AP Statistics Summer Assignment

Due: First Day of School

Textbook: *The Practice of Statistics: Sixth Ed. Updated Version*

You will need the textbook (attached) to complete this assignment.

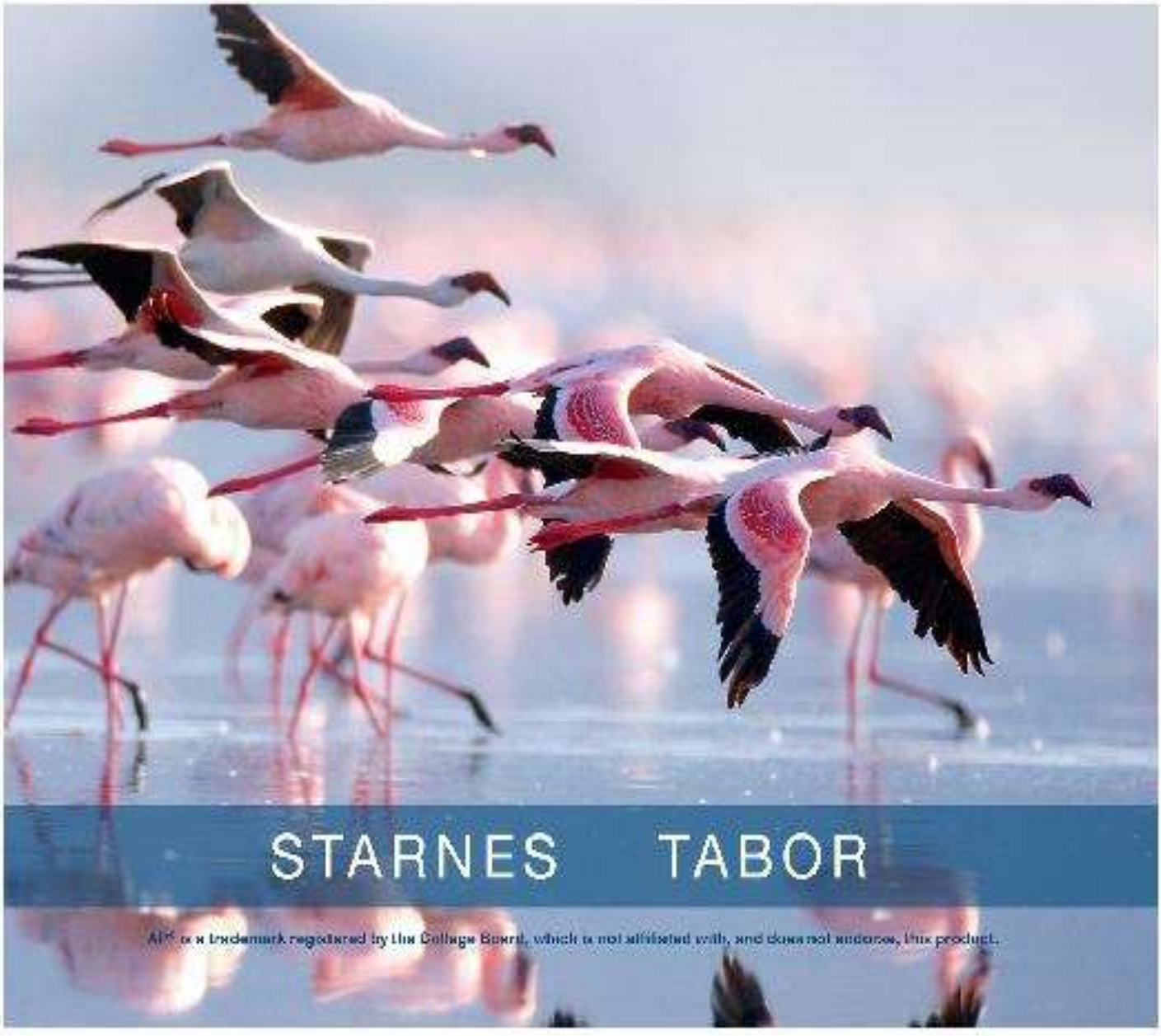
Assignment:

- Read the "Preface: To The Student" and the "Overview: What Is Statistics?"
    - Be prepared for a reading quiz covering the preface and the overview on the first day of class.
- Read the Chapter 1 Introduction
    - Complete Exercises #1-10
- Read Section 1.1: Analyzing Categorical Data
    - Complete Exercises #11, 13, 15, 17, 19, 21, 23, 24, 25, 27-39
    - You will have a quiz on Section 1.1 on Friday, August 11th.

for the AP® Exam

The Practice of Statistics

SIXTH EDITION



STARNES TABOR

To the Student

Statistical Thinking and You

The purpose of this book is to give you a working knowledge of the big ideas of statistics and of the methods used in solving statistical problems. Because data always come from a real-world context, doing statistics means more than just manipulating data. *The Practice of Statistics* (*TPS*), Sixth Edition, is full of data. Each set of data has some brief background to help you understand where the data come from. We deliberately chose contexts and data sets in the examples and exercises to pique your interest.

TPS 6e is designed to be easy to read and easy to use. This book is written by current high school AP® Statistics teachers, for high school students. We aimed for clear, concise explanations and a conversational approach that would encourage you to read the book. We also tried to enhance both the visual appeal and the book's clear organization in the layout of the pages.

Be sure to take advantage of all that *TPS* 6e has to offer. You can learn a lot by reading the text, but you will develop deeper understanding by doing the Activities and Projects and answering the Check Your Understanding questions along the way. The walkthrough guide on pages xiv–xx gives you an inside look at the important features of the text.

You learn statistics best by doing statistical problems. This book offers many different types of problems for you to tackle.

- **Section Exercises** include paired odd- and even-numbered problems that test the same skill or concept from that section. There are also some multiple-choice questions to help prepare you for the AP® Statistics exam. Recycle and Review exercises at the end of each exercise set involve material you studied in preceding sections.
- **Chapter Review Exercises** consist of free-response questions aligned to specific learning targets from the chapter. Go through the list of learning targets summarized in the Chapter Review and be sure you can say of each item on the list, "I can do that." Then prove it by solving some problems.
- The **AP® Statistics Practice Test** at the end of each chapter will help you prepare for in-class exams. Each test has about 10 multiple-choice questions and 3 free-response problems, very much in the style of the AP® Statistics exam.
- Finally, the **Cumulative AP® Practice Tests** after Chapters 4, 7, 10, and 12 provide challenging, cumulative multiple-choice and free-response questions like those you might find on a midterm, final, or the AP® Statistics exam.

The main ideas of statistics, like the main ideas of any important subject, took a long time to discover and thus take some time to master. The basic principle of learning them is to be persistent. Once you put it all together, statistics will help you make informed decisions based on data in your daily life.

TPS and AP® Statistics

The Practice of Statistics (*TPS*) was the first book written specifically for the Advanced Placement (AP®) Statistics course. Like the previous five editions, *TPS* 6e is organized to closely follow the AP® Statistics Course Description. Every item on the College Board's "Topic Outline" is covered thoroughly in the text. Visit the book's website at highschool.bfwpub.com/tps6e for a detailed alignment guide. The few topics in the book that go beyond the AP® Statistics syllabus are marked with an asterisk (*).

Most importantly, *TPS* 6e is designed to prepare you for the AP® Statistics exam. The author team has been involved in the AP® Statistics program since its early days. We have more than 40 years' combined experience teaching AP® Statistics and grading the AP® exam! Both of us have served as Question Leaders for more than 10 years, helping to write scoring rubrics for free-response questions. Including our Content Advisory Board and Supplements Team (page vi), we have extensive knowledge of how the AP® Statistics exam is developed and scored.

TPS 6e will help you get ready for the AP® Statistics exam throughout the course by:

- **Using terms, notation, formulas, and tables consistent with those found on the AP® Statistics exam.** Key terms are shown in bold in the text, and they are defined in the Glossary. Key terms also are cross-referenced in the Index. See page F-1 to find "Formulas for the AP® Statistics Exam," as well as Tables A, B, and C in the back of the book for reference.
- **Following accepted conventions from AP® Statistics exam rubrics when presenting model solutions.** Over the years, the scoring guidelines for free-response questions have become fairly consistent. We kept these guidelines in mind when writing the solutions that appear throughout *TPS* 6e. For example, the four-step State–Plan–Do–Conclude process that we use to complete inference problems in Chapters 8–12 closely matches the four-point AP® scoring rubrics.
- **Including AP® Exam Tips in the margin where appropriate.** We place exam tips in the margins as "on-the-spot" reminders of common mistakes and how to avoid them. These tips are collected and summarized in the About the AP® Exam and AP® Exam Tips appendix.
- **Providing over 1600 AP®-style exercises throughout the book.** Each chapter contains a mix of free-response and multiple-choice questions that are similar to those found on the AP® Statistics exam. At the start of each Chapter Wrap-Up, you will find a FRAPPY (Free Response AP® Problem, Yay!). Each FRAPPY gives you the chance to solve an AP®-style free-response problem based on the material in the chapter. After you finish, you can view

and critique two example solutions from the book's Student Site (highschool.bfwpub.com/tps6e). Then you can score your own response using a rubric provided by your teacher.

Turn the page for a tour of the text. See how to use the book to realize success in the course and on the AP® Statistics exam.

READ THE TEXT and use the book's features to help you grasp the big ideas.

SECTION 3.1 Scatterplots and Correlation

Read the **LEARNING TARGETS** at the beginning of each section. Focus on mastering these skills and concepts as you work through the chapter.

LEARNING TARGETS *By the end of the section, you should be able to:*

- Distinguish between explanatory and response variables for quantitative data.
- Make a scatterplot to display the relationship between two quantitative variables.
- Describe the direction, form, and strength of a relationship displayed in a scatterplot and identify unusual features.
- Interpret the correlation.
- Understand the basic properties of correlation, including how the correlation is influenced by outliers.
- Distinguish correlation from causation.

Most statistical studies examine data on more than one variable. Fortunately, analysis of relationships between two variables builds on the same tools we used to analyze one variable. The principles that guide our work also remain the same:

- Plot the data, then look for overall patterns and departures from those patterns.
- Add numerical summaries.
- When there's a regular overall pattern, use a simplified model to describe it.

Explanatory and Response Variables

In the "Candy grab" activity, the number of candies is the **response variable.** Hand span is the **explanatory variable** because we anticipate that knowing a student's hand span will help us predict the number of candies that student can grab.

DEFINITION **Response variable, Explanatory variable**

A **response variable** measures an outcome of a study. An **explanatory variable** may help predict or explain changes in a response variable.

It's important to learn the language of statistics. Take note of the green **DEFINITION** boxes that explain important vocabulary. Flip back to them to review key terms and their definitions, or turn to the Glossary/Glosario at the back of the book.

Scan the margins for the green notes, which represent the "voice of the teacher" giving helpful hints for being successful in the course.

You will often see explanatory variables called *independent variables* and response variables called *dependent variables*. Because the words *independent* and *dependent* have other meanings in statistics, we won't use them here.

It is easiest to identify explanatory and response variables when we initially specify the values of one variable to see how it affects another variable. For instance, to study the effect of alcohol on body temperature, researchers gave several different amounts of alcohol to mice. Then they measured the change in each mouse's body temperature 15 minutes later. In this case, amount of alcohol is the explanatory variable, and change in body temperature is the response variable. When we don't specify the values of either variable before collecting the data, there may or may not be a clear explanatory variable.

Read the **AP® EXAM TIPS.** They give advice on how to be successful on the AP® Statistics exam.

AP® EXAM TIP

When you are asked to *describe* the association shown in a scatterplot, you are expected to discuss the direction, form, and strength of the association, along with any unusual features, *in the context of the problem*. This means that you need to use both variable names in your description.

HOW TO DESCRIBE A SCATTERPLOT

To describe a scatterplot, make sure to address the following four characteristics in the context of the data:

- **Direction:** A scatterplot can show a positive association, nega association, or no association.
- **Form:** A scatterplot can show a linear form or a nonlinear for form is linear if the overall pattern follows a straight line. Oth form is nonlinear.
- **Strength:** A scatterplot can show a weak, moderate, or strong An association is strong if the points don't deviate much from identified. An association is weak if the points deviate quite a

Look for the boxes with the green bands. Some explain how to make graphs or set up calculations, while others recap important concepts.

Watch for **CAUTION ICONS.** They alert you to common mistakes that students often make.

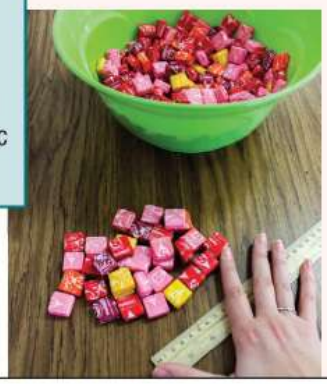
Few relationships are linear for all values of the explanatory variable. **Don't make predictions using values of x that are much larger or much smaller than those that actually appear in your data.**

LEARN STATISTICS BY *DOING* STATISTICS

Every chapter begins with a hands-on **ACTIVITY** that introduces the content of the chapter. Many of these activities involve collecting data and drawing conclusions from the data. In other activities, you'll use dynamic applets to explore statistical concepts.

ACTIVITY Candy grab

In this activity, you will investigate if students with a larger hand span can grab more candy than students with a smaller hand span.[1]



1. Measure the span of your dominant hand to the nearest half-centimeter (cm). Hand span is the distance from the tip of the thumb to the tip of the pinkie finger on your fully stretched-out hand.
2. One student at a time, go to the front of the class and use your dominant hand to grab as many candies as possible from the container. You must grab the candies with your fingers pointing down (no scooping!) and hold the candies for 2 seconds before counting them. After counting, put the candy back into the container.
3. On the board, record your hand span and number of candies in a table with the following headings:

Hand span (cm)	Number of candies

4. While other students record their values on the board, copy the table onto a piece of paper and make a graph. Begin by constructing a set of coordinate axes. Label the horizontal axis the vertical axis "Number of candies." scale for each axis and plot each point le as accurately as you can on the graph. ell you about the relationship between of candies? Summarize your observa- o.

ACTIVITY Guess the correlation

In this activity, we will have a class competition to see who can best guess the correlation.

1. Load the *Guess the Correlation* applet at www.rossmanchance.com/applets.
2. The teacher will press the "New Sample" button to see a "random" scatterplot. As a class, try to guess the correlation. Type the guess in the "Correlation guess" box and press "Check Guess" to see how the class did. Repeat several times to see more

Correlation Guessing Game

Number of points: 25
New Sample
Edit/Paste Data
Correlation guess: 0
Check Guess
Reset
Data
100

Chapters 1, 3, 4, 10, and 12 conclude with a **CHAPTER PROJECT.** Three of the projects (Chapters 1, 3, and 10) provide an opportunity to think like a statistician by analyzing larger data sets with multiple variables of interest. The other two (Chapters 4 and 12) are longer-term projects that require you to engage in the statistical problem-solving process: Ask Questions, Collect Data, Analyze Data, Interpret Results.

Chapter 3 Project Investigating Relationships in Baseball

What is a better predictor of the number of wins for a baseball team, the number of runs scored by the team or the number of runs they allow the other team to score? What variables can we use to predict the number of runs a team scores? To predict the number of runs it allows the other team to score? In this project, you will use technology to help answer these questions by exploring a large set of data from Major League Baseball.

Part 1

1. Download the "MLB Team Data 2012–2016" Excel file from the book's website, along with the "Glossary for MLB Team Data file," which explains each of the variables included in the data set.[44] Import the data into th cal software package you prefer.
2. Create a scatterplot to investigate the relationship runs scored per game (R/G) and wins (W). The late the equation of the least-squares regression standard deviation of the residuals, and r^2. *Note:* the section for hitting statistics and W is in the sc pitching statistics.
5. Because the number of wins a team has is dependent on both how many runs they score and how many runs they allow, we can use a combination of both variables to predict the number of wins. Add a column in your data table for a new variable, run differential. Fill in the values using the formula R/G – RA/G.
6. Create a scatterplot to investigate the relationship between run differential and wins. Then calculate the equation of the least-squares regression line, the standard deviation of the residuals, and r^2.
7. Is run differential a better predictor than the variable you chose in Question 4? Explain your reasoning.

CHECK YOUR UNDERSTANDING

In Exercises 3 and 7, we asked you to make and describe a scatterplot for the hiker data shown in the table.

Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28

1. Calculate the equation of the least-squares regression line.
2. Make a residual plot for the linear model in Question 1.
3. What does the residual plot indicate about the appropriateness of the linear model? Explain your answer.

CHECK YOUR UNDERSTANDING questions appear throughout the section. They help clarify definitions, concepts, and procedures. Be sure to check your answers in the back of the book.

EXAMPLES: Model statistical problems and how to solve them

Read through each **EXAMPLE**, and then try out the concept yourself by working the **FOR PRACTICE, TRY** exercise in the Section Exercises.

Need extra help? Examples and exercises marked with the **PLAY ICON** are supported by short video clips prepared by experienced AP® Statistics teachers. The video guides you through each step in the example and solution and provides additional explanation when you need it.

EXAMPLE Old Faithful and fertility

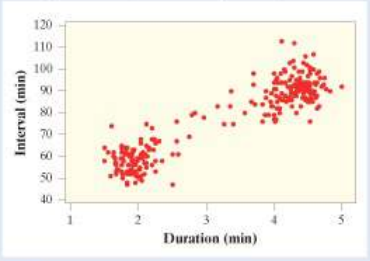
Describing a scatterplot

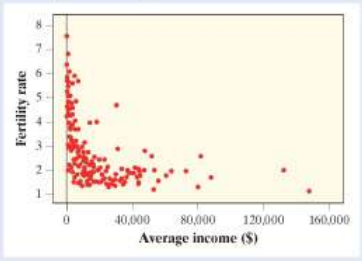
Bigshot73/iStock/Getty Images

PROBLEM: Describe the relationship in each of the following contexts.

(a) The scatterplot on the left shows the relationship between the duration (in minutes) of an eruption and the interval of time until the next eruption (in minutes) of Old Faithful during a particular month.

(b) The scatterplot on the right shows the relationship between the average income (gross domestic product per person, in dollars) and fertility rate (number of children per woman) in 187 countries.[4]



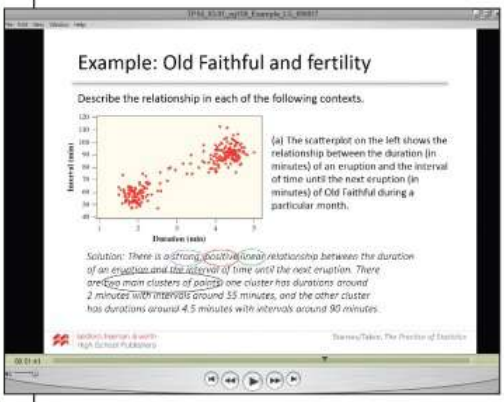


SOLUTION:

(a) There is a strong, positive linear relationship between the duration of an eruption and the interval of time until the next eruption. There are two main clusters of points: one cluster has durations around 2 minutes with in
has durations arc

(b) There is a modera
between average
is a potential out
fertility rate arou

Even with the clusters, the overall direction is still positive. In some cases, however, the points in a cluster



EXAMPLE Caffeine and pulse rates

How random assignment works

PROBLEM: A total of 20 students have agreed to participate in an experiment comparing the effects of caffeinated cola and caffeine-free cola on pulse rates. Describe how you would randomly assign 10 students to each of the two treatments:

(a) Using 20 identical slips of paper

(b) Using technology

(c) Using Table D

Zata Snively

SOLUTION:

(a) On 10 slips of paper, write the letter "A"; on the remaining 10 slips, write the letter "B." Shuffle the slips of paper and hand out one slip of paper to each volunteer. Students who get an "A" slip receive the cola with caffeine and students who get a "B" slip receive the cola without caffeine.

When describing a method of random assignment, don't stop after creating the groups. Make sure to identify which group gets which treatment.

(b) Label each student with a different integer from 1 to 20. Then randomly generate 10 different integers from 1 to 20. The students with these labels receive the cola with caffeine. The remaining 10 students receive the cola without caffeine.

When using a random number generator or a table of random digits to assign treatments, make sure to account for the possibility of repeated numbers when describing your method.

(c) Label each student with a different integer from 01 to 20. Go to a line of Table D and read two-digit groups moving from left to right. The first 10 different labels between 01 and 20 identify the 10 students who receive cola with caffeine. The remaining 10 students receive the caffeine-free cola. Ignore groups of digits from 21 to 00.

FOR PRACTICE, TRY EXERCISE 63

The **SOLUTION** is presented in a special font and models the style, steps, and language that you should use to earn full credit on the AP® Statistics exam.

THE VOICE OF THE TEACHER. Study the worked examples and pay special attention to the carefully placed **"Teacher Talk" comment boxes** that guide you step by step through the solution. These comments offer lots of good advice—as if your teacher is working directly with you to solve a problem.

The blue page number icon next to an exercise points you back to the page on which the model example appears.

63. Layoffs and "survivor guilt" Workers who survive a pg 251 layoff of other employees at their location may suffer from "survivor guilt." A study of survivor guilt and its effects used as subjects 120 students who were offered an opportunity to earn extra course credit by doing proofreading. Each subject worked in the same cubicle as another student, who was an accomplice of the experimenters. At a break midway through the work,

EXERCISES: Practice makes perfect!

Start by reading the **SECTION SUMMARY** to be sure that you understand the big ideas and key concepts.

Section 3.1 Summary

- A **scatterplot** displays the relationship between two quantitative variables measured on the same individuals. Mark values of one variable on the horizontal axis (x axis) and values of the other variable on the vertical axis (y axis). Plot each individual's data as a point on the graph.
- If we think that a variable x may help predict, explain, or even cause changes in another variable y, we call x an **explanatory variable** and y a **response variable.** Always plot the explanatory variable on the x axis of a scatterplot. Plot the response variable on the y axis.
- When describing a scatterplot, look for an overall pattern (direction, form, strength) and departures from the pattern (unusual features) and always answer in c
  - Directi values

Practice! Work the **EXERCISES** assigned by your teacher. Compare your answers to those in the Solutions appendix at the back of the book. Short solutions to the exercises numbered in red are found in the appendix.

Most of the **exercises are paired**, meaning that odd- and even-numbered exercises test the same skill or concept. If you answer an assigned exercise incorrectly, try to figure out your mistake. Then see if you can solve the paired exercise.

Look for **ICONS** that appear next to selected **EXERCISES**. They will guide you to

- the Example that models the exercise.
- videos that provide step-by-step instructions for solving the exercise.

Section 3.1 Exercises

1. **Coral reefs and cell phones** Identify the explanatory variable and the response variable for the following relationships, if possible. Explain your reasoning. pg 154

(a) The weight gain of corals in aquariums where the water temperature is controlled at different levels

(b) The number of text messages sent and the number of phone calls made in a sample of 100 students

2. **Teenagers and corn yield** Identify the explanatory variable and the response variable for the following relationships, if possible. Explain your reasoning.

(a) The height and arm span of a sample of 50 teenagers

(b) The yield of corn in bushels per acre and the amount of rain in the growing season

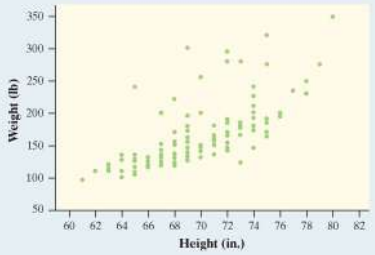
3. **Heavy backpacks** Ninth-grade students at the Webb Schools go on a backpacking trip each fall. Students are divided into hiking groups of size 8 by selecting names from a hat. Before leaving, students and their backpacks are weighed. The data here are from one hiking group. Make a scatterplot by hand that shows how backpack weight relates to body weight. pg 156

Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28

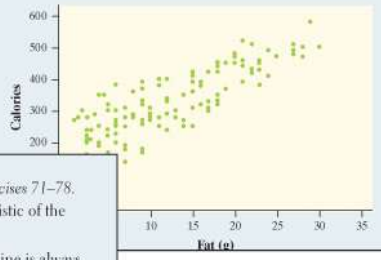
4. **Putting success** How well do professional golfers putt from various distances to the hole? The data show various distances to the hole (in feet) and the percent of putts mad
Make a scat
of putts mad

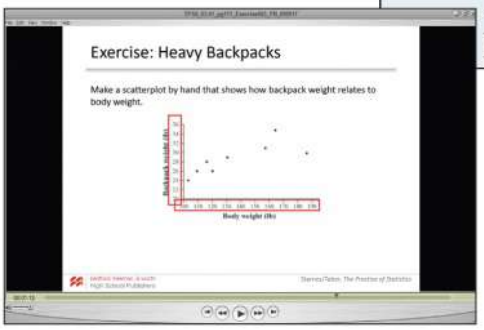
Distance (ft)	

Track and Field team.[10] Describe the relationship between height and weight for these athletes.



6. **Starbucks** The scatterplot shows the relationship between the amount of fat (in grams) and number of calories in products sold at Starbucks.[11] Describe the relationship between fat and calories for these products.





Multiple Choice: *Select the best answer for Exercises 71–78.*

71. Which of the following is *not* a characteristic of the least-squares regression line?

(a) The slope of the least-squares regression line is always between −1 and 1.

(b) The least-squares regression line always goes through the point $(\bar{x}, \bar{y})$.

(c) The least-squares regression line minimizes the sum of squared residuals.

(d) The slope of the least-squares regression line always have the same sign as the correlation.

(e) The least-squares regression line is not resistant

Recycle and Review

79. **Fuel economy** (2.2) In its recent *Fuel Economy Guide*, the Environmental Protection Agency (EPA) gives data on 1152 vehicles. There are a number of outliers, mainly vehicles with very poor gas mileage or hybrids with very good gas mileage. If we ignore the outliers, however, the combined city and highway gas mileage of the other 1120 or so vehicles is approximately Normal with mean 18.7 miles per gallon (mpg) and standard deviation 4.3 mpg.

(a) The Chevrolet Malibu with a four-cylinder engine has a combined gas mileage of 25 mpg. What percent of the 1120 vehicles have worse gas mileage than the Malibu?

Various types of problems in the Section Exercises let you practice solving many different types of questions, including AP®-style **multiple-choice** and **free-response**. The **Recycle and Review** exercises refer back to concepts and skills learned in an earlier section, noted in purple after the problem title.

REVIEW and PRACTICE for quizzes and tests and the AP® STATISTICS EXAM

Study the **CHAPTER REVIEW** to be sure that you understand the key concepts in each section.

Chapter 3 Wrap-Up

Chapter 3 Review

Section 3.1: Scatterplots and Correlation

In this section, you learned how to explore the relationship between two quantitative variables. As with distributions of a single variable, the first step is always to make a graph. A scatterplot is the appropriate type of graph to investigate relationships between two quantitative variables. To describe a scatterplot, be sure to discuss four characteristics: direction, form, strength, and unusual features. The direction of

Use the **WHAT DID YOU LEARN?** table that directs you to examples and exercises to verify your mastery of each **LEARNING TARGET**.

What Did You Learn?

Learning Target	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
Distinguish between explanatory and response variables for quantitative data.	3.1	154	R3.4
Make a scatterplot to display the relationship between two quantitative variables.	3.1	155	R3.4
Describe the direction, form, and strength of a relationship displayed in a scatterplot and identify unusual features.	3.1	158	R3.1 , R3.2
Interpret the correlation.	3.1	162	R3.3
Understand the basic properties of correlation, including how the correlation is influenced by outliers.	3.1	165, 169	R3.1, R3.2
Distinguish correlation from causation.	3		
Make predictions using regression lines, keeping in mind the dangers of extrapolation.	3		

SUMMARY TABLES in Chapters 8–11 review important details of each inference procedure, including conditions and formulas.

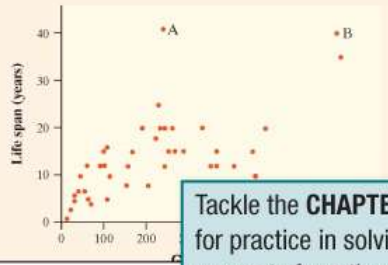
Comparing significance tests for a proportion and a mean

	Significance test for p	Significance test for μ
Name (TI-83/84)	One-sample z test for p (1-PropZTest)	One-sample t test for μ (TTest)
Formula	$z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$	$t = \dfrac{\bar{x} - \mu_0}{\dfrac{s_x}{\sqrt{n}}}$
	P-value from standard Normal distribution	P-value from t distribution with df $= n - 1$
Conditions	• **Random:** The data come from a random sample from the population of interest. ◦ **10%:** When sampling without replacement, $n < 0.10N$. • **Large Counts** Both np_0 and $n(1 - p_0)$ are at least 10. That is, the expected number of successes and the expected number of failures in the sample are both at least 10.	• **Random:** The data come from a random sample from the population of interest. ◦ **10%:** When sampling without replacement, $n < 0.10N$. • **Normal/Large Sample:** The population has a Normal distribution or the sample size is large ($n \geq 30$). If the population distribution has unknown shape and $n < 30$, use a graph of the sample data to assess the Normality of the population. Do not use t procedures if the graph shows strong skewness or outliers.

Chapter 3 Review Exercises

These exercises are designed to help you review the important ideas and methods of the chapter.

R3.1 **Born to be old?** Is there a relationship between the gestational period (time from conception to birth) of an animal and its average life span? The figure shows a scatterplot of the gestational period and average life span for 43 species of animals.[40]



(c) Point B is th
point have
squares regr
the residual

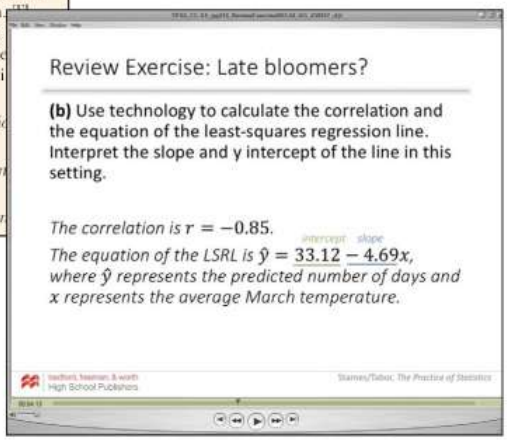
R3.2 **Penguins**
a relation
to seek fo
For all but the shallowest dives, there is an association between x = depth (in meters) and y = dive duration (in minutes) that is different for each penguin. study gives a scatterplot for one penguin titled Relation of Dive Duration (y) to Depth (x)." The terplot shows an association that is positive, li and strong.

(a) Explain the meaning of the term *positive associati* this context.

term *linear association*

term *strong association*

Tackle the **CHAPTER REVIEW EXERCISES** for practice in solving problems that test concepts from throughout the chapter. Need more help or just want additional insights before you take the practice test? Watch the **Chapter Review Exercise Videos**.



Chapter 3 AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question.*

T3.1 A school guidance counselor examines how many extracurricular activities students participate in and their grade point average. The guidance counselor says, "The evidence indicates that the correlation between the number of extracurricular activities a student participates in and his or her grade point average is close to 0." Which of the following is the most appropriate conclusion?

(a) Students involved in many extracurricular activities tend to be students with poor grades.

(b) Students with good grades tend to be students who are not involved in many extracurricular activities.

(c) Students involved in many extracurricular activities are just as likely to get good grades as bad grades.

(d) Students with good grades tend to be students who are involved in many extracurricular activities.

(e) No conclusion should be made based on the correlation without looking at a scatterplot of the data.

Questions T3.3–T3.5 refer to the following setting. Scientists examined the activity level of 7 fish at different temperatures. Fish activity was rated on a scale of 0 (no activity) to 100 (maximal activity). The temperature was measured in degrees Celsius. A computer regression printout and a residual plot are provided. Notice that the horizontal axis on the residual plot is labeled "Fitted value," which means the same thing as "predicted value."

Each chapter concludes with an **AP® STATISTICS PRACTICE TEST**. This test includes about 10 AP®-style multiple-choice questions and 3 free-response questions.

Cumulative AP® Practice Test 1

Section I: Multiple Choice *Choose the best answer for Questions AP1.1–AP1.14.*

AP1.1 You look at real estate ads for houses in Sarasota, Florida. Many houses have prices from $200,000 to $400,000. The few houses on the water, however, have prices up to $15 million. Which of the following statements best describes the distribution of home prices in Sarasota?

(a) The distribution is most likely skewed to the left, and the mean is greater than the median.

(b) The distribution is most likely skewed to the left, and the mean is less than the median.

(c) The distribution is roughly symmetric with a few high outliers, and the mean is approximately equal to the median.

(d) The distribution is most likely skewed to the right, and the mean is greater than the median.

(e) The distribution is most likely skewed to the right, and the mean is less than the median.

AP1.2 A child is 40 inches tall, which places her at the 90th percentile of all children of similar age. The heights for children of this age form an approximately Normal distribution with a mean of 38 inches. Based on this information, what is the standard deviation of the heights of all children of this age?

(a) 0.20 inch

(b) 0.31 inch

(c) 0.65 inch

(d) 1.21 inches

(e) 1.56 inches

Four **CUMULATIVE AP® PRACTICE TESTS** simulate the real exam. They are placed after Chapters 4, 7, 10, and 12. The tests expand in length and content coverage as you work through the book. The last test models a full AP® Statistics exam.

FRAPPY! FREE RESPONSE AP® PROBLEM, YAY!

The following problem is modeled after actual AP® Statistics exam free response questions. Your task is to generate a complete, concise response in 15 minutes.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

Two statistics students went to a flower shop and randomly selected 12 carnations. When they got home, the students prepared 12 identical vases with exactly the same amount of water in each vase. They put one tablespoon of sugar in 3 vases, two tablespoons of sugar in 3 vases, and three tablespoons of sugar in 3 vases. In the remaining 3 vases, they put no sugar. After the vases were prepared, the students randomly assigned 1 carnation to each vase and observed how many hours each flower continued to look fresh. A scatterplot of the data is shown below.

(a) Briefly describe the association shown in the scatterplot.

(b) The equation of the least-squares regression line for these data is $\hat{y} = 180.8 + 15.8x$. Interpret the slope of the line in the context of the study.

(c) Calculate and interpret the residual for the flower that had 2 tablespoons of sugar and looked fresh for 204 hours.

(d) Suppose that another group of students conducted a similar experiment using 12 flowers, but included different varieties in addition to carnations. Would you expect the value of r^2 for the second group's data to be greater than, less than, or about the same as the value of r^2 for the first group's data? Explain.

After you finish, you can view two example solutions on the book's website (highschool.bfwpub.com/tps6e). Determine whether you think each solution is "complete," "substantial," "developing," or "minimal." If the solution is not complete, what improvements would you suggest to the student who wrote it? Finally, your teacher will provide you with a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score.

Learn how to answer free response questions successfully by working the **FRAPPY!**—the Free Response AP® Problem, Yay!—that begins the Chapter Wrap-Up in every chapter.

Use TECHNOLOGY to discover and analyze

3. Technology Corner COMPUTING NUMERICAL SUMMARIES

TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/tps6e.

Let's find numerical summaries for the boys' shoes data from the example on page 64. We'll start by showing you how to compute summary statistics on the TI-83/84 and then look at output from computer software.

I. One-variable statistics on the TI-83/84

1. Enter the data in list L1.
2. Find the summary statistics for the shoe data.

- Press STAT (CALC); choose 1-VarStats.
  OS 2.55 or later: In the dialog box, press 2nd 1 (L1) and ENTER to specify L1 as the List. Leave FreqList blank. Arrow down to Calculate and press ENTER.
  Older OS: Press 2nd 1 (L1) and ENTER.
- Press ▼ to see the rest of the one-variable statistics.

II. Output from statistical software We used Minitab statistical software to calculate descriptive statistics for the boys' shoes data. Minitab allows you to choose which numerical summaries are included in the output.

```
Descriptive Statistics: Shoes

Variable  N   Mean  StDev  Minimum   Q1   Median   Q3   Maximum
Shoes    20  11.65  9.42     4.00   6.25   9.00  11.75   38.00
```

Note: The TI-83/84 gives the first and third quartiles of the boys' shoes distribution as $Q_1 = 6.5$ and $Q_3 = 11.5$. Minitab reports that $Q_1 = 6.25$ and $Q_3 = 11.75$. What happened? Minitab and some other software use slightly different rules for locating quartiles. Results from the various rules are usually close to each other. ating quartiles as they may affect more than just the *IQR*.

Use technology as a tool for discovery and analysis. **TECHNOLOGY CORNERS** give step-by-step instructions for using the TI-83/84 calculator. Instructions for the TI-Nspire and other calculators are on the book's Student Site (highschool.bfwpub.com/tps6e) and in the e-Book platform.

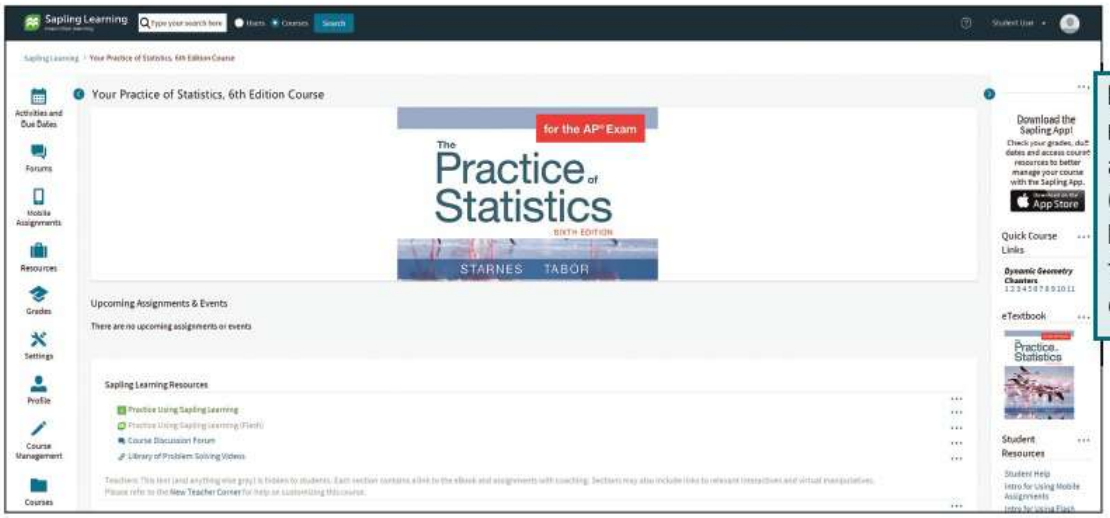
Technology Corner videos are also available to walk you through the key strokes needed to perform each analysis.

Although the Technology Corners focus on the TI-83/84 graphing calculator, output from multiple programs—including Minitab and JMP—is used in the book's Examples and Exercises to help you become familiar with reading and interpreting many different kinds of statistical summaries.

3.2 Technology Corners

TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/tps6e.

Find the Technology Corners easily by consulting the summary table at the end of each section or the complete table at the back of the book.



Read, practice, access the resources, and do homework assignments online with the new **Online Homework and e-Book Platform** that may be purchased to enhance your learning experience.

Overview: What Is Statistics?

Does listening to music while studying help or hinder learning? If an athlete fails a drug test, how sure can we be that she took a banned substance? Does having a pet help people live longer? How well do SAT scores predict college success? Do most people recycle? Which of two diets will help obese children lose more weight and keep it off? Can a new drug help people quit smoking? How strong is the evidence for global warming?

These are just a few of the questions that statistics can help answer. But what is statistics? And why should you study it?

Statistics Is the Science of Learning from Data

Data are usually numbers, but they are not "just numbers." *Data are numbers with a context.* The number 10.5, for example, carries no information by itself. But if we hear that a family friend's new baby weighed 10.5 pounds at birth, we congratulate her on the healthy size of the child. The context engages our knowledge about the world and allows us to make judgments. We know that a baby weighing 10.5 pounds is quite large, and that a human baby is unlikely to weigh 10.5 ounces or 10.5 kilograms. The context makes the number meaningful.

In your lifetime, you will be bombarded with data and statistical information. Poll results, television ratings, music sales, gas prices, unemployment rates, medical study outcomes, and standardized test scores are discussed daily in the media. Using data effectively is a large and growing part of most professions. A solid understanding of statistics will enable you to make sound, data-based decisions in your career and everyday life.

Data Beat Personal Experiences

It is tempting to base conclusions on your own experiences or the experiences of those you

know. But our experiences may not be typical. In fact, the incidents that stick in our memory are often the unusual ones.

Do Cell Phones Cause Brain Cancer?

Italian businessman Innocente Marcolini developed a brain tumor at age 60. He also talked on a cellular phone up to 6 hours per day for 12 years as part of his job. Mr. Marcolini's physician suggested that the brain tumor may have been caused by cell-phone use. So Mr. Marcolini decided to file suit in the Italian court system. A court ruled in his favor in October 2012.

Bloomberg via Getty Images

Several statistical studies have investigated the link between cell-phone use and brain cancer. One of the largest was conducted by the Danish Cancer Society. Over 350,000 residents of Denmark were included in the study. Researchers compared the brain-cancer rate for the cell-phone users with the rate in the general population. The result: no statistical difference in brain-cancer rates.[1] In fact, most studies have produced similar conclusions. In spite of the evidence, many people (like Mr. Marcolini) are still convinced that cell phones can cause brain cancer.

In the public's mind, the compelling story wins every time. A statistically literate person knows better. ***Data are more reliable than personal experiences because they systematically describe an overall picture, rather than focus on a few incidents.***

Where the Data Come from Matters

Are You Kidding Me?

The famous advice columnist Ann Landers once asked her readers, "If you had it to do over again, would you have children?" A few weeks later, her column was headlined "70% OF PARENTS SAY KIDS NOT WORTH IT." Indeed, 70% of the nearly 10,000 parents who wrote in said they would not have children if they could make the choice again. Do you believe that 70% of all parents regret having children?

You shouldn't. The people who took the trouble to write to Ann Landers are not representative of all parents. Their letters showed that many of them were angry with their children. All we know from these data is that there are some unhappy parents out there. A statistically designed poll, unlike Ann Landers's appeal, targets specific people chosen in a way that gives all parents the same chance to be asked. Such a poll showed that 91% of parents *would* have children again.

Where data come from matters a lot. If you are careless about how you get your data, you may announce 70% "No" when the truth is close to 90% "Yes."

Who Talks More—Women or Men?

According to Louann Brizendine, author of *The Female Brain*, women say nearly 3 times as many words per day as men. Skeptical researchers devised a study to test this claim. They used electronic devices to record the talking patterns of 396 university students from Texas, Arizona, and Mexico. The device was programmed to record 30 seconds of sound every 12.5 minutes without the carrier's knowledge. What were the results?

According to a published report of the study in *Scientific American*, "Men showed a slightly wider variability in words uttered. . . . But in the end, the sexes came out just about even in the daily averages: women at 16,215 words and men at 15,669."[2] When asked where she got her figures, Brizendine admitted that she used unreliable sources.[3]

The most important information about any statistical study is how the data were produced. Only carefully designed studies produce results that can be trusted.

Always Plot Your Data

Yogi Berra, a famous New York Yankees baseball player known for his unusual quotes, had this to say: "You can observe a lot just by watching." That's a motto for learning from data. ***A***

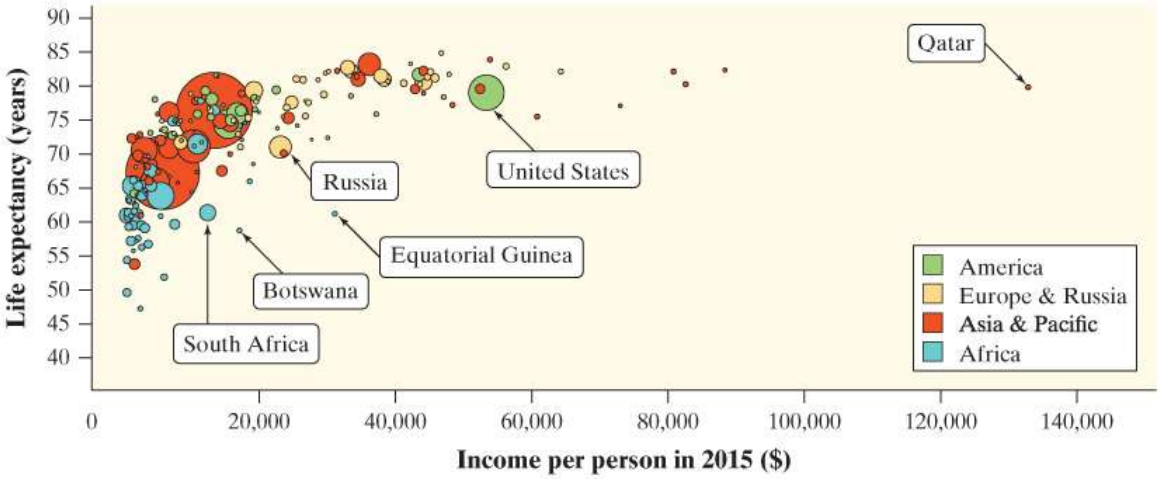
carefully chosen graph is often more instructive than a bunch of numbers.

Do People Live Longer in Wealthier Countries?

The Gapminder website, www.gapminder.org, provides loads of data on the health and well-being of the world's inhabitants. The graph below displays some data from Gapminder.[4] The individual points represent all the world's nations for which data are available. Each point shows the income per person and life expectancy for one country, along with the region (color of point) and population (size of point).

Hongqi Zhang/Alamy

We expect people in richer countries to live longer. The overall pattern of the graph does show this, but the relationship has an interesting shape. Life expectancy rises very quickly as personal income increases and then levels off. People in very rich countries like the United States live no longer than people in poorer but not extremely poor nations. In some less wealthy countries, people live longer than in the United States. Several other nations stand out in the graph. What's special about each of these countries?



Graph of the life expectancy of people in many nations against each nation's income per person in 2015.

Variation Is Everywhere

Individuals vary. Repeated measurements on the same individual vary. Chance outcomes—like spins of a roulette wheel or tosses of a coin—vary. Almost everything varies over time. Statistics provides tools for understanding variation.

Have Most Students Cheated on a Test?

Commercial Eye/Getty Images

Researchers from the Josephson Institute were determined to find out. So they surveyed about 23,000 students from 100 randomly selected schools (both public and private) nationwide. The question was: “How many times have you cheated during a test at school in the past year?” Fifty-one percent said they had cheated at least once.[5]

If the researchers had asked the same question of *all* high school students, would exactly 51% have answered “Yes”? Probably not. If the Josephson Institute had selected a different sample of about 23,000 students to respond to the survey, they would probably have gotten a different estimate. *Variation is everywhere!*

Fortunately, statistics provides a description of how the sample results will vary in relation to the actual population percent. Based on the sampling method that this study used, we can say that the estimate of 51% is very likely to be within 1% of the true population value. That is, we can be quite confident that between 50% and 52% of *all* high school students would say that they have cheated on a test.

Because variation is everywhere, conclusions are uncertain. Statistics gives us a language for talking about uncertainty that is understood by statistically literate people everywhere.

Chapter 1 Data Analysis

INTRODUCTION Statistics: The Science and Art of Data

LEARNING TARGETS *By the end of the section, you should be able to:*

- Identify the individuals and variables in a set of data.
- Classify variables as categorical or quantitative.

We live in a world of *data*. Every day, the media report poll results, outcomes of medical studies, and analyses of data on everything from stock prices to standardized test scores to global warming. The data are trying to tell us a story. To understand what the data are saying, you need to learn more about **statistics.**

DEFINITION Statistics

Statistics is the science and art of collecting, analyzing, and drawing conclusions from data.

A solid understanding of statistics will help you make good decisions based on data in your daily life.

Organizing Data

Every year, the U.S. Census Bureau collects data from over 3 million households as part of the American Community Survey (ACS). The table displays some data from the ACS in a recent year.

Rudy Sulgan/Corbis
Documentary/Getty Images

Household	Region	Number of people	Time in dwelling (years)	Response mode	Household income	Internet access?
425	Midwest	5	2–4	Internet	52,000	Yes
936459	West	4	2–4	Mail	40,500	Yes

50055	Northeast	2	10–19	Internet	481,000	Yes
592934	West	4	2–4	Phone	230,800	No
545854	South	9	2–4	Phone	33,800	Yes
809928	South	2	30+	Internet	59,500	Yes
110157	Midwest	1	5–9	Internet	80,000	Yes
999347	South	1	<1	Mail	8,400	No

Most data tables follow this format—each row describes an **individual** and each column holds the values of a **variable.**

DEFINITION **Individual, Variable**

An **individual** is an object described in a set of data. Individuals can be people, animals, or things.

A **variable** is an attribute that can take different values for different individuals.

Sometimes the individuals in a data set are called *cases* or *observational units*.

For the American Community Survey data set, the *individuals* are households. The *variables* recorded for each household are region, number of people, time in current dwelling, survey response mode, household income, and whether the dwelling has Internet access. Region, time in dwelling, response mode, and Internet access status are **categorical variables.** Number of people and household income are **quantitative variables.**

Note that household is *not* a variable. The numbers in the household column of the data table are just labels for the individuals in this data set. Be sure to look for a column of labels—names, numbers, or other identifiers—in any data table you encounter.

DEFINITION **Categorical variable, Quantitative variable**

A **categorical variable** assigns labels that place each individual into a particular group, called a category.

A **quantitative variable** takes number values that are quantities—counts or measurements.

Not every variable that takes number values is quantitative. Zip code is one example. Although zip codes are numbers, they are neither counts of anything, nor measurements of anything. They are simply labels for a regional location, making zip code a categorical variable. Some variables—such as gender, race, and occupation—are categorical by nature. Time in dwelling from the ACS data set is also a categorical variable because the values are recorded as intervals of time, such as 2–4 years. If time in dwelling had been recorded to

the nearest year for each household, this variable would be quantitative.

To make life simpler, we sometimes refer to *categorical data* or *quantitative data* instead of identifying the variable as categorical or quantitative.

EXAMPLE | Census At School

Individuals and Variables

Garry Black/Alamy

PROBLEM: Census At School is an international project that collects data about primary and secondary school students using surveys. Hundreds of thousands of students from Australia, Canada, Ireland, Japan, New Zealand, South Africa, South Korea, the United Kingdom, and the United States have taken part in the project. Data from the surveys are available online. We used the site's "Random Data Selector" to choose 10 Canadian students who completed the survey in a recent year. The table displays the data.

Province	Gender	Number of languages spoken	Handedness	Height (cm)	Wrist circumference (mm)	Preferred communication
Saskatchewan	Male	1	Right	175.0	180	In person
Ontario	Female	1	Right	162.5	160	In person
Alberta	Male	1	Right	178.0	174	Facebook
Ontario	Male	2	Right	169.0	160	Cell phone
Ontario	Female	2	Right	166.0	65	In person
Nunavut	Male	1	Right	168.5	160	Text messaging
Ontario	Female	1	Right	166.0	165	Cell phone
Ontario	Male	4	Left	157.5	147	Text messaging
Ontario	Female	2	Right	150.5	187	Text messaging
Ontario	Female	1	Right	171.0	180	Text messaging

a. Identify the individuals in this data set.

b. What are the variables? Classify each as categorical or quantitative.

SOLUTION:

a. 10 randomly selected Canadian students who participated in the Census At School survey.

We'll see in Chapter 4 why choosing at random, as we did in this example, is a good idea.

b. *Categorical: Province, gender, handedness, preferred communication method*
Quantitative: Number of languages spoken, height (cm), wrist circumference (mm)

There is at least one suspicious value in the data table. We doubt that the girl who is 166 cm tall really has a wrist circumference of 65 mm (about 2.6 inches). Always look to be sure the values make sense!

FOR PRACTICE, TRY EXERCISE 1

The proper method of data analysis depends on whether a variable is categorical or quantitative. For that reason, it is important to distinguish these two types of variables. The type of data determines what kinds of graphs and which numerical summaries are appropriate.

AP® EXAM TIP

If you learn to distinguish categorical from quantitative variables now, it will pay big rewards later. You will be expected to analyze categorical and quantitative variables correctly on the AP® exam.

ANALYZING DATA A variable generally takes values that vary (hence the name *variable*!). Categorical variables sometimes have similar counts in each category and sometimes don't. For instance, we might have expected similar numbers of males and females in the Census At School data set. But we aren't surprised to see that most students are right-handed. Quantitative variables may take values that are very close together or values that are quite spread out. We call the pattern of variation of a variable its **distribution.**

DEFINITION Distribution

The **distribution** of a variable tells us what values the variable takes and how often it takes those values.

Let's return to the data for the sample of 10 Canadian students from the preceding example. Figure 1.1(a) shows the distribution of preferred communication method for these students in a *bar graph.* We can see how many students chose each method from the heights of the bars: cell

phone (2), Facebook (1), in person (3), text messaging (4). Figure 1.1(b) shows the distribution of number of languages spoken in a *dotplot*. We can see that 6 students speak one language, 3 students speak two languages, and 1 student speaks four languages.

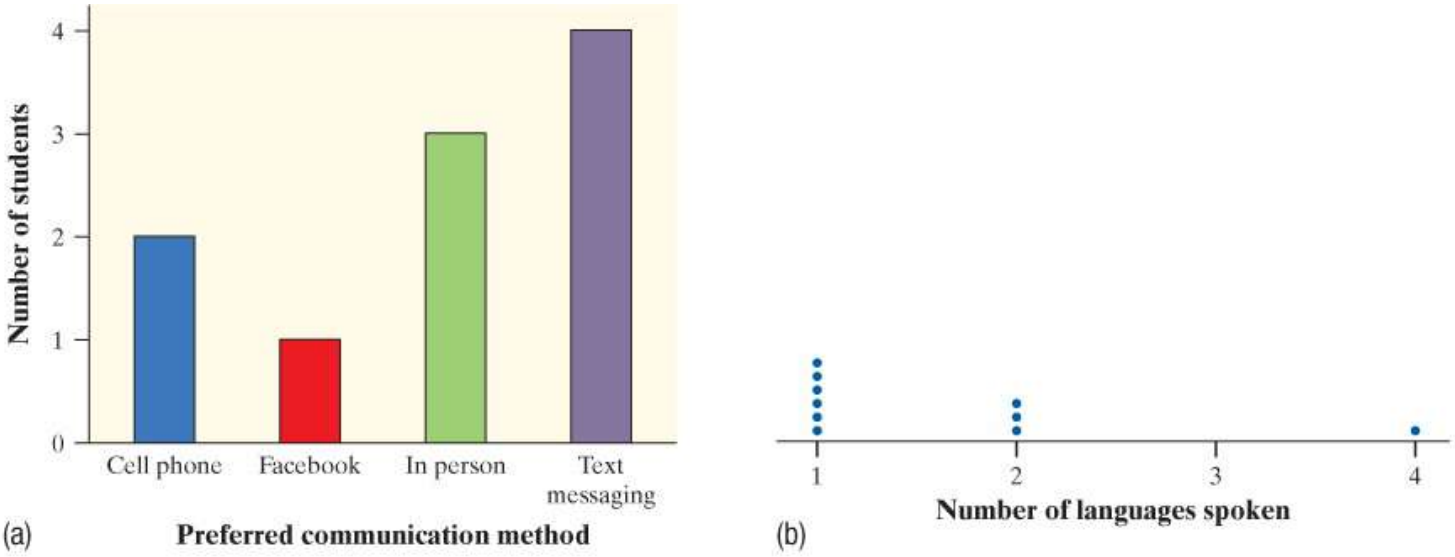


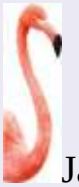
FIGURE 1.1 (a) Bar graph showing the distribution of preferred communication method for the sample of 10 Canadian students. (b) Dotplot showing the distribution of number of languages spoken by these students.

Section 1.1 begins by looking at how to describe the distribution of a single categorical variable and then examines relationships between categorical variables. Sections 1.2 and 1.3 and all of Chapter 2 focus on describing the distribution of a quantitative variable. Chapter 3 investigates relationships between two quantitative variables. In each case, we begin with graphical displays, then add numerical summaries for a more complete description.

HOW TO ANALYZE DATA

- Begin by examining each variable by itself. Then move on to study relationships among the variables.
- Start with a graph or graphs. Then add numerical summaries.

CHECK YOUR UNDERSTANDING



Jake is a car buff who wants to find out more about the vehicles that his classmates drive. He gets permission to go to the student parking lot and record some data. Later, he does some Internet research on each model of car he found. Finally, Jake makes a spreadsheet that includes each car's license plate, model, year, color, highway gas mileage, weight, and whether it has a navigation system.

1. Identify the individuals in Jake's study.

2. What are the variables? Classify each as categorical or quantitative.

From Data Analysis to Inference

Sometimes we're interested in drawing conclusions that go beyond the data at hand. That's the idea of *inference*. In the "Census At School" example, 9 of the 10 randomly selected Canadian students are right-handed. That's 90% of the *sample*. Can we conclude that exactly 90% of the *population* of Canadian students who participated in Census At School are right-handed? No.

If another random sample of 10 students were selected, the percent who are right-handed might not be exactly 90%. Can we at least say that the actual population value is "close" to 90%? That depends on what we mean by "close." The following activity gives you an idea of how statistical inference works.

ACTIVITY Hiring discrimination—it just won't fly!

An airline has just finished training 25 pilots—15 male and 10 female—to become captains. Unfortunately, only eight captain positions are available right now. Airline managers announce that they will use a lottery to determine which pilots will fill the available positions. The names of all 25 pilots will be written on identical slips of paper. The slips will be placed in a hat, mixed thoroughly, and drawn out one at a time until all eight captains have been identified.

Choja/Getty Images

A day later, managers announce the results of the lottery. Of the 8 captains chosen, 5 are female and 3 are male. Some of the male pilots who weren't selected suspect that the lottery was not carried out fairly. One of these pilots asks your statistics class for advice about whether to file a grievance with the pilots' union.

The key question in this possible discrimination case seems to be: *Is it plausible (believable) that these results happened just by chance*? To find out, you and your classmates will *simulate* the lottery process that airline managers said they used.

1. Your teacher will give you a bag with 25 beads (15 of one color and 10 of another) or 25 slips of paper (15 labeled "M" and 10 labeled "F") to represent the 25 pilots. Mix the beads/slips thoroughly. Without looking, remove 8 beads/slips from the bag. Count the number of female pilots selected. Then return the beads/slips to the bag.

2. Your teacher will draw and label a number line for a class *dotplot*. On the graph, plot the number of females you got in Step 1.
3. Repeat Steps 1 and 2 if needed to get a total of at least 40 simulated lottery results for your class.
4. Discuss the results with your classmates. Does it seem plausible that airline managers conducted a fair lottery? What advice would you give the male pilot who contacted you?

Our ability to do inference is determined by how the data are produced. Chapter 4 discusses the two main methods of data production—sampling and experiments—and the types of conclusions that can be drawn from each. As the activity illustrates, the logic of inference rests on asking, "What are the chances?" *Probability*, the study of chance behavior, is the topic of Chapters 5–7. We'll introduce the most common inference techniques in Chapters 8–12.

Introduction Summary

- **Statistics** is the science and art of collecting, analyzing, and drawing conclusions from data.
- A data set contains information about a number of **individuals.** Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables.** A variable describes some characteristic of an individual, such as a person's height, gender, or salary.
- A **categorical variable** assigns a label that places each individual in one of several groups, such as male or female. A **quantitative variable** has numerical values that count or measure some characteristic of each individual, such as number of siblings or height in meters.
- The **distribution** of a variable describes what values the variable takes and how often it takes them.

Introduction Exercises

The solutions to all exercises numbered in red may be found in the Solutions Appendix.

1. pg 3 **A class survey** Here is a small part of the data set that describes the students in an AP® Statistics class. The data come from anonymous responses to a questionnaire filled out on the first day of class.

Gender	Grade level	GPA	Children in family	Homework last night (min)	Android or iPhone?
F	9	2.3	3	0–14	iPhone
M	11	3.8	6	15–29	Android
M	10	3.1	2	15–29	Android

F	10	4.0	1	45–59	iPhone
F	10	3.4	4	0–14	iPhone
F	10	3.0	3	30–44	Android
M	9	3.9	2	15–29	iPhone
M	12	3.5	2	0–14	iPhone

a. Identify the individuals in this data set.

b. What are the variables? Classify each as categorical or quantitative.

2. Coaster craze Many people like to ride roller coasters. Amusement parks try to increase attendance by building exciting new coasters. The following table displays data on several roller coasters that were opened in a recent year.[1]

Roller coaster	Type	Height (ft)	Design	Speed (mph)	Duration (sec)
Wildfire	Wood	187.0	Sit down	70.2	120
Skyline	Steel	131.3	Inverted	50.0	90
Goliath	Wood	165.0	Sit down	72.0	105
Helix	Steel	134.5	Sit down	62.1	130
Banshee	Steel	167.0	Inverted	68.0	160
Black Hole	Steel	22.7	Sit down	25.5	75

a. Identify the individuals in this data set.

b. What are the variables? Classify each as categorical or quantitative.

3. Hit movies According to the Internet Movie Database, *Avatar* is tops based on box-office receipts worldwide as of January 2017. The following table displays data on several popular movies. Identify the individuals and variables in this data set. Classify each variable as categorical or quantitative.

Movie	Year	Rating	Time (min)	Genre	Box office ($)
Avatar	2009	PG-13	162	Action	2,783,918,982
Titanic	1997	PG-13	194	Drama	2,207,615,668
Star Wars: The Force Awakens	2015	PG-13	136	Adventure	2,040,375,795
Jurassic World	2015	PG-13	124	Action	1,669,164,161
Marvel's The Avengers	2012	PG-13	142	Action	1,519,479,547
Furious 7	2015	PG-13	137	Action	1,516,246,709
The Avengers: Age of Ultron	2015	PG-13	141	Action	1,404,705,868
Harry Potter and the Deathly Hallows: Part 2	2011	PG-13	130	Fantasy	1,328,111,219

Frozen	2013	PG	108	Animation	1,254,512,386
Iron Man 3	2013	PG-13	129	Action	1,172,805,920

4. Skyscrapers Here is some information about the tallest buildings in the world as of February 2017. Identify the individuals and variables in this data set. Classify each variable as categorical or quantitative.

Building	Country	Height (m)	Floors	Use	Year completed
Burj Khalifa	United Arab Emirates	828.0	163	Mixed	2010
Shanghai Tower	China	632.0	121	Mixed	2014
Makkah Royal Clock Tower Hotel	Saudi Arabia	601.0	120	Hotel	2012
Ping An Finance Center	China	599.0	115	Mixed	2016
Lotte World Tower	South Korea	554.5	123	Mixed	2016
One World Trade Center	United States	541.0	104	Office	2013
Taipei 101	Taiwan	509.0	101	Office	2004
Shanghai World Financial Center	China	492.0	101	Mixed	2008
International Commerce Center	China	484.0	118	Mixed	2010
Petronas Tower 1	Malaysia	452.0	88	Office	1998

5. Protecting wood What measures can be taken, especially when restoring historic wooden buildings, to help wood surfaces resist weathering? In a study of this question, researchers prepared wooden panels and then exposed them to the weather. Some of the variables recorded were type of wood (yellow poplar, pine, cedar); type of water repellent (solvent-based, water-based); paint thickness (millimeters); paint color (white, gray, light blue); weathering time (months). Classify each variable as categorical or quantitative.

6. Medical study variables Data from a medical study contain values of many variables for each subject in the study. Some of the variables recorded were gender (female or male); age (years); race (Asian, Black, White, or other); smoker (yes or no); systolic blood pressure (millimeters of mercury); level of calcium in the blood (micrograms per milliliter). Classify each variable as categorical or quantitative.

7. Ranking colleges Popular magazines rank colleges and universities on their "academic quality" in serving undergraduate students. Describe two categorical variables and two quantitative variables that you might record for each institution.

8. Social media You are preparing to study the social media habits of high school students. Describe two categorical variables and two quantitative variables that you might record for each student.

Multiple Choice *Select the best answer.*

Exercises 9 and 10 refer to the following setting. At the Census Bureau website

www.census.gov, you can view detailed data collected by the American Community Survey. The following table includes data for 10 people chosen at random from the more than 1 million people in households contacted by the survey. “School” gives the highest level of education completed.

Weight (lb)	Age (years)	Travel to work (min)	School	Gender	Income last year ($)
187	66	0	Ninth grade	1	24,000
158	66	n/a	High school grad	2	0
176	54	10	Assoc. degree	2	11,900
339	37	10	Assoc. degree	1	6000
91	27	10	Some college	2	30,000
155	18	n/a	High school grad	2	0
213	38	15	Master’s degree	2	125,000
194	40	0	High school grad	1	800
221	18	20	High school grad	1	2500
193	11	n/a	Fifth grade	1	0

9. The individuals in this data set are

a. households.

b. people.

c. adults.

d. 120 variables.

e. columns.

10. This data set contains

a. 7 variables, 2 of which are categorical.

b. 7 variables, 1 of which is categorical.

c. 6 variables, 2 of which are categorical.

d. 6 variables, 1 of which is categorical.

e. None of these.

SECTION 1.1 Analyzing Categorical Data

LEARNING TARGETS *By the end of the section, you should be able to:*

- Make and interpret bar graphs for categorical data.
- Identify what makes some graphs of categorical data misleading.
- Calculate marginal and joint relative frequencies from a two-way table.
- Calculate conditional relative frequencies from a two-way table.
- Use bar graphs to compare distributions of categorical data.
- Describe the nature of the association between two categorical variables.

Here are the data on preferred communication method for the 10 randomly selected Canadian students from the example on page 3:

In person	In person	Facebook	Cell phone	In person
Text messaging	Cell phone	Text messaging	Text messaging	Text messaging

We can summarize the distribution of this categorical variable with a **frequency table** or a **relative frequency table**.

DEFINITION Frequency table, Relative frequency table

A **frequency table** shows the number of individuals having each value.

A **relative frequency table** shows the proportion or percent of individuals having each value.

Some people use the terms *frequency distribution* and *relative frequency distribution* instead.

To make either kind of table, start by tallying the number of times that the variable takes each value. **Note that the frequencies and relative frequencies listed in these tables are not data.** The tables summarize the data by telling us how many (or what proportion or percent of) students in the sample said "Cell phone," "Facebook," "In person," and "Text messaging."

		Frequency table		Relative frequency table	
Preferred method	**Tally**	**Preferred method**	**Frequency**	**Preferred method**	**Relative frequency**
Cell phone	||	Cell phone	2	Cell phone	2/10 = 0.20 or 20%

Facebook	|	Facebook	1	Facebook	1/10 = 0.10 or 10%
In person	|||	In person	3	In person	3/10 = 0.30 or 30%
Text messaging	||||	Text messaging	4	Text messaging	4/10 = 0.40 or 40%

The same process can be used to summarize the distribution of a quantitative variable. Of course, it would be hard to make a frequency table or a relative frequency table for quantitative data that take many different values, like the ages of people attending a Major League Baseball game. We'll look at a better option for quantitative variables with many possible values in Section 1.2.

Displaying Categorical Data: Bar Graphs and Pie Charts

A frequency table or relative frequency table summarizes a variable's distribution with numbers. To display the distribution more clearly, use a graph. You can make a **bar graph** or a **pie chart** for categorical data.

DEFINITION **Bar graph, Pie chart**

A **bar graph** shows each category as a bar. The heights of the bars show the category frequencies or relative frequencies.

A **pie chart** shows each category as a slice of the "pie." The areas of the slices are proportional to the category frequencies or relative frequencies.

Bar graphs are sometimes called *bar charts.* Pie charts are sometimes called *circle graphs.*

Figure 1.2 shows a bar graph and a pie chart of the data on preferred communication method for the random sample of Canadian students. Note that the percents for each category come from the relative frequency table.

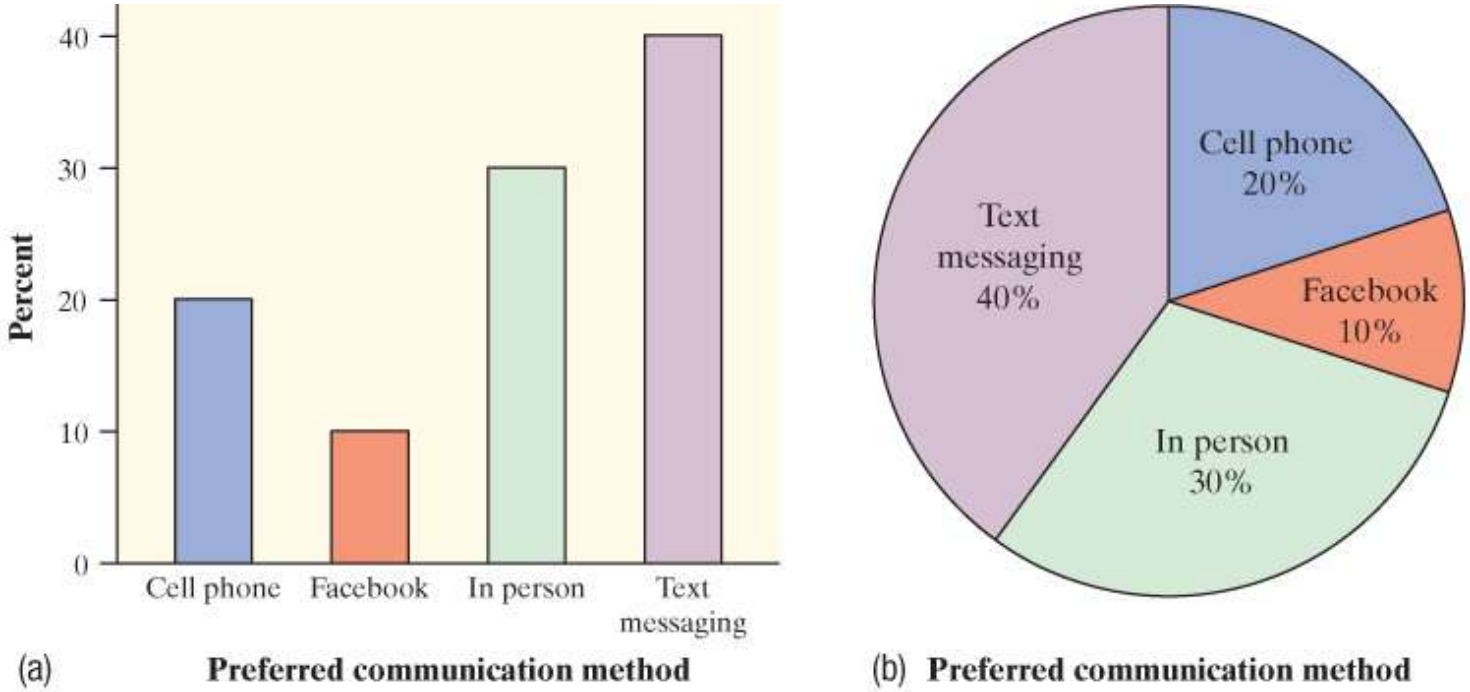


FIGURE 1.2 (a) Bar graph and (b) pie chart of the distribution of preferred communication method for a random sample of 10 Canadian students.

Relative frequency table

Preferred method	Relative frequency
Cell phone	2/10 = 0.20 or 20%
Facebook	1/10 = 0.10 or 10%
In person	3/10 = 0.30 or 30%
Text messaging	4/10 = 0.40 or 40%

It is fairly easy to make a bar graph by hand. Here's how you do it.

HOW TO MAKE A BAR GRAPH

- **Draw and label the axes.** Put the name of the categorical variable under the horizontal axis. To the left of the vertical axis, indicate whether the graph shows the frequency (count) or relative frequency (percent or proportion) of individuals in each category.
- **"Scale" the axes.** Write the names of the categories at equally spaced intervals under the horizontal axis. On the vertical axis, start at 0 and place tick marks at equal intervals until you exceed the largest frequency or relative frequency in any category.
- **Draw bars above the category names.** Make the bars equal in width and leave gaps between them. Be sure that the height of each bar corresponds to the frequency or relative frequency of individuals in that category.

Making a graph is not an end in itself. The purpose of a graph is to help us understand the data. When looking at a graph, always ask, "What do I see?" We can see from both graphs in Figure 1.2 that the most preferred communication method for these students is text messaging.

EXAMPLE | What's on the radio?

Making and interpreting bar graphs

PROBLEM: Arbitron, the rating service for radio audiences, categorizes U.S. radio stations in terms of the kinds of programs they broadcast. The frequency table summarizes the distribution of station formats in a recent year.[2]

Format	Number of stations
Adult contemporary	2536
All sports	1274
Contemporary hits	1012
Country	2893
News/talk/information	4077
Oldies	831
Religious	3884
Rock	1636
Spanish language	878
Variety	1579
Other formats	4852
Total	**25,452**

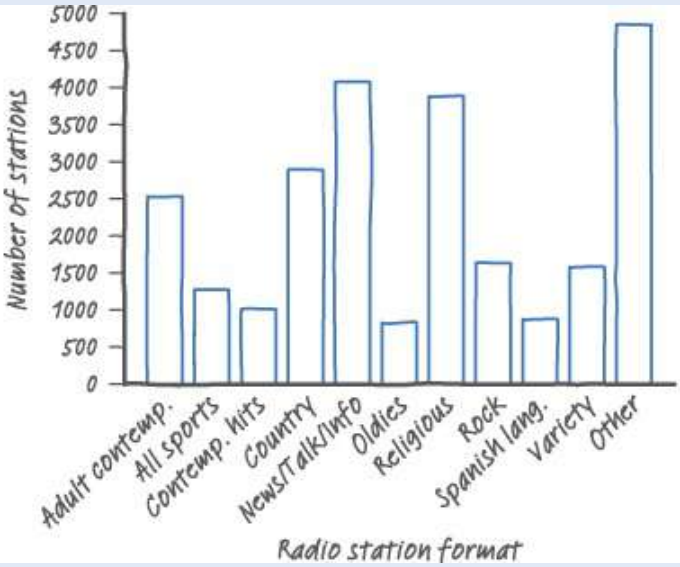
a. Identify the individuals in this data set.

b. Make a frequency bar graph of the data. Describe what you see.

SOLUTION:

a. U.S. radio stations

b.

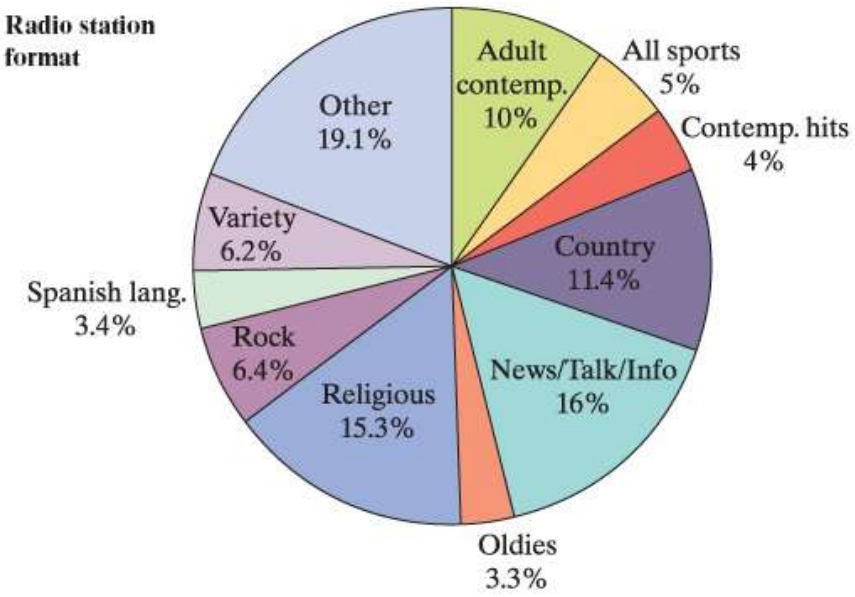


To make the bar graph:

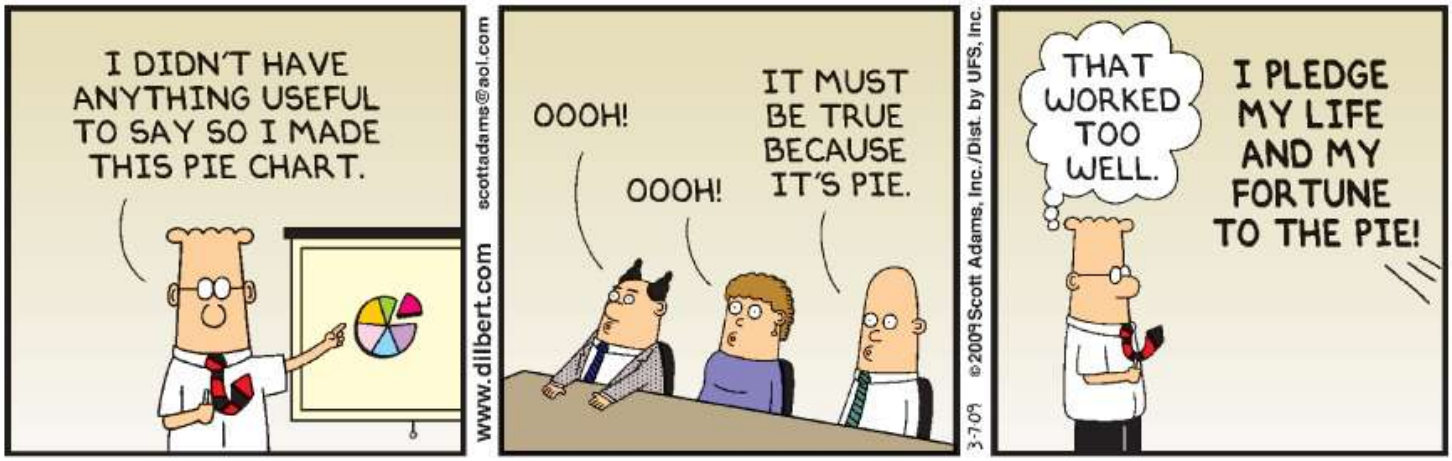
- **Draw and label the axes.**
- **"Scale" the axes.** The largest frequency is 4852. So we choose a vertical scale from 0 to 5000, with tick marks 500 units apart.
- **Draw bars above the category names.**

On U.S. radio stations, the most frequent formats are Other (4852), News/talk/information (4077), and Religious (3884), while the least frequent are Oldies (831), Spanish language (878), and Contemporary hits (1012).

FOR PRACTICE, TRY EXERCISE 11



Here is a pie chart of the radio station format data from the preceding example. You can use a pie chart when you want to emphasize each category's relation to the whole. Pie charts are challenging to make by hand, but technology will do the job for you. Note that **a pie chart must include all categories that make up a whole,** which might mean adding an "other" category, as in the radio station example.



CHECK YOUR UNDERSTANDING

The American Statistical Association sponsors a web-based project that collects data about primary and secondary school students using surveys. We used the site's "Random Sampler" to choose 40 U.S. high school students who completed the survey in a recent year.[3] One of the questions asked:

Which would you prefer to be? Select one.

_____Rich _____Happy _____Famous _____Healthy

Here are the responses from the 40 randomly selected students:

Famous	Healthy	Healthy	Famous	Happy	Famous	Happy	Happy	Famous
Rich	Happy	Happy	Rich	Happy	Happy	Happy	Rich	Happy
Famous	Healthy	Rich	Happy	Happy	Rich	Happy	Happy	Rich
Healthy	Happy	Happy	Rich	Happy	Happy	Rich	Happy	Famous
Famous	Happy	Happy	Happy					

Make a relative frequency bar graph of the data. Describe what you see.

Graphs: Good and Bad

Bar graphs are a bit dull to look at. It is tempting to replace the bars with pictures or to use special 3-D effects to make the graphs seem more interesting. Don't do it! Our eyes react to the area of the bars as well as to their height. When all bars have the same width, the area (width × height) varies in proportion to the height, and our eyes receive the right impression about the quantities being compared.

EXAMPLE | Who buys iMacs?

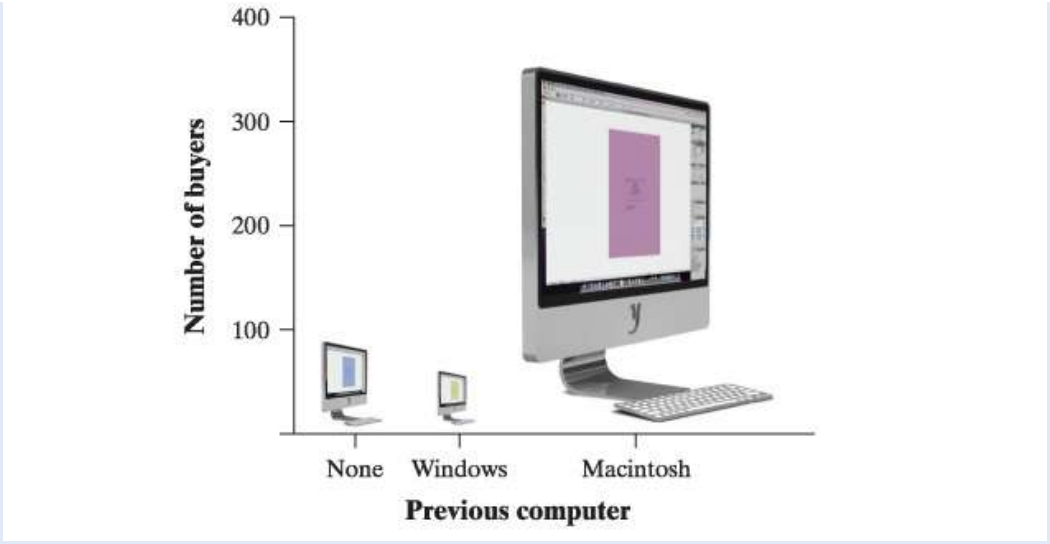
Beware the pictograph!

Justin Sullivan/Getty Images

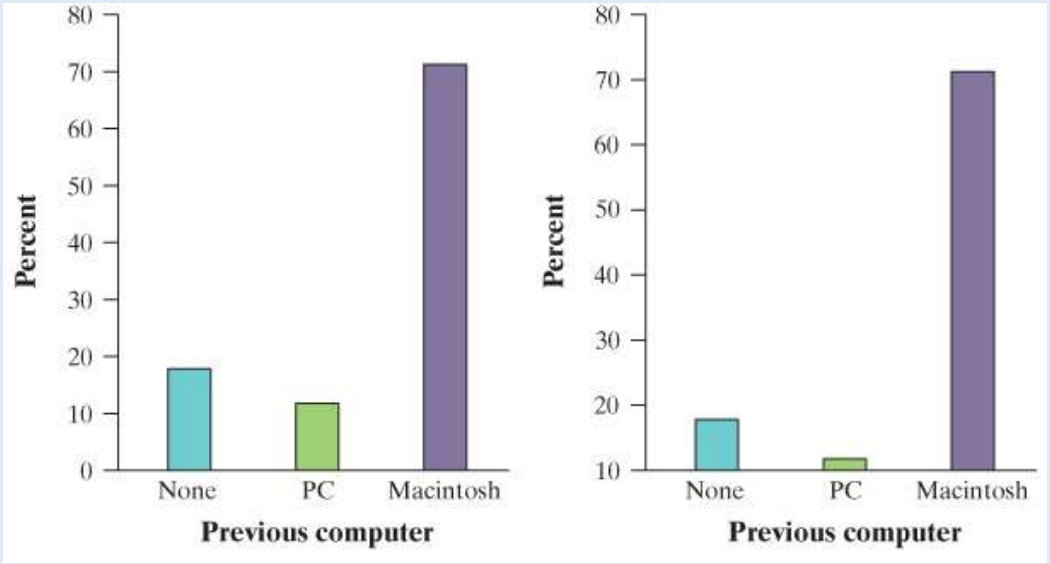
PROBLEM: When Apple, Inc., introduced the iMac, the company wanted to know whether this new computer was expanding Apple's market share. Was the iMac mainly being bought by previous Macintosh owners, or was it being purchased by first-time computer buyers and by previous PC users who were switching over? To find out, Apple hired a firm to conduct a survey of 500 iMac customers. Each customer was categorized as a new computer purchaser, a previous PC owner, or a previous Macintosh owner. The table summarizes the survey results.[4]

Previous ownership	Count	Percent (%)
None	85	17.0
PC	60	12.0
Macintosh	355	71.0
Total	**500**	**100.0**

a. Below is a clever graph of the data that uses pictures instead of the more traditional bars. How is this pictograph misleading?



b. Two possible bar graphs of the data are shown below. Which one could be considered deceptive? Why?



SOLUTION:

a. *The pictograph makes it look like the percentage of iMac buyers who are former Mac owners is at least 10 times larger than either of the other two categories, which isn't true.*

In part (a), the *heights* of the images are correct. But the *areas* of the images are misleading. The Macintosh image is about 6 times as tall as the PC image, but its area is about 36 times as large!

b. *The bar graph on the right is misleading. By starting the vertical scale at 10 instead of 0, it looks like the percentage of iMac buyers who previously owned a PC is less than half the percentage who are first-time computer buyers, which isn't true.*

FOR PRACTICE, TRY EXERCISE 19

There are two important lessons to be learned from this example: **(1) beware the pictograph, and (2) watch those scales.**

Analyzing Data on Two Categorical Variables

You have learned some techniques for analyzing the distribution of a single categorical variable. What should you do when a data set involves two categorical variables? For example, Yellowstone National Park staff surveyed a random sample of 1526 winter visitors to the park. They asked each person whether he or she belonged to an environmental club (like the Sierra Club). Respondents were also asked whether they owned, rented, or had never used a snowmobile. The data set looks something like the following:

Respondent	Environmental club?	Snowmobile use
1	No	Own
2	No	Rent
3	Yes	Never
4	Yes	Rent
5	No	Never
⋮	⋮	⋮

franz12/Shutterstock

The **two-way table** summarizes the survey responses.

		Environmental club member?	
		No	Yes
Snowmobile use	Never	445	212
	Rent	497	77
	Own	279	16

DEFINITION Two-way table

A **two-way table** is a table of counts that summarizes data on the relationship between two categorical variables for some group of individuals.

A two-way table is sometimes called a *contingency table.*

It's easier to grasp the information in a two-way table if row and column totals are included, like the one shown here.

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

Now we can quickly answer questions like:

- What percent of people in the sample are environmental club members?

3051526=0.200=20.0%$\frac{305}{1526} = 0.200 = 20.0\%$$\frac{305}{1526} = 0.200 = 20.0\%$

- What proportion of people in the sample never used a snowmobile?

6571526=0.431$\frac{657}{1526} = 0.431$$\frac{657}{1526} = 0.431$

These percents or proportions are known as **marginal relative frequencies** because they are calculated using values in the margins of the two-way table.

DEFINITION Marginal relative frequency

A **marginal relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable.

We could call this distribution the *marginal distribution* of environmental club membership.

We can compute marginal relative frequencies for the *column* totals to give the distribution of environmental club membership in the entire sample of 1526 park visitors:

No: 12211526=0.800 or 80.0% Yes: 3051526=0.200 or 20.0%

$\text{No}: \frac{1221}{1526} = 0.800 \text{ or } 80.0\%$ $\text{Yes}: \frac{305}{1526} = 0.200 \text{ or } 20.0\%$

$$\text{No}: \frac{1221}{1526} = 0.800 \text{ or } 80.0\% \quad \text{Yes}: \frac{305}{1526} = 0.200 \text{ or } 20.0\%$$

We can compute marginal relative frequencies for the *row* totals to give the distribution of snowmobile use for all the individuals in the sample:

Never: 6571526=0.431 or 43.1%Rent: 5741526=0.376 or 37.6%Own: 2951526=0.193 or 19.3%

$$\text{Never}: \frac{657}{1526} = 0.431 \text{ or } 43.1\% \quad \text{Rent}: \frac{574}{1526} = 0.376 \text{ or } 37.6\% \quad \text{Own}: \frac{295}{1526} = 0.193 \text{ or } 19.3\%$$

We could call this distribution the *marginal distribution* of snowmobile use.

Note that we could use a bar graph or a pie chart to display either of these distributions.

A marginal relative frequency tells you about only *one* of the variables in a two-way table. It won't help you answer questions like these, which involve values of *both* variables:

- What percent of people in the sample are environmental club members and own snowmobiles?

$$161526=0.010=1.0\% \frac{16}{1526} = 0.010 = 1.0\% \frac{16}{1526} = 0.010 = 1.0\%$$

- What proportion of people in the sample are not environmental club members and never use snowmobiles?

$$4451526=0.292 \frac{445}{1526} = 0.292 \frac{445}{1526} = 0.292$$

These percents or proportions are known as **joint relative frequencies**.

DEFINITION **Joint relative frequency**

A **joint relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable and a specific value for another categorical variable.

EXAMPLE **A *Titanic* disaster**

Calculating marginal and joint relative frequencies

Blank Archives/Getty Images

PROBLEM: In 1912 the luxury liner *Titanic*, on its first voyage across the Atlantic, struck an iceberg and sank. Some passengers got off the ship in lifeboats, but many died. The two-way table gives information about adult passengers who survived and who died, by class of travel.

		Class of travel		
		First	Second	Third
Survival status	Survived	197	94	151
	Died	122	167	476

a. What proportion of adult passengers on the *Titanic* survived?

b. Find the distribution of class of travel for adult passengers on the *Titanic* using relative frequencies.

c. What percent of adult *Titanic* passengers traveled in third class and survived?

SOLUTION:

Start by finding the marginal totals.

		Class of travel			
		First	Second	Third	Total
Survival status	Survived	197	94	151	442
	Died	122	167	476	765
	Total	319	261	627	1207

a. $4421207=0.366\frac{442}{1207}=0.366\frac{442}{1207}=0.366$

Remember that a distribution lists the possible values of a variable and how often

those values occur.

b.

First: 3191207=0.264=26.4%Second: 2611207=0.216=21.6%Third: 6271207=0.519=51.9%

$$\text{First}: \frac{319}{1207} = 0.264 = 26.4\%$$
$$\text{Second}: \frac{261}{1207} = 0.216 = 21.6\%$$
$$\text{Third}: \frac{627}{1207} = 0.519 = 51.9\%$$

$$\text{First}: \frac{319}{1207} = 0.264 = 26.4\%$$
$$\text{Second}: \frac{261}{1207} = 0.216 = 21.6\%$$
$$\text{Third}: \frac{627}{1207} = 0.519 = 51.9\%$$

Note that the three percentages for class of travel in part (b) do not add to exactly 100% due to roundoff error.

c. 1511207=0.125=12.5% $\frac{151}{1207} = 0.125 = 12.5\%$ $\frac{151}{1207} = 0.125 = 12.5\%$

FOR PRACTICE, TRY EXERCISE 23

CHECK YOUR UNDERSTANDING

An article in the *Journal of the American Medical Association* reports the results of a study designed to see if the herb St. John's wort is effective in treating moderately severe cases of depression. The study involved 338 patients who were being treated for major depression. The subjects were randomly assigned to receive one of three treatments: St. John's wort, Zoloft (a prescription drug), or placebo (an inactive treatment) for an 8-week period. The two-way table summarizes the data from the experiment.[5]

		Treatment		
		St. John's wort	Zoloft	Placebo
Change in depression	Full response	27	27	37
	Partial response	16	26	13
	No response	70	56	66

1. What proportion of subjects in the study were randomly assigned to take St. John's wort? Explain why this value makes sense.
2. Find the distribution of change in depression for the subjects in this study using relative frequencies.
3. What percent of subjects took Zoloft and showed a full response?

Relationships Between Two Categorical Variables

Let's return to the data from the Yellowstone National Park survey of 1526 randomly selected winter visitors. Earlier, we calculated marginal and joint relative frequencies from the two-way table. These values do not tell us much about the *relationship* between environmental club membership and snowmobile use for the people in the sample.

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

We can also use the two-way table to answer questions like:

- What percent of environmental club members in the sample are snowmobile owners?

$$16305=0.052=5.2\%\frac{16}{305}=0.052=5.2\%\frac{16}{305}=0.052=5.2\%$$

- What proportion of snowmobile renters in the sample are not environmental club members?

$$497574=0.866\frac{497}{574}=0.866\frac{497}{574}=0.866$$

These percents or proportions are known as **conditional relative frequencies**.

DEFINITION **Conditional relative frequency**

A **conditional relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable among individuals who share the same value of another categorical variable (the condition).

EXAMPLE | **A *Titanic* disaster**

Conditional relative frequencies

PROBLEM: In 1912 the luxury liner *Titanic*, on its first voyage across the Atlantic, struck an iceberg and sank. Some passengers made it off the ship in lifeboats, but many died. The two-way table gives information about adult passengers who survived and who died, by class of travel.

		Class of travel			
		First	Second	Third	Total
Survival status	Survived	197	94	151	442
	Died	122	167	476	765
	Total	319	261	627	1207

a. What proportion of survivors were third-class passengers?

b. What percent of first-class passengers survived?

SOLUTION:

a. $\frac{151}{442} = 0.342$ $151442=0.342$ $\frac{151}{442} = 0.342$

b. $\frac{197}{319} = 0.618 = 61.8\%$ $197319=0.618=61.8\%$ $\frac{197}{319} = 0.618 = 61.8\%$

Note that a proportion is always a number between 0 and 1, whereas a percent is a number between 0 and 100. To get a percent, multiply the proportion by 100.

FOR PRACTICE, TRY EXERCISE 27

We can study the snowmobile use habits of environmental club members by looking only at the "Yes" column in the two-way table.

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

It is easy to calculate the proportions or percents of environmental club members who never use, rent, and own snowmobiles:

Never: 212305=0.695 or 69.5%Rent: 77305=0.252 or 25.2%Own: 16305=0.052 or 5.2%

$$\text{Never: } \frac{212}{305} = 0.695 \text{ or } 69.5\%$$

$$\text{Rent: } \frac{77}{305} = 0.252 \text{ or } 25.2\%$$

$$\text{Own: } \frac{16}{305} = 0.052 \text{ or } 5.2\%$$

We could also refer to this distribution as the *conditional distribution* of snowmobile use among environmental club members.

This is the distribution of snowmobile use among environmental club members.

We can find the distribution of snowmobile use among the survey respondents who are not environmental club members in a similar way. The table summarizes the conditional relative frequencies for both groups.

Snowmobile use	Not environmental club members	En
Never	4451221=0.364 or 36.4%$\frac{445}{1221} = 0.364 \text{ or } 36.4\%$	212305=0.695 or 6
Rent	4971221=0.407 or 40.7%$\frac{497}{1221} = 0.407 \text{ or } 40.7\%$	77305=0.252 or 2
Own	2791221=0.229 or 22.9%$\frac{279}{1221} = 0.229 \text{ or } 22.9\%$	16305=0.052 or

AP® EXAM TIP

When comparing groups of different sizes, be sure to use relative frequencies (percents or proportions) instead of frequencies (counts) when analyzing categorical data. Comparing only the frequencies can be misleading, as in this setting. There are many more people who never use snowmobiles among the non-environmental club members in the sample (445) than among the environmental club members (212). However, the *percentage* of environmental club members who never use snowmobiles is much higher (69.5% to 36.4%). Finally, make sure to avoid statements like “More club members never use snowmobiles” when you mean “A greater percentage of club members never use snowmobiles.”

Figure 1.3 compares the distributions of snowmobile use for Yellowstone National Park visitors who are environmental club members and those who are not environmental club members with (a) a **side-by-side bar graph** and (b) a **segmented bar graph.** Notice that the segmented bar graph can be obtained by stacking the bars in the side-by-side bar graph for each of the two environmental club membership categories (no and yes).

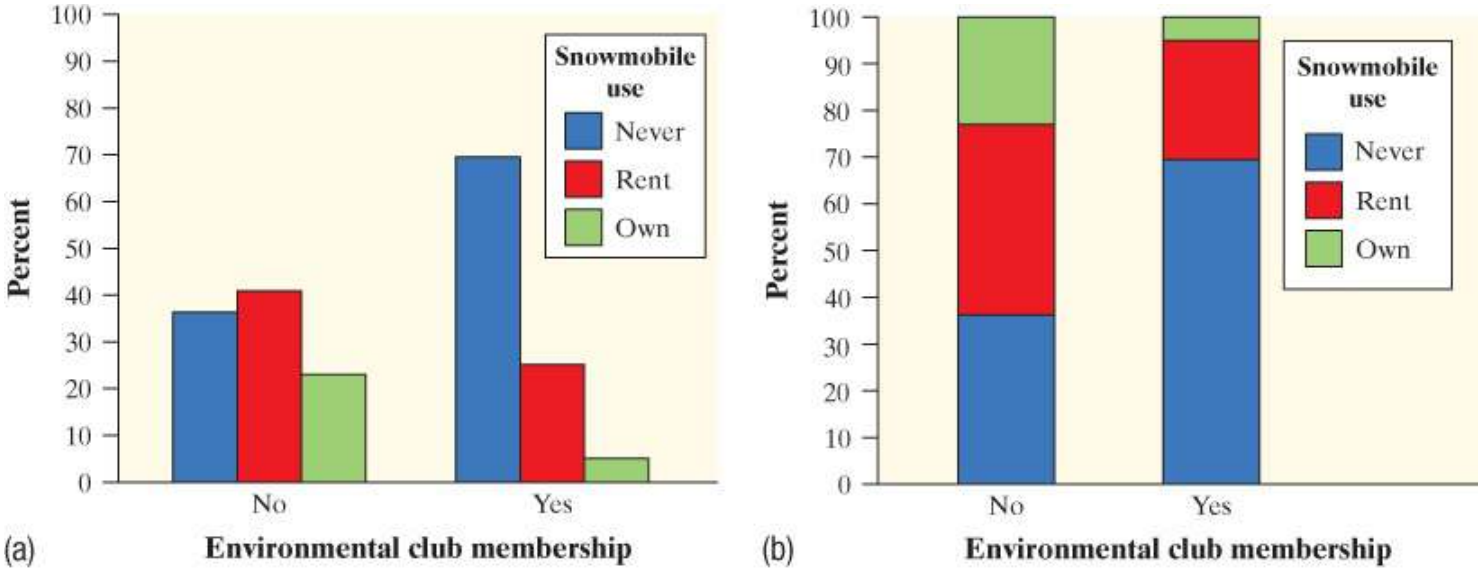


FIGURE 1.3 (a) Side-by-side bar graph and (b) segmented bar graph displaying the distribution of snowmobile use among environmental club members and among non-environmental club members from the 1526 randomly selected winter visitors to Yellowstone National Park.

DEFINITION Side-by side bar graph, Segmented bar graph

A **side-by-side bar graph** displays the distribution of a categorical variable for each value of another categorical variable. The bars are grouped together based on the values of one of the categorical variables and placed side by side.

A **segmented bar graph** displays the distribution of a categorical variable as segments of a rectangle, with the area of each segment proportional to the percent of individuals in the corresponding category.

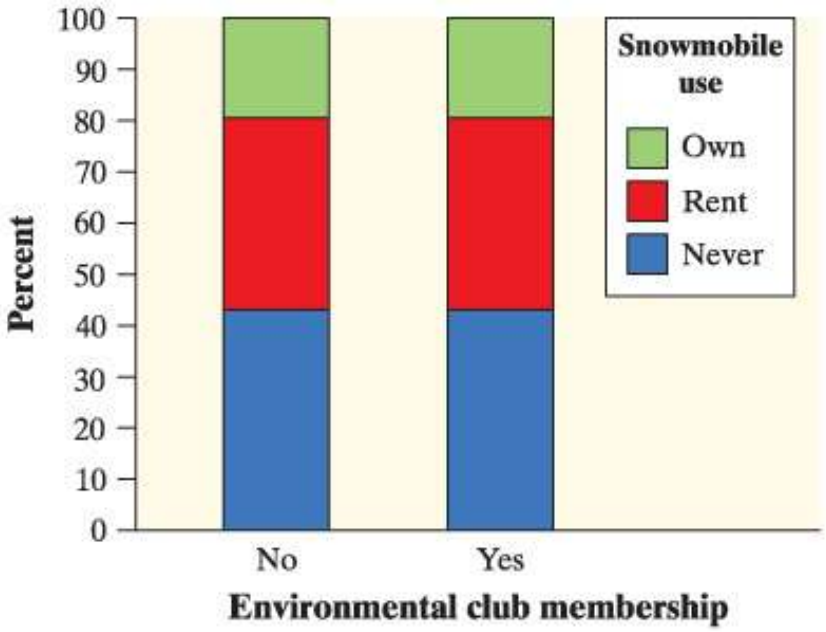
Both graphs in Figure 1.3 show a clear **association** between environmental club membership and snowmobile use in this random sample of 1526 winter visitors to Yellowstone National Park. The environmental club members were much less likely to rent (25.2% versus 40.7%) or own (5.2% versus 29.0%) snowmobiles than non-club-members and more likely to never use a snowmobile (69.5% versus 36.4%). Knowing whether or not a person in the sample is an environmental club member helps us predict that individual's snowmobile use.

DEFINITION Association

There is an **association** between two variables if knowing the value of one variable helps us predict the value of the other. If knowing the value of one variable does not help us predict the value of the other, then there is no association between the variables.

What would the graphs in Figure 1.3 look like if there was *no association* between environmental club membership and snowmobile use in the sample? The blue segments would

be the same height for both the “Yes” and “No” groups. So would the green segments and the red segments, as shown in the graph at left. In that case, knowing whether a survey respondent is an environmental club member would *not* help us predict his or her snowmobile use.



Which distributions should we compare? Our goal all along has been to analyze the relationship between environmental club membership and snowmobile use for this random sample of 1526 Yellowstone National Park visitors. We decided to calculate conditional relative frequencies of snowmobile use among environmental club members and among non-club-members. Why? Because we wanted to see if environmental club membership helped us predict snowmobile use. What if we had wanted to determine whether snowmobile use helps us predict whether a person is an environmental club member? Then we would have calculated conditional relative frequencies of environmental club membership among snowmobile owners, renters, and non-users. *In general, you should calculate the distribution of the variable that you want to predict for each value of the other variable.*

Can we say that there is an association between environmental club membership and snowmobile use in the *population* of all winter visitors to Yellowstone National Park? Making this determination requires formal inference, which will have to wait until Chapter 11.

EXAMPLE | **A *Titanic* disaster**

Conditional relative frequencies and association

Universal History Archive/Getty Images

PROBLEM: In 1912 the luxury liner *Titanic*, on its first voyage across the Atlantic, struck an iceberg and sank. Some passengers made it off the ship in lifeboats, but many died. The two-way table gives information about adult passengers who survived and who died, by class of travel.

		Class of travel			
		First	Second	Third	Total
Survival status	Survived	197	94	151	442
	Died	122	167	476	765
	Total	319	261	627	1207

a. Find the distribution of survival status for each class of travel. Make a segmented bar graph to compare these distributions.

b. Describe what the graph in part (a) reveals about the association between class of travel and survival status for adult passengers on the *Titanic*.

SOLUTION:

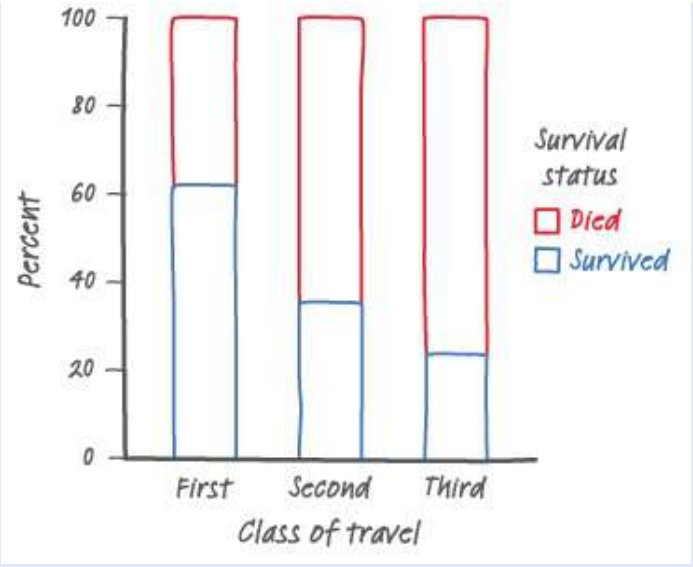
a. *First classSurvived: 197319=0.618=61.8%Died: 122319=0.382=38.2%Second classSurvived: 94261=0.36(*

First class Survived: $\frac{197}{319} = 0.618 = 61.8\%$ Died: $\frac{122}{319} = 0.382 = 38.2\%$

Second class Survived: $\frac{94}{261} = 0.360 = 36.0\%$ Died: $\frac{167}{261} = 0.640 = 64.0\%$

Third class Survived: $\frac{151}{627} = 0.241 = 24.1\%$ Died: $\frac{476}{627} = 0.759 = 75.9\%$

First class Survived: $\frac{197}{319} = 0.618 = 61.8\%$ Died: $\frac{122}{319} = 0.382 = 38.2\%$
Second class Survived: $\frac{94}{261} = 0.360 = 36.0\%$ Died: $\frac{167}{261} = 0.640 = 64.0\%$
Third class Survived: $\frac{151}{627} = 0.241 = 24.1\%$ Died: $\frac{476}{627} = 0.759 = 75.9\%$



To make the segmented bar graph:

- **Draw and label the axes.** Put class of travel on the horizontal axis and percent on the vertical axis.
- **"Scale" the axes.** Use a vertical scale from 0 to 100%, with tick marks every 20%.
- **Draw bars.** Make each bar have a height of 100%. Be sure the bars are equal in width and leave spaces between them. Segment each bar based on the conditional relative frequencies you calculated. Use different colors or shading patterns to represent the two possible statuses—survived and died. Add a key to the graph that tells us which color (or shading) represents which status.

b. Knowing a passenger's class of travel helps us predict his or her survival status. First class had the highest percentage of survivors (61.8%), followed by second class (36.0%), and then third class (24.1%).

FOR PRACTICE, TRY EXERCISE 29

Bar graphs can be used to compare any set of quantities that can be measured in the same units. See Exercises 33 and 34.

Because the variable "Survival status" has only two possible values, comparing the three distributions displayed in the segmented bar graph amounts to comparing the percent of passengers in each class of travel who survived. The bar graph in Figure 1.4 shows this

comparison. Note that the bar heights do *not* add to 100%, because each bar represents a different group of passengers on the *Titanic*.

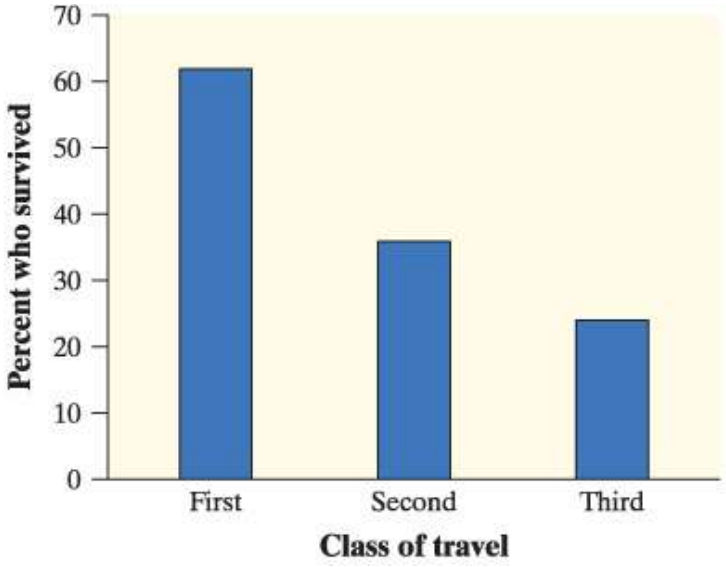


FIGURE 1.4 Bar graph comparing the percents of passengers who survived among each of the three classes of travel on the *Titanic*.

We offer a final caution about studying the relationship between two variables: **association does not imply causation.** It may be true that being in a higher class of travel on the *Titanic* increased a passenger's chance of survival. However, there isn't always a cause-and-effect relationship between two variables even if they are clearly associated. For example, a recent study proclaimed that people who are overweight are less likely to die within a few years than are people of normal weight. Does this mean that gaining weight will cause you to live longer? Not at all. The study included smokers, who tend to be thinner and also much more likely to die in a given period than non-smokers. Smokers increased the death rate for the normal-weight category, making it appear as if being overweight is better.[6] The moral of the story: *beware other variables!*

CHECK YOUR UNDERSTANDING

An article in the *Journal of the American Medical Association* reports the results of a study designed to see if the herb St. John's wort is effective in treating moderately severe cases of depression. The study involved 338 subjects who were being treated for major depression. The subjects were randomly assigned to receive one of three treatments: St. John's wort, Zoloft (a prescription drug), or placebo (an inactive treatment) for an 8-week period. The two-way table summarizes the data from the experiment.

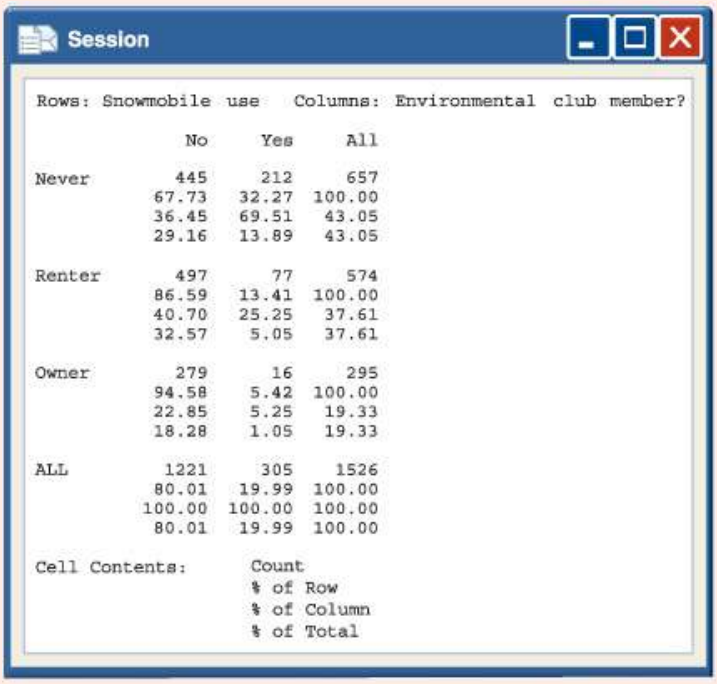
	Treatment		
	St. John's wort	Zoloft	Placebo
Full response	27	27	37

Change in depression	Partial response	16	26	13
	No response	70	56	66

1. What proportion of subjects who showed a full response took St. John's wort?
2. What percent of subjects who took St. John's wort showed no response?
3. Find the distribution of change in depression for the subjects receiving each of the three treatments. Make a segmented bar graph to compare these distributions.
4. Describe what the graph in Question 3 reveals about the association between treatment and change in depression for these subjects.

1. Technology Corner | ANALYZING TWO-WAY TABLES

Statistical software will provide marginal relative frequencies, joint relative frequencies, and conditional relative frequencies for data summarized in a two-way table. Here is output from Minitab for the data on snowmobile use and environmental club membership. Use the information on cell contents at the bottom of the output to help you interpret what each value in the table represents.



Section 1.1 Summary

- The distribution of a categorical variable lists the categories and gives the **frequency** (count) or **relative frequency** (percent or proportion) of individuals that fall in each category.
- You can use a **pie chart** or **bar graph** to display the distribution of a categorical variable. When examining any graph, ask yourself, "What do I see?"

- Beware of graphs that mislead the eye. Look at the scales to see if they have been distorted to create a particular impression. Avoid making graphs that replace the bars of a bar graph with pictures whose height and width both change.
- A **two-way table** of counts summarizes data on the relationship between two categorical variables for some group of individuals.
- You can use a two-way table to calculate three types of relative frequencies:
  - A **marginal relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable. Use the appropriate row total or column total in a two-way table when calculating a marginal relative frequency.
  - A **joint relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable and a specific value for another categorical variable. Use the value from the appropriate cell in the two-way table when calculating a joint relative frequency.
  - A **conditional relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable among individuals who share the same value of another categorical variable (the condition). Use conditional relative frequencies to compare distributions of a categorical variable for two or more groups.
- Use a **side-by-side bar graph** or a **segmented bar graph** to compare the distribution of a categorical variable for two or more groups.
- There is an **association** between two variables if knowing the value of one variable helps predict the value of the other. To see whether there is an association between two categorical variables, find the distribution of one variable for each value of the other variable by calculating an appropriate set of conditional relative frequencies.

1.1 Technology Corner

TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/tps6e.

1. Analyzing two-way tables Page 22

Section 1.1 Exercises

11. pg 11 **Birth days** The frequency table summarizes data on the numbers of babies born on each day of the week in the United States in a recent week.[7]

Day	Births
Sunday	7374

Monday	11,704
Tuesday	13,169
Wednesday	13,038
Thursday	13,013
Friday	12,664
Saturday	8459

a. Identify the individuals in this data set.

b. Make a frequency bar graph to display the data. Describe what you see.

12. Going up? As of 2015, there were over 75,000 elevators in New York City. The frequency table summarizes data on the number of elevators of each type.[8]

Type	Count
Passenger elevator	66,602
Freight elevator	4140
Escalator	2663
Dumbwaiter	1143
Sidewalk elevator	943
Private elevator	252
Handicap lift	227
Manlift	73
Public elevator	45

a. Identify the individuals in this data set.

b. Make a frequency bar graph to display the data. Describe what you see.

13. Buying cameras The brands of the last 45 digital single-lens reflex (SLR) cameras sold on a popular Internet auction site are listed here. Make a relative frequency bar graph for these data. Describe what you see.

Canon	Sony	Canon	Nikon	Fujifilm
Nikon	Canon	Sony	Canon	Canon
Nikon	Canon	Nikon	Canon	Canon
Canon	Nikon	Fujifilm	Canon	Nikon
Nikon	Canon	Canon	Canon	Canon
Olympus	Canon	Canon	Canon	Nikon
Olympus	Sony	Canon	Canon	Sony
Canon	Nikon	Sony	Canon	Fujifilm
Nikon	Canon	Nikon	Canon	Sony

14. Disc dogs Here is a list of the breeds of dogs that won the World Canine Disc Championships from 1975 through 2016. Make a relative frequency bar graph for these

data. Describe what you see.

Whippet	Mixed breed	Australian shepherd
Whippet	Australian shepherd	Australian shepherd
Whippet	Border collie	Australian shepherd
Mixed breed	Australian shepherd	Border collie
Mixed breed	Mixed breed	Border collie
Other purebred	Mixed breed	Australian shepherd
Labrador retriever	Mixed breed	Border collie
Mixed breed	Border collie	Border collie
Mixed breed	Border collie	Other purebred
Border collie	Australian shepherd	Border collie
Mixed breed	Border collie	Border collie
Mixed breed	Australian shepherd	Border collie
Labrador retriever	Border collie	Mixed breed
Labrador retriever	Mixed breed	Australian shepherd

15. **Cool car colors** The most popular colors for cars and light trucks change over time. Silver advanced past green in 2000 to become the most popular color worldwide, then gave way to shades of white in 2007. Here is a relative frequency table that summarizes data on the colors of vehicles sold worldwide in a recent year.[9]

Color	Percent of vehicles
Black	19
Blue	6
Brown/beige	5
Gray	12
Green	1
Red	9
Silver	14
White	29
Yellow/gold	3
Other	??

a. What percent of vehicles would fall in the “Other” category?

b. Make a bar graph to display the data. Describe what you see.

c. Would it be appropriate to make a pie chart of these data? Explain.

16. **Spam** Email spam is the curse of the Internet. Here is a relative frequency table that summarizes data on the most common types of spam:[10]

Type of spam	Percent
Adult	19

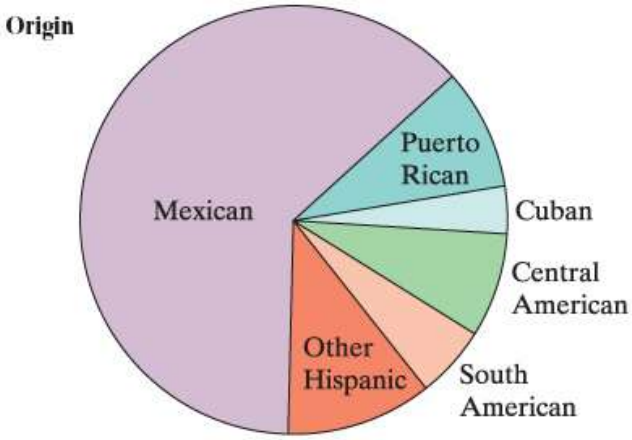
Financial	20
Health	7
Internet	7
Leisure	6
Products	25
Scams	9
Other	??

a. What percent of spam would fall in the "Other" category?

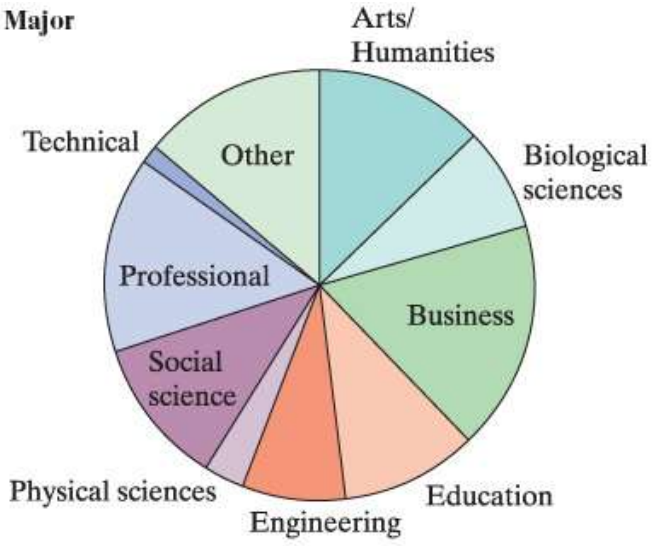
b. Make a bar graph to display the data. Describe what you see.

c. Would it be appropriate to make a pie chart of these data? Explain.

17. Hispanic origins Here is a pie chart prepared by the Census Bureau to show the origin of the more than 50 million Hispanics in the United States in 2010.[11] About what percent of Hispanics are Mexican? Puerto Rican?

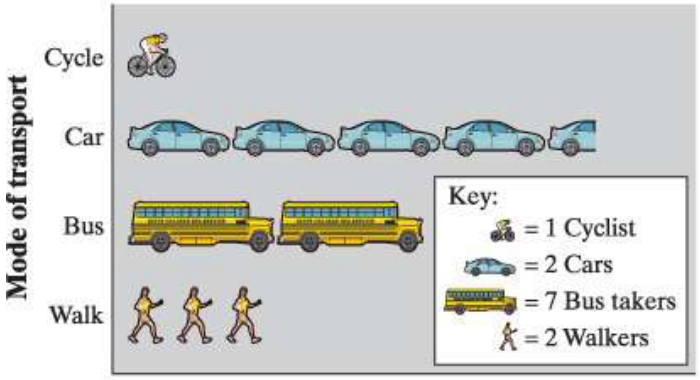


18. Which major? About 3 million first-year students enroll in colleges and universities each year. What do they plan to study? The pie chart displays data on the percent of first-year students who plan to major in several disciplines.[12] About what percent of first-year students plan to major in business? In social science?

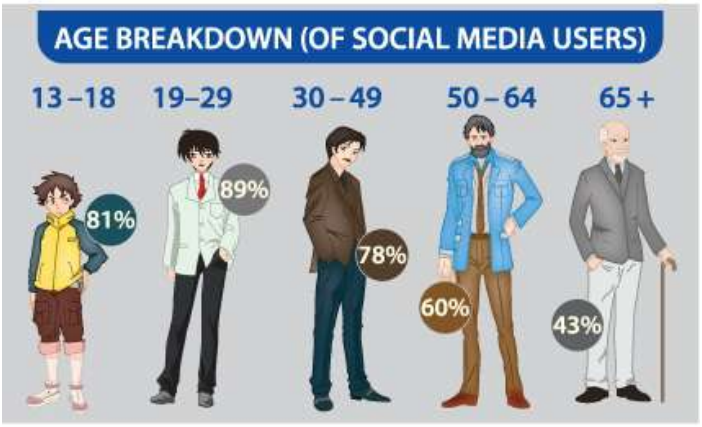


19. pg 12 **Going to school** Students in a high school statistics class were given data

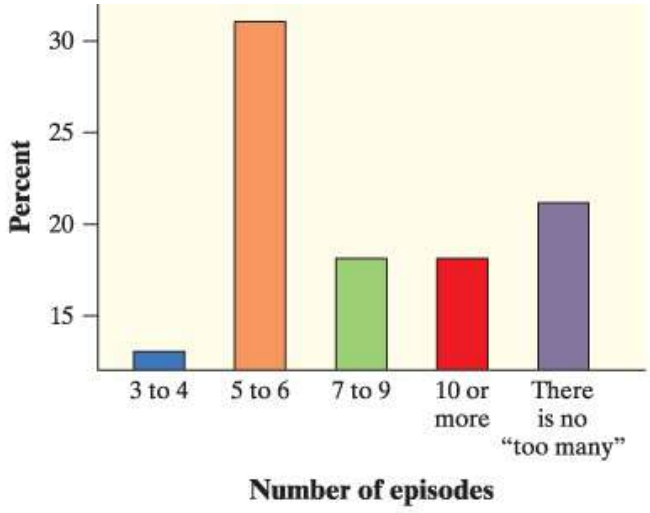
about the main method of transportation to school for a group of 30 students. They produced the pictograph shown. Explain how this graph is misleading.



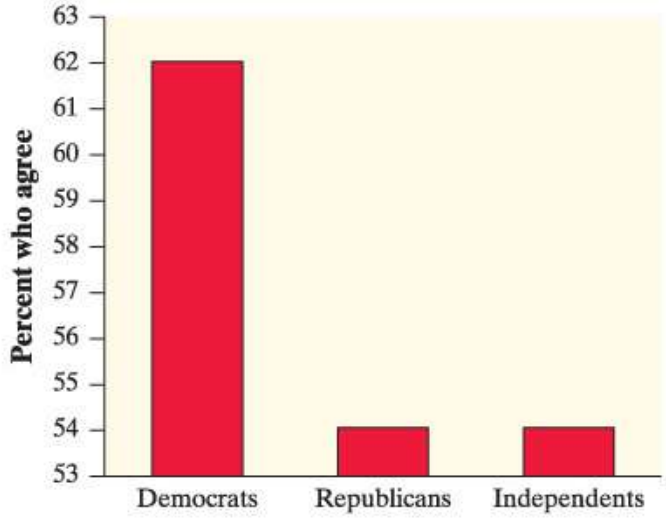
20. Social media The Pew Research Center surveyed a random sample of U.S. teens and adults about their use of social media. The following pictograph displays some results. Explain how this graph is misleading.



21. Binge-watching Do you "binge-watch" television series by viewing multiple episodes of a series at one sitting? A survey of 800 people who binge-watch were asked how many episodes is too many to watch in one viewing session. The results are displayed in the bar graph.[13] Explain how this graph is misleading.



22. Support the court? A news network reported the results of a survey about a controversial court decision. The network initially posted on its website a bar graph of the data similar to the one that follows. Explain how this graph is misleading. **(***Note:* When notified about the misleading nature of its graph, the network posted a corrected version.)



23. pg 15 **A smash or a hit?** Researchers asked 150 subjects to recall the details of a car accident they watched on video. Fifty subjects were randomly assigned to be asked, "About how fast were the cars going when they smashed into each other?" For another 50 randomly assigned subjects, the words "smashed into" were replaced with "hit." The remaining 50 subjects—the control group—were not asked to estimate speed. A week later, all subjects were asked if they saw any broken glass at the accident (there wasn't any). The table shows each group's response to the broken glass question.[14]

		Treatment		
		"Smashed into"	"Hit"	Control
Response	Yes	16	7	6
	No	34	43	44

a. What proportion of subjects were given the control treatment?

b. Find the distribution of responses about whether there was broken glass at the accident for the subjects in this study using relative frequencies.

c. What percent of the subjects were given the "smashed into" treatment and said they saw broken glass at the accident?

24. Superpowers A total of 415 children from the United Kingdom and the United States who completed a survey in a recent year were randomly selected. Each student's country of origin was recorded along with which superpower they would most like to have: the ability to fly, ability to freeze time, invisibility, superstrength, or telepathy (ability to read minds). The data are summarized in the following table.[15]

		Country	
		U.K.	U.S.
Superpower	Fly	54	45
	Freeze time	52	44
	Invisibility	30	37
	Superstrength	20	23

Telepathy	44	66

a. What proportion of students in the sample are from the United States?

b. Find the distribution of superpower preference for the students in the sample using relative frequencies.

c. What percent of students in the sample are from the United Kingdom and prefer telepathy as their superpower preference?1

25. Body image A random sample of 1200 U.S. college students was asked, "What is your perception of your own body? Do you feel that you are overweight, underweight, or about right?" The two-way table summarizes the data on perceived body image by gender.[16]

		Gender		
		Female	Male	Total
Body image	About right	560	295	855
	Overweight	163	72	235
	Underweight	37	73	110
	Total	760	440	1200

a. What percent of respondents feel that their body weight is about right?

b. What proportion of the sample is female?

c. What percent of respondents are males and feel that they are overweight or underweight?

26. Python eggs How is the hatching of water python eggs influenced by the temperature of the snake's nest? Researchers randomly assigned newly laid eggs to one of three water temperatures: hot, neutral, or cold. Hot duplicates the extra warmth provided by the mother python, and cold duplicates the absence of the mother. The two-way table summarizes the data on whether or not the eggs hatched.[17]

		Water temperature			
		Cold	Neutral	Hot	Total
Hatched?	Yes	16	38	75	129
	No	11	18	29	58
	Total	27	56	104	187

a. What percent of eggs were randomly assigned to hot water?

b. What proportion of eggs in the study hatched?

c. What percent of eggs in the study were randomly assigned to cold or neutral water and hatched?

27. pg 17 **A smash or a hit** Refer to Exercise 23.

a. What proportion of subjects who said they saw broken glass at the accident received the "hit" treatment?

b. What percent of subjects who received the "smashed into" treatment said they did not see broken glass at the accident?

28. Superpower Refer to Exercise 24.

a. What proportion of students in the sample who prefer invisibility as their superpower are from the United States?

b. What percent of students in the sample who are from the United Kingdom prefer superstrength as their superpower?

29. pg 20 **A smash or a hit** Refer to Exercise 23.

a. Find the distribution of responses about whether there was broken glass at the accident for each of the three treatment groups. Make a segmented bar graph to compare these distributions.

b. Describe what the graph in part (a) reveals about the association between response about broken glass at the accident and treatment received for the subjects in the study.

30. Superpower Refer to Exercise 24.

a. Find the distribution of superpower preference for the students in the sample from each country (i.e., the United States and the United Kingdom). Make a segmented bar graph to compare these distributions.

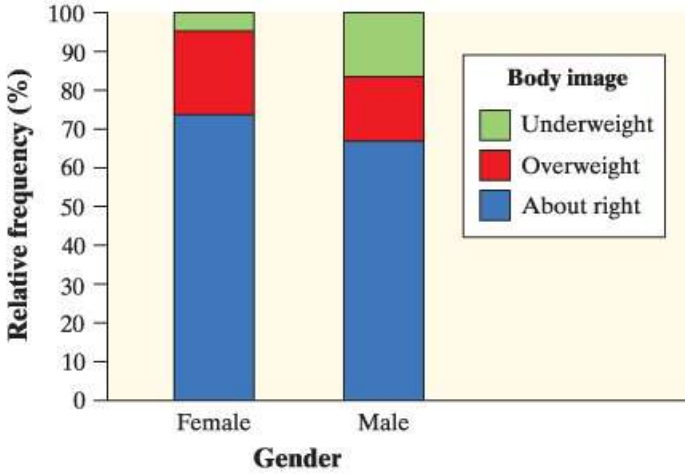
b. Describe what the graph in part (a) reveals about the association between country of origin and superpower preference for the students in the sample.

31. Body image Refer to Exercise 25.

a. Of the respondents who felt that their body weight was about right, what proportion were female?

b. Of the female respondents, what percent felt that their body weight was about right?

c. The segmented bar graph displays the distribution of perceived body image by gender. Describe what this graph reveals about the association between these two variables for the 1200 college students in the sample.

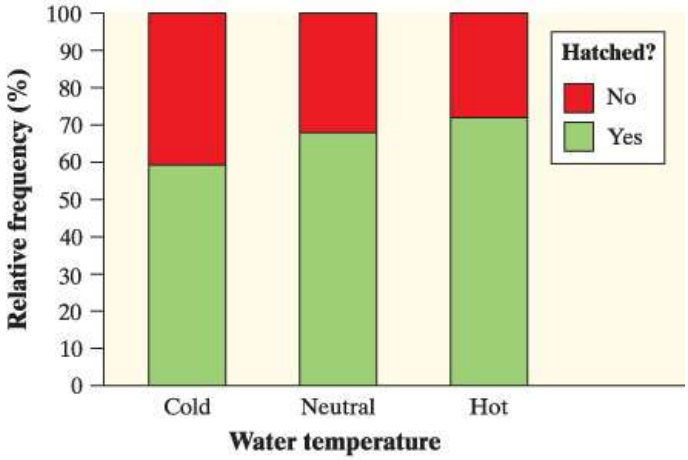


32. Python eggs Refer to Exercise 26.

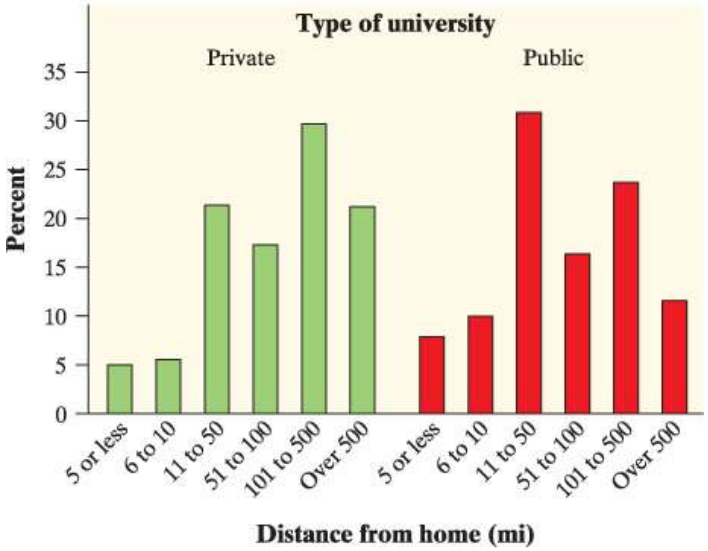
a. Of the eggs that hatched, what proportion were randomly assigned to hot water?

b. Of the eggs that were randomly assigned to hot water, what percent hatched?

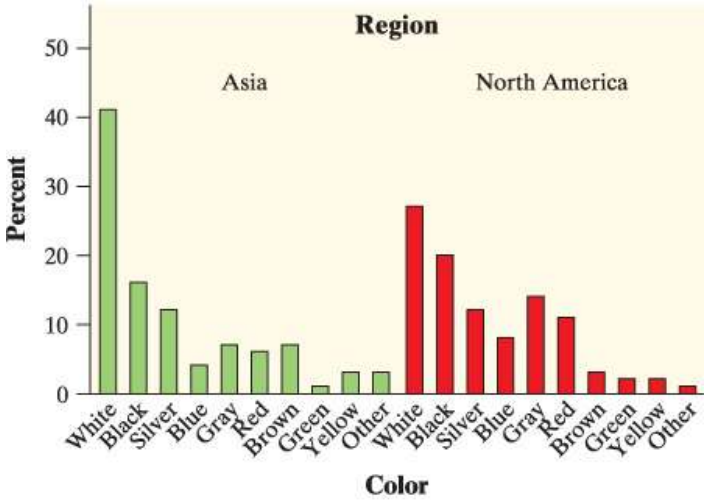
c. The segmented bar graph displays the distribution of hatching status by water temperature. Describe what this graph reveals about the association between these two variables for the python eggs in this experiment.



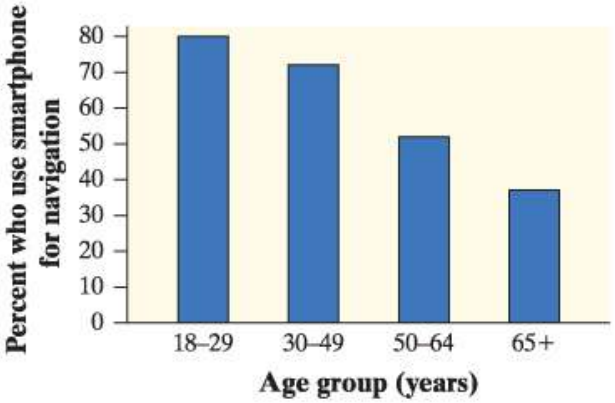
33. Far from home A survey asked first-year college students, "How many miles is this college from your permanent home?" Students had to choose from the following options: 5 or fewer, 6 to 10, 11 to 50, 51 to 100, 101 to 500, or more than 500. The side-by-side bar graph shows the percentage of students at public and private 4-year colleges who chose each option.[18] Write a few sentences comparing the distributions of distance from home for students from private and public 4-year colleges who completed the survey.



34. Popular car colors Favorite car colors may differ among countries. The side-by-side bar graph displays data on the most popular car colors in a recent year for North America and Asia. Write a few sentences comparing the distributions.[19]



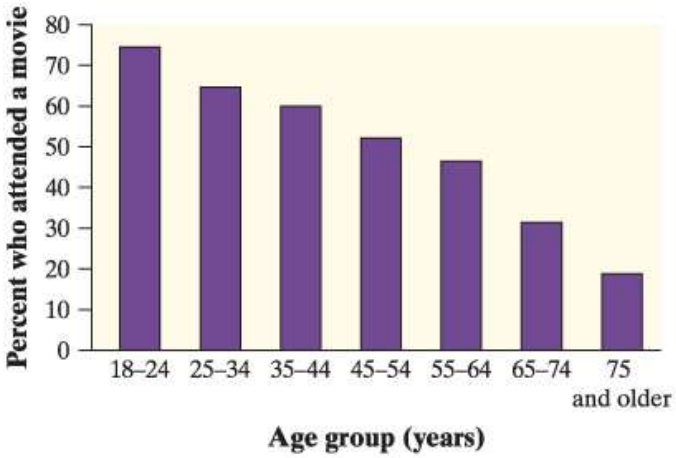
35. Phone navigation The bar graph displays data on the percent of smartphone owners in several age groups who say that they use their phone for turn-by-turn navigation.[20]



a. Describe what the graph reveals about the relationship between age group and use of smartphones for navigation.

b. Would it be appropriate to make a pie chart of the data? Explain.

36. Who goes to movies? The bar graph displays data on the percent of people in several age groups who attended a movie in the past 12 months.[21]



a. Describe what the graph reveals about the relationship between age group and movie attendance.

b. Would it be appropriate to make a pie chart of the data? Explain.

37. Marginal totals aren't the whole story Here are the row and column totals for a two-way table with two rows and two columns:

a	*b*	50
c	*d*	50
60	40	100

Find *two different* sets of counts *a, b, c,* and *d* for the body of the table that give these same totals. This shows that the relationship between two variables cannot be obtained from the two individual distributions of the variables.

38. Women and children first? Here's another table that summarizes data on survival status by gender and class of travel on the *Titanic*:

	Class of travel					
	First class		Second class		Third class	
Survival status	Female	Male	Female	Male	Female	Male
Survived	140	57	80	14	76	75
Died	4	118	13	154	89	387

a. Find the distributions of survival status for males and for females within each class of travel. Did women survive the disaster at higher rates than men? Explain.

b. In an earlier example, we noted that survival status is associated with class of travel. First-class passengers had the highest survival rate, while third-class passengers had the lowest survival rate. Does this same relationship hold for both males and females in all three classes of travel? Explain.

39. Simpson's paradox Accident victims are sometimes taken by helicopter from the accident scene to a hospital. Helicopters save time. Do they also save lives? The two-way table summarizes data from a sample of patients who were transported to the hospital by

helicopter or by ambulance.[22]

		Method of transport		
		Helicopter	**Ambulance**	**Total**
Survival status	Died	64	260	324
	Survived	136	840	976
	Total	200	1100	1300

a. What percent of patients died with each method of transport?

Here are the same data broken down by severity of accident:

Serious accidents

		Method of transport		
		Helicopter	**Ambulance**	**Total**
Survival status	Died	48	60	108
	Survived	52	40	92
	Total	100	100	200

Less serious accidents

		Method of transport		
		Helicopter	**Ambulance**	**Total**
Survival status	Died	16	200	216
	Survived	84	800	884
	Total	100	1000	1100

b. Calculate the percent of patients who died with each method of transport for the serious accidents. Then calculate the percent of patients who died with each method of transport for the less serious accidents. What do you notice?

c. See if you can explain how the result in part (a) is possible given the result in part (b).

Note: This is an example of *Simpson's paradox*, which states that an association between two variables that holds for each value of a third variable can be changed or even reversed when the data for all values of the third variable are combined.

Multiple Choice *Select the best answer for Exercises 40–43.*

40. For which of the following would it be inappropriate to display the data with a single pie chart?

a. The distribution of car colors for vehicles purchased in the last month

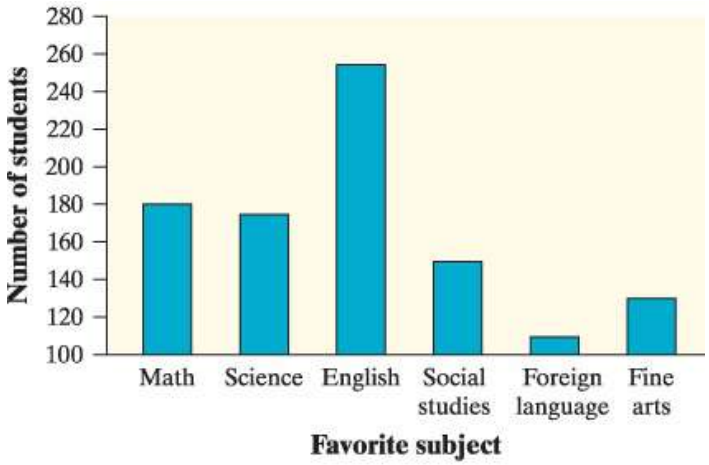
b. The distribution of unemployment percentages for each of the 50 states

c. The distribution of favorite sport for a sample of 30 middle school students

d. The distribution of shoe type worn by shoppers at a local mall

e. The distribution of presidential candidate preference for voters in a state

41. The following bar graph shows the distribution of favorite subject for a sample of 1000 students. What is the most serious problem with the graph?



a. The subjects are not listed in the correct order.

b. This distribution should be displayed with a pie chart.

c. The vertical axis should show the percent of students.

d. The vertical axis should start at 0 rather than 100.

e. The foreign language bar should be broken up by language.

42. The Dallas Mavericks won the NBA championship in the 2010–2011 season. The two-way table displays the relationship between the outcome of each game in the regular season and whether the Mavericks scored at least 100 points.

		Points scored		
		100 or more	Fewer than 100	Total
Outcome of game	Win	43	14	57
	Loss	4	21	25
	Total	47	35	82

Which of the following is the best evidence that there is an association between the outcome of a game and whether or not the Mavericks scored at least 100 points?

a. The Mavericks won 57 games and lost only 25 games.

b. The Mavericks scored at least 100 points in 47 games and fewer than 100 points in only 35 games.

c. The Mavericks won 43 games when scoring at least 100 points and only 14 games

when scoring fewer than 100 points.

d. The Mavericks won a higher proportion of games when scoring at least 100 points (43/47) than when they scored fewer than 100 points (14/35).

e. The combination of scoring 100 or more points and winning the game occurred more often (43 times) than any other combination of outcomes.

43. The following partially completed two-way table shows the marginal distributions of gender and handedness for a sample of 100 high school students.

		Gender		
		Male	Female	Total
Dominant hand	Right	*x*		90
	Left			10
	Total	40	60	100

If there is no association between gender and handedness for the members of the sample, which of the following is the correct value of *x*?

a. 20

b. 30

c. 36

d. 45

e. Impossible to determine without more information.

Recycle and Review

44. Hotels (Introduction) A high school lacrosse team is planning to go to Buffalo for a three-day tournament. The tournament's sponsor provides a list of available hotels, along with some information about each hotel. The following table displays data about hotel options. Identify the individuals and variables in this data set. Classify each variable as categorical or quantitative.

Hotel	Pool	Exercise room?	Internet ($/day)	Restaurants	Distance to site (mi)	Room service?	Room rate ($/day)
Comfort Inn	Out	Y	0.00	1	8.2	Y	149
Fairfield Inn & Suites	In	Y	0.00	1	8.3	N	119
Baymont Inn & Suites	Out	Y	0.00	1	3.7	Y	60
Chase Suite Hotel	Out	N	15.00	0	1.5	N	139
Courtyard	In	Y	0.00	1	0.2	Dinner	114
Hilton	In	Y	10.00	2	0.1	Y	156

Marriott	In	Y	9.95	2	0.0	Y	145