, it must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, usually notated by capital letters such as *X* and *Y*, is a characteristic or measurement that can be determined for each member of a population. Variables may describe values like weight in pounds or favorite subject in school. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let *X* equal the number of points earned by one math student at the end of a term, then *X* is a numerical variable. If we let *Y* be a person's party affiliation, then some examples of *Y* include Republican, Democrat, and Independent. *Y* is a categorical variable. We could do some math with values of *X*—calculate the average number of points earned, for example—but it makes no sense to do math with values of *Y*—calculating an average party affiliation makes no sense.

**Data** are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three. Your mean score would be 84.3 to one decimal place. If, in your math class, there are 40 students and 22 are

males and 18 females, then the proportion of men students is  $\frac{22}{40}$  and the proportion of women students is  $\frac{18}{40}$ . Mean and proportion are discussed in more detail in later chapters.

### NOTE

The words *mean* and *average* are often used interchangeably. In this book, we use the term *arithmetic mean* for mean.

## Example 1.1

Determine what the population, sample, parameter, statistic, variable, and data referred to in the following study.

We want to know the mean amount of extracurricular activities in which high school students participate. We randomly surveyed 100 high school students. Three of those students were in 2, 5, and 7 extracurricular activities, respectively.

### Solution 1.1

The **population** is all high school students.

The **sample** is the 100 high school students interviewed.

The **parameter** is the mean amount of extracurricular activities in which all high school students participate.

The **statistic** is the mean amount of extracurricular activities in which the sample of high school students participate.

The **variable** could be the amount of extracurricular activities by one high school student. Let  $X$  = the amount of extracurricular activities by one high school student.

The **data** are the number of extracurricular activities in which the high school students participate. Examples of the data are 2, 5, 7.

## Try It

**1.1** Find an article online or in a newspaper or magazine that refers to a statistical study or poll. Identify what each of the key terms—population, sample, parameter, statistic, variable, and data—refers to in the study mentioned in the article. Does the article use the key terms correctly?

## Example 1.2

Determine what the key terms refer to in the following study.

A study was conducted at a local high school to analyze the average cumulative GPAs of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1. Population \_\_\_\_ 2. Statistic \_\_\_\_ 3. Parameter \_\_\_\_ 4. Sample \_\_\_\_ 5. Variable \_\_\_\_ 6. Data \_\_\_\_

- a) all students who attended the high school last year
- b) the cumulative GPA of one student who graduated from the high school last year
- c) 3.65, 2.80, 1.50, 3.90
- d) a group of students who graduated from the high school last year, randomly selected
- e) the average cumulative GPA of students who graduated from the high school last year
- f) all students who graduated from the high school last year
- g) the average cumulative GPA of students in the study who graduated from the high school last year



**Solution 1.2**

1. f; 2. g; 3. e; 4. d; 5. b; 6. c

**Example 1.3**

Determine what the population, sample, parameter, statistic, variable, and data referred to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies (The Data and Story Library, n.d.). Here is the criterion they used.

Speed at which Cars Crashed	Location of <i>Driver</i> (i.e., dummies)
35 miles/hour	Front seat

**Table 1.1**

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver's seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

**Solution 1.3**

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies—if they had been real people—who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies—if they had been real people—who would have suffered head injuries in the sample.

The **variable**  $X$  = the number of driver dummies—if they had been real people—who would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

**Example 1.4**

Determine what the population, sample, parameter, statistic, variable, and data referred to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

**Solution 1.4**

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

The **variable**  $X$  = the number of medical doctors who have been involved in one or more malpractice suits.

The **data** are either: yes, was involved in one or more malpractice lawsuits; or no, was not.



## Collaborative Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average—mean—number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

## 1.2 | Data, Sampling, and Variation in Data and Sampling

Data may come from a population or from a sample. Lowercase letters like  $x$  or  $y$  generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

**Qualitative data** are the result of categorizing or describing attributes of a population. Qualitative data are also often called **categorical data**. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O–, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

**Quantitative data** are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

Data that are not only made up of counting numbers, but that may include fractions, decimals, or irrational numbers, are called **quantitative continuous data**. Continuous data are often the results of measurements like lengths, weights, or times. A list of the lengths in minutes for all the phone calls that you make in a week, with numbers like 2.4, 7.5, or 11.0, would be quantitative continuous data.

### Example 1.5 Data Sample of Quantitative Discrete Data

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books, 3, 4, 2, and 1, are the quantitative discrete data.

## Try It

**1.5** The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has 10 machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

### Example 1.6 Data Sample of Quantitative Continuous Data

The data are the weights of backpacks with books in them. You sample the same five students. The weights, in pounds, of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data.

### Try It $\Sigma$

**1.6** The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. ft., 160 sq. ft., 190 sq. ft., 180 sq. ft., and 210 sq. ft. What type of data is this?

### Example 1.7

You go to the supermarket and purchase three cans of soup (19 ounces tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces pistachio ice cream and 32 ounces chocolate chip cookies).

Name data sets that are quantitative discrete, quantitative continuous, and qualitative.

#### Solution 1.7

A possible solution

- One example of a quantitative discrete data set would be three cans of soup, two packages of nuts, four kinds of vegetables, and two desserts because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables, and desserts are qualitative data because they are categorical.

Try to identify additional data sets in this example.

### Example 1.8

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

### Try It $\Sigma$

**1.8** The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

#### NOTE

You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

### Example 1.9

Work collaboratively to determine the correct data type: quantitative or qualitative. Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words *the number of*.

- the number of pairs of shoes you own
- the type of car you drive
- the distance from your home to the nearest grocery store
- the number of classes you take per school year
- the type of calculator you use
- weights of sumo wrestlers
- number of correct answers on a quiz
- IQ scores (This may cause some discussion.)

#### Solution 1.9

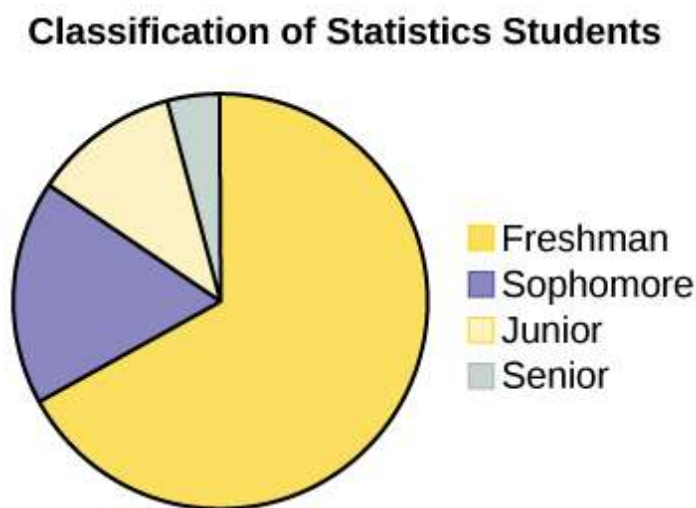
Items a, d, and g are quantitative discrete; items c, f, and h are quantitative continuous; items b and e are qualitative or categorical.

### Try It $\Sigma$

**1.9** Determine the correct data type, quantitative or qualitative, for the number of cars in a parking lot. Indicate whether quantitative data are continuous or discrete.

### Example 1.10

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart **Figure 1.2**. What type of data does this graph show?



**Figure 1.3**

**Solution 1.10**

This pie chart shows the students in each year, which is **qualitative or categorical data**.

## Try It

**1.10** A large school district keeps data of the number of students who receive test scores on an end of the year standardized exam. The data he collects are summarized in the histogram. The class boundaries are 50 to less than 60, 60 to less than 70, 70 to less than 80, 80 to less than 90, and 90 to less than 100.

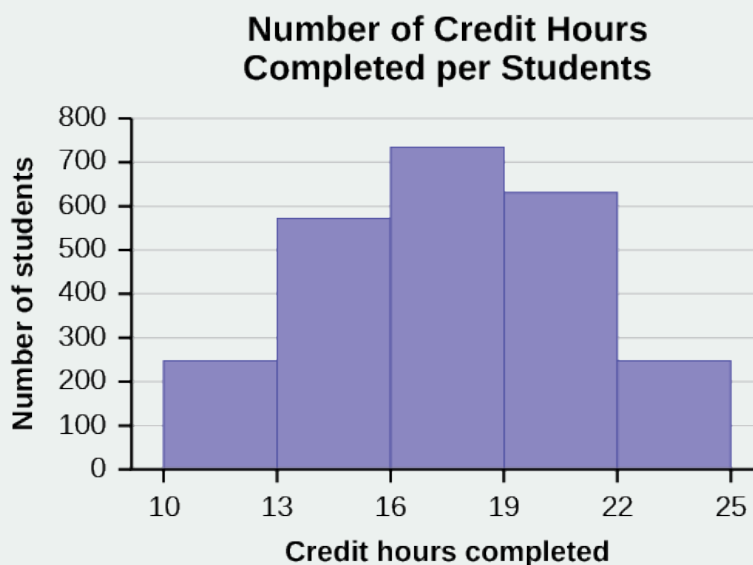


Figure 1.4

## Qualitative Data Discussion

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts, frequencies, and percentages or proportions, relative frequencies. For instance, to calculate the percentage of part time students at De Anza College, divide 9,200/22,496 to get .4089. Round to the nearest thousandth—third decimal place and then multiply by 100 to get the percentage, which is 40.9 percent.

So, the percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

De Anza College				Foothill College		
	Number	Percent			Number	Percent
Full-time	9,200	40.90%		Full-time	4,059	28.60%
Part-time	13,296	59.10%		Part-time	10,124	71.40%

Table 1.2 Fall Term 2007 (Census day)

De Anza College			Foothill College		
Total	22,496	100%	Total	14,183	100%

**Table 1.2 Fall Term 2007 (Census day)**

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data.

Two graphs that are used to display qualitative data are pie charts and bar graphs.

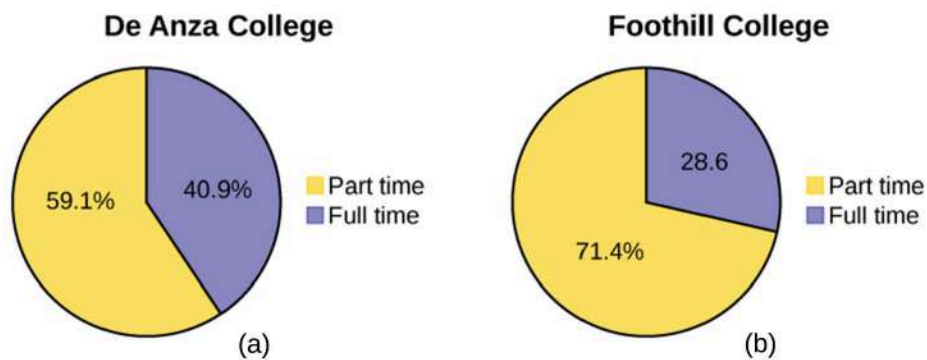
In a **pie chart**, categories of data are shown by wedges in a circle that represent the percent of individuals/items in each category. We use pie charts when we want to show parts of a whole.

In a **bar graph**, the length of the bar for each category represents the number or percent of individuals in each category. Bars may be vertical or horizontal. We use bar graphs when we want to compare categories or show changes over time.

A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at **Figure 1.5** and **Figure 1.6** and determine which graph (pie or bar) you think displays the comparisons better.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the *best* graph depending on the data and the context. Our choice also depends on what we are using the data for.



**Figure 1.5**

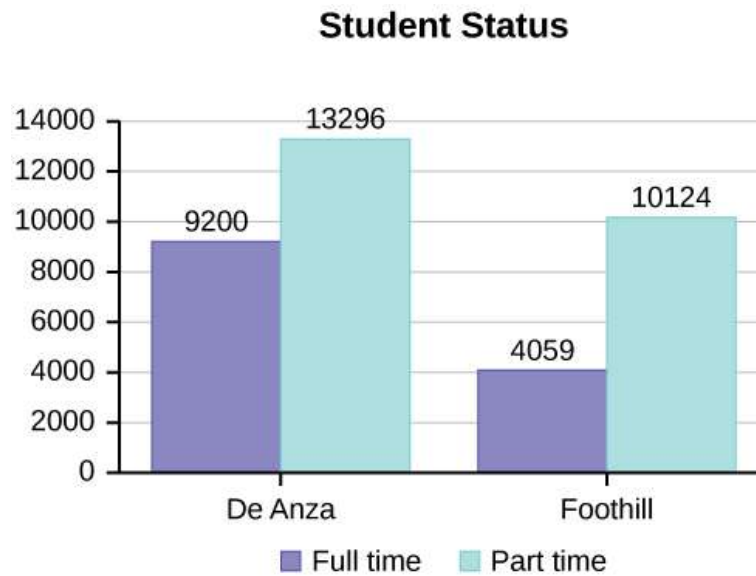


Figure 1.6

### Percentages That Add to More (or Less) Than 100 Percent

Sometimes percentages add up to be more than 100 percent (or less than 100 percent). In the graph, the percentages add to more than 100 percent because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100 percent.

Characteristic/Category	Percent
Students studying technical subjects	40.9%
Students studying non-technical subjects	48.6%
Students who intend to transfer to a four-year educational institutional	61.0%
TOTAL	150.5%

Table 1.3 De Anza College Year 2010

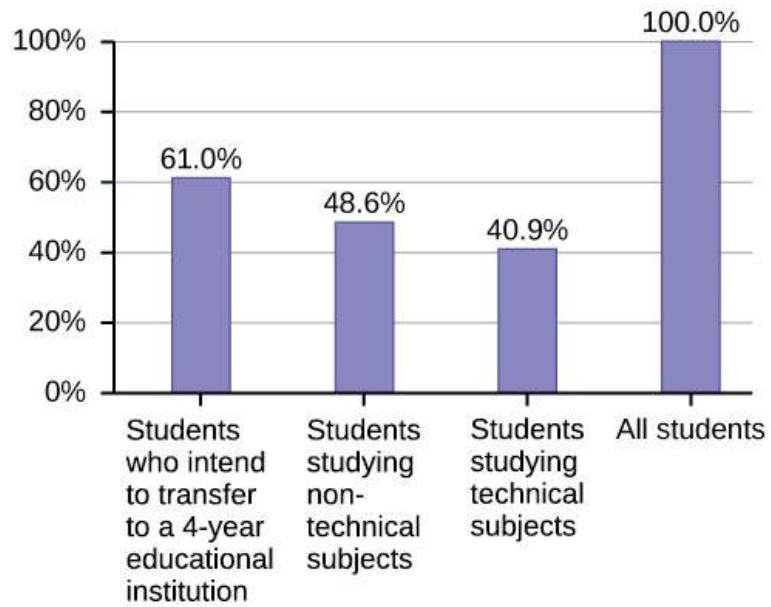


Figure 1.7

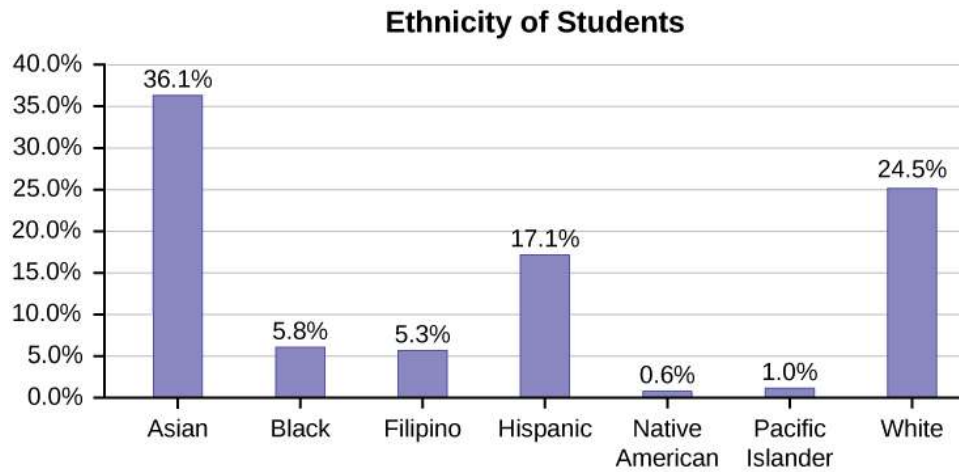
### Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the *Other/Unknown* category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

**Table 1.4 Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)**

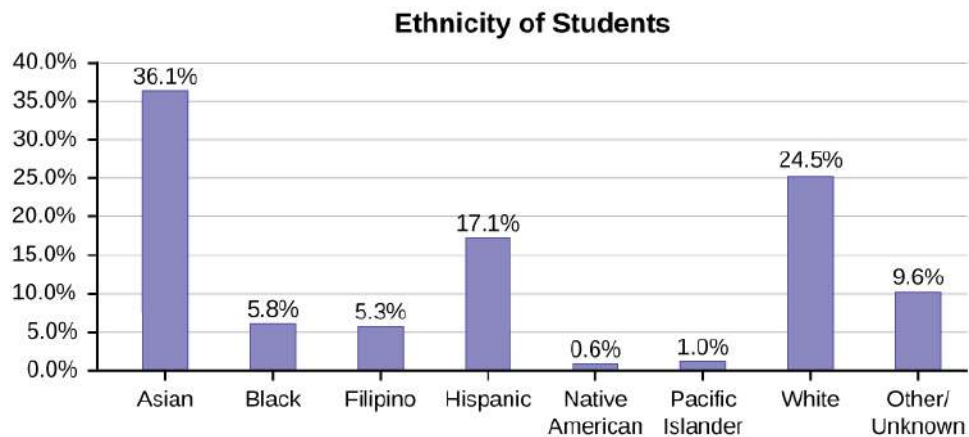




**Figure 1.8**

The following graph is the same as the previous graph but the *Other/Unknown* percent (9.6 percent) has been included. The *Other/Unknown* category is large compared to some of the other categories (Native American, .6 percent, Pacific Islander 1.0 percent). This is important to know when we think about what the data are telling us.

This particular bar graph in **Figure 1.9** can be difficult to understand visually. The graph in **Figure 1.10** is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.



**Figure 1.9 Bar Graph with Other/Unknown Category**

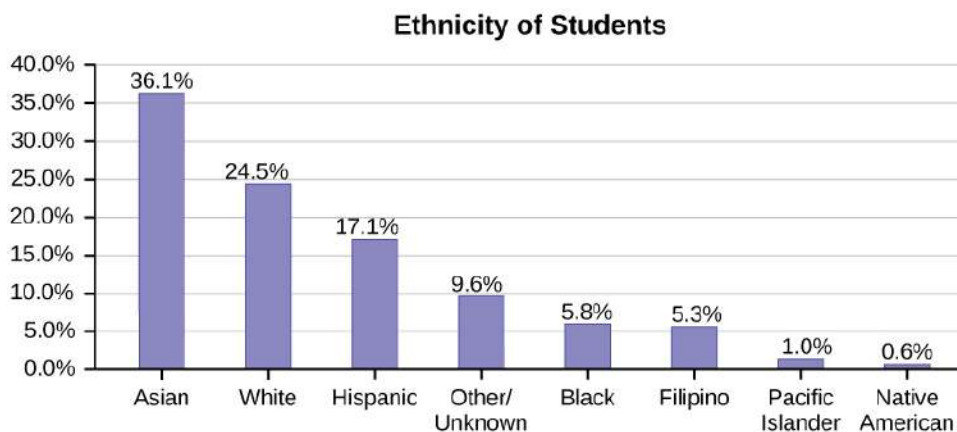


Figure 1.10 Pareto Chart With Bars Sorted by Size

### Pie Charts: No Missing Data

The following pie charts have the *Other/Unknown* category included since the percentages must add to 100 percent. The chart in **Figure 1.11b** is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in **Figure 1.11a**.

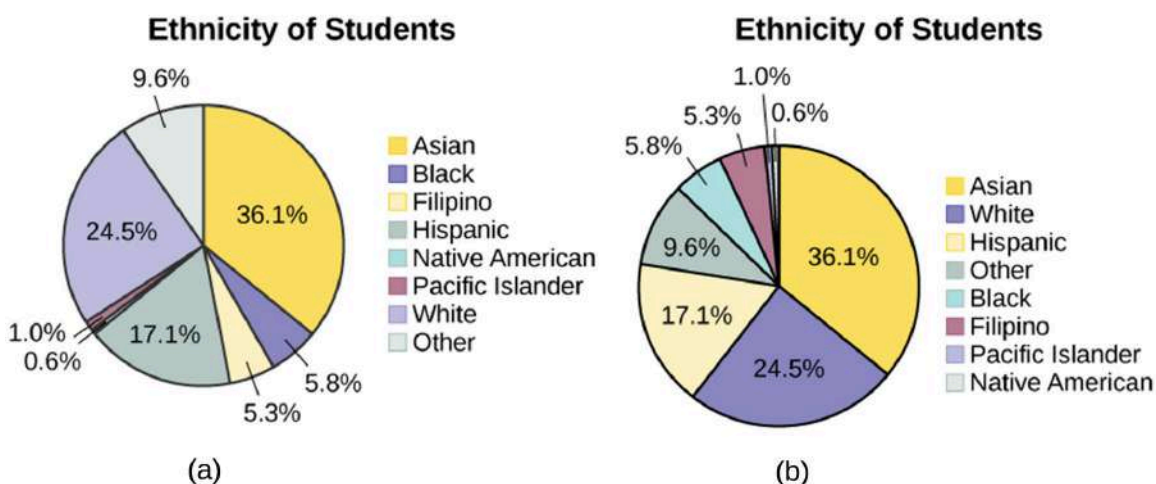


Figure 1.11

### Marginal Distributions in Two-Way Tables

Below is a *two-way table*, also called a contingency table, showing the favorite sports for 50 adults: 20 women and 30 men.

	Football	Basketball	Tennis	Total
Men	20	8	2	30
Women	5	7	8	20
Total	25	15	10	50

Table 1.5

This is a two-way table because it displays information about two categorical variables, in this case, gender and sports. Data

of this type (two variable data) are referred to as *bivariate data*. Because the data represent a count, or tally, of choices, it is a two-way *frequency* table. The entries in the total row and the total column represent marginal frequencies or *marginal distributions*. Note—The term marginal distributions gets its name from the fact that the distributions are found in the margins of frequency distribution tables. Marginal distributions may be given as a fraction or decimal: For example, the total for men could be given as .6 or  $3/5$  since  $30/50 = .6 = 3/5$ . Marginal distributions require bivariate data and only focus on *one* of the variables represented in the table. In other words, the reason 20 is a marginal frequency in this two-way table is because it represents the margin or portion of the total population that is women (20/50). The reason 25 is a marginal frequency is because it represents the portion of those sampled who favor football (25/50). Note: The values that make up the body of the table (e.g., 20, 8, 2) are called *joint frequencies*.

## Conditional Distributions in Two-Way Tables

The distinction between a marginal distribution and a *conditional distribution* is that the focus is on only a particular subset of the population (not the entire population). For example, in the table, if we focused only on the subpopulation of women who prefer football, then we could calculate the conditional distributions as shown in the two-way table below.

	Football	Basketball	Tennis	Total
Men	20	8	2	30
Women	5	7	8	20
Total	25	15	10	50

**Table 1.6**

To find the first sub-population of women who prefer football, read the value at the intersection of the Women row and Football column which is 5. Then, divide this by the total population of football players which is 25. So, the subpopulation of football players who are women is  $5/25$  which is .2.

Similarly, to find the subpopulation of women who play football, use the value of 5 which is the number of women who play football. Then, divide this by the total population of women which is 20. So, the subpopulation of women who play football is  $5/20$  which is .25.

## Presenting Data

After deciding which graph best represents your data, you may need to present your statistical data to a class or other group in an oral report or multimedia presentation. When giving an oral presentation, you must be prepared to explain exactly how you collected or calculated the data, as well as why you chose the categories, scales, and types of graphs that you are showing. Although you may have made numerous graphs of your data, be sure to use only those that actually demonstrate the stated intentions of your statistical study. While preparing your presentation, be sure that all colors, text, and scales are visible to the entire audience. Finally, make sure to allow time for your audience to ask questions and be prepared to answer them.

### Example 1.11

Suppose the guidance counselors at De Anza and Foothill need to make an oral presentation of the student data presented in Figures 1.5 and 1.6. Under what context should they choose to display the pie graph? When might they choose the bar graph? For each graph, explain which features they should point out and the potential display problems that might exist.

#### Solution 1.11

The guidance counselors should use the pie graph if the desired information is the percentage of each school's enrollment. They should use the bar graph if knowing the exact numbers of students and the relative sizes of each category at each school are important points to be made. For the pie graph, they should point out which color represents part-time students and which represents full-time students. They should also be sure that the numbers and colors are visible when displayed. For the bar graph, they should point out the scale and the total numbers for each category, and they should be sure that the numbers, colors, and scale marks are all displayed clearly.

## Try It

**1.11** Suppose you were asked to give an oral presentation of the data graphed in the pie chart in Figure 1.11(b). What features would you point out on the graph? What potential display problems with the graph should you check before giving your presentation?

## Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Each method has pros and cons. In a simple random sample, each group has the same chance of being selected. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in **Table 1.7**.

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	22	Roquero
01	Bautista	12	Legeny	23	Roth
02	Bayani	13	Lisa	24	Rowell
03	Cheng	14	Lundquist	25	Salangsang
04	Cuarismo	15	Macierz	26	Slade
05	Cunningham	16	Motogawa	27	Stratcher
06	Fontecha	17	Okimoto	28	Tallai
07	Hong	18	Patel	29	Tran
08	Hoobler	19	Price	30	Wai
09	Jiao	20	Quizon	31	Wood
10	Khan	21	Reyes		

**Table 1.7 Class Roster**

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. The most common random number generators are five digit numbers where each digit is a unique number from 0 to 9. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

.94360, .99832, .14669, .51470, .40581, .73381, .04399.

Lisa reads two-digit groups until she has chosen three class members (That is, she reads .94360 as the groups 94, 43, 36, 60.) Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The table below shows how Lisa reads two-digit numbers from each random number. Each two-digit number in the table would represent each student in the roster above in **Table 1.7**.

Random number	Numbers read by Lisa			
.94360	94	43	36	60
.99832	99	98	83	32
.14669	14	46	66	69
.51470	51	14	47	70
.40581	40	05	58	81
.73381	73	33	38	81
.04399	04	39	39	99

**Table 1.8** Lisa randomly generated the decimals in the Random Number column. She then used each consecutive number in each decimal to make the numbers she read. Some of the read numbers correspond with the ID numbers given to the students in her class (e.g., 14 = Lundquist in **Table 1.7**)

The random numbers .94360 and .99832 do not contain appropriate two digit numbers. However the third random number, .14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Lundquist, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa's group will consist of Lundquist, Cuningham, and Cuarismo.



Using the TI-83, 83+, 84, 84+ Calculator

To generate random numbers perform the following steps:

- Press MATH.
- Arrow over to PRB.
- Press 5:randInt(0, 30).
- Press ENTER for the first random number.
- Press ENTER two more times for the other two random numbers. If there is a repeat press ENTER again.

Note—randInt(0, 30, 3) will generate three random numbers.

```
randInt(0,30)
29
randInt(0,30)
28
randInt(0,30)
4
```

**Figure 1.12**

Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a stratified sample, divide the population into groups called strata and then the sample is selected by picking the

same number of values from each strata until the desired sample size is reached. For example, you could stratify (group) your high school student population by year (freshmen, sophomore, juniors, and seniors) and then choose a proportionate simple random sample from each stratum (each year) to get a stratified random sample. To choose a simple random sample from each year, number each student of the first year, number each student of the second year, and do the same for the remaining years. Then use simple random sampling to choose proportionate numbers of students from the first year and do the same for each of the remaining years. Those numbers picked from the first year, picked from the second year, and so on represent the students who make up the stratified sample.

To choose a cluster sample, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four homeroom classes from your student population, the four classes make up the cluster sample. Each class is a cluster. Number each cluster, and then choose four different numbers using random sampling. All the students of the four classes with those numbers are the cluster sample. So, unlike a stratified example, a cluster sample may not contain an equal number of randomly chosen students from each class.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased. They may favor a certain group. It is better for the person conducting the survey to select the sample respondents.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, a **sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others. Remember, each member of the population should have an equally likely chance of being chosen. When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied. For instance, if a survey of all students is conducted only during noon lunchtime hours is biased. This is because the students who do not have a noon lunchtime would not be included.

## Critical Evaluation

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include the following:

- Problems with samples: —A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not reliable. Reliability in statistical measures must also be considered when analyzing data. Reliability refers to the consistency of a measure. A measure is reliable when the same results are produced given the same circumstances.
- Self-selected samples—Responses only by people who choose to respond, such as internet surveys, are often unreliable.
- Sample size issues—: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples include crash testing cars or medical testing for rare conditions.
- Undue influence—: collecting data or asking questions in a way that influences the response.
- Non-response or refusal of subject to participate: —The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: —A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies—: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data—: These can be improperly displayed graphs, incomplete data, or lack of context.



## Collaborative Exercise

As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

1. To find the average GPA of all students in a high school, use all honor students at the university as the sample.
2. To find out the most popular cereal among young people under the age of 10, stand outside a large supermarket for three hours and speak to every twentieth child under age 10 who enters the supermarket.
3. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.
4. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
5. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

### Example 1.12

A study is done to determine the average tuition that private high school students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the fall semester. What is the type of sampling in each case?

- a. A sample of 100 high school students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior) and then selecting 25 students from each.
- b. A random number generator is used to select a student from the alphabetical listing of all high school students in the fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
- c. A completely random method is used to select 75 students. Each high school student in the fall semester has the same probability of being chosen at any stage of the sampling process.
- d. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.
- e. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the fall semester. Those 100 students are the sample.

#### Solution 1.12

a. stratified, b. systematic, c. simple random, d. cluster, e. convenience

## Try It

**1.12** You are going to use the random number generator to generate different types of samples from the data.

This table displays six sets of quiz scores (each quiz counts 10 points) for an elementary statistics class.

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

**Table 1.9** Scores for quizzes #1-6 for 10 students in a statistics class. Each quiz is out of 10 points.

Instructions: Use the Random Number Generator to pick samples.

- Create a stratified sample by column. Pick three quiz scores randomly from each column.
  - Number each row one through 10.
  - On your calculator, press Math and arrow over to PRB.
  - For column 1, Press 5:randInt( and enter 1,10). Press ENTER. Record the number. Press ENTER 2 more times (even the repeats). Record these numbers. Record the three quiz scores in column one that correspond to these three numbers.
  - Repeat for columns two through six.
  - These 18 quiz scores are a stratified sample.
- Create a cluster sample by picking two of the columns. Use the column numbers: one through six.
  - Press MATH and arrow over to the PRB function.
  - Press 5:randInt (“and then enter “1,6). Press ENTER.
  - Record the number the calculator displays into the first column. Then, press ENTER.
  - Record the next number the calculator displays into the second column.
  - Repeat steps (c) and (d) nine more times until there are a total of 20 quiz scores for the cluster sample.
- Create a simple random sample of 15 quiz scores.
  - Use the numbering one through 60.
  - Press MATH. Arrow over to PRB. Press 5:randInt(1, 60).
  - Press ENTER 15 times and record the numbers.
  - Record the quiz scores that correspond to these numbers.
  - These 15 quiz scores are the systematic sample.
- Create a systematic sample of 12 quiz scores.
  - Use the numbering one through 60.
  - Press MATH. Arrow over to PRB. Press 5:randInt(1, 60).
  - Press ENTER. Record the number and the first quiz score. From that number, count ten quiz scores and



record that quiz score. Keep counting ten quiz scores and recording the quiz score until you have a sample of 12 quiz scores. You may wrap around (go back to the beginning).

### Example 1.13

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

- A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
- A pollster interviews all human resource personnel in five different high tech companies.
- A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
- A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
- A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
- A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on average.

#### Solution 1.13

a. stratified b. cluster c. stratified d. systematic e. simple random f. convenience

### Try It

**1.13** Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

### Example 1.14

Suppose ABC high school has 10,000 upperclassman (junior and senior level) students (the population). We are interested in the average amount of money a upperclassmen spends on books in the fall term. Asking all 10,000 upperclassmen is an almost impossible task.

Suppose we take two different samples.

First, we use convenience sampling and survey ten upperclassman students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

\$128, \$87, \$173, \$116, \$130, \$204, \$147, \$189, \$93, \$153.

The second sample is taken using a list of seniors who take P.E. classes and taking every fifth seniors on the list, for a total of ten seniors. They spend the following:

\$50, \$40, \$36, \$15, \$50, \$100, \$40, \$53, \$22, \$22.

It is unlikely that any student is in both samples.

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

#### Solution 1.14

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average part-time student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

#### Solution 1.14

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines. Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:

\$180, \$50, \$150, \$85, \$260, \$75, \$180, \$200, \$200, \$150.

c. Is the sample biased?

#### Solution 1.14

c. The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is *good enough* to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

## Try It

**1.14** A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. Twenty-four people said they'd prefer more talk shows, and 176 people said they'd prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

## Variation in Data

**Variation** is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8, 16.1, 15.2, 14.8, 15.8, 15.9, 16.0, 15.5.

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose.

This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

## Variation in Samples

It was mentioned previously that two or more **samples** from the same **population**, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their high school sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples, that is, the number of data values is increased, their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This is called **sampling variability**. In other words, it refers to how much a statistic varies from sample to sample within a population. The larger the sample size, the smaller the variability between samples will be. So, the large sample size makes for a better, more reliable statistic.

## Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200–1,500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, internet surveys are invariably biased, because people choose to respond or not.



## Collaborative Exercise

Divide into groups of two, three, or four. Your instructor will give each group one six-sided die. Try this experiment twice. Roll one fair die (six-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get in **Table 1.10** and **Table 1.11** (*frequency* is the number of times a particular face of the die occurs)

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

**Table 1.10 First Experiment  
(20 rolls)**

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

**Table 1.11 Second  
Experiment (20 rolls)**

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? Why or why not?

Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

## 1.3 | Frequency, Frequency Tables, and Levels of Measurement

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as possible.

### Answers and Rounding Off

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round off only the final answer. Do not round off any intermediate results, if possible. If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer. Expect that some of your answers will vary from the text due to rounding errors.

It is not necessary to reduce most fractions in this course. Especially in **Probability Topics**, the chapter on probability, it

is more helpful to leave an answer as an unreduced fraction.

## Levels of Measurement

The way a set of data is measured is called its **level of measurement**. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are as follows (from lowest to highest level):

- **Nominal scale level**
- **Ordinal scale level**
- **Interval scale level**
- **Ratio scale level**

Data that is measured using a **nominal scale** is **qualitative (categorical)**. Categories, colors, names, labels, and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. The data are the names of the companies that make smartphones, but there is no agreed upon order of these brands, even though people may have personal preferences. Nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data but there is a big difference. The ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.

Another example of using the ordinal scale is a cruise survey where the responses to questions about the cruise are *excellent*, *good*, *satisfactory*, and *unsatisfactory*. These responses are ordered from the most desired response to the least desired. But the differences between two pieces of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements,  $40^\circ$  is equal to  $100^\circ$  minus  $60^\circ$ . Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like  $-10^\circ\text{F}$  and  $-15^\circ\text{C}$  exist and are colder than 0.

Interval level data can be used in calculations, but one type of comparison cannot be done.  $80^\circ\text{C}$  is not four times as hot as  $20^\circ\text{C}$  (nor is  $80^\circ\text{F}$  four times as hot as  $20^\circ\text{F}$ ). There is no meaning to the ratio of 80 to 20 (or four to one).

Data that is measured using the **ratio scale** takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams are machine-graded.

The data can be put in order from lowest to highest 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points. Ratios can be calculated. The smallest score is 0. So 80 is four times 20. The score of 80 is four times better than the score of 20.

## Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3.

**Table 1.12** lists the different data values in ascending order and their frequencies.

DATA VALUE	FREQUENCY
2	3
3	5

**Table 1.12** Frequency Table of Student Work Hours

DATA VALUE	FREQUENCY
4	3
5	6
6	2
7	1

**Table 1.12** Frequency Table of Student Work Hours

A **frequency** is the number of times a value of the data occurs. According to **Table 1.12**, there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample, in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or .15
3	5	$\frac{5}{20}$ or .25
4	3	$\frac{3}{20}$ or .15
5	6	$\frac{6}{20}$ or .30
6	2	$\frac{2}{20}$ or .10
7	1	$\frac{1}{20}$ or .05

**Table 1.13** Frequency Table of Student Work Hours with Relative Frequencies

The sum of the values in the relative frequency column of **Table 1.13** is  $\frac{20}{20}$ , or 1.

**Cumulative relative frequency** is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in **Table 1.14**.

In the first row, the cumulative frequency is simply .15 because it is the only one. In the second row, the relative frequency was .25, so adding that to .15, we get a relative frequency of .40. Continue adding the relative frequencies in each row to get the rest of the column.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or .15	.15

**Table 1.14** Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	5	$\frac{5}{20}$ or .25	.15 + .25 = .40
4	3	$\frac{3}{20}$ or .15	.40 + .15 = .55
5	6	$\frac{6}{20}$ or .30	.55 + .30 = .85
6	2	$\frac{2}{20}$ or .10	.85 + .10 = .95
7	1	$\frac{1}{20}$ or .05	.95 + .05 = 1.00

**Table 1.14** Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

#### NOTE

Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

**Table 1.15** represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100} = .05$	.05
61.95–63.95	3	$\frac{3}{100} = .03$	.05 + .03 = .08
63.95–65.95	15	$\frac{15}{100} = .15$	.08 + .15 = .23
65.95–67.95	40	$\frac{40}{100} = .40$	.23 + .40 = .63
67.95–69.95	17	$\frac{17}{100} = .17$	.63 + .17 = .80
69.95–71.95	12	$\frac{12}{100} = .12$	.80 + .12 = .92
71.95–73.95	7	$\frac{7}{100} = .07$	.92 + .07 = .99
73.95–75.95	1	$\frac{1}{100} = .01$	.99 + .01 = 1.00
	<b>Total = 100</b>	<b>Total = 1.00</b>	

**Table 1.15** Frequency Table of Soccer Player Height

The data in this table have been **grouped** into the following intervals:

- 59.95–61.95 inches
- 61.95–63.95 inches
- 63.95–65.95 inches
- 65.95–67.95 inches
- 67.95–69.95 inches
- 69.95–71.95 inches
- 71.95–73.95 inches
- 73.95–75.95 inches

#### NOTE

This example is used again in **Descriptive Statistics**, where the method used to compute the intervals will be explained.

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

### Example 1.15

From **Table 1.15**, find the percentage of heights that are less than 65.95 inches.

#### Solution 1.15

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are  $5 + 3 + 15 = 23$  players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then  $\frac{23}{100}$  or 23 percent. This percentage is the cumulative relative frequency entry in the third row.



## Try It

**1.15** Table 1.16 shows the amount, in inches, of annual rainfall in a sample of towns.

Rainfall (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
2.95–4.97	6	$\frac{6}{50} = .12$	.12
4.97–6.99	7	$\frac{7}{50} = .14$	.12 + .14 = .26
6.99–9.01	15	$\frac{15}{50} = .30$	.26 + .30 = .56
9.01–11.03	8	$\frac{8}{50} = .16$	.56 + .16 = .72
11.03–13.05	9	$\frac{9}{50} = .18$	.72 + .18 = .90
13.05–15.07	5	$\frac{5}{50} = .10$	.90 + .10 = 1.00
	Total = 50	Total = 1.00	

**Table 1.16**

From Table 1.16, find the percentage of rainfall that is less than 9.01 inches.

### Example 1.16

From Table 1.15, find the percentage of heights that fall between 61.95 and 65.95 inches.

#### Solution 1.16

Add the relative frequencies in the second and third rows:  $.03 + .15 = .18$  or 18 percent.

## Try It

**1.16** From Table 1.16, find the percentage of rainfall that is between 6.99 and 13.05 inches.

### Example 1.17

Use the heights of the 100 male semiprofessional soccer players in Table 1.15. Fill in the blanks and check your answers.

- The percentage of heights that are from 67.95–71.95 inches is \_\_\_\_\_.
- The percentage of heights that are from 67.95–73.95 inches is \_\_\_\_\_.
- The percentage of heights that are more than 65.95 inches is \_\_\_\_\_.
- The number of players in the sample who are between 61.95 and 71.95 inches tall is \_\_\_\_\_.
- What kind of data are the heights?

- f. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

### Solution 1.17

- 29 percent
- 36 percent
- 77 percent
- 87
- quantitative continuous
- get rosters from each team and choose a simple random sample from each

## Try It

**1.17** From **Table 1.16**, find the number of towns that have rainfall between 2.95 and 9.01 inches.



## Collaborative Exercise

In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

- What percentage of the students in your class have no siblings?
- What percentage of the students have from one to three siblings?
- What percentage of the students have fewer than three siblings?

## Example 1.18

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10. **Table 1.17** was produced.

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	.1579
4	1	$\frac{1}{19}$	.2105

**Table 1.17** Frequency of Commuting Distances

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
5	3	$\frac{3}{19}$	.1579
7	2	$\frac{2}{19}$	.2632
10	3	$\frac{4}{19}$	.4737
12	2	$\frac{2}{19}$	.7895
13	1	$\frac{1}{19}$	.8421
15	1	$\frac{1}{19}$	.8948
18	1	$\frac{1}{19}$	.9474
20	1	$\frac{1}{19}$	1.0000

Table 1.17 Frequency of Commuting Distances

- Is the table correct? If it is not correct, what is wrong?
- True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
- What fraction of the people surveyed commute five or seven miles?
- What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

**Solution 1.18**

- No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
- False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read 1052, 01579, 02105, 03684, 04737, 06316, 07368, 07895, 08421, 09474, 1.0000.
- $\frac{5}{19}$
- $\frac{7}{19}$ ,  $\frac{12}{19}$ ,  $\frac{7}{19}$

**Try It**

**1.18** Table 1.16 represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

## Example 1.19

**Table 1.18** contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

**Table 1.18**

Answer the following questions:

- What is the frequency of deaths measured from 2006 through 2009?
- What percentage of deaths occurred after 2009?
- What is the relative frequency of deaths that occurred in 2003 or earlier?
- What is the percentage of deaths that occurred in 2004?
- What kind of data are the numbers of deaths?
- The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

### Solution 1.19

- 97,118 (11.8 percent)
- 41.6 percent
- $67,092/823,356$  or 0.081 or 8.1 percent
- 27.8 percent
- quantitative discrete
- quantitative continuous

## Try It

**1.19** Table 1.19 contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994–2011.

Year	Total Number of Crashes	Year	Total Number of Crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

**Table 1.19**

Answer the following questions:

- What is the frequency of deaths measured from 2000 through 2004?
- What percentage of deaths occurred after 2006?
- What is the relative frequency of deaths that occurred in 2000 or before?
- What is the percentage of deaths that occurred in 2011?
- What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

## 1.4 | Experimental Design and Ethics

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as speeding? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. In an experiment, there is the **explanatory variable** which affects the **response variable**. In a randomized experiment, the researcher manipulates the explanatory variable and then observes the response variable. Each value of the explanatory variable used in an experiment is called a **treatment**.

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments

randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

Confounding occurs when the effects of multiple factors on a response cannot be separated, for instance, if a student guesses on the even-numbered questions on an exam and sits in a favorite spot on exam day. Why does the student get a high test scores on the exam? It could be the increased study time or sitting in the favorite spot or both. Confounding makes it difficult to draw valid conclusions about the effect of each factor on the outcome. The way around this is to test several outcomes with one method (treatment). This way, we know which treatment really works.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing substances, researchers noted the following:

*Results showed that believing one had taken the substance resulted in [performance] times almost as fast as those associated with consuming the substance itself. In contrast, taking the substance without knowledge yielded no significant performance increment.<sup>[1]</sup>*

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment, a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill that contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment designed to reduce bias by hiding information. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

Sometimes, it is neither possible nor ethical for researchers to conduct experimental studies. For example, if you want to investigate whether malnutrition affects elementary school performance in children, it would not be appropriate to assign an experimental group to be malnourished. In these cases, **observational studies** or **surveys** may be used. In an observational study, the researcher does not directly manipulate the independent variable. Instead, he or she takes recordings and measurements of naturally occurring phenomena. By sorting these data into control and experimental conditions, the relationship between the dependent and independent variables can be drawn. In a survey, a researcher's measurements consist of questionnaires that are answered by the research participants.

### Example 1.20

Researchers want to investigate whether taking aspirin regularly reduces the risk of a heart attack. 400 men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.

Identify the following values for this study: population, sample, experimental units, explanatory variable, response variable, treatments.

#### Solution 1.20

The **population** is men aged 50 to 84.

The **sample** is the 400 men who participated.

The **experimental units** are the individual men in the study.

The **explanatory variable** is oral medication.

The **treatments** are aspirin and a placebo.

The **response variable** is whether a subject had a heart attack.

1. McClung, M. and Collins, D. (2007 June). "Because I know it will!" Placebo effects of an ergogenic aid on athletic performance. *Journal of Sport & Exercise Psychology*, 29(3), 382-94.

### Example 1.21

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

- Describe the explanatory and response variables in this study.
- What are the treatments?
- Identify any lurking variables that could interfere with this study.
- Is it possible to use blinding in this study?

#### Solution 1.21

- The explanatory variable is scent, and the response variable is the time it takes to complete the maze.
- There are two treatments: a floral-scented mask and an unscented mask.
- All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
- Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded, though. The researcher who is observing a subject will not know which mask is being worn.

### Example 1.22

A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

#### Solution 1.22

The explanatory variable is birth order. You cannot randomly assign a person's birth order. Random assignment eliminates the impact of lurking variables. When you cannot assign subjects to treatment groups at random, there will be differences between the groups other than the explanatory variable.

### Try It

**1.22** You are concerned about the effects of texting on driving performance. Design a study to test the response time of drivers while texting and while driving only. How many seconds does it take for a driver to respond when a leading car hits the brakes?

- Describe the explanatory and response variables in the study.
- What are the treatments?
- What should you consider when selecting participants?
- Your research partner wants to divide participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?
- Identify any lurking variables that could interfere with this study.
- How can blinding be used in this study?

## Ethics

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that “numbers don’t lie,” but the people who use numbers to support their claims often do.

A recent investigation of famous social psychologist, Diederik Stapel, has led to the retraction of his articles from some of the world’s top journals including, *Journal of Experimental Social Psychology*, *Social Psychology*, *Basic and Applied Social Psychology*, *British Journal of Social Psychology*, and the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Over the past two years, an extensive investigation involving three universities where Stapel has worked concluded that the psychologist is guilty of fraud on a colossal scale. Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

*Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct results that journals found attractive. “It was a quest for aesthetics, for beauty—instead of the truth,” he said. He described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud.*<sup>[2]</sup>

The committee investigating Stapel concluded that he is guilty of several practices including

- creating datasets, which largely confirmed the prior expectations,
- altering data in existing datasets,
- changing measuring instruments without reporting the change, and
- misrepresenting the number of experimental subjects.

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot.

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel’s fraud states that, “statistical flaws frequently revealed a lack of familiarity with elementary statistics.”<sup>[3]</sup> Many of Stapel’s co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don’t want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees known as **Institutional Review Boards (IRB)**. All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.

These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant’s name from the data record sufficient to protect privacy? Perhaps the person’s identity could be discovered from the data that remains. What happens if the study does not proceed as planned and risks arise that were not anticipated? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your cholesterol level. Once the sample has been tested, you expect the lab to dispose of the remaining blood. At that point the blood becomes biological waste. Does a

2. Bhattacharjee, Y. (2013, April 26). The mind of a con man. *The New York Times*. Retrieved from

[http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?\\_r=3&src=dayp&](http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?_r=3&src=dayp&)

3. Tilburg University. (2012, Nov. 28). Flawed science: the fraudulent research practices of social psychologist Diederik Stapel. Retrieved from [https://www.tilburguniversity.edu/upload/3ff904d7-547b-40ae-85fe-bea38e05a34a\\_Final%20report%20Flawed%20Science.pdf](https://www.tilburguniversity.edu/upload/3ff904d7-547b-40ae-85fe-bea38e05a34a_Final%20report%20Flawed%20Science.pdf).



researcher have the right to take it for use in a study?

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a [website \(http://openstaxcollege.org//40introone\)](http://openstaxcollege.org//40introone) dedicated to cataloging retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics will empower you to analyze statistical studies critically.

### Example 1.23

Describe the unethical behavior in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A researcher is collecting data in a community.

- She selects a block where she is comfortable walking because she knows many of the people living on the street.
- No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try to find residents at home.
- She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the neighborhood.

#### Solution 1.23

- By selecting a convenient sample, the researcher is intentionally selecting a sample that could be biased. Claiming that this sample represents the community is misleading. The researcher needs to select areas in the community at random.
- Intentionally omitting relevant data will create bias in the sample. Suppose the researcher is gathering information about jobs and child care. By ignoring people who are not home, she may be missing data from working families that are relevant to her study. She needs to make every effort to interview all members of the target sample.
- It is never acceptable to fake data. Even though the responses she uses are *real* responses provided by other participants, the duplication is fraudulent and can create bias in the data. She needs to work diligently to interview everyone on her route.

### Try It

**1.23** Describe the unethical behavior, if any, in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A study is commissioned to determine the favorite brand of fruit juice among teens in California.

- The survey is commissioned by the seller of a popular brand of apple juice.
- There are only two types of juice included in the study: apple juice and cranberry juice.
- Researchers allow participants to see the brand of juice as samples are poured for a taste test.
- Twenty-five percent of participants prefer Brand X, 33 percent prefer Brand Y and 42 percent have no preference between the two brands. Brand X references the study in a commercial saying “Most teens like Brand X as much as or more than Brand Y.”

## 1.5 | Data Collection Experiment

## 1.1 Data Collection Experiment

### Student Learning Outcomes

- The student will demonstrate the systematic sampling technique.
- The student will construct relative frequency tables.
- The student will interpret results and their differences from different data groupings.

### Movie Survey

Get a class roster/list. Randomly mark a person's name, and then mark every fourth name on the list until you get 12 names. You may have to go back to the start of the list. For each name marked, record the number of movies they saw at the theater last month.

### Order the Data

Complete the two relative frequency tables below using your class data.

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0			
1			
2			
3			
4			
5			
6			
7+			

**Table 1.20** Frequency of Number of Movies Viewed

Number of Movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0–1			
2–3			
4–5			
6–7+			

**Table 1.21** Frequency of Number of Movies Viewed

1. Using the tables, find the percent of data that is at most two. Which table did you use and why?
2. Using the tables, find the percent of data that is at most three. Which table did you use and why?
3. Using the tables, find the percent of data that is more than two. Which table did you use and why?
4. Using the tables, find the percent of data that is more than three. Which table did you use and why?

### Discussion Questions

1. Is one of the tables *more correct* than the other? Why or why not?
2. In general, how could you group the data differently? Are there any advantages to either way of grouping the data?
3. Why did you switch between tables, if you did, when answering the question above?

## 1.6 | Sampling Experiment

## 1.2 Sampling Experiment

### Student Learning Outcomes

- The student will demonstrate the simple random, systematic, stratified, and cluster sampling techniques.
- The student will explain the details of each procedure used.

In this lab, you will be asked to pick several random samples of restaurants. In each case, describe your procedure briefly, including how you might have used the random number generator, and then list the restaurants in the sample you obtained.

#### NOTE

The following section contains restaurants stratified by city into columns and grouped horizontally by entree cost (clusters).

### Restaurants Stratified by City and Entree Cost

Entree Cost	Under \$10	\$10 to under \$15	\$15 to under \$20	Over \$20
San Jose	El Abuelo Taq, Pasta Mia, Emma's Express, Bamboo Hut	Emperor's Guard, Creekside Inn	Agenda, Gervais, Miro's	Blake's, Eulipia, Hayes Mansion, Germania
Palo Alto	Senor Taco, Tuscan Garden, Taxi's	Ming's, P.A. Joe's, Stickney's	Scott's Seafood, Poolside Grill, Fish Market	Sundance Mine, Maddalena's, Sally's
Los Gatos	Mary's Patio, Mount Everest, Sweet Pea's, Andele Taqueria	Lindsey's, Willow Street	Toll House	Charter House, La Maison Du Cafe
Mountain View	Maharaja, New Ma's, Thai-Rific, Garden Fresh	Amber Indian, La Fiesta, Fiesta del Mar, Dawit	Austin's, Shiva's, Mazeh	Le Petit Bistro
Cupertino	Hobees, Hung Fu, Samrat, China Express	Santa Barb. Grill, Mand. Gourmet, Bombay Oven, Kathmandu West	Fontana's, Blue Pheasant	Hamasushi, Helios
Sunnyvale	Chekijababi, Taj India, Full Throttle, Tia Juana, Lemon Grass	Pacific Fresh, Charley Brown's, Cafe Cameroon, Faz, Aruba's	Lion & Compass, The Palace, Beau Sejour	
Santa Clara	Rangoli, Armadillo Willy's, Thai Pepper, Pasand	Arthur's, Katie's Cafe, Pedro's, La Galleria	Birk's, Truya Sushi, Valley Plaza	Lakeside, Mariani's

Table 1.22 Restaurants Used in Sample

### A Simple Random Sample

Pick a **simple random sample** of 15 restaurants.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____
2. _____	7. _____	12. _____
3. _____	8. _____	13. _____
4. _____	9. _____	14. _____
5. _____	10. _____	15. _____

Table 1.23

## A Systematic Sample

Pick a **systematic sample** of 15 restaurants.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____
2. _____	7. _____	12. _____
3. _____	8. _____	13. _____
4. _____	9. _____	14. _____
5. _____	10. _____	15. _____

Table 1.24

## A Stratified Sample

Pick a **stratified sample**, by city, of 20 restaurants. Use 25 percent of the restaurants from each stratum. Round to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____
2. _____	7. _____	12. _____	17. _____
3. _____	8. _____	13. _____	18. _____
4. _____	9. _____	14. _____	19. _____
5. _____	10. _____	15. _____	20. _____

Table 1.25

## A Stratified Sample

Pick a **stratified sample**, by entree cost, of 21 restaurants. Use 25 percent of the restaurants from each stratum. Round

to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____
2. _____	7. _____	12. _____	17. _____
3. _____	8. _____	13. _____	18. _____
4. _____	9. _____	14. _____	19. _____
5. _____	10. _____	15. _____	20. _____
			21. _____

**Table 1.26**

## A Cluster Sample

Pick a **cluster sample** of restaurants from two cities. The number of restaurants will vary.

1. Describe your procedure.
2. Complete the table with your sample.

1. _____	6. _____	11. _____	16. _____	21. _____
2. _____	7. _____	12. _____	17. _____	22. _____
3. _____	8. _____	13. _____	18. _____	23. _____
4. _____	9. _____	14. _____	19. _____	24. _____
5. _____	10. _____	15. _____	20. _____	25. _____

**Table 1.27**

## KEY TERMS

**average** also called mean; a number that describes the central tendency of the data

**blinding** not telling participants which treatment a subject is receiving

**categorical variable** variables that take on values that are names or labels

**cluster sampling** a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters; every individual in the chosen clusters is included in the sample

**continuous random variable** a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV

**control group** a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups

**convenience sampling** a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data

**cumulative relative frequency** the term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value

**data** a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

**discrete random variable** a random variable (RV) whose outcomes are counted

**double-blinding** the act of blinding both the subjects of an experiment and the researchers who work with the subjects

**experimental unit** any individual or object to be measured

**explanatory variable** the independent variable in an experiment; the value controlled by researchers

**frequency** the number of times a value of the data occurs

**informed consent** any human subject in a research study must be cognizant of any risks or costs associated with the study; the subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits; consent must be given freely by an informed, fit participant

**institutional review board** a committee tasked with oversight of research programs that involve human subjects

**lurking variable** a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

**mathematical models** a description of a phenomenon using mathematical concepts, such as equations, inequalities, distributions, etc.

**nonsampling error** an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis

**numerical Variable** variables that take on values that are indicated by numbers

**observational study** a study in which the independent variable is not manipulated by the researcher

**parameter** a number that is used to represent a population characteristic and that generally cannot be determined easily

**placebo** an inactive treatment that has no real effect on the explanatory variable

**population** all individuals, objects, or measurements whose properties are being studied

**probability** a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

**proportion** the number of successes divided by the total number in the sample

**qualitative data** see **data**

**quantitative data** see **data**

**random assignment** the act of organizing experimental units into treatment groups using random methods

**random sampling** a method of selecting a sample that gives every member of the population an equal chance of being selected

**relative frequency** the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes

**reliability** the consistency of a measure; a measure is reliable when the same results are produced given the same circumstances

**representative sample** a subset of the population that has the same characteristics as the population

**response variable** the dependent variable in an experiment; the value that is measured for change at the end of an experiment

**sample** a subset of the population studied

**sampling bias** not all members of the population are equally likely to be selected

**sampling error** the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error

**sampling with replacement** once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual

**sampling without replacement** a member of the population may be chosen for inclusion in a sample only once; if chosen, the member is not returned to the population before the next selection

**simple random sampling** a straightforward method for selecting a random sample; give each member of the population a number  
Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample

**statistic** a numerical characteristic of the sample; a statistic estimates the corresponding population parameter

**statistical models** a description of a phenomenon using probability distributions that describe the expected behavior of the phenomenon and the variability in the expected observations

**stratified sampling** a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum

**survey** a study in which data is collected as reported by individuals.

**systematic sampling** a method for selecting a random sample; list the members of the population  
Use simple random sampling to select a starting point in the population. Let  $k = (\text{number of individuals in the population})/(\text{number of individuals needed in the sample})$ . Choose every  $k$ th individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample

**treatments** different values or components of the explanatory variable applied in an experiment

**validity** refers to how much a measure or conclusion accurately reflects real world



**variable** a characteristic of interest for each person or object in a population

## CHAPTER REVIEW

### 1.1 Definitions of Statistics, Probability, and Key Terms

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

### 1.2 Data, Sampling, and Variation in Data and Sampling

Data are individual items of information that come from a population or sample. Data may be classified as qualitative (categorical), quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

### 1.3 Frequency, Frequency Tables, and Levels of Measurement

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth. Expect that some of your answers will vary from the text due to rounding errors.

In addition to rounding your answers, you can measure your data using the following four levels of measurement:

- **Nominal scale level** data that cannot be ordered nor can it be used in calculations
- **Ordinal scale level** data that can be ordered; the differences cannot be measured
- **Interval scale level** data with a definite ordering but no starting point; the differences can be measured, but there is no such thing as a ratio
- **Ratio scale level** data with a starting point that can be ordered; the differences have meaning and ratios can be calculated

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these.

### 1.4 Experimental Design and Ethics

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

“An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule.”<sup>[4]</sup> Ethical violations in statistics are not always easy to spot. Professional

---

4. Gelman, A. (2013, May 1). Open data and open methods. *Ethics and Statistics*. Retrieved from <http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics1.pdf>.

associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

## PRACTICE

### 1.1 Definitions of Statistics, Probability, and Key Terms

1. Below is a two-way table showing the types of college sports played by men and women.

	Soccer	Basketball	Lacrosse	Total
Women	8	8	4	20
Men	4	12	4	20
Total	12	20	8	40

**Table 1.28**

Given these data, calculate the marginal distributions of college sports for the people surveyed.

2. Below is a two-way table showing the types of college sports played by men and women.

	Soccer	Basketball	Lacrosse	Total
Women	8	8	4	20
Men	4	12	4	20
Total	12	20	8	40

**Table 1.29**

Given these data, calculate the conditional distributions for the subpopulation of women who play college sports.

*Use the following information to answer the next five exercises.* Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new viral antibody drug is currently under study. It is given to patients once the virus's symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with the viral disease from the start of treatment until their deaths. The following data (in months) are collected.

#### Researcher A

3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

#### Researcher B

3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

Determine what the key terms refer to in the example for Researcher A.

3. population
4. sample
5. parameter
6. statistic
7. variable

## 1.2 Data, Sampling, and Variation in Data and Sampling

**8.** *Number of times per week* is what type of data?

- a. qualitative (categorical); b. quantitative discrete; c. quantitative continuous

*Use the following information to answer the next four exercises:* A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Antonio, Texas. The first house in the neighborhood around the park was selected randomly, and then the resident of every eighth house in the neighborhood around the park was interviewed.

**9.** The sampling method was

- a. simple random; b. systematic; c. stratified; d. cluster

**10.** *Duration (amount of time)* is what type of data?

- a. qualitative (categorical); b. quantitative discrete; c. quantitative continuous

**11.** The colors of the houses around the park are what kind of data?

- a. qualitative (categorical); b. quantitative discrete; c. quantitative continuous

**12.** The population is \_\_\_\_\_.

**13. Table 1.30** contains the total number of deaths worldwide as a result of earthquakes from 2000–2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
<b>Total</b>	<b>823,856</b>

**Table 1.30**

Use **Table 1.30** to answer the following questions.

- What is the proportion of deaths between 2007–2012?
- What percent of deaths occurred before 2001?
- What is the percent of deaths that occurred in 2003 or after 2010?
- What is the fraction of deaths that happened before 2012?
- What kind of data is the number of deaths?
- Earthquakes are quantified according to the amount of energy they produce (examples are 2.1, 5.0, 6.7). What type of data is that?
- What contributed to the large number of deaths in 2010? In 2004? Explain.
- If you were asked to present these data in an oral presentation, what type of graph would you choose to present and why? Explain what features you would point out on the graph during your presentation.

*For the following four exercises, determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).*

- A group of test subjects is divided into twelve groups; then four of the groups are chosen at random.
- A market researcher polls every tenth person who walks into a store.
- The first 50 people who walk into a sporting event are polled on their television preferences.
- A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.

*Use the following information to answer the next seven exercises:* Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new viral antibody drug is currently under study. It is given to patients once the virus's symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 patients with the viral disease from the start of treatment until their deaths. The following data (in months) are collected:

**Researcher A:** 3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

**Researcher B:** 3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

**18.** Complete the tables using the data provided.

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			
24.5–30.5			
30.5–36.5			
36.5–42.5			
42.5–48.5			

**Table 1.31 Researcher A**

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			
24.5–30.5			
30.5–36.5			
36.5–45.5			

**Table 1.32 Researcher B**

**19.** Determine what the key term data refers to in the above example for Researcher A.

**20.** List two reasons why the data may differ.

**21.** Can you tell if one researcher is correct and the other one is incorrect? Why?

**22.** Would you expect the data to be identical? Why or why not?

**23.** Suggest at least two methods the researchers might use to gather random data.

**24.** Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?

**25.** Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?

*Use the following data to answer the next five exercises:* Two researchers are gathering data on hours of video games played by school-aged children and young adults. They each randomly sample different groups of 150 students from the same school. They collect the following data:

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	26	.17	.17
2–4	30	.20	.37
4–6	49	.33	.70
6–8	25	.17	.87
8–10	12	.08	.95
10–12	8	.05	1

Table 1.33 Researcher A

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	48	.32	.32
2–4	51	.34	.66
4–6	24	.16	.82
6–8	12	.08	.90
8–10	11	.07	.97
10–12	4	.03	1

Table 1.34 Researcher B

- 26.** Give a reason why the data may differ.
- 27.** Would the sample size be large enough if the population is the students in the school?
- 28.** Would the sample size be large enough if the population is school-aged children and young adults in the United States?
- 29.** Researcher A concludes that most students play video games between four and six hours each week. Researcher B concludes that most students play video games between two and four hours each week. Who is correct?
- 30.** Suppose you were asked to present the data from researchers A and B in an oral presentation. When would a pie graph be appropriate? When would a bar graph more desirable? Explain which features you would point out on each type of graph and what potential display problems you would try to avoid.
- 31.** As part of a way to reward students for participating in the survey, the researchers gave each student a gift card to a video game store. Would this affect the data if students knew about the award before the study?

Use the following data to answer the next five exercises: A pair of studies was performed to measure the effectiveness of a new software program designed to help stroke patients regain their problem-solving skills. Patients were asked to use the software program twice a day, once in the morning, and once in the evening. The studies observed 200 stroke patients recovering over a period of several weeks. The first study collected the data in Table 1.35. The second study collected the data in Table 1.36.

Group	Showed Improvement	No Improvement	Deterioration
Used program	142	43	15
Did not use program	72	110	18

Table 1.35

Group	Showed Improvement	No Improvement	Deterioration
Used program	105	74	19
Did not use program	89	99	12

**Table 1.36**

32. Given what you know, which study is correct?
33. The first study was performed by the company that designed the software program. The second study was performed by the American Medical Association. Which study is more reliable?
34. Both groups that performed the study concluded that the software works. Is this accurate?
35. The company takes the two studies as proof that their software causes mental improvement in stroke patients. Is this a fair statement?
36. Patients who used the software were also a part of an exercise program whereas patients who did not use the software were not. Does this change the validity of the conclusions from **Exercise 1.34**?
37. Is a sample size of 1,000 a reliable measure for a population of 5,000?
38. Is a sample of 500 volunteers a reliable measure for a population of 2,500?
39. A question on a survey reads: "Do you prefer the delicious taste of Brand X or the taste of Brand Y?" Is this a fair question?
40. Is a sample size of two representative of a population of five?
41. Is it possible for two experiments to be well run with similar sample sizes to get different data?

### 1.3 Frequency, Frequency Tables, and Levels of Measurement

42. What type of measure scale is being used? Nominal, ordinal, interval or ratio.
- High school soccer players classified by their athletic ability: superior, average, above average
  - Baking temperatures for various main dishes: 350, 400, 325, 250, 300
  - The colors of crayons in a 24-crayon box
  - Social security numbers
  - Incomes measured in dollars
  - A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied
  - Preferred TV shows: comedy, drama, science fiction, sports, news
  - Time of day on an analog watch
  - The distance in miles to the closest grocery store
  - The dates 1066, 1492, 1644, 1947, and 1944
  - The heights of 21–65-year-old women
  - Common letter grades: A, B, C, D, and F

### 1.4 Experimental Design and Ethics

43. Design an experiment. Identify the explanatory and response variables. Describe the population being studied and the experimental units. Explain the treatments that will be used and how they will be assigned to the experimental units. Describe how blinding and placebos may be used to counter the power of suggestion.
44. Discuss potential violations of the rule requiring informed consent.
- Inmates in a correctional facility are offered good behavior credit in return for participation in a study.
  - A research study is designed to investigate a new children's allergy medication.
  - Participants in a study are told that the new medication being tested is highly promising, but they are not told that only a small portion of participants will receive the new medication. Others will receive placebo treatments and traditional treatments.

## HOMEWORK

### 1.1 Definitions of Statistics, Probability, and Key Terms

**45.** For each of the following situations, indicate whether it would be best modeled with a mathematical model or a statistical model. Explain your answers.

- driving time from New York to Florida
- departure time of a commuter train at rush hour
- distance from your house to school
- temperature of a refrigerator at any given time
- weight of a bag of rice at the store

*For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.*

**46.** A fitness center is interested in the mean amount of time a client exercises in the center each week.

**47.** Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

**48.** A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

**49.** Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

**50.** A politician is interested in the proportion of voters in his district who think he is doing a new good job.

**51.** A marriage counselor is interested in the proportion of clients she counsels who stay married.

**52.** Political pollsters may be interested in the proportion of people who will vote for a particular cause.

**53.** A marketing company is interested in the proportion of people who will buy a particular product.

*Use the following information to answer the next three exercises:* A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

**54.** What is the population she is interested in?

- all Lake Tahoe Community College students
- all Lake Tahoe Community College English students
- all Lake Tahoe Community College students in her classes
- all Lake Tahoe Community College math students

**55.** Consider the following

$X$  = number of days a Lake Tahoe Community College math student is absent.

In this case,  $X$  is an example of which of the following?

- variable
- population
- statistic
- data

**56.** The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of which of the following?

- parameter
- data
- statistic
- variable

### 1.2 Data, Sampling, and Variation in Data and Sampling

*For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.*

**57.** number of tickets sold to a concert



- 58. percent of body fat
- 59. favorite baseball team
- 60. time in line to buy groceries
- 61. number of students enrolled at Evergreen Valley College
- 62. most-watched television show
- 63. brand of toothpaste
- 64. distance to the closest movie theatre
- 65. age of executives in Fortune 500 companies
- 66. number of competing computer spreadsheet software packages

*Use the following information to answer the next two exercises:* A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

- 67. *Number of times per week* is what type of data?
  - a. qualitative
  - b. quantitative discrete
  - c. quantitative continuous
- 68. *Duration (amount of time)* is what type of data?
  - a. qualitative
  - b. quantitative discrete
  - c. quantitative continuous
- 69. Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.
  - a. Using complete sentences, list three things wrong with the way the survey was conducted.
  - b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.
- 70. Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.
- 71. Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.
- 72. List some practical difficulties involved in getting accurate results from a telephone survey.
- 73. List some practical difficulties involved in getting accurate results from a mailed survey.
- 74. With your classmates, brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.
- 75. The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is which of the following?
  - a. cluster sampling
  - b. stratified sampling
  - c. simple random sampling
  - d. convenience sampling
- 76. A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was which of the following?
  - a. simple random
  - b. systematic
  - c. stratified
  - d. cluster

**77.** Name the sampling method used in each of the following situations:

- A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.
- A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.
- The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.
- The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.
- A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

**78.** A random survey was conducted of 3,274 people of the *microprocessor generation*—people born since 1971, the year the microprocessor was invented. It was reported that 48 percent of those individuals surveyed stated that if they had \$2,000 to spend, they would use it for computer equipment. Also, 66 percent of those surveyed considered themselves relatively savvy computer users.

- Do you consider the sample size large enough for a study of this type? Why or why not?
- Based on your *gut feeling*, do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?  
Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."
- With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
- With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

**79.** The Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below.

Identify the type of data obtained from each question used in this survey: qualitative, quantitative discrete, or quantitative continuous.

- Do you have any health problems that prevent you from doing any of the things people your age can normally do?
- During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
- In the last seven days, on how many days did you exercise for 30 minutes or more?
- Do you have health insurance coverage?

**80.** In advance of the 1936 presidential election, a magazine released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.

- Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
- What effect does the low response rate have on the reliability of the sample?
- Are these problems examples of sampling error or nonsampling error?
- During the same year, another pollster conducted a poll of 30,000 prospective voters. These researchers used a method they called *quota sampling* to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

**81.** Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI's *Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates.

Which of the potential problems with samples discussed in **Data, Sampling, and Variation in Data and Sampling** could explain this connection?

**82.** A website that allows anyone to create and respond to polls had a question posted on April 15 which asked:

“Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?”<sup>[5]</sup>

As of April 25, 11 people responded to this question. Each participant answered “NO!”

Which of the potential problems with samples discussed in this module could explain this connection?

**83.** A scholarly article about response rates begins with the following quote:

“Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research.”<sup>[6]</sup>

The Pew Research Center for People and the Press admits

“The percentage of people we interview—out of all we try to interview—has been declining over the past decade or more.”<sup>[7]</sup>

- What are some reasons for the decline in response rate over the past decade?
- Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.

### 1.3 Frequency, Frequency Tables, and Levels of Measurement

**84.** Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below.

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	.6	
2	15		
3			

**Table 1.37** Part-time Student Course Loads

- Fill in the blanks in **Table 1.37**.
- What percent of students take exactly two courses?
- What percent of students take one or two courses?

5. lastbaldeagle. Retrieved from <http://www.youpolls.com/details.aspx?id=12328>.

6. Keeter, S., et al. (2006). Gauging the impact of growing nonresponse on estimates from a national RDD telephone survey. *Public Opinion Quarterly*, 70(5). Retrieved from <http://hbanaszak.mjr.uw.edu.pl/TempTxt/Links/GAUGING%20THE%20IMPACT%20OF%20GROWING.pdf>.

7. Pew Research Center. (n.d.). Frequently asked questions. Retrieved from <http://www.pewresearch.org/methodology/u-s-survey-research/frequently-asked-questions/#dont-you-have-trouble-getting-people-to-answer-your-polls>.

**85.** Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in **Table 1.38**.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	.4500	
1	18		
3			.9333
6	3	.0500	
7	1	.0167	

**Table 1.38 Flossing Frequency for Adults with Gum Disease**

- Fill in the blanks in **Table 1.38**.
- What percent of adults flossed six times per week?
- What percent flossed at most three times per week?

**86.** Nineteen immigrants to the United States were asked how many years, to the nearest year, they have lived in the United States. The data are as follows: 2, 5, 7, 2, 2, 10, 20, 15, 0, 7, 0, 20, 5, 12, 15, 12, 4, 5, 10.

**Table 1.39** was produced.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
0	2	$\frac{2}{19}$	.1053
2	3	$\frac{3}{19}$	.2632
4	1	$\frac{1}{19}$	.3158
5	3	$\frac{3}{19}$	.4737
7	2	$\frac{2}{19}$	.5789
10	2	$\frac{2}{19}$	.6842
12	2	$\frac{2}{19}$	.7895
15	1	$\frac{1}{19}$	.8421
20	1	$\frac{1}{19}$	1.0000

**Table 1.39 Frequency of Immigrant Survey Responses**

- Fix the errors in **Table 1.39**. Also, explain how someone might have arrived at the incorrect number(s).
- Explain what is wrong with this statement: “47 percent of the people surveyed have lived in the United States for 5 years.”
- Fix the statement in **b** to make it correct.
- What fraction of the people surveyed have lived in the United States five or seven years?
- What fraction of the people surveyed have lived in the United States at most 12 years?
- What fraction of the people surveyed have lived in the United States fewer than 12 years?
- What fraction of the people surveyed have lived in the United States from five to 20 years, inclusive?

**87.** How much time does it take to travel to work? **Table 1.40** shows the mean commute time by state for workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

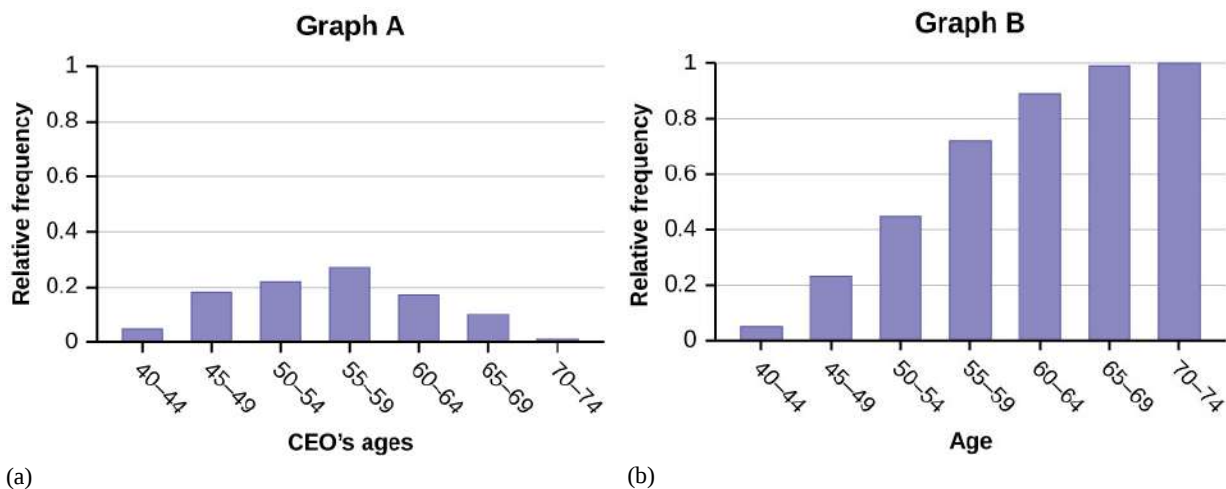
**Table 1.40**

**88.** A business magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least \$5 per share, and have reported annual revenue between \$5 million and \$1 billion. **Table 1.41** shows the ages of the chief executive officers for the first 60 ranked firms.

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

**Table 1.41**

- What is the frequency for CEO ages between 54 and 65?
- What percentage of CEOs are 65 years or older?
- What is the relative frequency of ages under 50?
- What is the cumulative relative frequency for CEOs younger than 55?
- Which graph shows the relative frequency and which shows the cumulative relative frequency?



**Figure 1.13**

Use the following information to answer the next two exercises: **Table 1.42** contains data on hurricanes that have made direct hits on the United States. Between 1851-2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Category	Number of Direct Hits	Relative Frequency	Cumulative Frequency
1	109	.3993	.3993
2	72	.2637	.6630
3	71	.2601	
4	18		.9890
5	3	.0110	1.0000
	<b>Total = 273</b>		

**Table 1.42 Frequency of Hurricane Direct Hits**

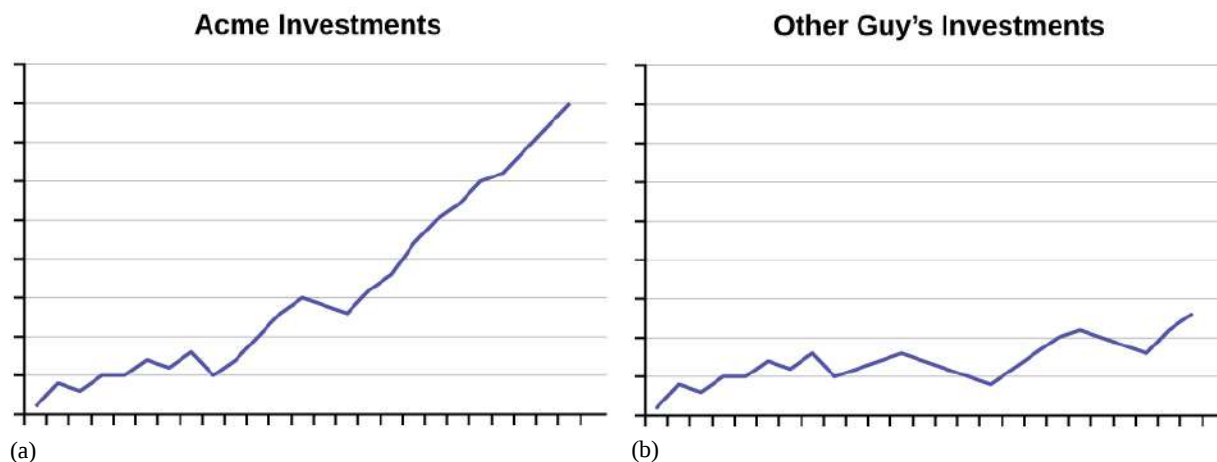
- 89.** What is the relative frequency of direct hits that were category 4 hurricanes?
- .0768
  - .0659
  - .2601
  - not enough information to calculate
- 90.** What is the relative frequency of direct hits that were AT MOST a category 3 storm?
- .3480
  - .9231
  - .2601
  - .3370

## 1.4 Experimental Design and Ethics

**91.** How does sleep deprivation affect your ability to drive? A recent study measured the effects on 19 professional drivers. Each driver participated in two experimental sessions: one after normal sleep and one after 27 hours of total sleep deprivation. The treatments were assigned in random order. In each session, performance was measured on a variety of tasks including a driving simulation.

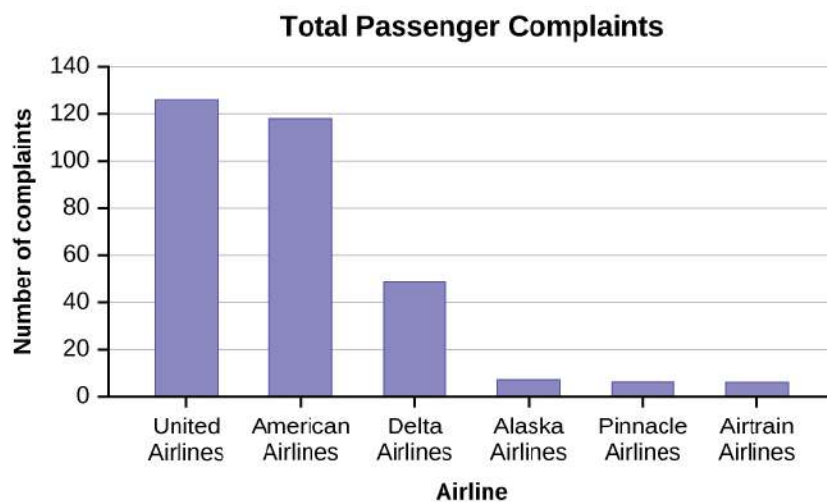
Use key terms from this module to describe the design of this experiment.

**92.** An advertisement for Acme Investments displays the two graphs in **Figure 1.14** to show the value of Acme's product in comparison with the Other Guy's product. Describe the potentially misleading visual effect of these comparison graphs. How can this be corrected?



**Figure 1.14** As the graphs show, Acme consistently outperforms the Other Guys!

**93.** The graph in **Figure 1.15** shows the number of complaints for six different airlines as reported to the U.S. Department of Transportation in February 2013. Alaska, Pinnacle, and Airtrain Airlines have far fewer complaints reported than American, Delta, and United. Can we conclude that American, Delta, and United are the worst airline carriers since they have the most complaints?



**Figure 1.15**

**94.** An epidemiologist is studying the spread of the common cold among college students. He is interested in how the temperature of the dorm room correlates with the incidence of new infections. How can he design an observational study to answer this question? If he chooses to use surveys in his measurements, what type of questions should he include in the survey?

## BRINGING IT TOGETHER: HOMEWORK

**95.** Seven hundred and seventy-one distance learning students at Long Beach City College responded to surveys in the 2010–11 academic year. Highlights of the summary report are listed in **Table 1.43**.

Have computer at home	96%
Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%
Live at least 16 miles from campus	13%
Took DL courses to fulfill transfer requirements	71%

**Table 1.43 LBCC Distance Learning Survey Results**

- What percent of the students surveyed do not have a computer at home?
- About how many students in the survey live at least 16 miles from campus?
- If the same survey were done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

**96.** Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the Internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the college term. If the book is not available, then a student would not be able to get the textbook at all, or might get a delayed delivery if the book is back ordered.

A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook for each of the following seven subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these seven textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers.

Write an analysis of his study that addresses the following issues: Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

## REFERENCES

### 1.1 Definitions of Statistics, Probability, and Key Terms

The Data and Story Library. Retrieved from <http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html>.

### 1.2 Data, Sampling, and Variation in Data and Sampling

Gallup. Retrieved from <http://www.well-beingindex.com/>.

Gallup. Retrieved from <http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4>.

Gallup. Retrieved from <http://www.gallup.com/175196/gallup-healthways-index-methodology.aspx>.

Data from <http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President>.

LBCC Distance Learning (DL) Program. Retrieved from <http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus>.

Lusinchi, D. (2012). “President” Landon and the 1936 Literary Digest poll: Were automobile and telephone owners to blame? *Social Science History* 36(1), 23-54. Retrieved from <https://muse.jhu.edu/article/471582/pdf>.

The Data and Story Library. Retrieved from <http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html>.



The Mercury News. Retrieved from <http://www.mercurynews.com/>.

Virtual Laboratories in Probability and Statistics. Retrieved from <http://www.math.uah.edu/stat/data/LiteraryDigest.html>.

The Mercury News. Retrieved from <http://www.mercurynews.com/>.

### 1.3 Frequency, Frequency Tables, and Levels of Measurement

Levels of Measurement. Retrieved from <http://cnx.org/content/m10809/latest/>.

National Hurricane Center. Retrieved from <http://www.nhc.noaa.gov/gifs/table5.gif>.

ThoughtCo. Retrieved from <https://www.thoughtco.com/levels-of-measurement-in-statistics-3126349>.

U.S. Census Bureau. Retrieved from <https://www.census.gov/quickfacts/table/PST045216/00>.

Levels of measurement. Retrieved from [https://www.cos.edu/Faculty/georgew/Tutorial/Data\\_Levels.htm](https://www.cos.edu/Faculty/georgew/Tutorial/Data_Levels.htm).

### 1.4 Experimental Design and Ethics

Econoclass.com. Retrieved from <http://www.econoclass.com/misleadingstats.html>.

Bloomberg Businessweek. Retrieved from [www.businessweek.com](http://www.businessweek.com).

Ethics in statistics. Retrieved from <http://cnx.org/content/m15555/latest/>.

Forbes. Retrieved from [www.forbes.com](http://www.forbes.com).

Forbes. <http://www.forbes.com/best-small-companies/list/>.

Harvard School of Public Health. Retrieved from <https://www.hsph.harvard.edu/nutritionsource/vitamin-e/>.

Jacskon, M.L., et al. (2013). Cognitive components of simulated driving performance: Sleep loss effect and predictors. Accident Analysis and Prevention Journal 50, 438-44. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/22721550>.

International Business Times. Retrieved from <http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443>.

National Highway Traffic Safety Administration. Retrieved from <http://www-fars.nhtsa.dot.gov/Main/index.aspx>.

Athleteinme.com. Retrieved from <http://www.athleteinme.com/ArticleView.aspx?id=1053>.

The Data and Story Library. Retrieved from <http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html>.

U.S. Department of Health and Human Services. Retrieved from <https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html>.

U.S. Department of Transportation. Retrieved from <http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report>.

U.S. Geological Survey. Retrieved from <http://earthquake.usgs.gov/earthquakes/eqarchives/year/>.

## SOLUTIONS

1 soccer =  $12/40 =$  ; basketball =  $20/40 =$  ; lacrosse =  $8/40 = 0.2$

2 women who play soccer =  $8/20 =$  ; women who play basketball =  $8/20 =$  ; women who play lacrosse =  $4/20 =$  ;

3 patients with the virus

5 The average length of time (in months) patients live after treatment.

7  $X$  = the length of time (in months) patients live after treatment

9 b

11 a

13

a. .5242

b. .03 percent

- c. 6.86 percent
  - d.  $\frac{823,088}{823,856}$
  - e. quantitative discrete
  - f. quantitative continuous
  - g. In both years, underwater earthquakes produced massive tsunamis.
  - h. Answers may vary. Sample answer: A bar graph with one bar for each year, in order, would be best since it would show the change in the number of deaths from year to year. In my presentation, I would point out that the scale of the graph is in thousands, and I would discuss which specific earthquakes were responsible for the greatest numbers of deaths in those years.
- 15** systematic
- 17** simple random
- 19** values for  $X$ , such as 3, 4, 11, and so on
- 21** No, we do not have enough information to make such a claim.
- 23** Take a simple random sample from each group. One way is by assigning a number to each patient and using a random number generator to randomly select patients.
- 25** This would be convenience sampling and is not random.
- 27** Yes, the sample size of 150 would be large enough to reflect a population of one school.
- 29** Even though the specific data support each researcher's conclusions, the different results suggest that more data need to be collected before the researchers can reach a conclusion.
- 30** Answers may vary. Sample answer: A pie graph would be best for showing the percentage of students that fall into each *Hours Played* category. A bar graph would be more desirable if knowing the total numbers of students in each category is important. I would be sure that the colors used on the two pie graphs are the same for each category and are clearly distinguishable when displayed. The percentages should be legible, and the pie graph should be large enough to show the smaller sections clearly. For the bar graph, I would display the bars in chronological order and make sure that the colors used for each researcher's data are clearly distinguishable. The numbers and the scale should be legible and clear when the bar graph is displayed.
- 32** There is not enough information given to judge if either one is correct or incorrect.
- 34** The software program seems to work because the second study shows that more patients improve while using the software than not. Even though the difference is not as large as that in the first study, the results from the second study are likely more reliable and still show improvement.
- 36** Yes, because we cannot tell if the improvement was due to the software or the exercise; the data is confounded, and a reliable conclusion cannot be drawn. New studies should be performed.
- 38** No, even though the sample is large enough, the fact that the sample consists of volunteers makes it a self-selected sample, which is not reliable.
- 40** No, even though the sample is a large portion of the population, two responses are not enough to justify any conclusions. Because the population is so small, it would be better to include everyone in the population to get the most accurate data.
- 42**
- a. ordinal
  - b. interval
  - c. nominal
  - d. nominal
  - e. ratio
  - f. ordinal
  - g. nominal

- h. interval
- i. ratio
- j. interval
- k. ratio
- l. ordinal

**44**

- a. Inmates may not feel comfortable refusing participation, or may feel obligated to take advantage of the promised benefits. They may not feel truly free to refuse participation.
- b. Parents can provide consent on behalf of their children, but children are not competent to provide consent for themselves.
- c. All risks and benefits must be clearly outlined. Study participants must be informed of relevant aspects of the study in order to give appropriate consent.

**45**

- a. statistical model: The time any journey takes from New York to Florida is variable and depends on traffic and other driving conditions.
- b. statistical model: Although trains try to leave on time, the exact time of departure differs slightly from day to day.
- c. mathematical model: The distance from your house to school is the same every day and can be precisely determined.
- d. statistical model: The temperature of a refrigerator fluctuates as the compressor turns on and off.
- e. statistical model: The fill weight of a bag of rice is different for each bag. Manufacturers spend considerable effort to minimize the variance from bag to bag.

**47**

- a. all children who take ski or snowboard lessons
- b. a group of these children
- c. the population mean age of children who take their first snowboard lesson
- d. the sample mean age of children who take their first snowboard lesson
- e.  $X$  = the age of one child who takes his or her first ski or snowboard lesson
- f. values for  $X$ , such as 3, 7, and so on

**49**

- a. the clients of the insurance companies
- b. a group of the clients
- c. the mean health costs of the clients
- d. the mean health costs of the sample
- e.  $X$  = the health costs of one client
- f. values for  $X$ , such as 34, 9, 82, and so on

**51**

- a. all the clients of this counselor
- b. a group of clients of this marriage counselor
- c. the proportion of all her clients who stay married
- d. the proportion of the sample of the counselor's clients who stay married
- e.  $X$  = the number of couples who stay married
- f. yes, no

**53**

- a. all people (maybe in a certain geographic area, such as the United States)
- b. a group of the people
- c. the proportion of all people who will buy the product
- d. the proportion of the sample who will buy the product
- e.  $X$  = the number of people who will buy it
- f. buy, not buy

**55** a**57** quantitative discrete, 150**59** qualitative, Oakland A's**61** quantitative discrete, 11,234 students**63** qualitative, Crest**65** quantitative continuous, 47.3 years**67** b**69**

- a. The survey was conducted using six similar flights.  
The survey would not be a true representation of the entire population of air travelers.  
Conducting the survey on a holiday weekend will not produce representative results.
- b. Conduct the survey during different times of the year.  
Conduct the survey using flights to and from various locations.  
Conduct the survey on different days of the week.

**71** Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

**73** Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can't be sure who is responding. In addition, mailing lists can be incomplete.

**75** b**77** convenience; cluster; stratified ; systematic; simple random**79**

- a. qualitative
- b. quantitative discrete
- c. quantitative discrete
- d. qualitative

**81** Causality: The fact that two variables are related does not guarantee that one variable is influencing the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate. Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.

**83**

- a. Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed
- b. When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the

subject of the survey.

**85**

a.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	.4500	.4500
1	18	.3000	.7500
3	11	.1833	.9333
6	3	.0500	.9833
7	1	.0167	1

**Table 1.44**

b. 5.00 percent

c. 93.33 percent

**87** The sum of the travel times is 1,173.1. Divide the sum by 50 to calculate the mean value: 23.462. Because each state's travel time was measured to the nearest tenth, round this calculation to the nearest hundredth: 23.46.

**89** b

**91** Explanatory variable: amount of sleep

Response variable: performance measured in assigned tasks

Treatments: normal sleep and 27 hours of total sleep deprivation

Experimental Units: 19 professional drivers

Lurking variables: none – all drivers participated in both treatments

Random assignment: treatments were assigned in random order; this eliminated the effect of any *learning* that may take place during the first experimental session

Control/Placebo: completing the experimental session under normal sleep conditions

Blinding: researchers evaluating subjects' performance must not know which treatment is being applied at the time

**93** You cannot assume that the numbers of complaints reflect the quality of the airlines. The airlines shown with the greatest number of complaints are the ones with the most passengers. You must consider the appropriateness of methods for presenting data; in this case displaying totals is misleading.

**94** He can observe a population of 100 college students on campus. He can collect data about the temperature of their dorm rooms and track how many of them catch a cold. If he uses a survey, the temperature of the dorm rooms can be determined from the survey. He can also ask them to self-report when they catch a cold.

**96** Answers will vary. Sample answer: The sample is not representative of the population of all college textbooks. Two reasons why it is not representative are that he only sampled seven subjects and he only investigated one textbook in each subject. There are several possible sources of bias in the study. The seven subjects that he investigated are all in mathematics and the sciences; there are many subjects in the humanities, social sciences, and other subject areas, for example: literature, art, history, psychology, sociology, business, that he did not investigate at all. It may be that different subject areas exhibit different patterns of textbook availability, but his sample would not detect such results. He also looked only at the most popular textbook in each of the subjects he investigated. The availability of the most popular textbooks may differ from the availability of other textbooks in one of two ways:

- The most popular textbooks may be more readily available online, because more new copies are printed, and more students nationwide are selling back their used copies
- The most popular textbooks may be harder to find available online, because more student demand exhausts the supply more quickly.

In reality, many college students do not use the most popular textbook in their subject, and this study gives no useful information about the situation for those less popular textbooks. He could improve this study by

- expanding the selection of subjects he investigates so that it is more representative of all subjects studied by college