AP Statistics Summer Requirements

Welcome to AP Statistics! Next year, you will be expected to study hard, which includes doing assigned readings from the book and homework problems. Much of this course requires analysis. It is important to get into the habit of stating what needs to be said without redundancy.

The AP Statistics course uses the book *The Practice of Statistics, 7th edition.* We'll get started by completing the following two assignments, which are due the first day of school:

1. *Get the Most from this Program Assignment:* Please read the book introduction on pages xvii – xxiii. You will be familiarizing yourself with the format and features of the textbook. We will be using the book as a basis for our course, so you should know what to expect.

2. *Unit 1 Part 1 Assignment* (TO BE TURNED IN AND GRADED)
   a. Read the attached unit thoroughly. Pay attention to the "Tech Corner" boxes that discuss the use of technology (TI-83 Plus or TI-84 is required for the course) and the examples/exercises that report summary output from software packages.

   b. You are responsible for completing and turning in the exercises that are listed below.

      It is expected that your work is neat and legible. If not, it cannot be graded, and therefore will be given a zero. You may complete your work using a word processer or by hand. You will be expected to draw diagrams for some problems. Your answers should be concise, yet thorough.

      Section 1.A (page 8 – 10) # 2, 4, 6, 8, 9
      Section 1.B (page 16 – 20) # 2, 4, 6, 8, 10, 12, 14, 18, 19
      Section 1.C (page 38 – 47) # 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 24, 28, 30, 32, 40, 41, 42, 43, 44, 45
      Section 1.D (page 72 – 78) # 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 24, 28, 30, 32, 36, 40, 43, 44, 45, 46

Should you have any questions at any time throughout the summer, please contact either Mr. Brodie or Mrs. Chang ASAP. If you email us, we can try to make arrangements to work together on the assignments. Good luck!!! ☺

Mr. Brodie: tbrodie@wsdweb.org

Mrs. Chang: rchang@wsdweb.org

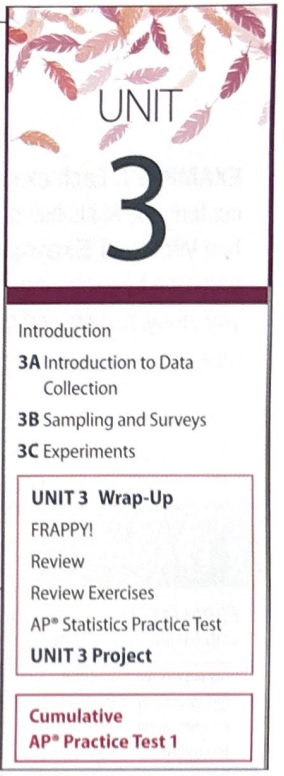
Get the Most from This Program

Your AP® Statistics adventure begins here! This program was created with YOU in mind. It is packed with features and resources to help you learn effectively and to do well on the AP® Statistics exam.

Written to support your AP® Statistics journey.

Your book matches the course both in organization and in scope. Units, Parts, and Sections focus on the concepts and skills that are central to the **AP® STATISTICS COURSE.** Daily practice is built in to support you at every step.

UNIT 1	**Exploring One-Variable Data**	1
PART I	Data Analysis	1
PART II	Modeling Distributions of Quantitative Data	87
UNIT 2	**Exploring Two-Variable Data**	141
UNIT 3	**Collecting Data**	239
	Cumulative AP® Practice Test 1	313
UNIT 4	**Probability, Random Variables, and Probability Distributions**	317
PART I	Probability	317
PART II	Random Variables and Probability Distributions	379
UNIT 5	**Sampling Distributions**	451
	Cumulative AP® Practice Test 2	524



AP® EXAM TIP

Students often do not get full credit on the AP® Statistics exam because they use option (ii) with "calculator-speak" to show their work on normal distribution calculation questions — for example, normalcdf(−1000,6,6.84,1.55). This is *not* considered clear communication. To get full credit, be sure to carefully label each of the inputs in the calculator command if you use technology in Step 2: normalcdf(lower: −1000, upper: 6, mean: 6.84, SD: 1.55).

AP® EXAM TIP
Formula sheet

The regression line is defined as $\hat{y} = a + bx$ on the formula sheet provided on both sections of the AP® Statistics exam.

AP® EXAM TIP
AP® Daily Videos

Review the content of this section and get extra help by watching the AP® Daily Videos for Topics 1.1–1.3, which are available in AP® Classroom.

Pay attention to the Integrated **AP® EXAM TIPS** that support you in three ways:

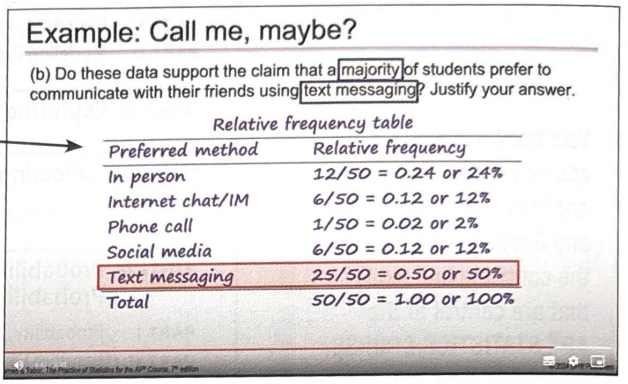
- By providing advice from experts about how to be successful on the AP® Statistics exam
- By offering insights into the AP® Statistics formula sheet so you know where to focus your study
- By guiding you to the relevant AP® Daily Video that corresponds to each section in the text

Examples Are Your Friends

Study the examples for detailed instruction, with text and video guidance, for every concept and skill in the course.

Examples with paired exercises are the heart of the program.

EXAMPLES. Each example title is followed by the content and skills being taught and a "play button" link to a **Worked Example Video.** These videos were prepared by experienced AP® Statistics teachers to help you study, review, and get extra support when needed.



EXAMPLE Call me, maybe? Summarizing data with tables — Skills 2.A, 2.B

PROBLEM: Here are the data on preferred method of communicating with friends for all 50 students in the Census at School sample:

Internet chat/IM	In person	In person	Internet chat/IM	Phone call
Text messaging	Text messaging	Text messaging	Social media	Social media
Text messaging	Text messaging	Text messaging	In person	In person
Text messaging	Text messaging	Text messaging	In person	In person
Text messaging	Text messaging	Internet chat/IM	Social media	Text messaging
Text messaging	Text messaging	In person	Text messaging	Internet chat/IM
Text messaging	Social media	Text messaging	Social media	Text messaging
Text messaging	In person	Social media	In person	Text messaging
Text messaging	Text messaging	Text messaging	Internet chat/IM	In person
Text messaging	In person	Internet chat/IM	Text messaging	In person


SDI Productions/E+/Getty Images

(a) Make a frequency table and a relative frequency table to summarize the distribution of preferred communication method.

(b) Do these data support the claim that a majority of students prefer to communicate with their friends using text messaging? Justify your answer.

SOLUTION:

(a)

Frequency table

Preferred method	Frequency
In person	12
Internet chat/IM	6
Phone call	1
Social media	6
Text messaging	25
Total	50

The frequency table shows the *number* of students who prefer each communication method. To create the frequency table, count how many students said "In person," how many said "Internet chat or instant messenger," and so on.

Relative frequency table

Preferred method	Relative frequency
In person	12/50 = 0.24 or 24%
Internet chat/IM	6/50 = 0.12 or 12%
Phone call	1/50 = 0.02 or 2%
Social media	6/50 = 0.12 or 12%
Text messaging	25/50 = 0.50 or 50%
Total	50/50 = 1.00 or 100%

The relative frequency table shows the *proportion* or *percentage* of students who prefer each communication method. Note that in statistics, a proportion is a value between 0 and 1 that is equivalent to a percentage.

(b) No. Exactly 50% of the students in the sample said that they prefer to communicate with their friends via text messaging, but a majority is more than half.

FOR PRACTICE, TRY EXERCISE 5

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

THE VOICE OF THE TEACHER. Pay special attention to the carefully placed "teacher talk" boxes that guide you through the solution. These comments offer lots of good advice — as if your teacher is working directly with you to solve a problem.

PRACTICE! Take time to work the paired problem in the Section Exercises. Practicing a similar problem after studying the worked example helps you know whether or not you've "got it."

Exercises: Practice Makes Perfect!

Work the exercises assigned by your teacher. Watch the videos when you need extra help.

Time spent working exercises is key to success in this course.

SECTION EXERCISES ORGANIZED BY TOPIC
Use the handy headings to guide you to topics on which you need practice.

THE TPS "BACKWARD/FORWARD" NAVIGATION SYSTEM

Do you read the text first, or do you start with the assigned exercises? Either way, *TPS* is designed to support you.

Look for the icons that appear next to selected exercises. They will guide you

- Back to the **example** that models the exercise so that you can review the worked solution, if you are confused.
- To a short **video** that provides step-by-step instruction for solving the exercise.

Summarizing Data with Tables

5. pg 6 New York squirrels How do squirrels in New York City's Central Park respond to humans? The Squirrel Census project enlisted volunteers to record data on all squirrels observed in Central Park in a recent year. One of the variables measures how the squirrels responded to humans. We selected a random sample of 75 of the more than 3000 squirrels observed. Here are the data on their reactions to humans: [7]

Approach	Run from	Indifferent	Indifferent	Run from
Indifferent	Approach	Indifferent	Indifferent	Indifferent
Approac				
Indiffere				

Ex 5: New York squirrels

Frequency Table

Response	Frequency
Approach	II
Run	II
Indifferent	IIII IIII I
Total	

Approach	Run from	Indifferent	Indifferent	Run from
Indifferent	Approach	Indifferent	Indifferent	Indifferent
Approach	Indifferent	Approach	Run from	Run from
Indifferent	Indifferent	Indifferent	Run from	Indifferent
Indifferent	Indifferent	Run from	Indifferent	Indifferent
Indifferent	Run from	Run from	Indifferent	Indifferent
Run from	Run from	Approach	Approach	Run from
Indifferent	Approach	Indifferent	Indifferent	Indifferent
Indifferent	Approach	Indifferent	Indifferent	Indifferent
Indifferent	Run from	Approach	Approach	Indifferent
Indifferent	Indifferent	Run from	Approach	Indifferent
Indifferent	Approach	Indifferent	Run from	Approach
Run from	Approach	Run from	Indifferent	Indifferent
Indifferent	Approach	Indifferent	Run from	Indifferent
Indifferent	Indifferent	Approach	Approach	Indifferent

For Investigation *Apply the skills from the section in a new context or nonroutine way.*

38. **Population pyramids** The histograms show the age distribution of the population in Australia and Ethiopia in a recent ye
tional databas
population py

Multiple Choice *Select the best answer for each question.*

40. Here are the amounts of money (cents) in coins carried by 10 students in a statistics class: 50, 35, 0, 46, 86, 0, 5, 47, 23, 65. To ma
would use stems

(A) 0, 2, 3, 4, 5, 6,

(B) 0, 1, 2, 3, 4, 5,

Recycle and Review *Practice what you learned in previous sections.*

46. **Marital Status (1B)** The side-by-side bar graph compares the marital status of U.S. adult residents (18 years old or older) in 1980 and 2020.[63] Compare the distributions of marital status for these two years.

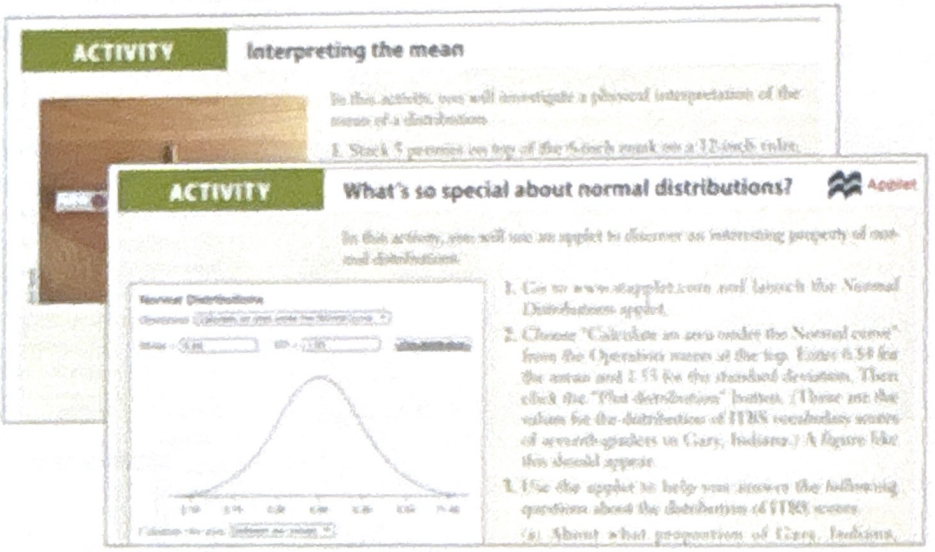
Additional problem types in the Section Exercises let you practice solving many different types of questions, including:

- **For Investigation** exercises that assess multiple skills and concepts, and let you apply your knowledge in nonroutine ways.
- **AP®-style Multiple Choice** questions that give you regular practice for the AP® Statistics exam.
- **Recycle and Review** problems that refer back to concepts and skills learned in an earlier section, as noted in the references in parentheses (1B).

Learn Statistics by Doing Statistics

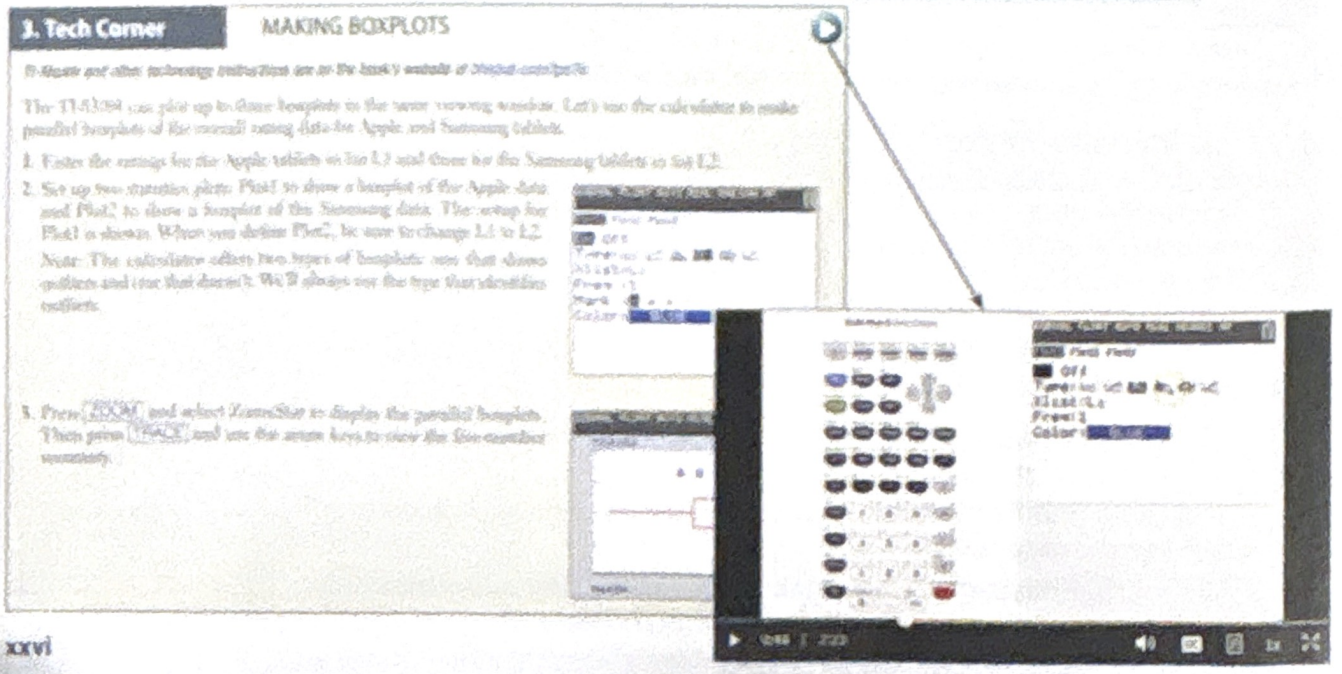
Experience statistical principles with simulations and technology tools.

ACTIVITIES Every unit includes a hands-on activity in the first few pages to introduce the content, with other activities appearing later in the unit. Many of the 30 activities use dynamic **applets** that help you experience the process of collecting data and drawing conclusions from those data. All of the applets are available to you whether you do the activity in class or on your own.



TECHNOLOGY Use technology as a tool for discovery and analysis. The 30 **Tech Corners,** placed strategically throughout the book at the optimal point of use, give step-by-step instructions for using the TI-83/84 calculator. Instructions for additional calculators are available on the book's website.

Supporting **Tech Corner Videos** are available to walk you through the keystrokes needed to perform each analysis.

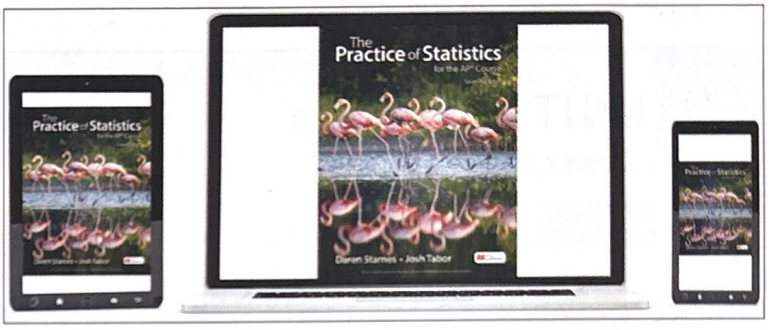


Achieve bfw publishers Your Goals

Access everything you need for success through the ACHIEVE platform.

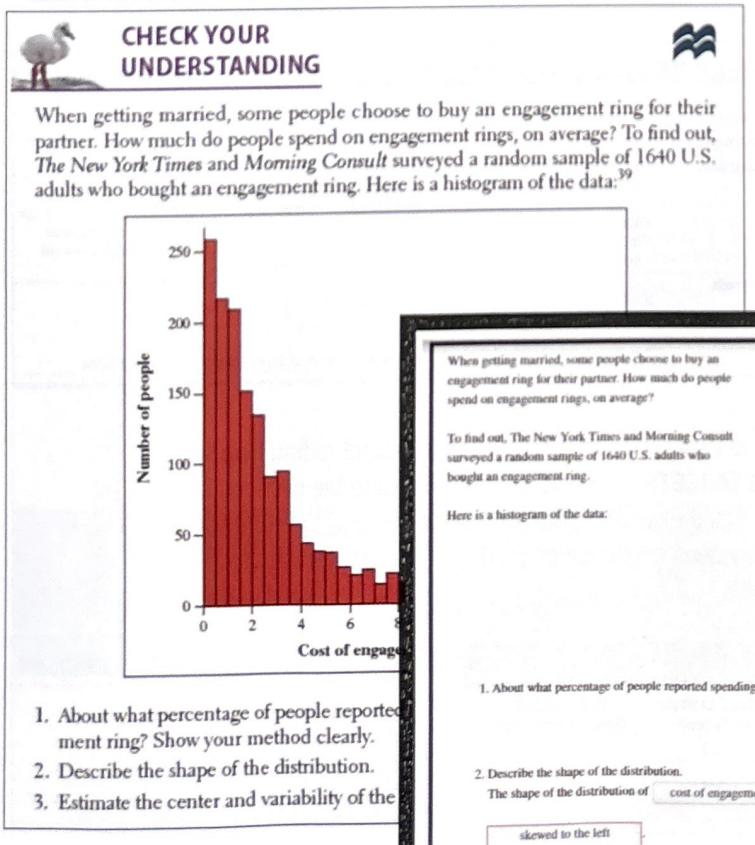
Read the text, watch videos, and receive guided feedback on your homework.

EASY ACCESS TO RESOURCES WITH THE E-BOOK Interactive and mobile-ready, the e-book allows you to read and reference the text when you are online and offline. All offline highlights and notes sync when you connect to the internet. What's more, clicking on ▶ gives you instant access to each video while you are online.



The **online homework** system helps you learn with targeted feedback based on common misconceptions, so you learn even if you get the answer wrong!

Hundreds of practice problems from the text are available, including the **Check Your Understanding** problems, selected **Section Exercises**, all **Unit/Part Review exercises,** and each **Unit/Part AP® Statistics Practice Test.**

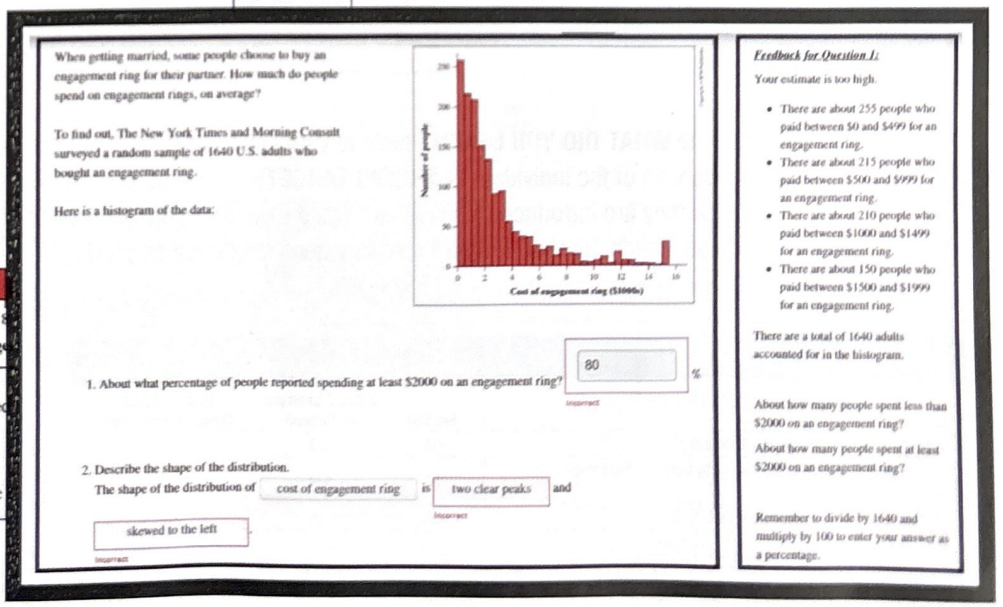


CHECK YOUR UNDERSTANDING

When getting married, some people choose to buy an engagement ring for their partner. How much do people spend on engagement rings, on average? To find out, *The New York Times* and *Morning Consult* surveyed a random sample of 1640 U.S. adults who bought an engagement ring. Here is a histogram of the data:[39]

1. About what percentage of people reported ment ring? Show your method clearly.
2. Describe the shape of the distribution.
3. Estimate the center and variability of the

Sprinkled throughout the text after key concepts and skills are presented, **Check Your Understanding** questions do just that. Working them in Achieve or on paper will help you stay on track.



When getting married, some people choose to buy an engagement ring for their partner. How much do people spend on engagement rings, on average?

To find out, The New York Times and Morning Consult surveyed a random sample of 1640 U.S. adults who bought an engagement ring.

Here is a histogram of the data:

1. About what percentage of people reported spending at least $2000 on an engagement ring? 80 % Incorrect

2. Describe the shape of the distribution.
The shape of the distribution of cost of engagement ring is two clear peaks and Incorrect
skewed to the left. Incorrect

Feedback for Question 1:

Your estimate is too high.

- There are about 255 people who paid between $0 and $499 for an engagement ring.
- There are about 215 people who paid between $500 and $999 for an engagement ring.
- There are about 210 people who paid between $1000 and $1499 for an engagement ring.
- There are about 150 people who paid between $1500 and $1999 for an engagement ring.

There are a total of 1640 adults accounted for in the histogram.

About how many people spent less than $2000 on an engagement ring?

About how many people spent at least $2000 on an engagement ring?

Remember to divide by 1640 and multiply by 100 to enter your answer as a percentage.

Review and Practice for Quizzes and Tests

Take advantage of the review tools included at the end of every Unit and Part.

UNIT 6 Wrap-Up
PART I

FRAPPY! Free Response AP® Problem, Yay!

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.

Members at a popular fitness club currently pay a $40 per month membership fee. The owner of the club wants to raise the fee to $50 but is concerned that some members will leave the gym if the fee increases. To investigate, the owner plans to survey a random sample of the club members and construct a 95% confidence interval for the proportion of all members who would quit if the fee was raised to $50.

DO THE FRAPPY! Learn how to answer FRQs successfully by working the FRAPPY! — the Free Response AP® Problem, Yay! — that begins each Unit/Part Wrap-up.

SNAPSHOT REVIEW Study the Unit Review, which gives a short summary of each section, to be sure you understand the key concepts in each section.

SUMMARY TABLES in Units 5–9 review important details of each sampling distribution and inference procedure, including conditions and formulas.

UNIT 6, PART I REVIEW

SECTION 6A Confidence Intervals: The Basics

In this section, you learned that a **point estimate** is the single best guess for the value of a population parameter. You also learned that a **confidence interval,** also known as an **interval estimate,** provides an interval of plausible values for a parameter based on sample data. To interpret a confidence interval, say, "We are C% confident that the interval

size from the same population and used them to construct C% confidence intervals, about C% of those intervals would capture the [parameter in context]."

SECTION 6B Confidence Intervals for a Population Proportion

In this section, you learned how to construct and interpret

Inference for a Population Proportion

	Confidence Interval for p	Significance Test for p
Name (TI-83/84)	One-sample z interval for p (1-PropZInt)	One-sample z test for p (1-PropZTest)
Null Hypothesis	*Not applicable.*	$H_0: p = p_0$
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 $n\hat{p}$ and $n(1-\hat{p})$ are at least 10. That is, the number of successes and the 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, $n < 0.10N$. • **Large Counts:** Both np_0 and $n(1-p_0)$ are at least 10, where p_0 is the proportion specified by the null hypothesis.
Formula	$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
	Critical value z^* from the standard normal distribution.	P-value from the standard normal distribution.

Use the **WHAT DID YOU LEARN?** table to verify your mastery of each topic and to find help when needed. All of the individual **LEARNING TARGETS** are listed with references to the sections in which they are introduced, as well as related examples and relevant Unit Review Exercises. And, of course, watch the Unit Review Exercise Videos to get tips on solving these multifaceted problems.

What Did You Learn?

Learning Target	Section	Related Example on Page(s)	Relevant Unit Review Exercise(s)
Interpret a confidence interval in context.	6A	533	R1, R2
Use a confidence interval to make a decision about the value of a parameter.	6A	534	R1, R5
Interpret a confidence level in context.	6A	536	R1

UNIT 6, PART I REVIEW EXERCISES

These exercises are designed to help you review the important concepts and skills of the unit.

R1 **Sports fans (6A)** Are you a sports fan? That's the question the Gallup polling organization asked a random sample of 1527 U.S. adults.[57] Gallup reported that a 95% confidence interval for the proportion of all U.S. adults who are sports fans is 0.565 to 0.615.

(a) Interpret the confidence interval.

(b) Interpret the confidence level.

(c) Based on the interval, is there convincing evidence that a majority of U.S. adults are sports fans? Explain your answer.

... and for the AP® Statistics Exam

Consistent daily practice will help you avoid cramming before the AP® Statistics exam.

Each Unit/Part concludes with an **AP® STATISTICS PRACTICE TEST** that usually consists of 10 multiple-choice questions and 3 free-response questions.

UNIT 6, PART I AP® STATISTICS PRACTICE TEST

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

T1 A confidence interval for a proportion is 0.272 to 0.314. What are the point estimate and the margin of error for this interval?

(A) 0.293, 0.021
(B) 0.293, 0.042
(C) 0.293, 0.084
(D) 0.586, 0.021
(E) 0.586, 0.042

(A) 0.05
(B) 0.10
(C) 0.20
(D) 0.40
(E) 0.90

T5 Bags of a certain brand of tortilla chips are labeled as having a net weight of 14 ounces. A representative of a consumer advocacy group wishes to see if there

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

T11 The U.S. Forest Service is considering additional restrictions on the number of vehicles allowed to enter Yellowstone National Park. To assess public reaction, the service asks a random sample of 150 visitors if they favor the proposal. Of these, 89 say "Yes."

(a) Construct and interpret a 99% confidence interval for the proportion of all visitors to Yellowstone who favor the restrictions.

(b) Interpret the confidence level.

(a) Do these data provide convincing evidence for the Pew Research Center's claim?

(b) Which type of error, Type I or Type II, could you have made in part (a)? Explain your answer.

T13 A mattress manufacturer has designed a new production process that is intended to reduce the proportion of mattresses with defects in the stitching. Before the change, 15% of the mattresses had this type of defect. After the new production process is implemented, 100

Cumulative AP® Practice Test 1

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

AP1 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.

AP4 For a biology project, you measure the weight in grams (g) and the tail length in millimeters (mm) of a group of mice. The equation of the least-squares line for predicting tail length from weight is

$$\text{predicted tail length} = 20 + 3 \times \text{weight}$$

Which of the following is *not* correct?

(A) A mouse's predicted tail length increases by about 3 mm for each additional gram of weight.
(B) The pred... 38 g is 1...
(C) The corr... positive.
(D) If you ha... ters inste... sion line...
(E) Mice tha... length 2...

Section II: Part A *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

AP16 The manufacturer of exercise machines for fitness centers has designed two new elliptical machines that are meant to increase cardiovascular fitness. The two machines are being tested on 30 volunteers at a fitness center near the company's headquarters. The volunteers are randomly assigned to one of the machines and use it daily for two months. A measure of cardiovascular fitness is administered at the start of the experiment and again at the end. The following stemplot contains the differences (After − Before) in the two scores for subjects using the two machines. Note that greater differences indicate larger gains in fitness.

(a) Write a few sentences comparing the distributions of cardiovascular fitness gains from the two elliptical machines.

(b) Which machine should be chosen if the company wants to advertise it as achieving the highest overall gain in cardiovascular fitness? Explain your reasoning.

(c) Which machine should be chosen if the company wants to advertise it as achieving the most consistent gain in cardiovascular fitness? Explain your reasoning.

(d) What is the largest population to which we can apply the results of this study?

AP17 A national tire company wants to compare two different types of incentives for increasing sales: awarding cash bonuses or awarding noncash prizes such as vacations. The company has 60 retail sales districts of various sizes across the country and will randomly

Four **CUMULATIVE AP® PRACTICE TESTS** simulate the real exam. Placed after Units 3, 5, 7, and 9, these tests expand in coverage of content and skills as you work through the book. The last test models a full AP® Statistics exam.

Strategies for Success on the AP® Statistics Exam

ABOUT THE AP® STATISTICS EXAM

The AP® Statistics exam consists of two distinct sections: Multiple Choice and Free Response. Here are specific details about the structure and scoring of the exam, as of Fall 2023.

AP® STATISTICS EXAM STRUCTURE		
Section I: Multiple Choice *40 multiple choice questions, each with 5 answer choices*	90 minutes	50% of exam score

The composition of the Multiple Choice section is based on both content and skills. Here are the number of questions for each unit and skill category.

Consult the **STRATEGIES FOR SUCCESS ON THE AP® STATISTICS EXAM** appendix for insight into the AP® Statistics exam and advice on how to take it with confidence.

UNIT

1

Exploring One-Variable Data

PART I
Data Analysis

Introduction

1A Statistics: The Language of Variation

1B Displaying and Describing Categorical Data

1C Displaying Quantitative Data with Graphs

1D Describing Quantitative Data with Numbers

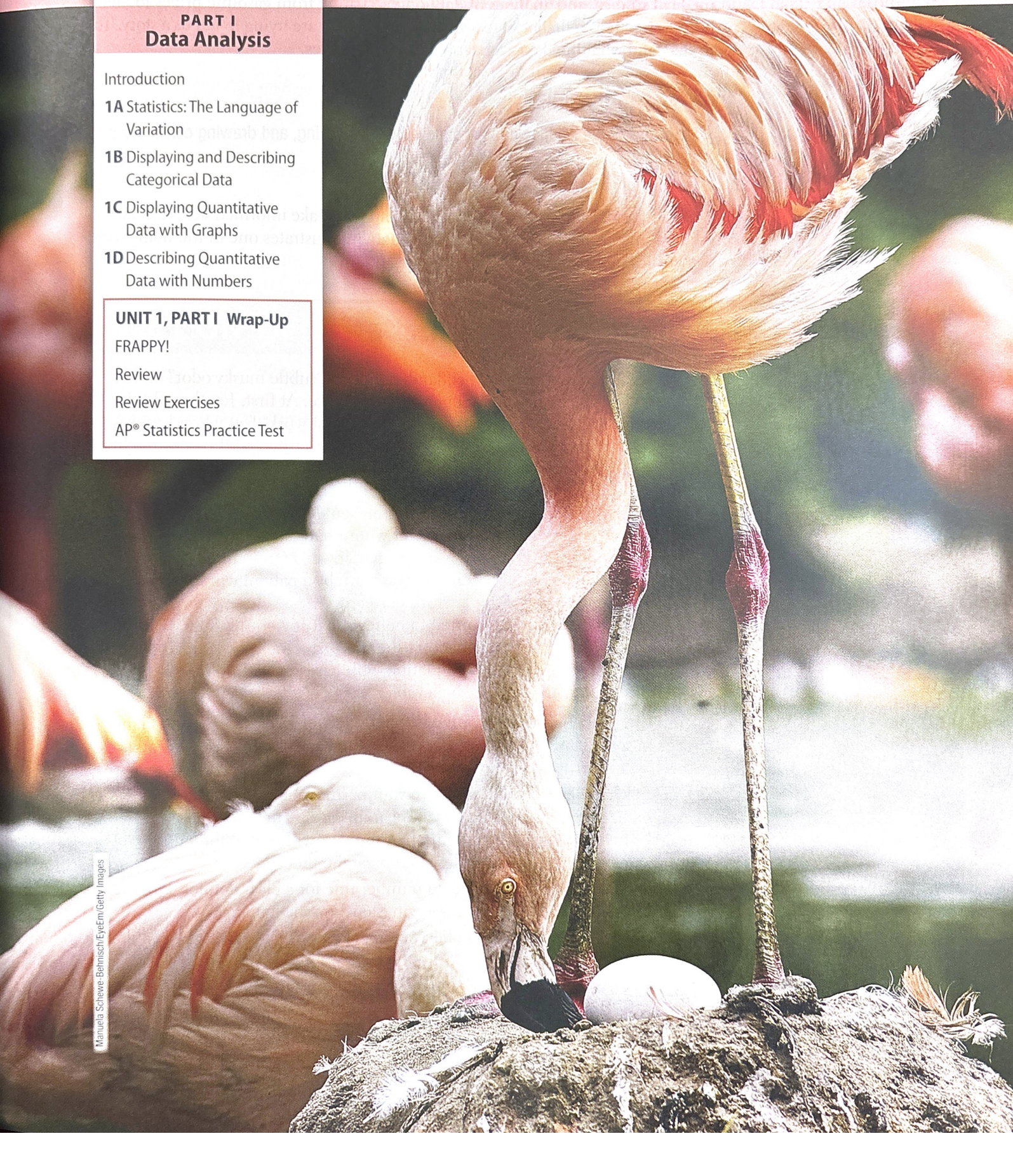
UNIT 1, PART I Wrap-Up

FRAPPY!

Review

Review Exercises

AP® Statistics Practice Test



Manuela Schewe-Behnisch/EyeEm/Getty Images

Introduction

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 gasoline prices to elections to weather to the latest technology. These data are trying to tell us a story. To understand what the data are saying, we use **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 informed decisions based on data in your daily life. The following activity illustrates one of the many uses of statistics in the real world.

ACTIVITY Smelling Parkinson's

Applet

Joy Milne, a retired nurse from Scotland, noticed a "subtle musky odor" on her husband Les that she had never encountered before. At first, Joy thought the smell might be from Les's sweat after long hours of work. But when Les was diagnosed with Parkinson's disease 6 years later, Joy suspected the odor might be a result of the disease.

Scientists were intrigued by Joy's claim and designed an experiment to test her ability to "smell Parkinson's." Joy was presented with 12 different shirts, each worn by a different person, some of whom had Parkinson's disease and some of whom did not. The shirts were given to Joy in a random order, and she had to decide whether or not each shirt was worn by a patient with Parkinson's disease. Joy identified 11 of the 12 shirts correctly.

Although the researchers wanted to believe that Joy could detect Parkinson's disease by smell, it is possible that she was just a lucky guesser. You and your classmates will perform a simulation with cards or technology to determine which explanation is more believable.



© Chris Watt/Telegraph Media Group Limited 2017

1. Your teacher will hand each pair of students a set of 12 cards (shirts). On the back of some cards is "Parkinson's" and on the back of others is "No Parkinson's." Shuffle the cards thoroughly.
2. Decide who will guess first and have your partner act as the researcher. For each card, guess "Parkinson's" or "No Parkinson's." The researcher will not reveal whether each guess is right or wrong, but will record the number of correct guesses. Now switch roles and repeat the process.
3. Your teacher will draw and label a number line for a class dotplot. Plot the number of correct guesses you made in Step 2 on the graph.
4. Repeat Steps 1 and 2 until you have a total of at least 50 trials of the simulation for your class.
5. How often were 11 or more shirts correctly identified by chance alone? Based on this result, which seems more believable: Joy is just a lucky guesser, or Joy really can smell Parkinson's disease? Explain your reasoning.

The Smelling Parkinson's activity outlined the steps in the *statistical problem-solving process*: (1) Ask questions, (2) collect data, (3) analyze the data, and (4) interpret the results.[1] Researchers began by identifying the question to be answered: "Can Joy Milne smell Parkinson's disease?" To answer this question, the researchers designed and carried out an appropriate plan to collect data. The resulting data consisted of Joy's 11 correct and 1 incorrect shirt identifications. When analyzing the data, researchers had to consider the possibility that Joy was just guessing and correctly identified 11 of the 12 shirts by chance alone. After careful analysis, the researchers concluded that Joy Milne could actually smell Parkinson's disease. To make the case even stronger, the researchers later discovered that Joy had correctly identified all 12 shirts. Her one "mistake" was a person who was diagnosed with Parkinson's disease a few months later. That's pretty amazing!

In AP® Statistics, you will be asked to solve a variety of statistical problems. Your success will depend on steadily developing *skills* in four categories: selecting statistical methods, analyzing data, using probability and simulation, and making statistical arguments.[2] We'll emphasize both content and skills throughout the book.

SECTION 1A Statistics: The Language of Variation

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

- Identify the individuals and variables in a set of data, and classify the variables as categorical or quantitative.
- Make and interpret a frequency table or a relative frequency table for a distribution of data.

People's opinions vary on all kinds of issues. Data on physical characteristics of people, animals, and plants vary from one individual to another. Chance outcomes — like tosses of a coin or guesses about which shirts were worn by patients with Parkinson's disease — vary. Statistics is the language of variation.

This section starts by showing you how to organize and classify data. Then, you will learn how to represent data in tabular form.

Individuals and Variables

Every year, the U.S. Census Bureau collects data from more than 3.5 million randomly selected U.S. households as part of the American Community Survey (ACS). The table displays some data from 10 households included in the ACS in a recent year.[3]

Household	Region	People in household	Time in dwelling (years)	Response mode	Household income ($)	Internet access?
1	South	3	20–29	Phone	272,000	Yes
2	Midwest	3	30+	Internet	54,600	Yes
5	South	1	10–19	Mail	49,900	Yes
6	South	2	2–4	Phone	297,000	Yes
7	Northeast	4	5–9	Internet	130,000	Yes
11	Midwest	3	2–4	Internet	82,000	Yes
14	Midwest	3	10–19	Internet	57,000	Yes
16	West	3	2–4	Mail	36,800	Yes
17	South	2	2–4	Mail	133,000	Yes
25	Midwest	1	20–29	Mail	80,000	No



Bruce Leighty/Photodisc/Getty Images

Most data tables follow this format: each row describes an **individual,** and each column contains the values of a **variable.** Sometimes the individuals in a data set are called *cases* or *observational units.*

DEFINITION Individual, Variable

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

A **variable** is a characteristic that can take different values for different individuals.

For the ACS data set, the *individuals* are households. It is important to note that household is *not* a variable in this data set — the numbers in that column of the data table are just labels for the individuals. The *variables* recorded for each household are region, people in household, time in their 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.** People in household and household income (in dollars) are **quantitative variables.**

DEFINITION Categorical variable, Quantitative variable

A **categorical variable** takes values that are labels, which 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 nor measurements of anything. Instead, they are simply labels for a regional location, making zip code a categorical variable. 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.

The variable "year" is often treated as categorical. But it depends on how the data are being used. Consider a data set about cars, in which one of the variables recorded is model year. If we want to know what percentage of cars on the road are 2022 models, we treat year as categorical. If we want to know the average age of cars on the road, we would convert model year to age (in years) and treat this variable as quantitative.

EXAMPLE Census at school Individuals and variables

Skill 2.A



PROBLEM: Census at School is an international project that collects data about primary and secondary school students using surveys. Since its launch in 2000, students from Australia, Canada, Ireland, Japan, New Zealand, South Africa, the United Kingdom, and the United States have taken part in the project. We selected a random sample of 50 U.S. high school students who completed the survey in a recent year. The table displays data on some of the survey questions for the first 10 students in the sample.



EF Volart/Moment/Getty Images

State	Birth month	Age (years)	Handed	Height (cm)	Home occupants	Allergies	Preferred communication
WI	11	17	Right	175	4	Yes	Internet chat/IM
IN	6	16	Right	175.5	5	No	In person
NY	6	17	Right	157	5	Yes	In person
NC	6	17	Right	169	3	No	Internet chat/IM
MA	6	18	Right	169	3	Yes	Phone call
MO	10	18	Right	170	5	No	Text messaging
PA	5	14	Right	170	6	No	Text messaging
IA	1	17	Left	176	2	No	Text messaging
NC	5	17	Right	175	5	No	Social media
CA	2	17	Right	158	8	Yes	Social media
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(a) Identify the individuals and variables in this data set.
(b) Classify each variable as categorical or quantitative.

SOLUTION:

(a) Individuals: 50 randomly selected U.S. high school students who completed the Census at School survey. Variables: State, birth month, age (years), handedness, height (cm), number of home occupants, whether the student has allergies, preferred communication method.

We'll see in Unit 3 why choosing at random, as we did in this example, is a good idea.

(b) Categorical: State, birth month, handedness, whether the student has allergies, preferred communication method. Quantitative: Age (years), height (cm), and number of home occupants.

Note that birth month is categorical, even though the values listed are numbers.

FOR PRACTICE, TRY EXERCISE 1

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 data appropriately on the AP® Statistics exam.

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. Be sure to include any units of measurement for a quantitative variable (like centimeters for height). To make life simpler, we sometimes refer to *categorical data* or *quantitative data* instead of identifying the variable as categorical or quantitative.

Summarizing Data with Tables

A variable generally takes values that vary from one individual to another. That's why we call it a variable! The **distribution** of a variable describes the pattern of variation of the values.

DEFINITION Distribution

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

We can summarize a variable's distribution with a **frequency table** or a **relative frequency table.** To make either kind of table, start by tallying the number of times that the variable takes each value.

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 percentage of individuals having each value.

We can use a frequency table or a relative frequency table to help describe the distribution of a variable, as the following example illustrates.

EXAMPLE Call me, maybe? Summarizing data with tables

Skills 2.A, 2.B



SDI Productions/E+/Getty Images

PROBLEM: Here are the data on preferred method of communicating with friends for all 50 students in the Census at School sample:

Internet chat/IM	In person	In person	Internet chat/IM	Phone call
Text messaging	Text messaging	Text messaging	Social media	Social media
Text messaging	Text messaging	Text messaging	In person	In person
Text messaging	Text messaging	Text messaging	In person	In person
Text messaging	Text messaging	Internet chat/IM	Social media	Text messaging
Text messaging	Text messaging	In person	Text messaging	Internet chat/IM
Text messaging	Social media	Text messaging	Social media	Text messaging
Text messaging	In person	Social media	In person	Text messaging
Text messaging	Text messaging	Text messaging	Internet chat/IM	In person
Text messaging	In person	Internet chat/IM	Text messaging	In person

(a) Make a frequency table and a relative frequency table to summarize the distribution of preferred communication method.

(b) Do these data support the claim that a majority of students prefer to communicate with their friends using text messaging? Justify your answer.

SOLUTION:

(a)

Frequency table

Preferred method	Frequency
In person	12
Internet chat/IM	6
Phone call	1
Social media	6
Text messaging	25
Total	50

The frequency table shows the *number* of students who prefer each communication method. To create the frequency table, count how many students said "In person," how many said "Internet chat or instant messenger," and so on.

Relative frequency table

Preferred method	Relative frequency
In person	12/50 = 0.24 or 24%
Internet chat/IM	6/50 = 0.12 or 12%
Phone call	1/50 = 0.02 or 2%
Social media	6/50 = 0.12 or 12%
Text messaging	25/50 = 0.50 or 50%
Total	50/50 = 1.00 or 100%

The relative frequency table shows the *proportion* or *percentage* of students who prefer each communication method. Note that in statistics, a proportion is a value between 0 and 1 that is equivalent to a percentage.

(b) No. Exactly 50% of the students in the sample said that they prefer to communicate with their friends via text messaging, but a majority is more than half.

FOR PRACTICE, TRY EXERCISE 5



Note that the frequencies and relative frequencies listed in these tables are not data. The frequency and relative frequency tables summarize the data by telling us how many, or what proportion or percentage of, students in the Census at School sample prefer each method of communicating with friends.

The same process can be used to summarize the distribution of a quantitative variable. However, it would be hard to make a frequency table or a relative frequency table for quantitative data that take many different values, like height (cm) in the Census at School data set. We'll look at a better option for summarizing quantitative variables in Section 1C.

FROM DATA ANALYSIS TO INFERENCE

Sometimes we're interested in drawing conclusions that go beyond the data at hand. That's the idea of *statistical inference*. In the Census at School survey, 12 of the 50 randomly selected students prefer communicating in person with their friends. That's 24% of the *sample*. Can we conclude that exactly 24% of the *population* of students who completed the online survey prefer communicating in person with their friends? No.

If another random sample of 50 students who completed the survey were selected, the percentage who prefer communicating in person with their friends would probably not be exactly 24%. Can we at least say that the actual population value is "close" to 24%? As you will learn in future units, that depends on what we mean by "close."

Our ability to do statistical inference is determined by how the data are produced. Unit 3 discusses the two main methods of collecting data — sampling and experiments — and the types of conclusions that can be drawn from each. As the Smelling Parkinson's activity at the beginning of this unit illustrates, the logic of inference rests on asking, "What are the chances?" *Probability*, the study of chance behavior, is the focus of Units 4 and 5. We'll introduce the most common methods of statistical inference in Units 6–9.



CHECK YOUR UNDERSTANDING



Malena is a car buff who wants to find out more about the cars that high school students drive. The principal of a local high school gives Malena permission to go to the student parking lot and record some data. Later, Malena does some internet research on each model of car in the parking lot and makes a spreadsheet that includes each car's license plate, model, year, number of stickers on the car, color, weight (in kilograms), whether it has a navigation system, and highway gas mileage.

1. Identify the individuals and variables in Malena's study.
2. Classify each variable as categorical or quantitative.

A professor suspects that most students in an 8 A.M. introductory statistics class did not complete the pre-class reading assignment. To find out, the professor gave the students a 10-question pop quiz about the assigned reading at the beginning of class. Here are the number of correct answers on the pop quiz for the 50 students in the class:

9	8	6	7	7	8	4	7	7	8	8	8	6	7	8	8	7	7	6	8	9	7	6	5	7
8	8	7	9	6	6	6	8	9	5	8	7	7	7	7	2	4	8	3	6	5	5	8	7	3

3. Make a frequency table and a relative frequency table to summarize the distribution of number of correct answers on the pop quiz.
4. What proportion of students got fewer than 7 correct answers on the pop quiz? Does this result support the professor's belief that most students did not complete the pre-class reading assignment? Justify your answer.

SECTION 1A 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 is a characteristic that can take different values for different individuals.
- A **categorical variable** takes values that are labels, which place each individual into a particular group, called a category. A **quantitative variable** takes numerical values that count or measure some characteristic of each individual.
- The **distribution** of a variable describes what values the variable takes and how often it takes each value.
- To summarize the distribution of a variable, you can use a **frequency table** that shows the number of individuals having each value or a **relative frequency table** that shows the proportion or percentage of individuals having each value. You can use a frequency table or a relative frequency table to help describe the distribution of a variable.

AP® EXAM TIP
AP® Daily Videos

Review the content of this section and get extra help by watching the AP® Daily Videos for Topics 1.1–1.3, which are available in AP® Classroom.

SECTION 1A Exercises

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

Individuals and Variables

1. pg 4 **A class survey** Here is a small part of a data set that describes the students in a math class. The data come from anonymous responses to a questionnaire filled out on the second day of class.

Grade level	Dominant hand	GPA	Children in family	Homework last night (min)	Android or iPhone?
9	L	2.3	3	0–14	iPhone
11	R	3.8	6	15–29	Android
10	R	3.1	2	15–29	Android
10	R	4.0	1	45–59	iPhone
10	R	3.4	4	0–14	iPhone
10	L	3.0	3	30–44	Android
9	R	3.9	2	15–29	iPhone
12	R	3.5	2	0–14	iPhone

(a) Identify the individuals and variables in this data set.

(b) Classify each variable 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 from around the world.[4]

Roller coaster	Type	Height (ft)	Design	Speed (mph)	Duration (sec)
Copperhead Strike	Steel	82.0	Sit down	50.0	144
Eurostar	Steel	98.9	Inverted	50.2	140
Jungle Trailblazer	Wood	108.3	Sit down	54.1	150
Falcon	Steel	197.5	Wing	73.3	156
Olympia Looping	Steel	106.7	Sit down	49.7	105
Time Traveler	Steel	100.0	Sit down	50.3	117

(a) Identify the individuals and variables in this data set.

(b) Classify each variable as categorical or quantitative.

3. **Hit movies** *Avatar* was the top-earning movie based on box-office receipts worldwide as of December 2022. The following table displays data on several popular movies.[5] Identify the individuals and variables in this data set. Then classify each variable as categorical or quantitative.

Movie	Year	Rating	Time (min)	Genre	Box office ($)
Avatar	2009	PG-13	162	Action	2,910,284,102
Avengers: Endgame	2019	PG-13	181	Action	2,797,732,053
Titanic	1997	PG-13	194	Drama	2,207,986,545
Star Wars: The Force Awakens	2015	PG-13	136	Adventure	2,064,615,817
Avengers: Infinity War	2018	PG-13	156	Action	2,048,359,754
Spider Man: No Way Home	2021	PG-13	148	Action	1,910,041,582
Jurassic World	2015	PG-13	124	Action	1,669,963,641
The Lion King	2019	PG	118	Adventure	1,647,733,638
The Avengers	2012	PG-13	143	Action	1,515,100,211
Furious 7	2015	PG-13	137	Action	1,514,553,486
Top Gun: Maverick	2022	PG-13	131	Action	1,488,187,968
Frozen II	2019	PG	103	Adventure	1,444,735,491

4. **Skyscrapers** Here is some information about the tallest buildings in the world as of 2023.[6] Identify the individuals and variables in this data set. Then 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
Merdeka 118	Malaysia	678.9	118	Mixed	2023
Shanghai Tower	China	632.0	128	Mixed	2015
Makkah Royal Clock Tower Hotel	Saudi Arabia	601.0	120	Hotel	2012
Ping An Finance Center	China	599.1	115	Mixed	2017
Lotte World Tower	South Korea	554.5	123	Mixed	2017
One World Trade Center	United States	541.3	94	Office	2014
Guangzhou CTF Finance Center	China	530.0	111	Mixed	2016
Tianjin CTF Finance Center	China	530.0	97	Mixed	2019
China Zun Tower	China	527.7	101	Mixed	2019

Summarizing Data with Tables

5. **New York squirrels** How do squirrels in New York City's Central Park respond to humans? The Squirrel Census project enlisted volunteers to record data on all squirrels observed in Central Park in a recent year. One of the variables measures how the squirrels responded to humans. We selected a random sample of 75 of the more than 3000 squirrels observed. Here are the data on their reactions to humans:[7]

pg 6

Approach	Run from	Indifferent	Indifferent	Run from
Indifferent	Approach	Indifferent	Indifferent	Indifferent
Approach	Indifferent	Approach	Run from	Run from
Indifferent	Indifferent	Indifferent	Run from	Indifferent
Indifferent	Indifferent	Run from	Indifferent	Indifferent
Indifferent	Run from	Run from	Indifferent	Indifferent
Run from	Run from	Approach	Approach	Run from
Indifferent	Approach	Indifferent	Indifferent	Indifferent
Indifferent	Approach	Indifferent	Indifferent	Indifferent
Indifferent	Run from	Approach	Approach	Indifferent
Indifferent	Indifferent	Run from	Approach	Indifferent
Indifferent	Approach	Indifferent	Run from	Approach
Run from	Approach	Run from	Indifferent	Indifferent
Indifferent	Approach	Indifferent	Run from	Indifferent
Indifferent	Indifferent	Approach	Approach	Indifferent

(a) Make a frequency table and a relative frequency table to summarize the distribution of response to humans.

(b) Were these Central Park squirrels more likely to approach humans, run from humans, or be indifferent to humans? Justify your answer.

6. **Disc dogs** Here is a list of the breeds of dogs that won the World Canine Disc Championships over a 45-year period.[8]

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

(a) Make a frequency table and a relative frequency table to summarize the distribution of dog breed among World Canine Disc Championship winners.

(b) Which breeds of dog were the most frequent winners? Justify your answer.

7. **CubeSat missions** A CubeSat is a miniature satellite used for space research that can be easily deployed from a launch vehicle or from the International Space Station. More than 1000 CubeSats have been launched since specifications for these satellites were jointly developed by Cal Poly San Luis Obispo and Stanford University in 1999. The relative frequency table summarizes data on the type of mission for each CubeSat launched in November and December 2019.[9]

Mission type	Proportion
Science	0.18
Technology	0.28
Imaging	0.34
Communications	0.12
Education	0.08

(a) There were a total of 50 CubeSats launched during these two months. How many had a communications mission type?

(b) A report states that a majority of these CubeSat launches had an imaging mission type. Explain why this statement is incorrect.

Multiple Choice *Select the best answer for each question.*

Exercises 8 and 9 refer to the following setting. A realtor in Hilton Head, South Carolina, collects data on all of the homes sold in the town during 2023. The resulting data set lists each home's address, along with the following information: zip code, number of bedrooms, primary building material (wood, brick, etc.), square footage, whether it has a pool, age of home, and sales price.

8. The individuals in this data set are

(A) the Hilton Head realtor who collected the data.

(B) all Hilton Head homeowners who sold their homes during 2023.

(C) the addresses of all homes sold in Hilton Head during 2023.

(D) all homes sold in Hilton Head during 2023.

(E) all homes in Hilton Head during 2023.

9. This data set contains

(A) 7 variables, 2 of which are categorical.

(B) 7 variables, 3 of which are categorical.

(C) 7 variables, 4 of which are categorical.

(D) 8 variables, 3 of which are categorical.

(E) 8 variables, 4 of which are categorical.

SECTION 1B Displaying and Describing Categorical Data

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

- Make and interpret bar graphs of categorical data.
- Compare distributions of categorical data.
- Identify what makes some graphs of categorical data misleading.

In this section, you will learn how to display and describe the distribution of a categorical variable. Sections 1C, 1D, 1E, and 1F focus on displaying and describing the distribution of a quantitative variable. Unit 2 examines relationships between two categorical or two quantitative variables. This process of exploratory data analysis is known as *descriptive statistics*.

Displaying Categorical Data: Bar Graphs

A frequency table or relative frequency table summarizes a variable's distribution with numbers. To display the distribution more clearly, use a graph. The most

common way to display categorical data is with a **bar graph** (also known as a *bar chart*).

DEFINITION Bar graph

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

Figure 1.1 shows a bar graph of the data on preferred communication method for the random sample of 50 U.S. high school students who completed the Census at School survey. Note that the percentages for each category come from the relative frequency table.

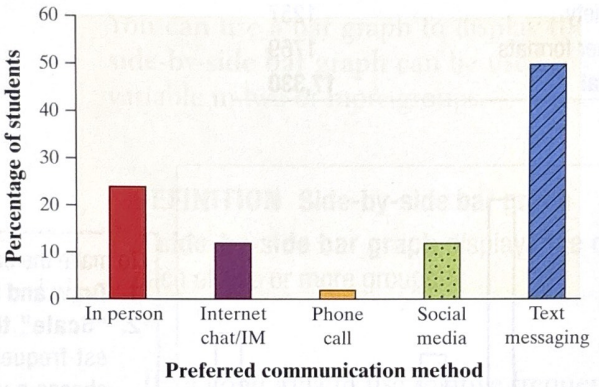


FIGURE 1.1 Bar graph of the distribution of preferred communication method for a random sample of 50 U.S. high school students.

Relative frequency table

Preferred method	Relative frequency
In person	12/50 = 0.24 or 24%
Internet chat/IM	6/50 = 0.12 or 12%
Phone call	1/50 = 0.02 or 2%
Social media	6/50 = 0.12 or 12%
Text messaging	25/50 = 0.50 or 50%

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

HOW TO MAKE A BAR GRAPH

1. **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 (percentage or proportion) of individuals in each category.
2. **"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 equal or exceed the largest frequency or relative frequency in any category.
3. **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 real purpose of a graph is to help us interpret the data. When you look at a graph, always ask, "What do I see?" In Figure 1.1, the bar graph reveals that the most preferred method of communicating with friends for these high school students is text messaging (50%). The next most preferred method is talking with their friends in person (24%). Social media and internet chat/instant messenger are somewhat less popular (12% each) methods of communicating with friends. By far the least preferred method of communication for these students is a phone call (2%).

EXAMPLE What's on the radio? Displaying categorical data: Bar graphs

Skills 2.A, 2.B

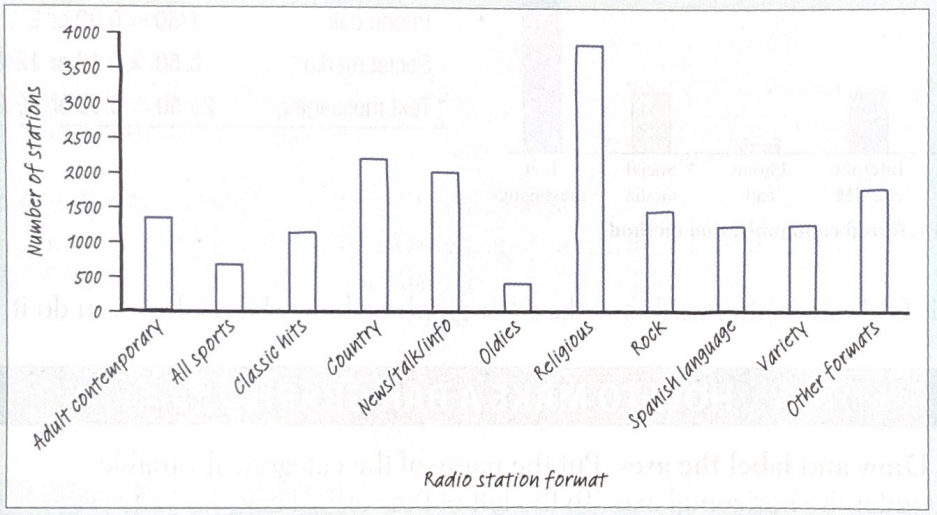


ACORN 1/Alamy Stock Photo

PROBLEM: Nielsen Audio, the rating service for radio audiences, places U.S. radio stations into categories that describe the kinds of programs they broadcast. The frequency table summarizes the distribution of station format in a recent year.[10] Make a frequency bar graph to display the data. Describe what you see.

Format	Number of stations	Format	Number of stations
Adult contemporary	1357	Religious	3837
All sports	669	Rock	1466
Classic hits	1140	Spanish language	1228
Country	2200	Variety	1257
News/talk/information	2002	Other formats	1769
Oldies	405	**Total**	**17,330**

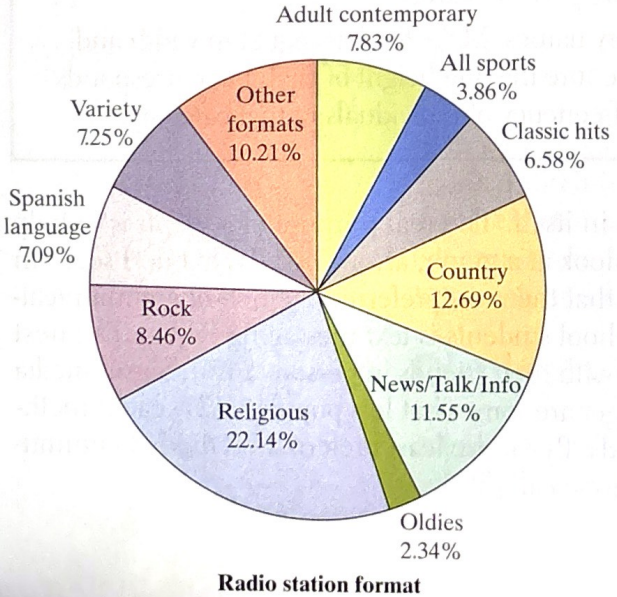
SOLUTION:



To make the bar graph:
1. **Draw and label the axes.**
2. **"Scale" the axes.** The largest frequency is 3837, so we choose a vertical scale from 0 to 4000, with tick marks 500 units apart.
3. **Draw bars** above the category names.

On U.S. radio stations, the most common formats are Religious (3837), Country (2200), and News/Talk/Information (2002), while the least common formats are Oldies (405) and All Sports (669). Moderately common formats offered by a similar number of stations include Classic Hits (1140), Spanish Language (1228), Variety (1257), Adult Contemporary (1357), and Rock (1466). There are 1769 stations with Other formats.

FOR PRACTICE, TRY EXERCISE 1



PIE CHARTS

Here is a **pie chart** of the radio station format data from the preceding example. Notice that this graph shows the *relative frequency* for each response category. For instance, the "Spanish language" slice makes up 7.09% of the graph because the relative frequency for this category is 1228/17,330 = 0.0709 = 7.09%.

DEFINITION Pie chart

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.

Use a pie chart when you want to emphasize each category's relation to the whole. Each slice of the pie shows the count or percentage of individuals in that category. Pie charts are challenging to make by hand, but technology will do the job for you.



A bar graph or pie chart that displays a distribution of categorical data must include *all* individuals in the data set. This might require including an "Other" category, as in the radio station example. Bar graphs are more flexible than pie charts. Both types of graphs can display the distribution of a categorical variable, but a bar graph can also compare any set of quantities that are measured in the same units.

Comparing Distributions of Categorical Data

You can use a bar graph to display the distribution of a categorical variable. A **side-by-side bar graph** can be used to compare the distribution of a categorical variable in two or more groups.

DEFINITION Side-by-side bar graph

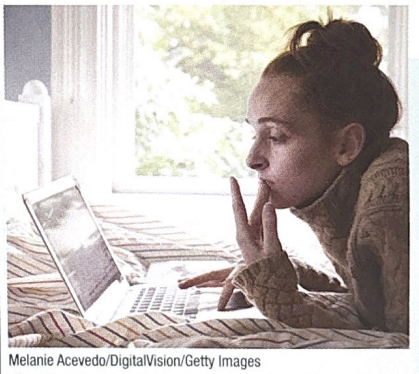
A **side-by-side bar graph** displays the distribution of one categorical variable in each of two or more groups.

It's a good idea to use relative frequencies when comparing data for multiple groups, especially if the groups have different sizes.

EXAMPLE

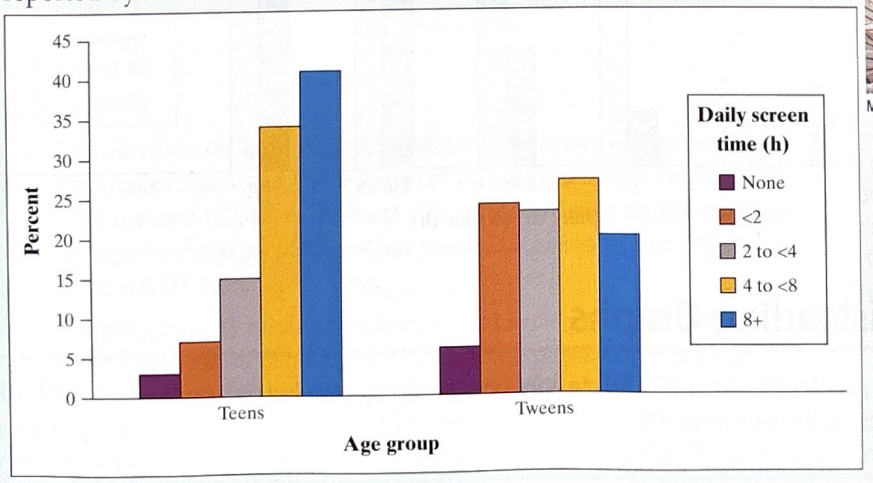
Screen time for teens and tweens
Comparing distributions of categorical data

Skill 2.D



Melanie Acevedo/DigitalVision/Getty Images

PROBLEM: How much time do tweens (ages 8–12) and teens (ages 13–18) spend using digital devices each day? Researchers surveyed a random sample of more than 1300 U.S. 8- to 18-year-olds to find out. The side-by-side bar graph summarizes the data on daily screen time reported by the 560 tweens and 746 teens in the sample.[11]



Compare the distributions of daily screen time for teens and tweens.

SOLUTION:

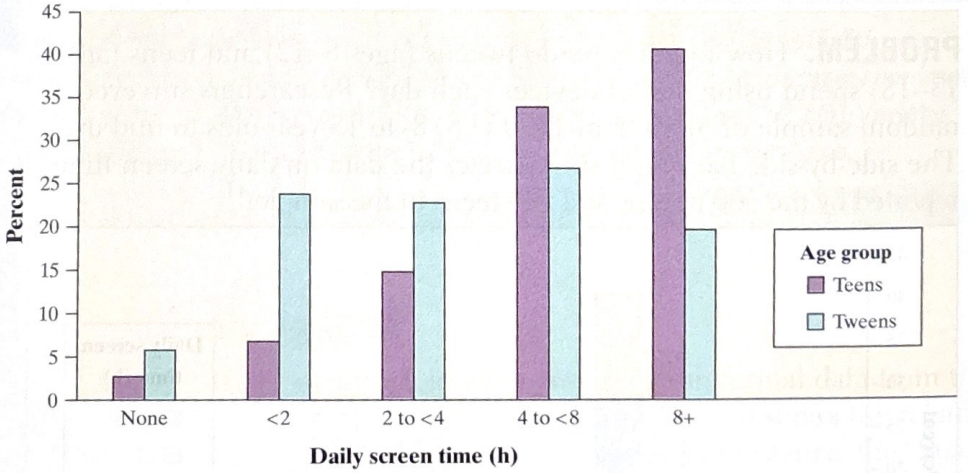
Tweens tend to report less daily screen time than teens. Tweens are more likely than teens to spend no time (6% versus 3%), less than 2 hours (24% versus 7%), and from 2 to less than 4 hours (23% versus 15%) on digital devices per day. Teens are more likely than tweens to spend from 4 to less than 8 hours (34% versus 27%) and 8 or more hours (41% versus 20%) on digital devices per day. More than half (53%) of tweens report less than 4 hours of screen time per day, while 75% of teens report 4 or more hours per day of screen time.

FOR PRACTICE, TRY EXERCISE 7

The following relative frequency table summarizes the data on daily screen time from the example by age group. As a result, the percentages in each row add to 100%. Note that we could have also compared the distributions with this relative frequency table.

		Daily time spent on digital devices				
		None	<2 hours	2 to <4 hours	4 to <8 hours	8+ hours
Age group	Teens	3%	7%	15%	34%	41%
	Tweens	6%	24%	23%	27%	20%

In the example, we grouped the bars by age group in the side-by-side bar graph. This arrangement clearly shows the two distributions of daily screen time — one for the teens and one for the tweens in the sample. It is also possible to group the bars by daily screen time, as in the following graph. This arrangement makes category-by-category comparisons for the daily time spent on digital devices easier — but now it's harder to see the individual distributions of daily screen time for the teens and for the tweens. Whichever way you organize the bars, be sure to include a key that describes what each color or type of shading on the side-by-side bar graph represents.



Misleading Graphs

Bar graphs can be a bit dull to look at. It is tempting to replace the bars with pictures or to use special three-dimensional (3D) 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? Misleading graphs

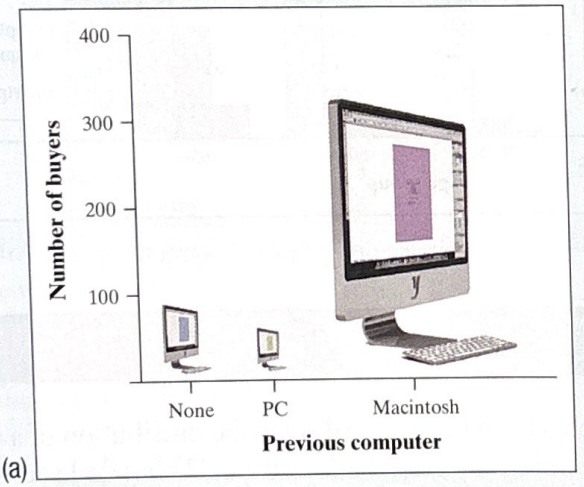
Skill 2.A

PROBLEM: When Apple, Inc., first introduced the iMac in 1998, the company wanted to know whether this new computer was expanding its market share. (The iMac has enjoyed great success ever since!) 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 randomly selected iMac customers. The firm categorized each customer as a new computer purchaser, a previous PC owner, or a previous Macintosh owner. The table summarizes the survey data.[12]

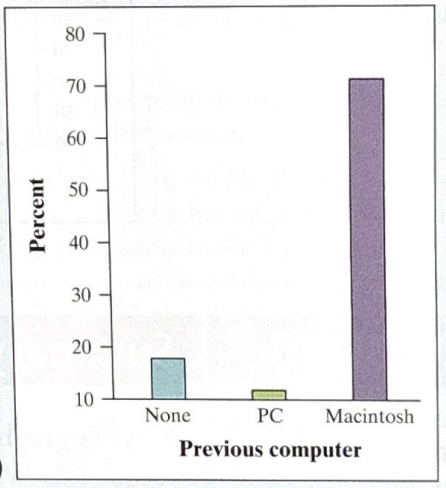
Previous ownership	Count	Percentage (%)
None	85	17.0
PC	60	12.0
Macintosh	355	71.0
Total	**500**	**100.0**

(a) A graph of the data that uses pictures instead of the more traditional bars is shown in figure (a). How is this pictograph misleading?

(b) A bar graph of the data is shown in figure (b). Explain why this graph could be considered deceptive.



(a)



(b)

SOLUTION:

(a) The pictograph is misleading because the areas of the computers make it look like the number of iMac buyers who are former Mac owners (355) is at least 10 times as large as the number of buyers in either of the other two categories (None: 85, PC owner: 60), which isn't true.

(b) The bar graph is misleading because starting the vertical scale at 10 instead of 0 makes it look like the percentage of iMac buyers who previously owned a PC (12.0%) is less than half the percentage who are first-time computer buyers (17.0%), 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!

FOR PRACTICE, TRY EXERCISE 11

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

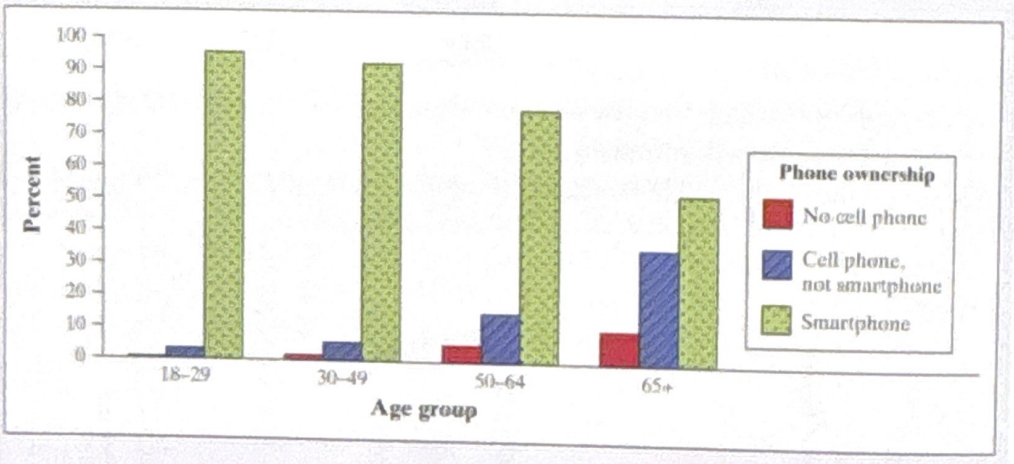


CHECK YOUR UNDERSTANDING

The Pew Research Center asked a random sample of 1502 U.S. adults about their cell phone ownership. The frequency table summarizes their responses.[13]

Type of cell phone	Frequency
None	73
Cell phone, not smartphone	212
Smartphone	1217
Total	**1502**

1. Make a relative frequency bar graph to display the distribution of cell phone ownership for the 1502 people in the sample. Describe what you see.
2. The following graph displays the distribution of cell phone ownership for each of four age groups in the sample. Compare the distributions.



SECTION 1B Summary

AP® EXAM TIP
AP® Daily Videos

Review the content of this section and get extra help by watching the AP® Daily Videos for Topic 1.4, which are available in AP® Classroom.

- A **bar graph** or **pie chart** can be used to display the distribution of a categorical variable. When examining any graph, ask yourself, "What do I see?"
- You can use a **side-by-side bar graph** to compare the distribution of one categorical variable for two or more groups. Be sure to use relative frequencies when comparing groups of different sizes.
- 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.

SECTION 1B Exercises

Displaying Categorical Data: Bar Graphs

1. **Birth days** The frequency table summarizes data on the numbers of babies born on each day of a single week in the United States.[14] pg 12

Day	Births
Sunday	7374
Monday	11,704
Tuesday	13,169
Wednesday	13,038
Thursday	13,013
Friday	12,664
Saturday	8459

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

2. **Going up?** The website fivethirtyeight.com published an interesting article about the more than 75,000 elevators in New York City. The frequency table summarizes data on the number of elevators of each type at that time.[15]

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

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

3. **Cool car colors** The popularity of colors for cars and light trucks changes over time. Silver passed green in 2000 to become the most popular color worldwide, then gave way to shades of white in 2007. The relative frequency table summarizes data on the colors of vehicles sold worldwide in 2020.[16]

Color	Percentage of vehicles
Black	19
Blue	7
Brown	3
Gray	15
Green	1
Red	5
Silver	9
White	38
Yellow	2
Other	??

(a) What percentage 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 your answer.

4. **Spam** Email spam is very annoying. The relative frequency table summarizes data on the most common types of spam.[17]

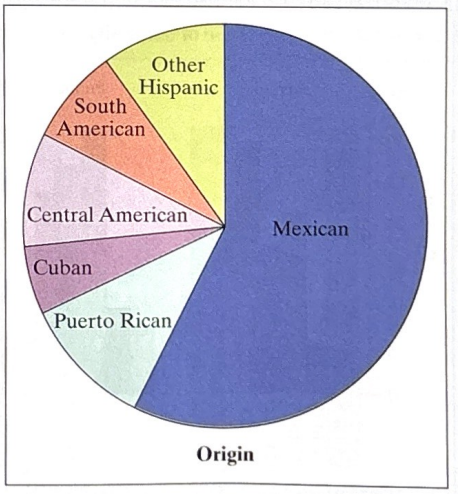
Type of spam	Percentage
Adult	19
Financial	20
Health	7
Internet	7
Leisure	6
Products	25
Scams	9
Other	??

(a) What percentage of spam would fall in the "Other" category?

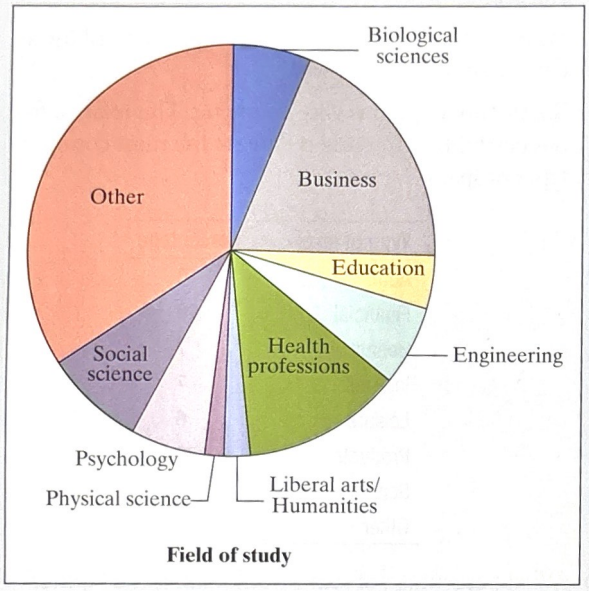
(b) Make a bar graph to display the data. Describe what you see.

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

5. **Family origins** Here is a pie chart of U.S. Census Bureau data showing the origin of more than 62 million Hispanic people in the United States.[18] About what percentage of Hispanic people are of Mexican origin? Puerto Rican origin?



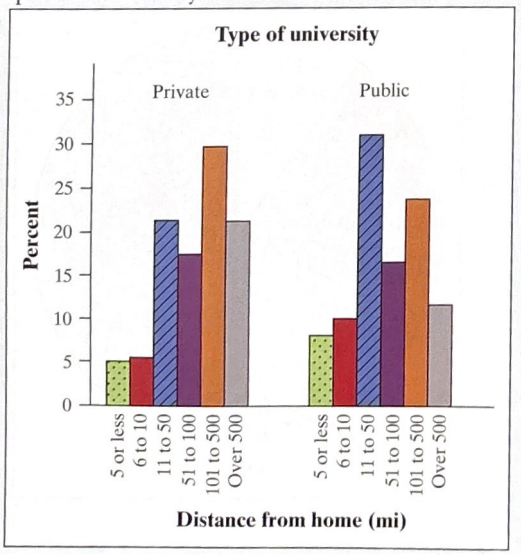
6. **Which major?** More than 2 million students earn bachelor's degrees in U.S. colleges and universities each year. The pie chart displays data on bachelor's degrees earned by field of study in a recent year.[19] About what percentage of students earned their bachelor's degrees in business? In health professions? In social science?



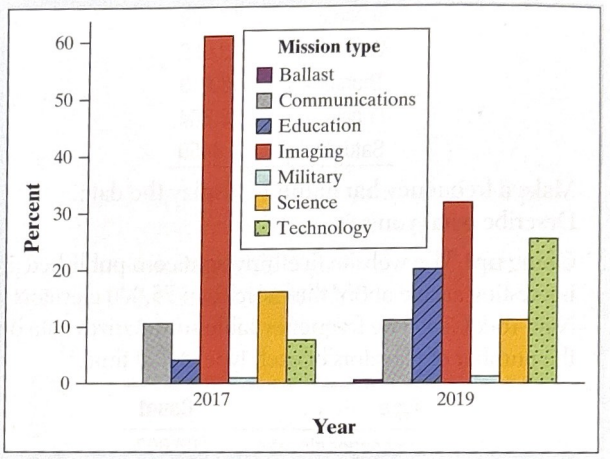
Comparing Distributions of Categorical Data

7. **Far from home** A survey asked first-year college students, "How many miles is this college from your permanent home?" Students selected 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.[20] Compare the distributions of distance from home for students from private and public 4-year colleges who completed the survey.

pg 13



8. **More CubeSat missions** A CubeSat is a miniature satellite used for space research that can be easily deployed from a launch vehicle or from the International Space Station. More than 1000 CubeSats have been launched since specifications for these satellites were jointly developed by Cal Poly San Luis Obispo and Stanford University in 1999. The side-by-side bar graph displays data on the types of missions undertaken by CubeSats launched in 2017 and 2019. Compare the distributions of mission type for these two years.[21]



9. **Reading for pleasure** Researchers surveyed a random sample of more than 1300 U.S. tweens (ages 8–12) and teens (ages 13–18) about how often they read for pleasure. The following table summarizes the responses in each age group.[22]

How often they read for pleasure	Age group: Teens	Tweens
Every day	21%	34%
At least once a week	27%	29%
At least once a month	14%	12%
Less than once a month	20%	12%
Never	18%	13%

(a) Make a side-by-side bar graph to compare the distributions of how often teens and tweens read for pleasure.

(b) Describe similarities and differences between the two distributions.

10. **Popular car colors** Favorite car colors may differ among countries in different parts of the world. The following table summarizes data on the most popular car colors in a recent year for the countries in North America and Asia.[23]

Color	Continent: North America	Continent: Asia
White	27%	45%
Gray	21%	13%
Black	20%	18%
Blue	10%	5%
Silver	10%	7%
Red	8%	4%
Brown	2%	3%
Green	1%	2%
Yellow	0.5%	2%
Other	0.5%	1%

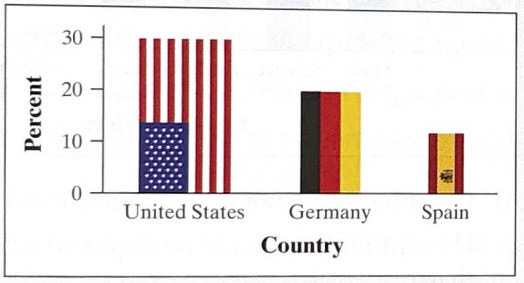
(a) Make a side-by-side bar graph to compare the distributions of most popular car color in North America and Asia.

(b) Describe similarities and differences between the two distributions.

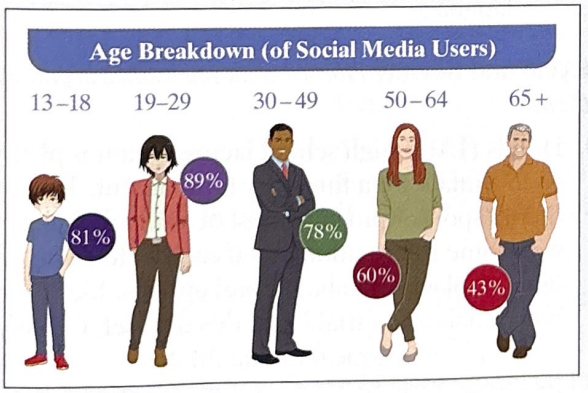
Misleading Graphs

11. **Who dislikes shopping?** Harris Interactive asked adults from several countries if they like or dislike shopping for clothes. The following pictograph displays the percentage of adults in the United States, Germany, and Spain who said that they dislike shopping for clothes. Explain how this graph is misleading.[24]

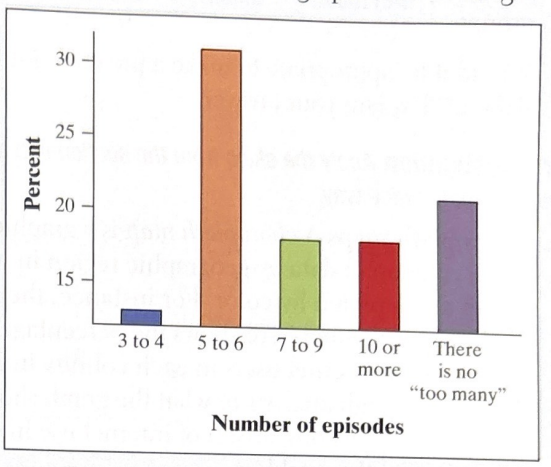
pg 15



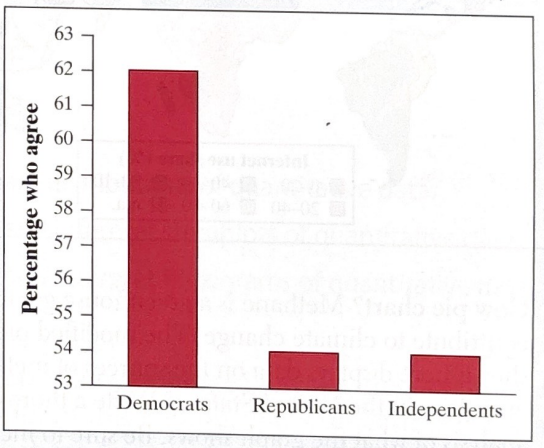
12. **Social media** The Pew Research Center surveyed a random sample of U.S. teens and adults about their use of social media. The following pictograph displays the percentage of people in various age groups who report using social media. Explain how this graph is misleading.



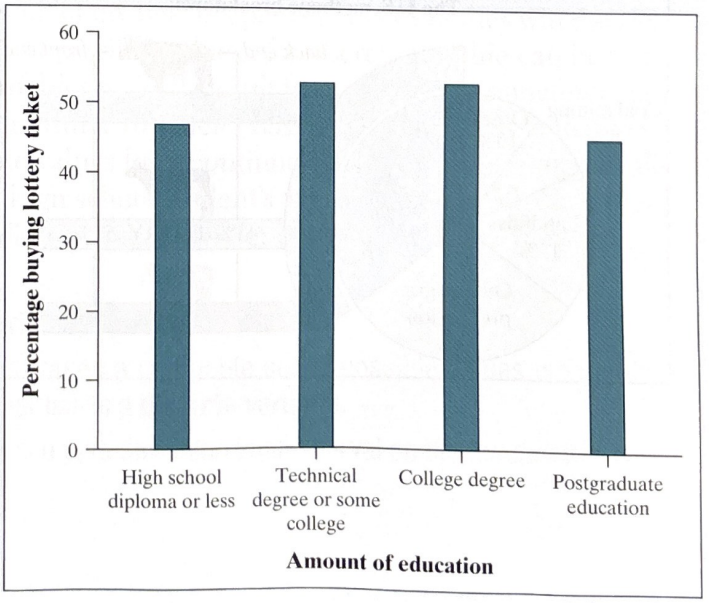
13. **Binge-watching** Do you "binge-watch" television series by viewing multiple episodes of a series at one sitting? A survey of 800 people who binge-watch were asked how many episodes is too many to watch in one viewing session. The results are displayed in the bar graph.[25] Explain how this graph is misleading.



14. **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 shown here. Explain how this graph is misleading. (*Note:* When notified about the misleading nature of its graph, the network posted a corrected version.)



15. **Lotteries and education** A Gallup Poll asked respondents about their highest level of education and whether they had bought a state lottery ticket in the last 12 months.[26] Here is a bar graph of the data:

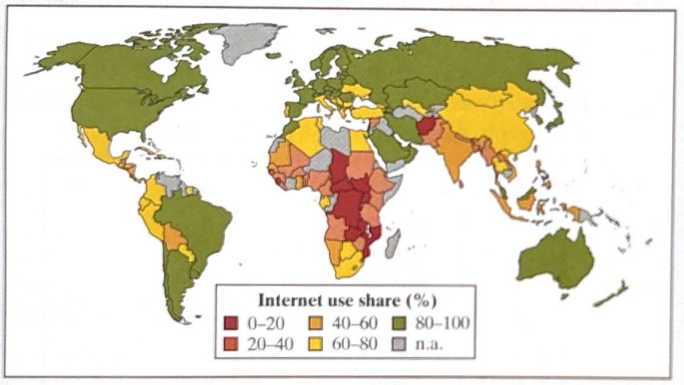


(a) Describe what this graph reveals about lottery ticket buying habits among the different education groups.

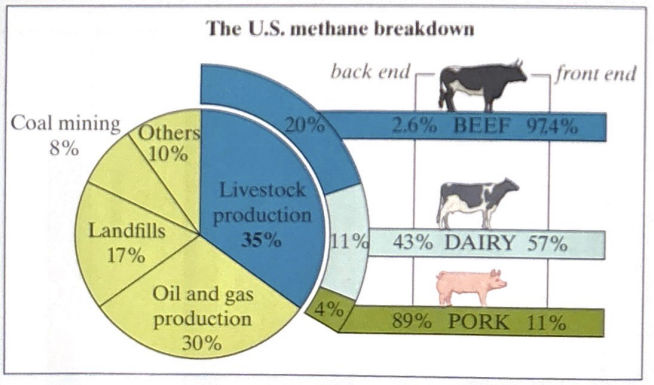
(b) Would it be appropriate to make a pie chart for this data set? Explain your answer.

For Investigation *Apply the skills from the section in a new context or nonroutine way.*

16. **Choropleth maps** A *choropleth map* is a graphical representation of data by geographic region in which values are depicted by color. For instance, the choropleth map presented here shows the percentage of people who were internet users in each country in 2020.[27] Write a thorough analysis of what the graph shows. Be sure to include a comparison of internet use in different regions of the world.



17. **Cow pie chart?** Methane is a greenhouse gas that may contribute to climate change. The modified pie chart shown here displays data on the sources of methane emissions in the United States.[28] Write a thorough analysis of what the graph shows. Be sure to include a comparison of methane emissions by beef cows, dairy cows, and pigs.



Multiple Choice *Select the best answer for each question.*

18. 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

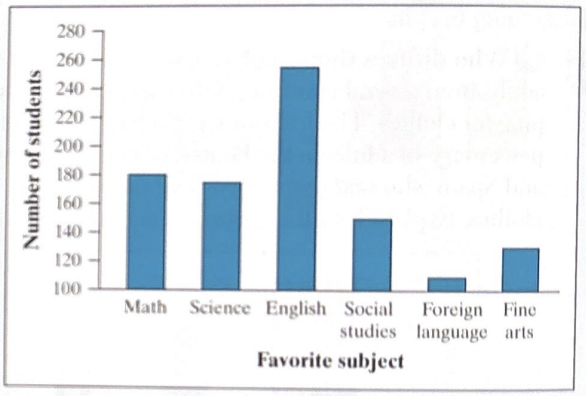
(B) The distribution of unemployment percentages for each of the 50 states

(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

19. 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 percentage of students.

(D) The vertical axis should start at 0 rather than 100.

(E) The foreign language bar should be broken up by language.

Recycle and Review *Practice what you learned in previous sections.*

20. **Hotels (1A)** 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.

Table for Exercise #20

Hotel	Pool	Exercise room?	Internet cost ($/day)	Restaurants	Distance to site (mi)	Room service?	Room rate ($/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 only	114
Hilton	In	Y	10.00	2	0.1	Y	156
Marriott	In	Y	9.95	2	0.0	Y	145

SECTION 1C Displaying Quantitative Data with Graphs

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

- Make and interpret dotplots of quantitative data.
- Describe the shape of a distribution of quantitative data.
- Describe the distribution of a quantitative variable.
- Compare distributions of quantitative data.
- Make and interpret stemplots of quantitative data.
- Make and interpret histograms of quantitative data.

As you learned in the previous section, bar graphs are typically used to display the distribution of a categorical variable. In this section, you will learn to make and interpret several types of graphs that can be used to display the distribution of a quantitative variable: *dotplots*, *stemplots*, and *histograms*.

There are two types of quantitative variables: *discrete* and *continuous*. Most **discrete variables** result from counting something, like the number of people in a household or the number of lottery tickets a person buys until they win the jackpot. Note that the number of possible values of a discrete variable can be finite or infinite. **Continuous variables** typically result from measuring something, like height (in inches) or time to run a 100-meter dash (in seconds). Note that there are infinitely many possible values for a continuous variable. Age is technically a continuous variable — a high school student's age might be 17.30162... years. But it is often treated as a discrete variable — for example, age = 17 years.

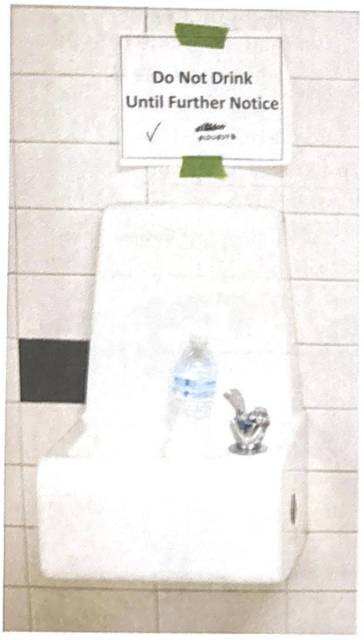
DEFINITION Discrete variable, Continuous variable

A quantitative variable that takes a countable set of possible values with gaps between them on the number line is a **discrete variable.**

A quantitative variable that can take any value in an interval on the number line is a **continuous variable.**

For a discrete variable, countable means that we can number the possible values of the quantitative variable using whole numbers. As an example, the number of lottery tickets a person buys until they win the jackpot can take any of the values 1, 2, 3, and so on.

Displaying Quantitative Data: Dotplots



JIM WATSON/AFP/Getty Images

In April 2014, managers for the city of Flint, Michigan, decided to save money by using water from the Flint River rather than continuing to buy water sourced from Lake Huron. Soon after, Flint residents noticed that the water coming out of their taps looked, tasted, and smelled bad. Some residents developed rashes, hair loss, and itchy skin. Authorities insisted that drinking water from the Flint River was safe.

As part of its regular testing program, city officials measured lead levels in water samples collected from 71 randomly selected Flint dwellings between January and June 2015. Here are the data (in parts per billion, ppb).[29] The U.S. Environmental Protection Agency (EPA) requires action if a water system's lead level exceeds 15 parts per billion.

0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	2	2	2	2	2	2	2
2	2	2	2	3	3	3	3	3	3	3	3
3	3	3	4	4	5	5	5	5	5	5	5
5	6	6	6	6	7	7	7	8	8	9	10
10	11	13	18	20	21	22	29	42	42	104	

Figure 1.2 is a **dotplot** of these data.

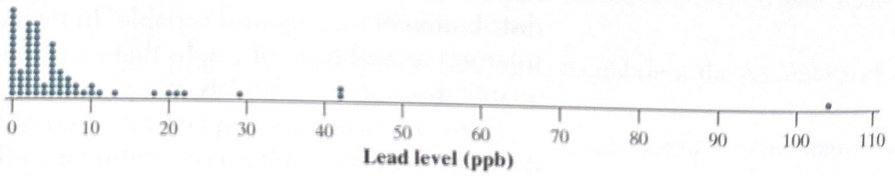


FIGURE 1.2 Dotplot of lead levels in water samples from 71 randomly selected Flint, Michigan, dwellings in January to June, 2015.

DEFINITION Dotplot

A **dotplot** shows each data value as a dot above its location on a number line.

A dotplot is the simplest graph for displaying the distribution of a quantitative variable. It is fairly easy to make a dotplot by hand for small sets of quantitative data.

HOW TO MAKE A DOTPLOT

1. **Draw and label the axis.** Draw a horizontal axis and put the name of the quantitative variable underneath it. Be sure to include units of measurement if appropriate.
2. **Scale the axis.** Find the smallest and largest values in the data set. Start the horizontal axis at a convenient number equal to or less than the smallest value and place tick marks at equal intervals until you equal or exceed the largest value.
3. **Plot the values.** Mark a dot above the location on the horizontal axis corresponding to each data value. Try to make all the dots the same size and space them out equally as you stack them.

Remember what we said in Section 1B: making a graph is not an end in itself. When you look at a graph, always ask, "What do I see?" The dotplot in Figure 1.2 shows that $8/71 = 0.113 = 11.3\%$ of the Flint water samples had lead levels that exceeded 15 parts per billion. Does that mean Flint's water system required action based on the EPA guideline? We'll answer that question — and tell you the rest of the story about the Flint water crisis — in Section 1E.

EXAMPLE

Eating healthy

Displaying quantitative data: Dotplots

Skills 2.A, 2.B

PROBLEM: How healthy is plant-based yogurt? The *Nutrition Action Healthletter* provided data on calories, saturated fat, protein, calcium, and total sugars for many popular brands of yogurt. Here are the data on the amount of added sugar, in teaspoons (tsp), in single-serving containers for all of the plant-based yogurt brands:[30]

0.0	1.0	1.0	2.5	0.0	0.0	1.0	1.5	2.0	2.0	3.0	3.5	2.5
3.0	5.0	3.0	3.5	3.5	3.5	2.0	4.0	3.5	2.0	3.5	1.5	2.0



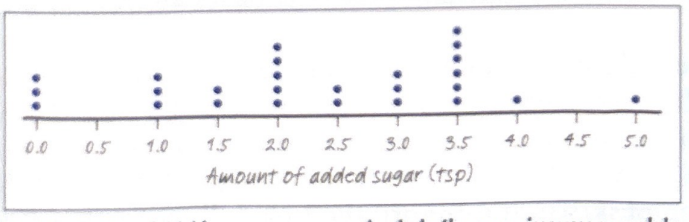
Clare Gainey/Alamy Stock Photo

(a) Make a dotplot of these data.

(b) The American Heart Association (AHA) recommends a maximum of 6 teaspoons of added sugar per day for women.[31] What percentage of these plant-based yogurt brands have less than half of the AHA's daily recommendation?

SOLUTION:

(a)



To make the dotplot:
1. **Draw and label the axis.** Be sure to include units along with the variable name.
2. **Scale the axis.** The smallest value is 0.0 and the largest value is 5.0. So we choose a scale from 0 to 5 with tick marks 0.5 unit apart.
3. **Plot the values.**

(b) Half of the AHA's recommended daily maximum would be 3 teaspoons of added sugar, and $15/26 = 0.577 = 57.7\%$ of these plant-based yogurt brands have less than 3 teaspoons of added sugar.

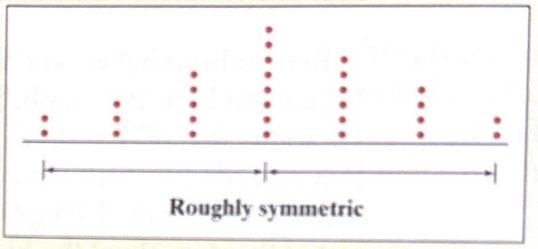
FOR PRACTICE, TRY EXERCISE 3

Describing Shape

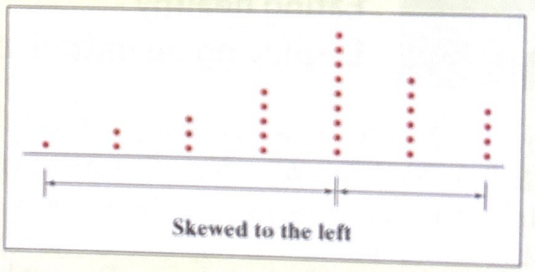
When you describe the shape of a dotplot or another graph of quantitative data, focus on the main features. Look for clear *peaks*, not for minor ups and downs in the graph. Look for *clusters* of values and obvious *gaps*. Decide if the distribution is **roughly symmetric, skewed to the left,** or **skewed to the right.**

DEFINITION Roughly symmetric, Skewed to the left, Skewed to the right

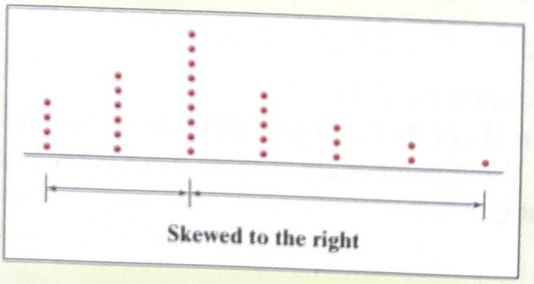
A distribution is **roughly symmetric** if the right side of the graph (containing the half of the observations with the largest values) is approximately a mirror image of the left side.

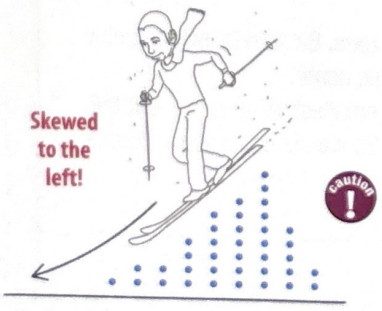


A distribution is **skewed to the left** if the left side of the graph is much longer than the right side.



A distribution is **skewed to the right** if the right side of the graph is much longer than the left side.





We could also describe a distribution with a long tail to the left as "skewed toward negative values" or "negatively skewed," and a distribution with a long right tail as "positively skewed." For ease, we sometimes say "left-skewed" instead of "skewed to the left" and "right-skewed" instead of "skewed to the right." **The direction of skewness is toward the long tail, not in the direction where most observations are clustered.** The drawing is a cute but corny way to help you keep this straight. To avoid danger, Mr. Starnes skis on the gentler slope — in the direction of the skewness.

EXAMPLE Quiz scores and die rolls
Describing shape

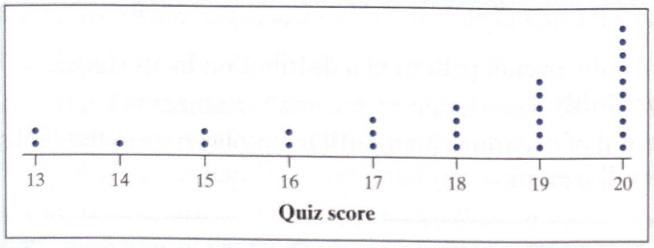
Skill 2.A



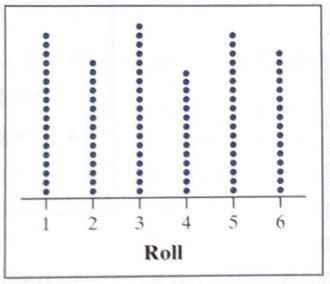
malerapaso/Getty Images

PROBLEM: The dotplots display two different sets of quantitative data. Graph (a) shows the scores on a 20-point quiz for each of the 30 students in a statistics class. Graph (b) shows the results of 100 rolls of a six-sided die. Describe the shape of each distribution.

(a)



(b)



SOLUTION:

(a) The distribution of statistics quiz scores is skewed to the left, with a single peak at 20 (a perfect score).

(b) The distribution of die rolls is roughly symmetric. It has no clear peak.

FOR PRACTICE, TRY EXERCISE 7

We can describe the shape of the distribution in part (b) of the example as **approximately uniform** because the frequencies are about the same for all possible rolls.

DEFINITION Approximately uniform

A distribution in which the frequency (relative frequency) of each possible value is about the same is **approximately uniform.**

Some people refer to graphs with a single peak, like the dotplot of quiz scores in part (a) of the example, as *unimodal*. Figure 1.3 is a dotplot of the duration (in minutes) of 263 eruptions of the Old Faithful geyser in July 1995, when the Starnes family made its first trip to Yellowstone National Park. We describe this graph as *bimodal* because it has two clear peaks: one at about 2 minutes and one at about 4.5 minutes. Although we could continue the pattern with "trimodal" for three peaks, and so on, it's more common to refer to distributions with more than two clear peaks as *multimodal*. When you examine a graph of quantitative data, describe any pattern you see as clearly as you can.

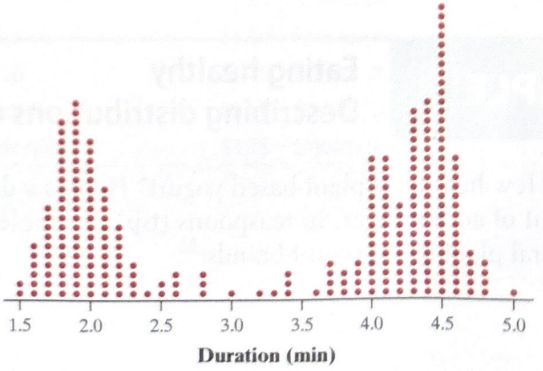


FIGURE 1.3 Dotplot displaying the duration (in minutes) of 263 eruptions of the Old Faithful geyser in July 1995. This distribution has two main clusters of data and two clear peaks—one near 2 minutes and the other near 4.5 minutes.

Describing Distributions of Quantitative Data

Here is a general strategy for describing the distribution of a quantitative variable.

HOW TO DESCRIBE THE DISTRIBUTION OF A QUANTITATIVE VARIABLE

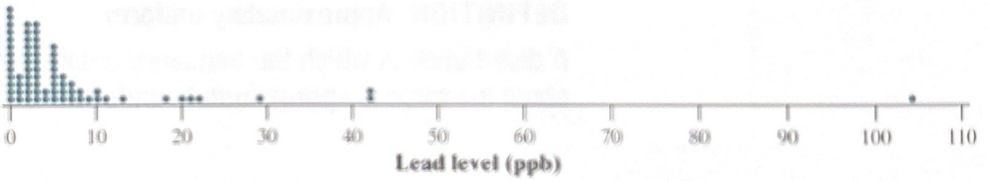
In any graph, look for the *overall pattern* and for clear *departures* from that pattern.

- You can describe the overall pattern of a distribution by its **shape, center,** and **variability.**
- An important kind of departure is an **outlier,** an observation that falls outside the overall pattern.

Variability is sometimes referred to as *spread*. We prefer variability because students often think that spread refers only to the distance between the maximum and minimum values of a quantitative data set (the *range*).

We will discuss more formal ways to measure center and variability and to identify outliers in Section 1D. For now, just use the middle value in the ordered data set (what you may have learned as the *median* in previous math classes) when describing center and the *minimum* and *maximum* values when describing variability.

Let's practice with the dotplot of lead levels in water samples from 71 randomly selected Flint, Michigan, dwellings that you saw in Figure 1.2.



Shape: The distribution of lead level is skewed to the right, with a single peak at 0 ppb.

Outliers: There is one obvious outlier at 104 ppb. The two lead levels of 42 ppb may also be outliers.

Center: The middle value (median) is a lead level of 3 ppb.

Variability: The lead levels vary from 0 to 104 ppb.

When describing a distribution of quantitative data, don't forget: Statistical Opinions Can Vary (Shape, Outliers, Center, Variability).

EXAMPLE

Eating healthy

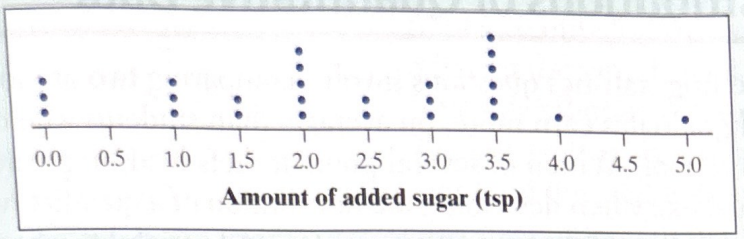
Skill 2.A

Describing distributions of quantitative data

PROBLEM: How healthy is plant-based yogurt? Here is a dotplot of data on the amount of added sugar, in teaspoons (tsp), in single-serving containers of several plant-based yogurt brands:[32]



Sara Stathas/Alamy Stock Photo



Describe the distribution.

SOLUTION:

Shape: The distribution of amount of added sugar for these plant-based yogurt brands is roughly symmetric, with two clear peaks at 2 tsp and 3.5 tsp of added sugar. There are small gaps at 0.5 tsp and 4.5 tsp.

Outliers: The three brands of yogurt with 0 tsp of added sugar and the one brand with 5 tsp of added sugar are possible outliers.

Center: The middle value is between 2 and 2.5 tsp of added sugar (median rating = 2.25 tsp).

Variability: The amount of added sugar varies from 0 to 5 tsp.

AP® EXAM TIP

Always be sure to include context when you are asked to describe a distribution. This means using the variable name (e.g., amount of added sugar), not just the units the variable is measured in (tsp).

FOR PRACTICE, TRY EXERCISE 9



CHECK YOUR UNDERSTANDING

Knoebels Amusement Park in Elysburg, Pennsylvania, has earned acclaim for being an affordable, family-friendly entertainment venue. Knoebels does not charge for general admission or parking, but it does charge customers for each ride they take. How much do the rides cost at Knoebels? The table shows the cost for each ride in a sample of 22 rides in a recent year.[33]

Name	Cost	Name	Cost
Merry Mixer	$1.50	Looper	$1.75
Italian Trapeze	$1.50	Flying Turns	$3.00
Satellite	$1.50	Flyer	$1.50
Galleon	$1.50	The Haunted Mansion	$1.75
Whipper	$1.25	StratosFear	$2.00
Skooters	$1.75	Twister	$2.50
Ribbit	$1.25	Cosmotron	$1.75
Roundup	$1.50	Paratrooper	$1.50
Paradrop	$1.25	Downdraft	$1.50
The Phoenix	$2.50	Rockin' Tug	$1.25
Gasoline Alley	$1.75	Sklooosh!	$1.75

1. Make a dotplot of the data.
2. Describe the distribution.

Comparing Distributions of Quantitative Data

Some of the most interesting statistics questions involve comparing two or more groups. Do high school graduates earn more, on average, than students who do not graduate from high school? Which of several popular diets leads to greater long-term weight loss? Just like when describing the distribution of a quantitative variable, you should always discuss shape, outliers, center, and variability whenever you compare distributions of quantitative data.

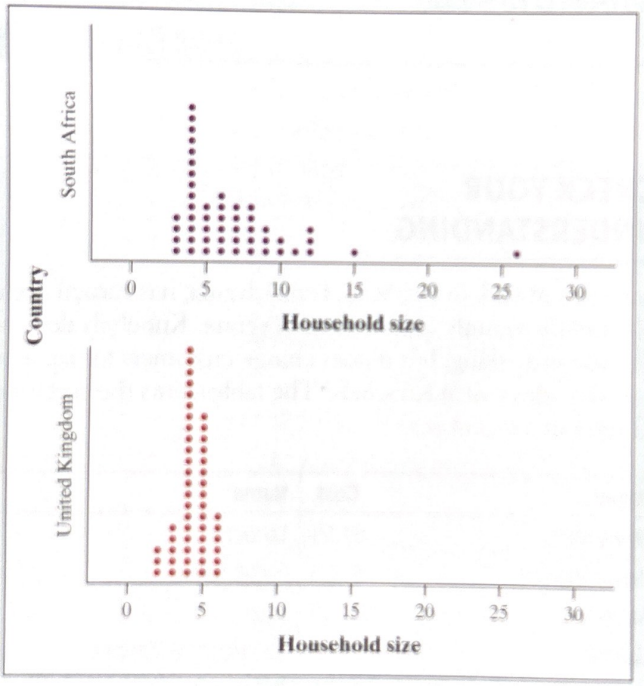
EXAMPLE

Household size in South Africa and the United Kingdom

Comparing distributions of quantitative data

Skill 2.D

PROBLEM: How do the numbers of people living in households in the United Kingdom (U.K.) and South Africa compare? To help answer this question, we selected separate random samples of 50 households from each country.[34] Here are dotplots of the number of people in each household. Compare the distributions of household size for these two countries.





FrankvandenBergh/Getty Images

SOLUTION:

Shape: The distribution of household size for the U.K. sample is roughly symmetric, with a single peak at 4 people. The distribution of household size for the South Africa sample is skewed to the right, with a single peak at 4 people and a clear gap between 15 and 26 people.

Outliers: There are no apparent outliers in the U.K. dotplot. The two largest values in the South Africa dotplot — households with 15 and 26 people — appear to be outliers.

Center: Household sizes for the South Africa sample tend to be larger (median = 6 people) than for the U.K. sample (median = 4 people).

Variability: The household sizes for the South Africa sample vary more (from 3 to 26 people) than for the U.K. sample (from 2 to 6 people).

AP® EXAM TIP

When comparing distributions of quantitative data, it's not enough just to list values for the center and variability of each distribution. You have to *compare* these values explicitly, using words like "greater than," "less than," or "about the same as. And be sure to include context!

FOR PRACTICE, TRY EXERCISE 13

Notice that in the preceding example, we discussed the distributions of household size only for the two *samples* of 50 households. We might be interested in whether the sample data give us convincing evidence of a difference in the *population* distributions of household size for South Africa and the United Kingdom. We'll have to wait several units to decide whether we can reach such a conclusion, but our ability to make such an inference later will be helped by the fact that the households in our samples were chosen at random.

Displaying Quantitative Data: Stemplots

Another simple type of graph for displaying quantitative data is a **stemplot**, also called a *stem-and-leaf plot*.

DEFINITION Stemplot

A **stemplot** shows each data value separated into two parts: a *stem*, which consists of the leftmost digits, and a *leaf*, consisting of the final digit. The stems are ordered from least to greatest and arranged in a vertical column. The leaves are arranged in increasing order out from the appropriate stems.

Here are data on the resting pulse rates (beats per minute, bpm) of 19 middle school students:

71	104	76	88	78	71	68	86	70	90
74	76	69	68	88	96	68	82	120	

Figure 1.4 shows a stemplot of these data.

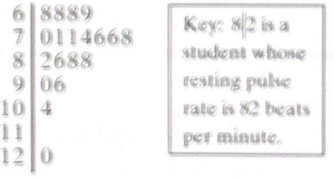


FIGURE 1.4 Stemplot of the resting pulse rates of 19 middle school students.

According to the American Heart Association, a resting pulse rate greater than 100 bpm is considered high for this age group. We can see that $2/19 = 0.105 = 10.5\%$ of these students have high resting pulse rates by this standard. Also, the distribution of pulse rates for these 19 students is skewed to the right (toward the larger values).

Stemplots give us a quick picture of a distribution that includes the individual data values in the graph. It is fairly easy to make a stemplot by hand for small sets of quantitative data.

HOW TO MAKE A STEMPLOT

1. **Make stems.** Separate each data value into a stem (all but the final digit) and a leaf (the final digit). Write the stems in a vertical column from smallest to largest. Draw a vertical line at the right of this column. Do not skip any stems, even if there is no data value for a particular stem.
2. **Add leaves.** Write each leaf in the row to the right of its stem.
3. **Order the leaves.** Arrange the leaves in increasing order out from the stem.
4. **Add a key.** Provide a key that explains in context what the stems and leaves represent.

EXAMPLE Preventing concussions
Displaying quantitative data: Stemplots

Skills 2.A, 2.B

PROBLEM: Many athletes (and their parents) worry about the risk of concussions when playing sports. A youth football coach plans to obtain specially made helmets for the players that are designed to reduce their chance of getting a concussion. Here are the measurements of head circumference (in inches) for the 30 players on the team:[35]

23.0	22.2	21.7	22.0	22.3	22.6	22.7	21.5	22.7	25.6
20.8	23.0	24.2	23.5	20.8	24.0	22.7	22.6	23.9	22.5
23.1	21.9	21.0	22.4	23.5	22.5	23.9	23.4	21.6	23.3



Pete Saloutos/AGE Fotostock

(a) Make a stemplot of these data.
(b) Describe the shape of the distribution. Are there any obvious outliers?

SOLUTION:

(a)

20	88
21	05679
22	02345566777
23	001345599
24	02
25	6

Key: 23|5 is a player with a head circumference of 23.5 inches.

To make the stemplot:
1. **Make stems.** The smallest head circumference is 20.8 inches and the largest is 25.6 inches. We use the first two digits as the stem and the final digit as the leaf. So we need stems from 20 to 25.
2. **Add leaves.** For the player with a head circumference of 23.0 inches, place a 0 on the 23 stem. For the player with a head circumference of 22.2 inches, place a 2 on the 22 stem. Continue in this way until you have added the data for all the players.
3. **Order the leaves.**
4. **Add a key.**

(b) The distribution of head circumference for the 30 players on the youth football team is roughly symmetric, with a single peak on the 22-inch stem. There are no obvious outliers.

FOR PRACTICE, TRY EXERCISE 17

We can get a better picture of the head circumference data by *splitting stems*. In Figure 1.5(a), the leaves from 0 to 9 are placed on the same stem. Figure 1.5(b) shows another stemplot of the same data. This time, leaves 0 through 4 are placed on one stem, while leaves 5 through 9 are placed on another stem. Now we can see the shape of the distribution more clearly — including the possible outlier at 25.6 inches.

FIGURE 1.5 Two stemplots showing the head circumference data. The graph in (b) improves on the graph in (a) by splitting stems.

(a)

20	88
21	05679
22	02345566777
23	001345599
24	02
25	6

Key: 23|5 is a player with a head circumference of 23.5 inches.

(b)

20	88
21	0
21	5679
22	0234
22	5566777
23	00134
23	5599
24	02
24	
25	
25	6

Be sure to include these stems even though they include no data.

Here are a few tips to consider before making a stemplot:

- There is no magic number of stems to use. Too few or too many stems will make it difficult to see the distribution's shape. Five stems is a good minimum.
- If you split stems, make sure that each stem is assigned an equal number of possible leaf digits.
- When the data have too many digits, you can get more flexibility by rounding or truncating the data.

COMPARING DISTRIBUTIONS WITH STEMPLOTS

You can use a *back-to-back stemplot* with common stems to compare the distribution of a quantitative variable in two groups. The leaves on each side are placed in order leading out from the common stem. For example, Figure 1.6 shows a back-to-back stemplot of the 19 middle school students' resting pulse rates and their pulse rates after 5 minutes of running.

Resting		After exercise
9888	6	
8664110	7	
8862	8	6788
60	9	02245899
4	10	044
	11	8
0	12	44
	13	
	14	6

Key: 8|2 is a student whose pulse rate is 82 beats per minute.

FIGURE 1.6 Back-to-back stemplot of 19 middle school students' resting pulse rates and their pulse rates after 5 minutes of running.

Let's compare the distributions. As always, we discuss shape, outliers, center, and variability.

Shape: The distribution of resting pulse rates is skewed to the right, with a single peak on the 70s stem. The distribution of after-exercise pulse rates is also skewed to the right, with a single peak on the 90s stem.

Outliers: The student whose resting pulse rate was 120 bpm appears to be an outlier in that group. The student whose after-exercise pulse rate was 146 bpm appears to be an outlier in that group.

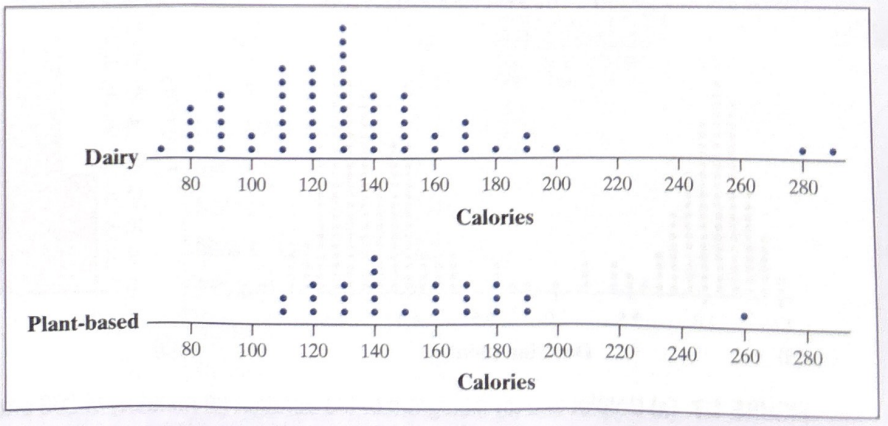
Center: The students' resting pulse rates tended to be lower (median = 76 bpm) than their after-exercise pulse rates (median = 98 bpm).

Variability: The after-exercise pulse rates vary more (from 86 bpm to 146 bpm) than the resting pulse rates (from 68 bpm to 120 bpm).



CHECK YOUR UNDERSTANDING

1. Which is healthier: dairy yogurt or plant-based yogurt? The *Nutrition Action Healthletter* provided data on calories, saturated fat, protein, calcium, and total sugars for many popular brands of yogurt.[36] Here are parallel dotplots that compare the calories in single-serving containers of 57 brands of dairy yogurt and 26 brands of plant-based yogurt. Describe similarities and differences in the two distributions.



Here is a stemplot of the percentage of residents aged 65 and older in the 50 U.S. states:[37]

11	18
12	6
13	9
14	23
15	44445677889999
16	23444557999
17	0112225567
18	12477
19	39
20	56

Key: 19|3 means 19.3% of the state's residents are aged 65 and older.

2. The largest data value is for Maine. What percentage of Maine residents are 65 or older? (Florida has the second largest data value.)
3. Make another stemplot of the data by splitting stems. What does this new graph reveal that is not apparent from the original stemplot?

Displaying Quantitative Data: Histograms

You can use a dotplot or a stemplot to display quantitative data. Both of these types of graphs show every individual data value. However, for large data sets, this can make it difficult to see the overall pattern in the graph. We often get a cleaner picture of the distribution by grouping nearby values together. Doing so allows us to make a new type of graph: a **histogram**.

DEFINITION Histogram

A **histogram** shows each interval as a bar. The heights of the bars show the frequencies or relative frequencies of values in each interval.

Figure 1.7 shows a dotplot and a histogram of the durations (in minutes) of 263 eruptions of the Old Faithful geyser in July 1995. Notice how the histogram groups nearby values together into equal-width intervals.

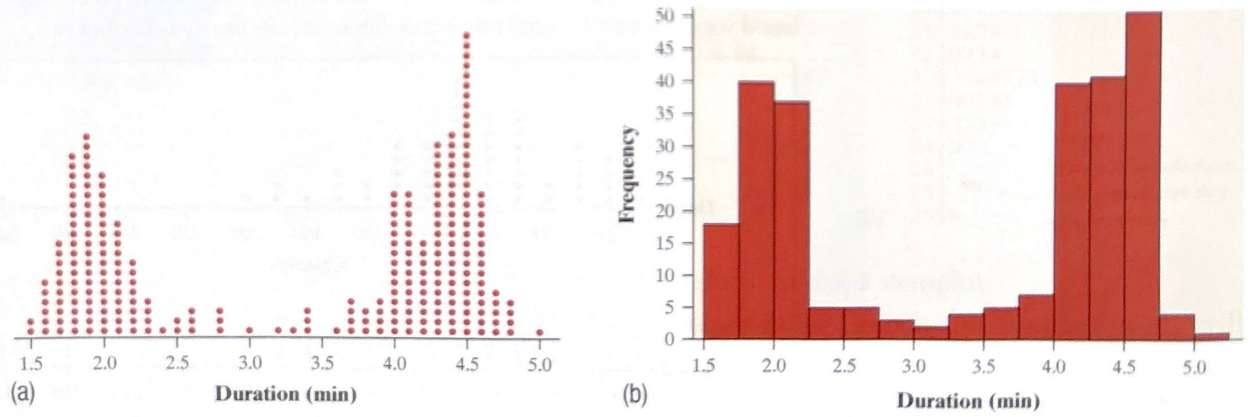


FIGURE 1.7 (a) Dotplot and (b) histogram of the duration (in minutes) of 263 eruptions of the Old Faithful geyser in July 1995.

You can make a histogram by hand, even for fairly large sets of quantitative data. For details on making histograms with technology, see the Tech Corner at the end of this section.

HOW TO MAKE A HISTOGRAM

1. **Choose equal-width intervals** that span the data. Five intervals is a good minimum.
2. **Make a table** that shows the frequency (count) or relative frequency (percentage or proportion) of individuals in each interval.
3. **Draw and label the axes.** Draw horizontal and vertical axes. Put the name of the quantitative variable under the horizontal axis. To the left of the vertical axis, indicate whether the graph shows the frequency (count) or relative frequency (percentage or proportion) of individuals in each interval.
4. **Scale the axes.** Place equally spaced tick marks at the smallest value in each interval along the horizontal axis until you equal or exceed the largest data value. On the vertical axis, start at 0 and place equally spaced tick marks until you equal or exceed the largest frequency or relative frequency in any interval.
5. **Draw bars** above the intervals. Make the bars equal in width and leave no gaps between them. Be sure that the height of each bar corresponds to the frequency or relative frequency of data values in that interval. An interval with no data values will appear as a bar of height 0 on the graph.

It is possible to choose intervals of unequal widths when making a histogram, but such graphs are beyond the scope of this book.

EXAMPLE

Foreign-born residents
Displaying quantitative data: Histograms

Skills 2.A, 2.B



Sarah Reingewirtz/MediaNews Group/Los Angeles Daily News/ Getty Images

PROBLEM: How does the percentage of foreign-born residents in each U.S. state compare to the rest of the country? The table presents the data for all 50 states in a recent year.[38]

State	Percentage	State	Percentage	State	Percentage
Alabama	3.6	Georgia	10.3	Maine	3.9
Alaska	8.0	Hawaii	19.3	Maryland	15.4
Arizona	13.4	Idaho	5.8	Massachusetts	17.3
Arkansas	5.1	Illinois	13.9	Michigan	7.0
California	26.7	Indiana	5.3	Minnesota	8.4
Colorado	9.5	Iowa	5.6	Mississippi	2.1
Connecticut	14.8	Kansas	7.2	Missouri	4.3
Delaware	10.0	Kentucky	4.4	Montana	2.3
Florida	21.1	Louisiana	4.2	Nebraska	7.4

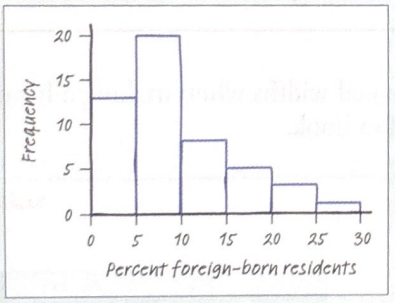
State	Percentage	State	Percentage	State	Percentage
Nevada	19.8	Oklahoma	6.1	Utah	8.6
New Hampshire	6.4	Oregon	9.7	Vermont	4.7
New Jersey	23.4	Pennsylvania	7.0	Virginia	12.7
New Mexico	9.6	Rhode Island	13.7	Washington	14.9
New York	22.4	South Carolina	5.6	West Virginia	1.6
North Carolina	8.4	South Dakota	4.1	Wisconsin	5.1
North Dakota	4.1	Tennessee	5.5	Wyoming	3.1
Ohio	4.8	Texas	17.1		

(a) Make a frequency histogram to display the data.
(b) What proportion of states have less than 10% foreign-born residents?

SOLUTION:

(a)

Interval	Frequency
0 to < 5	13
5 to < 10	20
10 to < 15	8
15 to < 20	5
20 to < 25	3
25 to < 30	1



To make the histogram:

1. **Choose equal-width intervals** that span the data. The data vary from 1.6% to 26.7%. We choose intervals of width 5, beginning at 0: 0 to < 5, 5 to < 10, and so on. This choice results in more than the minimum of five intervals.
2. **Make a table.** Record the number of states in each interval when making a frequency histogram.
3. **Draw and label the axes.**
4. **Scale the axes.** The scale on the horizontal axis matches the intervals we chose in Step 1. The highest frequency in an interval is 20, so we scale the vertical axis from 0 to 20, placing tick marks every 5 units.
5. **Draw bars.**

(b) $(13 + 20)/50 = 33/50 = 0.66$ of states have less than 10% foreign-born residents.

FOR PRACTICE, TRY EXERCISE 25

From the histogram, we can see that the distribution of the percentage of foreign-born residents in the 50 U.S. states is skewed to the right. The graph does not show any obvious outliers. How should we describe the center and variability of the distribution? Because a histogram does not show individual data values, we can only give *estimates* of the center and variability using the intervals on the graph. With 50 data values, the middle value (median) falls between the 25th and 26th values in the ordered data set, so the median percentage of foreign-born residents is in the 5% to < 10% interval. The data vary from at least 0% to at most 29.9% foreign-born residents. Using the raw data from the example, we can confirm that the median is 7.3% and that the data vary from 1.6% to 26.7%.

Figure 1.8 shows two different histograms of the foreign-born resident data. The one on the left (a) uses the intervals of width 5 from the example. The distribution has a single peak in the 5 to <10 interval. The one on the right (b) uses

intervals half as wide: 0 to <2.5, 2.5 to <5, and so on. Now we see the shape of the distribution more clearly. **The choice of intervals in a histogram can affect the appearance of a distribution.** Histograms with more intervals show more detail but may have a less clear overall pattern.

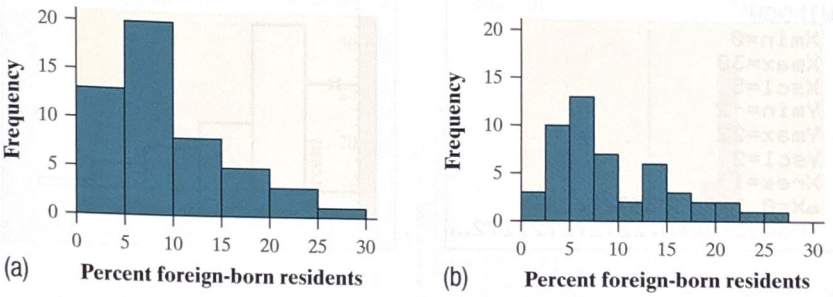


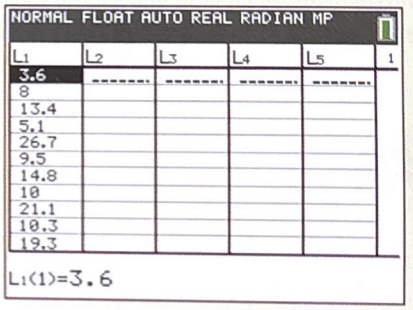
FIGURE 1.8 (a) Frequency histogram of the percentage of foreign-born residents in the 50 states with intervals of width 5, from the previous example. (b) Frequency histogram of the data with intervals of width 2.5.

1. Tech Corner MAKING HISTOGRAMS

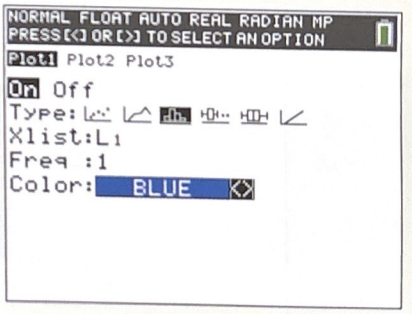
TI-Nspire and other technology instructions are on the book's website at bfwpub.com/tps7e.

You can use technology to make a histogram. The technology's default choice of intervals is a good starting point, but you should adjust the intervals to fit with common sense. We'll illustrate using data on the percentage of foreign-born residents in the states from the preceding example.

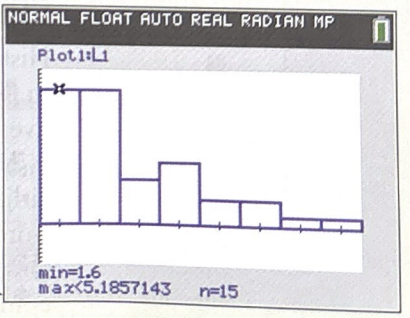
1. Enter the percentage of foreign-born residents in each state in your statistics/list editor. Press STAT and choose Edit..., then type the values into list L1.



2. Set up a histogram in the statistics plots menu. Press 2nd Y= (STAT PLOT), press ENTER or 1 to go into Plot1, and adjust the settings as shown.

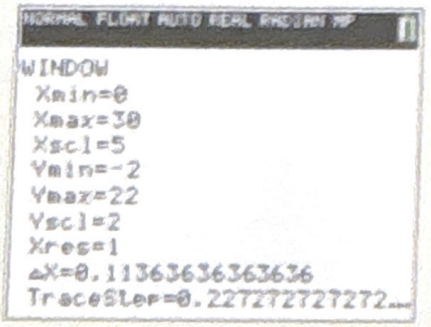


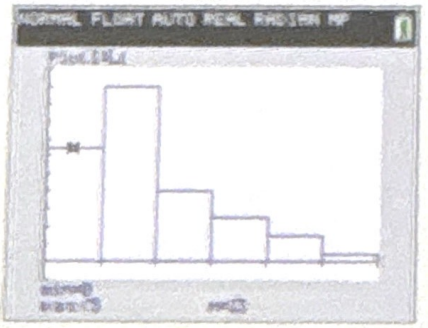
3. Use ZoomStat to let the calculator choose intervals and make a histogram. Press ZOOM and choose ZoomStat. Press TRACE and use the left and right arrow keys to examine the intervals.



Note the calculator's unusual choice of intervals.

4. Adjust the intervals to match those in Figure 1.8(a), then graph the histogram. Press WINDOW and enter the values shown for Xmin, Xmax, Xscl, Ymin, Ymax, and Yscl. Then press GRAPH. Press TRACE and use the left and right arrow keys to examine the intervals.



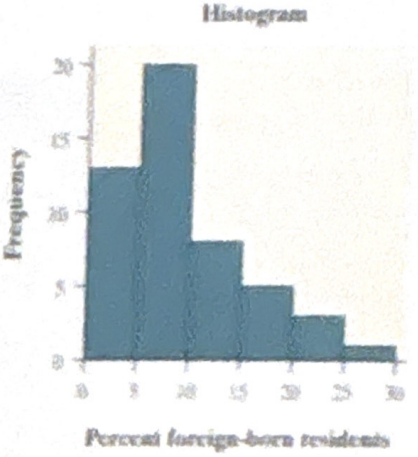


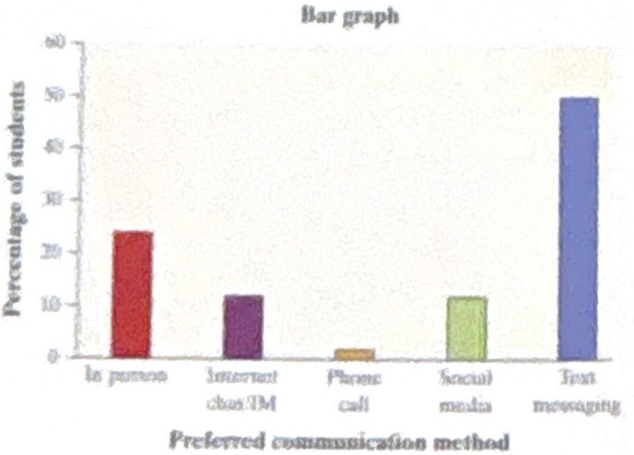
AP® EXAM TIP

If you're asked to make a graph on a free-response question, be sure to label and scale your axes. Don't just transfer what you see on a TI-83/84 calculator screen to your paper and expect to earn full credit.

5. See if you can modify the graph so that it matches the histogram in Figure 1.8(b).

Don't confuse histograms and bar graphs. Although histograms resemble bar graphs, their details and uses are different. A histogram displays the distribution of a *quantitative variable*. Its horizontal axis identifies intervals of values that the variable can take. A bar graph displays the distribution of a *categorical variable*. Its horizontal axis identifies the categories. Be sure to draw bar graphs with blank space between the bars to separate the categories. Draw histograms with no space between the bars for adjacent intervals. For comparison, here is one of each type of graph from earlier examples:





COMPARING DISTRIBUTIONS WITH HISTOGRAMS

Histograms can also be used to compare the distribution of a quantitative variable in two or more groups. Be sure to use the same intervals when making comparative histograms, so the graphs can be drawn using a common horizontal axis scale.

Mazie is interested in comparing the reading levels of a biology journal and an airline magazine. To do so, Mazie counts the number of letters in a sample of 400 words from an article in the journal, and then counts the number of letters in a sample of 100 words from an article in the airline magazine. Figure 1.9(a) displays comparative histograms of the data. This figure is misleading — it compares frequencies, but the two samples were of very different sizes (400 and 100).

Using the same data, we produced the histograms in Figure 1.9(b). By using relative frequencies, this figure makes a valid comparison of word lengths in the two samples. The moral of the story: **use relative frequencies (percentages or proportions) when comparing distributions, especially if the groups have different sizes.**

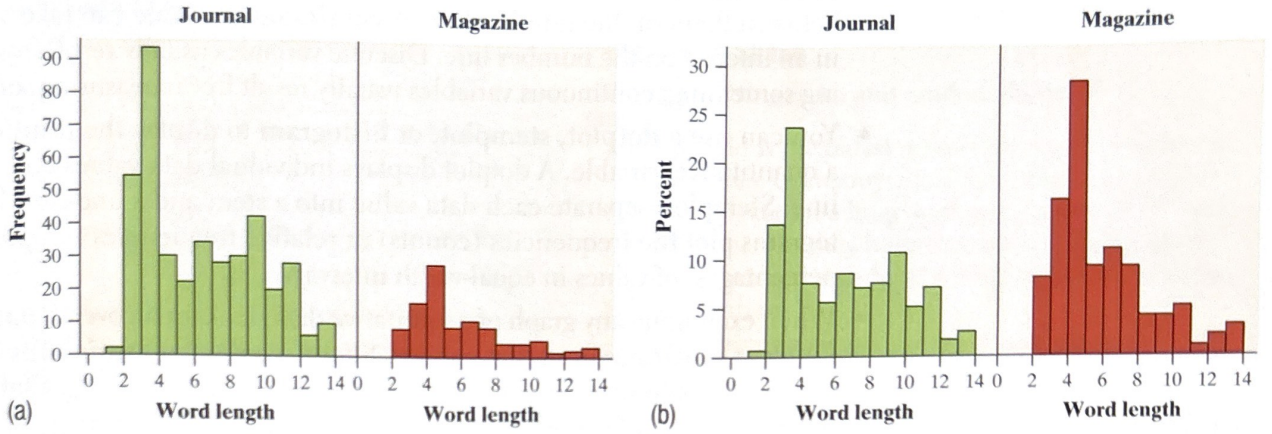
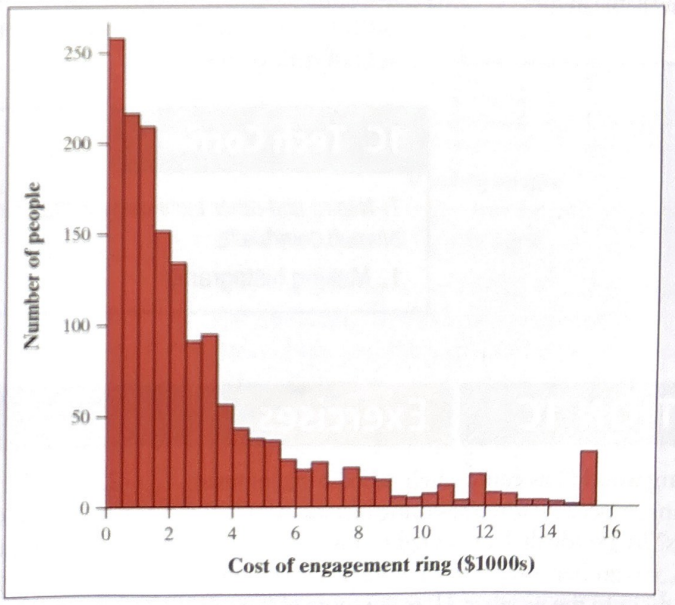


FIGURE 1.9 Two sets of histograms comparing word lengths in articles from a biology journal and from an airline magazine. In graph (a), the vertical scale uses frequencies. Graph (b) fixes the problem of different sample sizes by using percentages (relative frequencies) on the vertical scale.



CHECK YOUR UNDERSTANDING

When getting married, some people choose to buy an engagement ring for their partner. How much do people spend on engagement rings, on average? To find out, *The New York Times* and *Morning Consult* surveyed a random sample of 1640 U.S. adults who bought an engagement ring. Here is a histogram of the data:[39]



1. About what percentage of people reported spending at least $2000 on an engagement ring? Show your method clearly.
2. Describe the shape of the distribution.
3. Estimate the center and variability of the distribution.

SECTION 1C Summary

- There are two types of quantitative variables: discrete and continuous. A **discrete variable** can take a countable set of possible numeric values with gaps between them on the number line. A **continuous variable** can take any value in an interval on the number line. Discrete variables usually result from counting something; continuous variables usually result from measuring something.
- You can use a **dotplot, stemplot,** or **histogram** to display the distribution of a quantitative variable. A dotplot displays individual data values on a number line. Stemplots separate each data value into a stem and a one-digit leaf. Histograms plot the frequencies (counts) or relative frequencies (proportions or percentages) of values in equal-width intervals.
- When examining any graph of quantitative data, look for an *overall pattern* and for clear *departures* from that pattern. **Shape, center,** and **variability** describe the overall pattern of the distribution of a quantitative variable. **Outliers** are observations that lie outside the overall pattern of a distribution.
- Some distributions have simple shapes, such as **roughly symmetric, skewed to the left,** or **skewed to the right.** A distribution in which the frequency (relative frequency) of each possible value is about the same is **approximately uniform.**
- The number of peaks is another aspect of overall shape. So are distinct clusters and gaps. A single-peaked graph is sometimes called *unimodal*, and a double-peaked graph is sometimes called *bimodal*.
- Dotplots, back-to-back stemplots, and histograms can also be used to compare distributions of quantitative data. When comparing the distribution of a quantitative variable in two or more groups, be sure to compare shape, outliers, center, and variability.
- Histograms are for quantitative data; bar graphs are for categorical data. In both types of graphs, be sure to use relative frequencies when comparing data sets of different sizes.

AP® EXAM TIP
AP® Daily Videos

Review the content of this section and get extra help by watching the AP® Daily Videos for Topics 1.5 and 1.6, which are available in AP® Classroom.

1C Tech Corner

TI-Nspire and other technology instructions are on the book's website at bfwpub.com/tps7e.

1. Making histograms Page 35

SECTION 1C Exercises

1. **Protecting wood** How can we help wood surfaces resist weathering, especially when restoring historic wooden buildings? In a study that attempted to answer this question, researchers prepared wooden panels and then exposed them to the weather. Here are some of the variables the researchers recorded: type of wood, paint thickness, paint color, weathering time (1, 2, or 3 months), and number of blemishes. Classify each variable as categorical, quantitative (discrete), or quantitative (continuous).

2. **Social media** A social media company records data on each of its users for several variables: internet provider, age (in years), how many times they visited the site, total time spent on the site, country where they live, and how long since they created a member profile. Classify each variable as categorical, quantitative (discrete), or quantitative (continuous).

Displaying Quantitative Data: Dotplots

3. **Women's soccer** How good was the 2019 U.S. women's soccer team? With players like Carli Lloyd, Alex Morgan, and Megan Rapinoe, the team put on an impressive showing en route to winning the 2019 Women's World Cup. Here are data on the number of goals scored by the team in games played in the 2019 season:[40]

pg 23

1	1	2	2	1	5	6	3	5	3	13	3	2	2	2	2	2	3	4	3	2	1	3	6

(a) Make a dotplot of these data.

(b) In what proportion of games did the team score 4 or more goals?

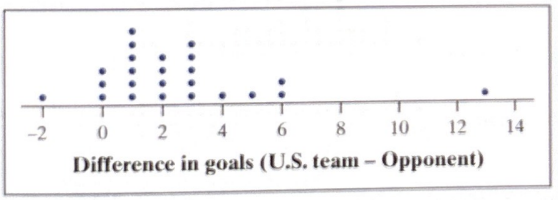
4. **Fuel efficiency** The Environmental Protection Agency (EPA) is in charge of determining and reporting fuel economy ratings for cars. Here are the EPA estimates of highway gas mileage in miles per gallon (mpg) for a sample of 21 model-year 2022 midsize cars:[41]

25	30	27	31	38	26	28	40	25	28	30
31	30	30	34	30	31	31	32	48	31	

(a) Make a dotplot of these data.

(b) What percentage of the car models in the sample get more than 35 mpg on the highway?

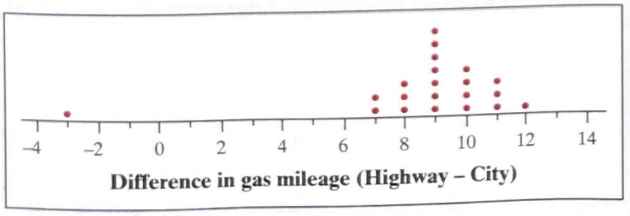
5. **More women's soccer** The following dotplot shows the difference in the number of goals scored (U.S. women's team – Opponent) in each game from Exercise 3.



(a) Explain what the dot above −2 represents.

(b) What does the graph tell us about how well the team did in 2019? Be specific.

6. **Better fuel efficiency** The following dotplot shows the difference in EPA mileage ratings (Highway – City) for each of the 21 model-year 2022 midsize cars from Exercise 4.



(a) The dot above −3 is for the Toyota Prius. Explain what this dot represents.

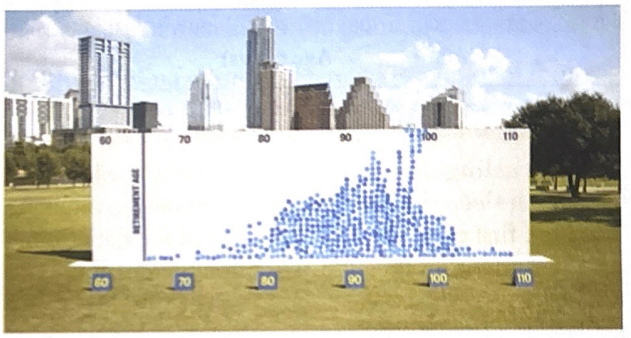
(b) What does the graph tell us about fuel economy in the city versus on the highway for these car models? Be specific.

Describing Shape

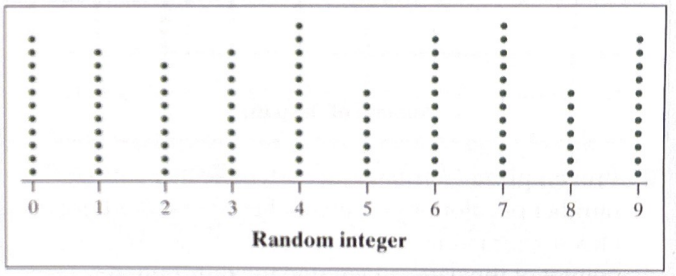
7. **Getting older and random digits**

pg 25

(a) How old is the oldest person you know? Prudential Insurance Company asked 400 people to place a blue sticker on a huge wall next to the age of the oldest person they have ever known. An image of the graph is shown here. Describe the shape of the distribution.

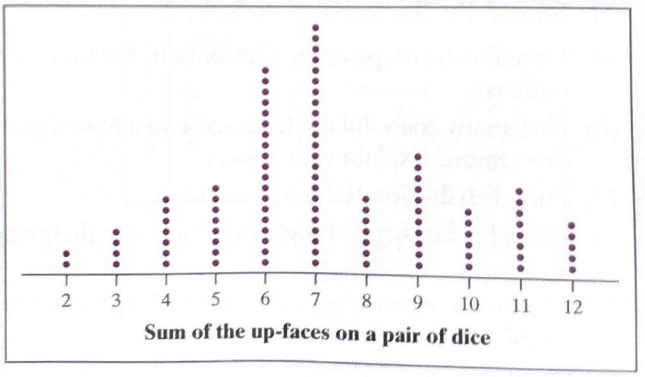


(b) The dotplot displays the results of using a random number generator to produce 100 digits between 0 and 9, inclusive. Describe the shape of the distribution.

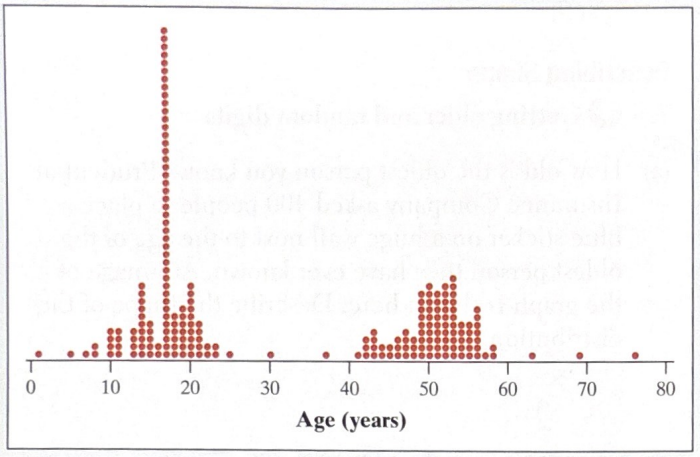


8. **Roll the dice and family ages**

(a) The dotplot shows the results of rolling a pair of fair, six-sided dice and finding the sum of the up-faces 100 times. Describe the shape of the distribution.



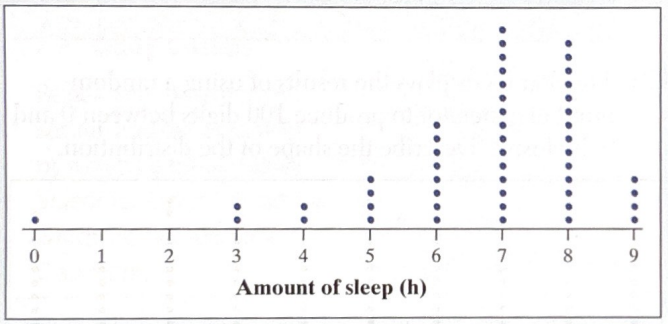
(b) Statistics instructor Paul Myers collected data on the ages (in years) of family members for each student in his class. The dotplot displays the data. Describe the shape of the distribution.



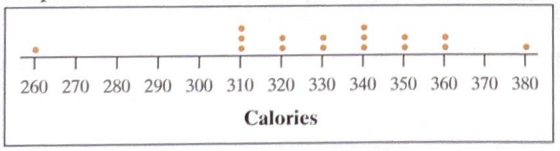
Describing Distributions of Quantitative Data

9. **Feeling sleepy?** A statistics professor asked how much sleep (in hours) students got on the night prior to their first exam. Here is a dotplot of the data from the 50 students in the class. Describe the distribution.

pg 26



10. **Frozen pizza** *Consumer Reports* collected data on the number of calories per serving for 16 brands of frozen cheese pizza as part of its product reviews. Here is a dotplot of the data.[42] Describe the distribution.



11. **Soccer distribution** Refer to Exercise 3.

(a) Describe the shape of the distribution. Are there any outliers?

(b) How many goals did the team score in a typical game that season? Explain your answer.

12. **Fuel distribution** Refer to Exercise 4.

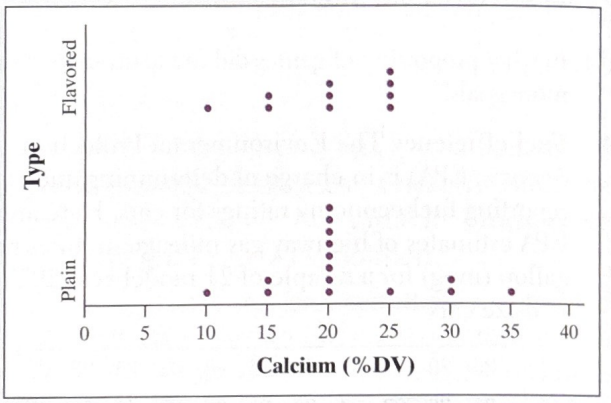
(a) Describe the shape of the distribution. Are there any outliers?

(b) What is the typical fuel efficiency of the 21 cars in the sample? Explain your answer.

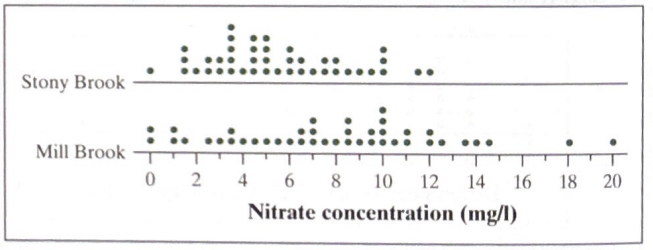
Comparing Distributions of Quantitative Data

13. **Calcium in yogurt** Which dairy yogurt is healthier: plain or flavored? The *Nutrition Action Healthletter* provided data on calories, saturated fat, protein, calcium, and total sugars for many popular brands of yogurt.[43] Here are parallel dotplots that display the data on calcium content as a percentage of the recommended daily value (%DV) of 1000 milligrams for several brands of plain and flavored dairy yogurt. Compare the distributions of calcium content in these two types of yogurt.

pg 28

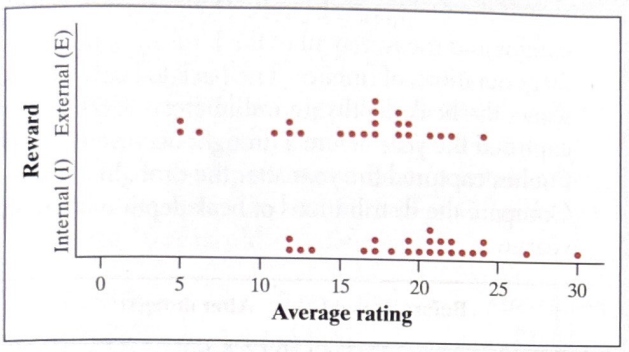


14. **Healthy streams** Nitrates are organic compounds that are a main ingredient in fertilizers. When fertilizers run off into streams, the nitrates can have a toxic effect on fish. An ecologist studying nitrate pollution in two streams measures nitrate concentrations at 42 places in Stony Brook and 42 places in Mill Brook. The parallel dotplots display the data. Compare the distributions of nitrate concentration in these two streams.



15. **Enhancing creativity** Do external rewards — like money, praise, fame, and grades — promote creativity? Researcher Teresa Amabile recruited 47 experienced creative writers who were college students and divided them at random into two groups. The students in one group were given a list of statements about external reasons (E) for writing, such as public recognition, making money, or pleasing their parents. Students in the other group were given a list of statements about internal reasons (I) for writing, such as expressing yourself and enjoying playing with words. Both groups were then instructed to write a poem about laughter. Each student's poem was rated separately by 12 different poets using a creativity scale.[44] These ratings were averaged to obtain an overall creativity score for each poem.

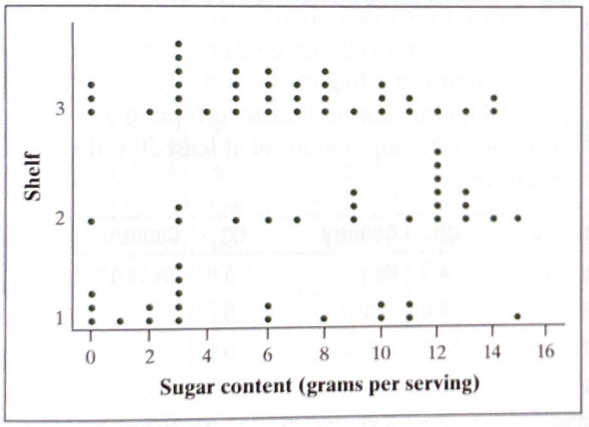
Parallel dotplots of the two groups' creativity scores are shown here.



(a) Is the variability in creativity scores similar or different for the two groups? Justify your answer.

(b) Do the data suggest that external rewards promote creativity? Justify your answer.

16. **Healthy cereal?** Researchers collected data on 76 brands of cereal at a local supermarket.[45] For each brand, the sugar content (grams per serving) and the shelf in the store on which the cereal was located (1 = bottom, 2 = middle, 3 = top) were recorded. Here are parallel dotplots of the data:



(a) Is the variability in sugar content of the cereals on the three shelves similar or different? Justify your answer.

(b) Critics claim that supermarkets tend to put sugary cereals on the middle shelf, where kids can better see them. Do the data from this study support this claim? Justify your answer.

Displaying Quantitative Data: Stemplots

17. **Snickers® are fun!** Here are the weights (in grams) of 17 Snickers Fun Size bars from a single bag:

pg 30

17.1	17.4	16.6	17.4	17.7	17.1	17.3	17.7	17.8
19.2	16.0	15.9	16.5	16.8	16.5	17.1	16.7	

(a) Make a stemplot of these data. What interesting feature does the graph reveal?

(b) The advertised weight of a Snickers Fun Size bar is 17 grams. What percentage of candy bars in this sample weigh less than advertised?

18. **Eat your beans!** Beans and other legumes are a great source of protein. The following data give the protein content of 30 different varieties of beans, in grams per 100 grams of cooked beans.[46]

7.5	8.2	8.9	9.3	7.1	8.3	8.7	9.5	8.2	9.1
9.0	9.0	9.7	9.2	8.9	8.1	9.0	7.8	8.0	7.8
7.0	7.5	13.5	8.3	6.8	10.6	8.3	7.6	7.7	8.1

(a) Make a stemplot of these data. What interesting feature does the graph reveal?

(b) What proportion of these bean varieties contain more than 9 grams of protein per 100 grams of cooked beans?

19. **South Carolina counties** Here is a stemplot of the areas of the 46 counties in South Carolina. Note that the data have been rounded to the nearest 10 square miles (mi^2).

3	9999
4	0116689
5	01115566778
6	47899
7	01245579
8	0011
9	13
10	8
11	233
12	2

Key: 6|4 represents a county with an area of 635 to 644.99 square miles.

(a) What is the area of the largest South Carolina county?

(b) Describe the distribution of area for the 46 South Carolina counties.

20. **Shopping spree** The stemplot displays data on the amount spent by 50 shoppers at a grocery store. Note that the values have been rounded to the nearest dollar.

0	399
1	1345677889
2	000123455668888
3	25699
4	1345579
5	0359
6	1
7	0
8	366
9	3

Key: 9|3 = $92.50 to $93.49 spent

(a) What was the smallest amount spent by any of the shoppers?

(b) Describe the distribution of amount spent by these 50 shoppers.

21. **Arizona heat** Here is a stemplot of the high temperature readings (in degrees Fahrenheit) for Phoenix, Arizona, for each day in July in a recent year:[47]

8	4
8	
9	3
9	799
10	011223444
10	556667788999
11	0113
11	5

(a) Why did we split the stems?
(b) Give an appropriate key for this stemplot.
(c) Describe the shape of the distribution. Are there any outliers?

22. **Watch that caffeine!** The U.S. Food and Drug Administration (FDA) limits the amount of caffeine in a 12-ounce can of carbonated beverage to 72 milligrams. That translates to a maximum of 48 milligrams of caffeine per 8-ounce serving. Data on the caffeine content of popular soft drinks (in milligrams per 8-ounce serving) are displayed in the stemplot.

1	556
2	033344
2	55667778888899
3	113
3	55567778
4	33
4	77

(a) Why did we split the stems?
(b) Give an appropriate key for this graph.
(c) Describe the shape of the distribution. Are there any outliers?

23. **Acorns and oak trees** Of the many species of oak trees in the United States, 28 grow on the Atlantic Coast and 11 grow in California. The back-to-back stemplot displays data on the average volume of acorns (in cubic centimeters) for these 39 oak species. Compare the distributions of acorn size for the oak trees in these two regions.[48]

Atlantic Coast		California
998643	0	4
88864211111	1	06
50	2	06
6640	3	
8	4	1
	5	59
8	6	0
	7	1
1	8	
1	9	
5	10	
	11	
	12	
	13	
	14	
	15	
	16	
	17	1

Key: 2|6 = An oak species whose acorn volume is 2.6 cm^3.

24. **Finch evolution** Biologists Peter and Rosemary Grant spent many years collecting data on finches in a remote part of the Galápagos Islands. Their research team caught and measured all of the birds in more than 20 generations of finches. The back-to-back stemplot shows the beak depths (in millimeters) of 89 finches captured the year before a drought occurred and 89 finches captured the year after the drought occurred.[49] Compare the distributions of beak depth in these two years.

Before drought		After drought
2	6	
8	6	
411	7	1
98	7	9
44420	8	044
9999977765555	8	778
444321111100000	9	0011123344
99999888887777655	9	5666666777789999
444433322111111000	10	000222223333334444
8766655555	10	55555566666777778999
440	11	0000111134444
7	11	5667

KEY: 10 | 0 = 10.0 mm beak depth

Displaying Quantitative Data: Histograms

25. **Carbon dioxide emissions** Burning fuels in power plants and motor vehicles emits carbon dioxide (CO_2), which contributes to global warming. The table displays CO_2 emissions in metric tons per person from 48 countries with populations of at least 20 million in a recent year.[50]

pg 33

Country	CO_2	Country	CO_2	Country	CO_2
Algeria	4.0	Italy	5.6	South Africa	8.2
Argentina	4.0	Japan	8.7	Spain	5.4
Australia	16.3	Kenya	0.3	Sudan	0.5
Bangladesh	0.6	Korea, North	1.5	Tanzania	0.2
Brazil	2.2	Korea, South	11.9	Thailand	4.1
Canada	15.4	Malaysia	7.8	Turkey	4.9
China	7.1	Mexico	3.4	Ukraine	5.1
Colombia	2.0	Morocco	2.0	United Kingdom	5.5
Congo	0.6	Myanmar	0.5	United States	16.1
Egypt	2.5	Nepal	0.5	Uzbekistan	3.3
Ethiopia	0.2	Nigeria	0.7	Venezuela	4.1
France	5.0	Pakistan	1.2	Vietnam	2.6
Germany	8.4	Peru	1.7		
Ghana	0.5	Philippines	1.3		
India	1.9	Poland	8.5		
Indonesia	2.3	Romania	3.9		
Iran	9.4	Russia	11.5		
Iraq	5.6	Saudi Arabia	17.0		

(a) Make a histogram of the data using intervals of width 2, starting at 0.

(b) What proportion of countries had CO_2 emissions of at least 10 metric tons per person?

26. **Traveling to work** How long do people travel each day to get to work? The following table gives the average travel times to work (in minutes) in a recent year for workers in each state and the District of Columbia who are at least 16 years old and don't work at home.[51]

State	Travel time to work (min)	State	Travel time to work (min)	State	Travel time to work (min)
AL	24.9	KY	23.6	ND	17.3
AK	19.1	LA	25.7	OH	23.7
AZ	25.7	ME	24.2	OK	21.9
AR	21.7	MD	33.2	OR	23.9
CA	29.8	MA	30.2	PA	27.2
CO	25.8	MI	24.6	RI	25.2
CT	26.6	MN	23.7	SC	25.0
DE	26.3	MS	24.8	SD	17.2
DC	30.8	MO	23.9	TN	25.2
FL	27.8	MT	18.3	TX	26.6
GA	28.8	NE	18.8	UT	21.9
HI	27.5	NV	24.6	VT	23.3
ID	21.1	NH	27.5	VA	28.7
IL	29.2	NJ	32.2	WA	28.0
IN	23.8	NM	22.3	WV	25.9
IA	19.3	NY	33.6	WI	22.2
KS	19.4	NC	24.8	WY	17.9

(a) Make a histogram to display the travel time data using intervals of width 2 minutes, starting at 16 minutes.

(b) In what proportion of states plus the District of Columbia is the average travel time at least 20 minutes?

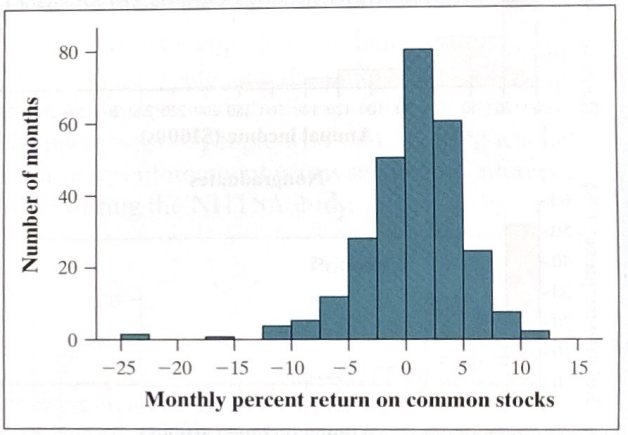
27. **Home runs** Here are the number of home runs hit by each of the 30 Major League Baseball teams in a single season. Make a histogram that effectively displays the distribution of number of home runs hit. Describe the distribution.

220	249	213	245	256	182	227	223	224	149
288	162	220	279	146	250	307	242	306	257
215	163	219	239	167	210	217	223	247	231

28. **Country music** Here are the lengths, in minutes, of 50 songs by country artist Dierks Bentley. Make a histogram that effectively displays the distribution of song length. Describe the distribution.

4.2	4.0	3.9	3.8	3.7	4.7
3.4	4.0	4.4	5.0	4.6	3.7
4.6	4.4	4.1	3.0	3.2	4.7
3.5	3.7	4.3	3.7	4.8	4.4
4.2	4.7	6.2	4.0	7.0	3.9
3.4	3.4	2.9	3.3	4.0	4.2
3.2	3.4	3.7	3.5	3.4	3.7
3.9	3.7	3.8	3.1	3.7	3.6
4.5	3.7				

29. **Returns on common stocks** The return on a stock is the change in its market price plus any dividend payments made. Return is usually expressed as a percentage of the beginning price. The figure shows a histogram of the distribution of monthly percent return for the U.S. stock market (total return on all common stocks) in 273 consecutive months.[52]

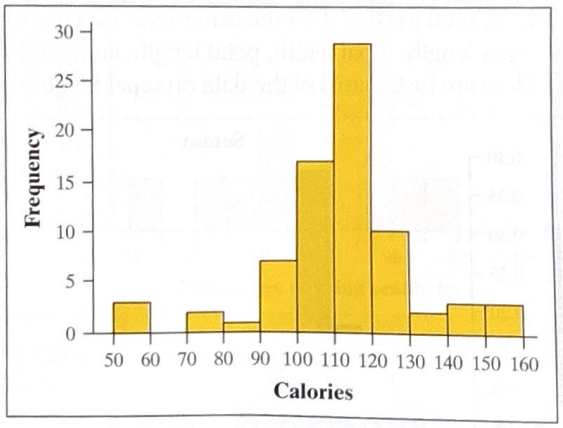


(a) A return less than zero means that stocks lost value in that month. About what percentage of all months had returns less than zero?

(b) Describe the shape of the distribution. Are there any outliers?

(c) Estimate the center and variability of the distribution.

30. **More healthy cereal?** Researchers collected data on calories per serving for 77 brands of breakfast cereal. The histogram displays the data.[53]

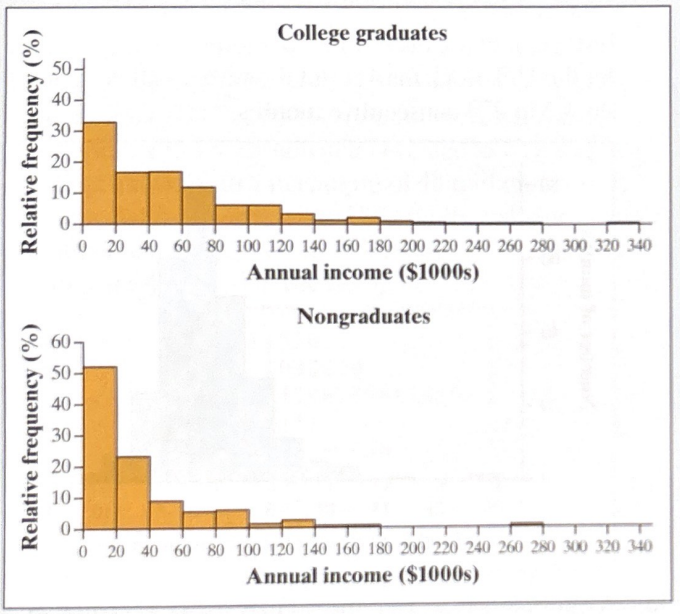


(a) About what percentage of the cereal brands have 130 or more calories per serving?

(b) Describe the shape of the distribution. Are there any outliers?

(c) Estimate the center and variability of the distribution.

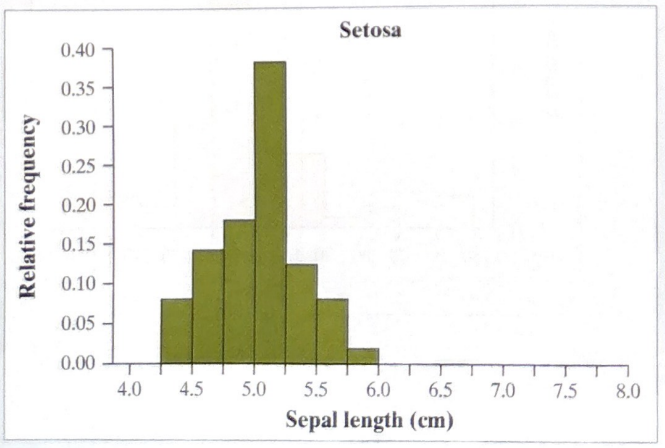
31. **Value of a college degree** Is it true that students who earn an associate's degree or a bachelor's degree make more money than students who attend college but do not earn a degree? To find out, we selected a random sample of 500 U.S. residents aged 18 and older who had attended college from a recent Current Population Survey.[54] The educational attainment and annual income of each person were recorded. Here are relative frequency histograms of the income data for the 327 college graduates and the 173 nongraduates:

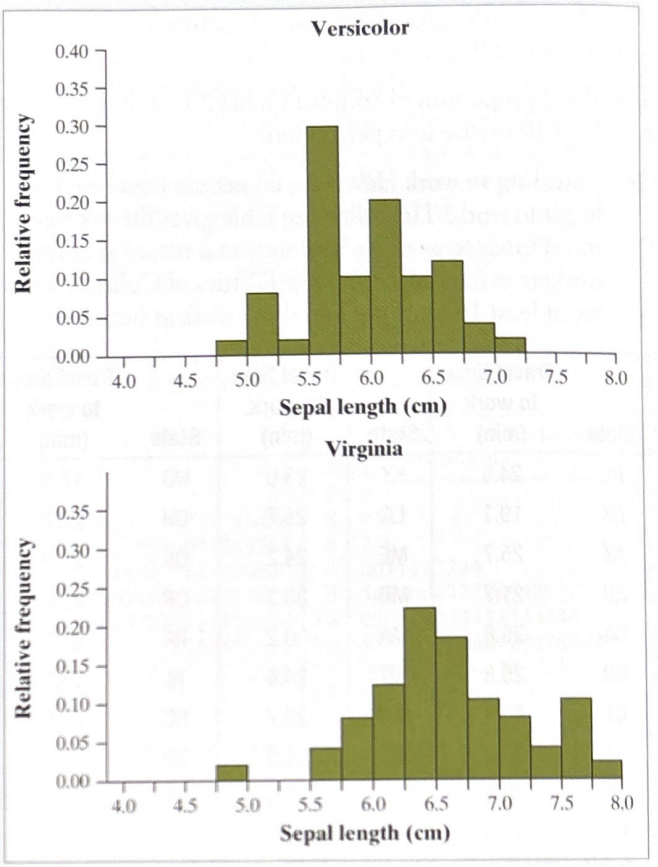


(a) Compare the distributions of annual income for the two groups.

(b) Would it be appropriate to use frequency histograms instead of relative frequency histograms in this setting? Explain why or why not.

32. **Irises** U.S. botanist Edgar Anderson collected data on 50 iris flowers from each of three different species: *Setosa*, *Versicolor*, and *Virginia*.[55] British statistician Ronald Fisher used these data to develop a way of classifying an individual iris based on measurements (in centimeters) of sepal length, sepal width, petal length, and petal width. Here are histograms of the data on sepal length:

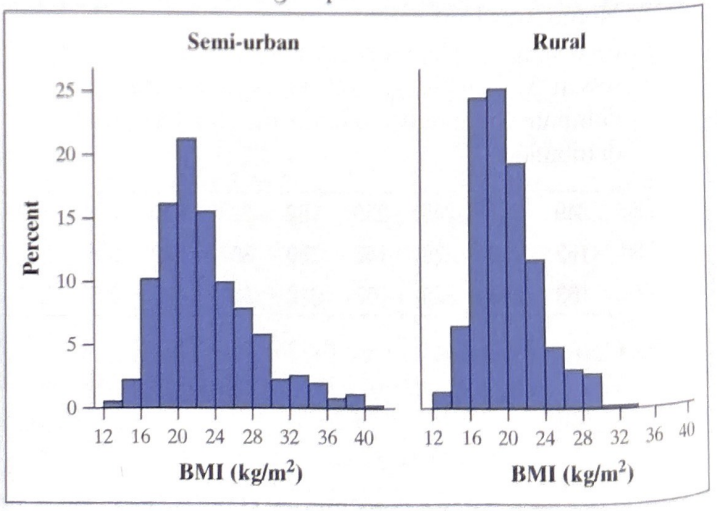




(a) Compare the distributions of sepal length for the three species.

(b) Would it be appropriate to use frequency histograms instead of relative frequency histograms in this setting? Explain why or why not.

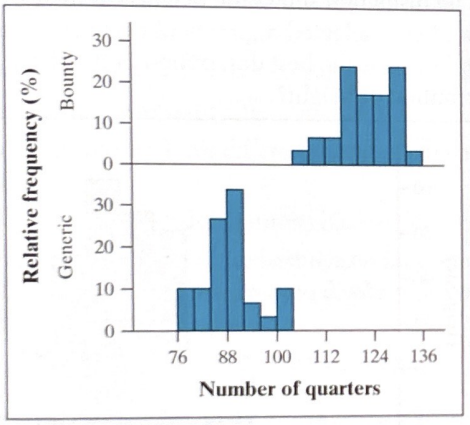
33. **Body mass index** Researchers wanted to investigate the relationship between blood pressure and body mass index (BMI) in predominantly lean African populations. As part of the research, they collected data on BMI from a sample of 338 rural and 290 semi-urban women in Ghana. Here are comparative histograms of the data for the two groups:[56]



(a) Describe similarities and differences between the two distributions.

(b) According to the World Health Organization, a person with a BMI less than 18 is considered underweight. Compare the percentages of rural and semi-urban women in the sample who are underweight.

34. **Paper towels** In commercials for Bounty paper towels, the manufacturer claims that they are the "quicker picker-upper"—but are they also the stronger picker-upper? Two statistics students decided to investigate. They selected a random sample of 30 Bounty paper towels and a random sample of 30 generic paper towels and measured their strength when wet. To do this, the students uniformly soaked each paper towel with 4 ounces of water, held two opposite edges of the paper towel, and counted how many quarters each paper towel could hold until ripping. The data are displayed in the relative frequency histograms.[57]



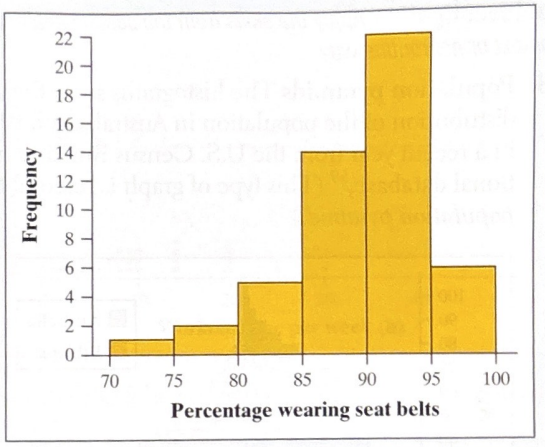
(a) Describe similarities and differences between the two distributions.

(b) What conclusion should the students make from their research? Explain your answer.

35. **Birth months** Imagine asking a random sample of 60 students from your school about their birth months. Draw a plausible (believable) graph of the distribution of birth month. Should you use a bar graph or a histogram to display the data?

36. **Die rolls** Imagine rolling a fair, six-sided die 60 times. Draw a plausible (believable) graph of the distribution of die rolls. Should you use a bar graph or a histogram to display the data?

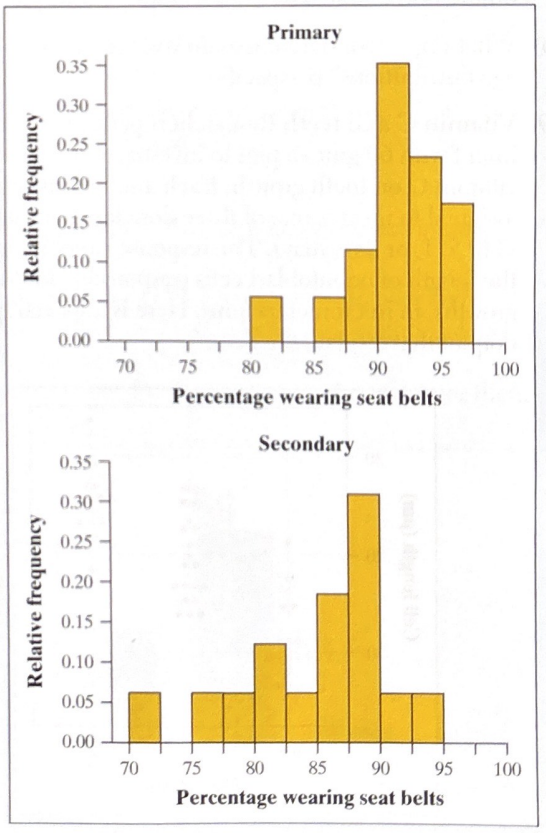
37. **Seat belts** Each year, the National Highway Traffic Safety Administration (NHTSA) conducts an observational study of seat belt use in all 50 states. Trained observers station themselves at randomly selected locations along roadways in each state, and then record data on seat belt use by people in passing vehicles. Here is a histogram of the percentage of people who were wearing seat belts in each state during a recent year:[58]



(a) In what proportion of states was at least 90% seat belt use observed?

(b) Describe the shape of the distribution.

Thirty-four states enforce a seat belt violation as a primary offense, while 16 states enforce a seat belt violation only as a secondary offense. The histograms show the percentage of people who were wearing seat belts in primary enforcement versus secondary enforcement states during the NHTSA study.

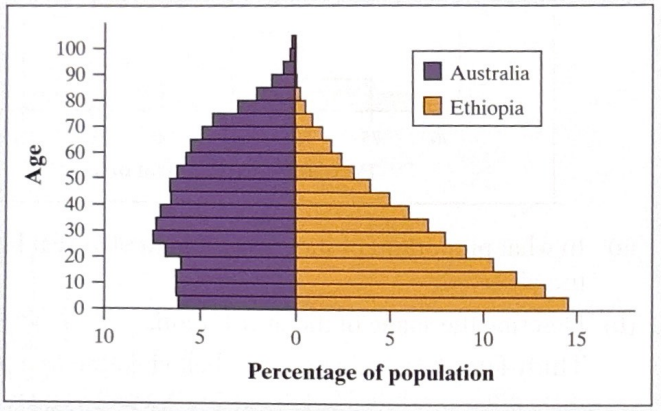


(c) Why did we use relative frequency instead of frequency when making these histograms?

(d) Compare the distributions of seat belt use in primary and secondary enforcement states.

For Investigation *Apply the skills from the section in a new context or nonroutine way.*

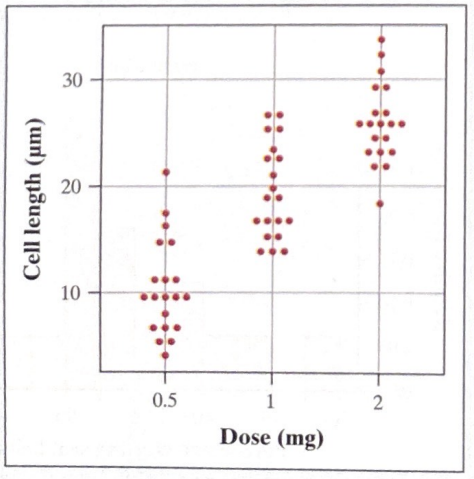
38. **Population pyramids** The histograms show the age distribution of the population in Australia and Ethiopia in a recent year from the U.S. Census Bureau's international database.[59] (This type of graph is referred to as a *population pyramid.*)



(a) The total population of Australia at this time was about 26 million. Ethiopia's population was about 111 million. Why did we use percentages rather than counts on the horizontal axis?

(b) What important differences do you see between the age distributions? Be specific.

39. **Vitamin C and teeth** Researchers performed an experiment with 60 guinea pigs to investigate the effect of vitamin C on tooth growth. Each animal was randomly assigned to receive one of three dose levels of vitamin C (0.5, 1, or 2 mg/day). The response variable was the length of odontoblast cells (responsible for tooth growth), in micrometers (μm). Here is a special type of dotplot that displays the data:[60]



(a) How does this dotplot differ from the ones shown in this section?

(b) Do these data provide evidence that vitamin C helps teeth grow in guinea pigs similar to the ones used in this experiment? Justify your answer.

Multiple Choice *Select the best answer for each question.*

40. Here are the amounts of money (cents) in coins carried by 10 students in a statistics class: 50, 35, 0, 46, 86, 0, 5, 47, 23, 65. To make a stemplot of these data, you would use stems

(A) 0, 2, 3, 4, 5, 6, 8.

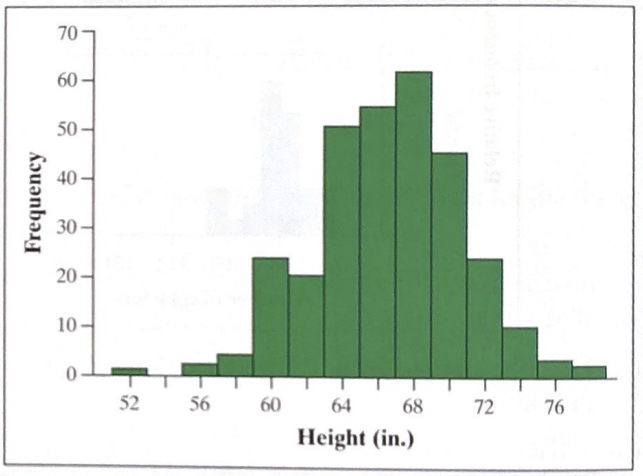
(B) 0, 1, 2, 3, 4, 5, 6, 7, 8.

(C) 0, 3, 5, 6, 7.

(D) 00, 10, 20, 30, 40, 50, 60, 70, 80, 90.

(E) None of these.

41. The histogram shows the heights (in inches) of 300 randomly selected high school students. Which of the following is the best description of the shape of the distribution of height?



(A) Roughly symmetric and single-peaked (unimodal)

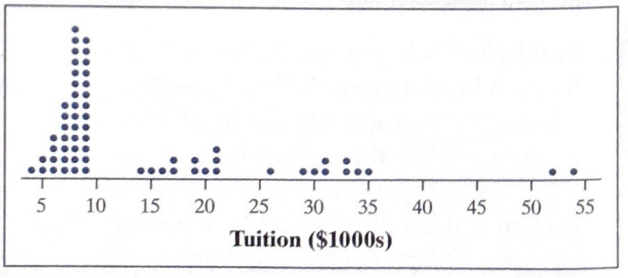
(B) Roughly symmetric and double-peaked (bimodal)

(C) Roughly symmetric and multi-peaked (multimodal)

(D) Skewed to the left

(E) Skewed to the right

42. The dotplot shows the tuition (to the nearest $1000) for the 63 largest colleges and universities in North Carolina in a recent year.[61]



Which of the following statements about the dotplot is *not* correct?

(A) The center (median) of the distribution is \$8000.

(B) There are more North Carolina colleges and universities with tuitions less than \$10,000 than with tuitions greater than \$10,000.

(C) The tuitions vary from about \$4000 to about \$54,000.

(D) There are two obvious outliers — institutions with tuitions greater than \$50,000.

(E) The distribution of tuition is skewed to the right.

43. When comparing the distribution of a quantitative variable for two groups, it would be best to use relative frequency histograms rather than frequency histograms when

(A) the distributions have different shapes.

(B) the distributions have different amounts of variability.

(C) the distributions have different centers.

(D) the distributions have different numbers of data values.

(E) at least one of the distributions has outliers.

44. Which of the following is the best reason for choosing a stemplot rather than a histogram to display the distribution of a quantitative variable?

(A) Stemplots allow you to split stems; histograms don't.

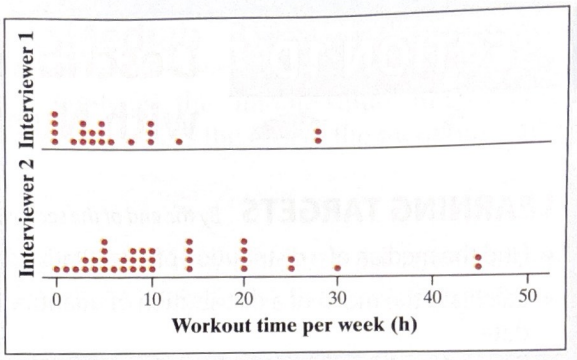
(B) Stemplots allow you to see individual data values.

(C) Stemplots are better for displaying very large sets of data.

(D) Stemplots never require rounding of values.

(E) Stemplots make it easier to determine the shape of a distribution.

45. For their final project, two statistics students designed an experiment to investigate the effect of the interviewer on people's responses to the question, "How many hours a week do you work out?" Interviewer 1 wore jeans and a school T-shirt, while Interviewer 2 wore athletic attire with the school logo. The two students were each randomly assigned three 15-minute time slots during a 90-minute period to serve as the interviewer. They asked the question of people who passed by their location near campus. Here are parallel dotplots of the data obtained by each interviewer:[62]



Which of the following statements about the graph is correct?

(A) For both interviewers, the distribution of workout time per week is roughly symmetric.

(B) There is a similar amount of variability in people's reported workout times for both interviewers.

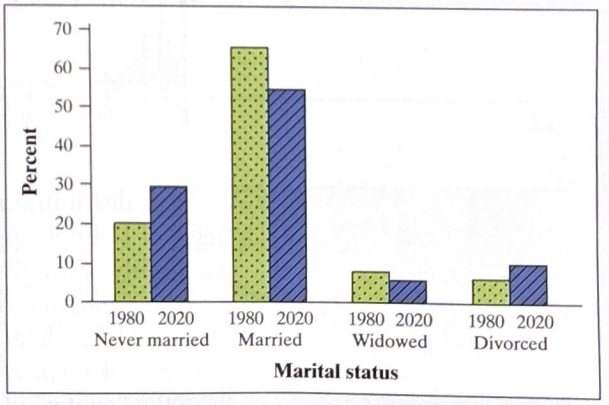
(C) Interviewer 1's distribution of workout time per week has outliers; Interviewer 2's distribution of workout time per week has no outliers.

(D) People who responded to Interviewer 2 tended to report a higher amount of workout time per week than people who responded to Interviewer 1.

(E) For both interviewers, more than 75% of people reported working out 10 or less hours per week.

Recycle and Review *Practice what you learned in previous sections.*

46. **Marital Status (1B)** The side-by-side bar graph compares the marital status of U.S. adult residents (18 years old or older) in 1980 and 2020.[63] Compare the distributions of marital status for these two years.



SECTION 1D Describing Quantitative Data with Numbers

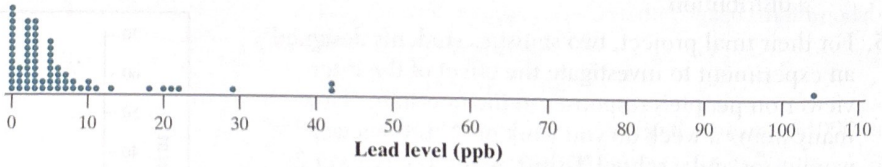
LEARNING TARGETS *By the end of the section, you should be able to:*

- Find the median of a distribution of quantitative data.
- Calculate the mean of a distribution of quantitative data.
- Find the range of a distribution of quantitative data.
- Calculate and interpret the standard deviation of a distribution of quantitative data.
- Find the interquartile range (*IQR*) of a distribution of quantitative data.
- Choose appropriate measures of center and variability to summarize a distribution of quantitative data.
- Identify outliers in a distribution of quantitative data.
- Make and interpret boxplots of quantitative data.
- Use boxplots and summary statistics to compare distributions of quantitative data.

In Section 1C, you learned how to display quantitative data with graphs. You also explored how to use these graphs to describe and compare distributions of a quantitative variable. In this section, we'll focus on numerical summaries of quantitative data.

Let's return to a familiar context from the preceding section. Recall that city managers in Flint, Michigan, switched the city's water source from Lake Huron to the Flint River to save money. Here once again are the lead levels (in parts per billion, ppb) in 71 water samples collected from randomly selected Flint dwellings after the switch, along with a dotplot of the data.[64]

0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	2	2	2	2	2	2	2
2	2	2	2	3	3	3	3	3	3	3	3
3	3	3	4	4	5	5	5	5	5	5	5
5	6	6	6	6	7	7	7	8	8	9	10
10	11	13	18	20	21	22	29	42	42	104	



This distribution is right-skewed and single-peaked. The dwelling with a lead level of 104 ppb appears to be an outlier. How should we describe the center and variability of this distribution?

The *mode* of a distribution is the most frequently occurring data value. For the lead level data, the mode is 0 ppb. But that value isn't representative of how much lead a typical Flint dwelling has in its water. We want to report a value that is in the "center" of the distribution. The mode is often not a good measure of the center because it can fall anywhere in a distribution. Also, a distribution can have multiple modes, or no mode at all.

Measuring Center: The Median

In Section 1C, we advised you to simply use the "middle value" in an ordered quantitative data set to describe its center. That's the idea of the **median.**

DEFINITION Median

The **median** is the midpoint of a distribution—the number such that about half the observations are smaller and about half are larger. To find the median, arrange the data values from smallest to largest.

- If the number n of data values is odd, the median is the middle value in the ordered list.
- If the number n of data values is even, use the average of the two middle values in the ordered list as the median.

You can find the median by hand for small sets of data. For instance, here are data on the population density (in number of people per square kilometer) for all seven countries in Central America:[65]

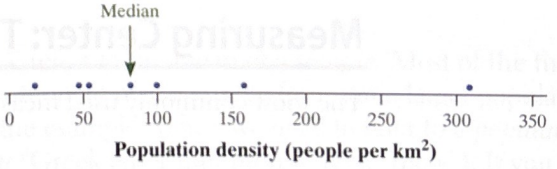
Country	Belize	Costa Rica	El Salvador	Guatemala	Honduras	Nicaragua	Panama
Population density (people per km^2)	17	100	308	158	82	48	52

To find the median, start by sorting the data values from smallest to largest:

17 48 52 (82) 100 158 308

Because there are $n = 7$ data values (an odd number), the median is the middle value in the ordered list: 82.

Here is a dotplot of the population density data. You can confirm that the median is 82 by "counting inward" from the minimum and maximum values.



EXAMPLE

More chips please
Measuring center: The median

Skill 2.C

PROBLEM: Have you ever noticed that bags of chips seem to contain lots of air and not enough chips? A group of chip enthusiasts collected data on the percentage of air in a sample of 14 popular brands of chips. Here are their data:[66]



Ann Heath

Brand	Percent air	Brand	Percent air
Cape Cod	46	Popchips	45
Cheetos	59	Pringles	28
Doritos	48	Ruffles	50
Fritos	19	Stacy's Pita Chips	50
Kettle Brand	47	Sun Chips	41
Lays	41	Terra	49
Lays Baked	39	Tostitos Scoops	34

Find the median.

SOLUTION:

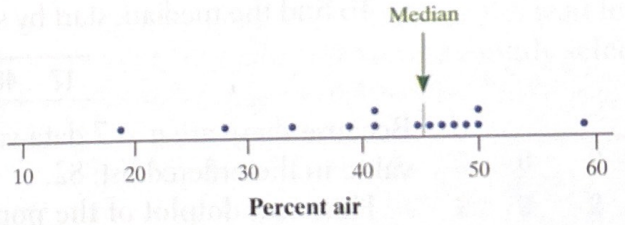
19 28 34 39 41 41 (45 | 46) 47 48 49 50 50 59

The median is $\frac{45+46}{2} = 45.5$ % air.

Sort the data values from smallest to largest. Because there are $n = 14$ data values (an even number), use the average of the middle two values in the ordered list as the median.

FOR PRACTICE, TRY EXERCISE 1

Here is a dotplot of the air in chips data from the preceding example with the median of 45.5 marked. How should we interpret the median? About half of the chip brands have less than 45.5% air, and about half have more.



Measuring Center: The Mean

The most commonly used measure of center is the **mean.**

DEFINITION The mean

The **mean** of a distribution of quantitative data is the average of all the individual data values. To find the mean, add all the values and divide by the total number of data values.

If the n data values are $x_1, x_2, \ldots, x_n$, the *sample mean* $\bar{x}$ (pronounced "x-bar") is given by the formula

$$\bar{x} = \frac{\text{sum of data values}}{\text{number of data values}} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

AP® EXAM TIP
Formula sheet

The formula for the sample mean $\bar{x}$ is included on the formula sheet provided on both sections of the AP® Statistics exam.

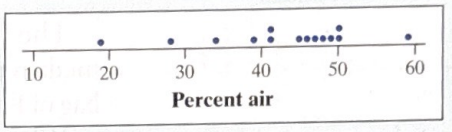
The Σ (capital Greek letter sigma) in the formula is short for "add them all up." The subscripts on the observations x_i are just a way of keeping the n data values distinct. They do not necessarily indicate the order or any other special facts about the data.

EXAMPLE More chips please
Measuring center: The mean

Skill 2.C

PROBLEM: Here are the data on percent air in the sample of 14 bags of chips from the preceding example, along with a dotplot:

Brand	Percent air	Brand	Percent air
Cape Cod	46	Popchips	45
Cheetos	59	Pringles	28
Doritos	48	Ruffles	50
Fritos	19	Stacy's Pita Chips	50
Kettle Brand	47	Sun Chips	41
Lays	41	Terra	49
Lays Baked	39	Tostitos Scoops	34



(a) Calculate the mean percent air in the bag for these 14 brands of chips.
(b) The bag of Fritos chips, with only 19% air, is a possible outlier. Calculate the mean percent air in the bag for the other 13 brands of chips. What do you notice?

SOLUTION:

(a) $\bar{x} = \dfrac{46+59+48+19+47+\ldots+34}{14} = \dfrac{596}{14} = 42.57\%$ air

(b) $\bar{x} = \dfrac{46+59+48+47+\ldots+34}{13} = \dfrac{577}{13} = 44.38\%$ air

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

The bag of Fritos decreased the mean percent air by 1.81 percentage points.

FOR PRACTICE, TRY EXERCISE 3

The notation $\bar{x}$ refers to the mean of a *sample*. Most of the time, the data we encounter can be thought of as a sample from some larger population, like the 14 bags of chips in the example. When we need to refer to a *population mean*, we'll use the symbol μ (Greek letter mu, pronounced "mew"). If you have the entire population of data available, then you can calculate μ in just the way you'd expect: add the values of all the observations, and divide by the number of observations. For instance, the population mean density in all seven South American countries is

$$\mu = \frac{17+48+52+82+100+158+308}{7} = 109.286 \text{ people per square kilometer}$$

We call $\bar{x}$ a **statistic** and μ a **parameter.** Remember **s** and **p**: statistics come from samples and parameters come from populations. In later units, you will learn how to use sample statistics to make conclusions about population parameters (known as *inferential statistics*).

DEFINITION Statistic, Parameter

A **statistic** is a number that describes some characteristic of a sample.

A **parameter** is a number that describes some characteristic of a population.

PROPERTIES OF THE MEAN

The preceding example illustrates an important weakness of the mean as a measure of center: The mean is not **resistant** to extreme values, such as outliers. The bag of Fritos, with only 19% air, decreases the mean by 1.81 percentage points.

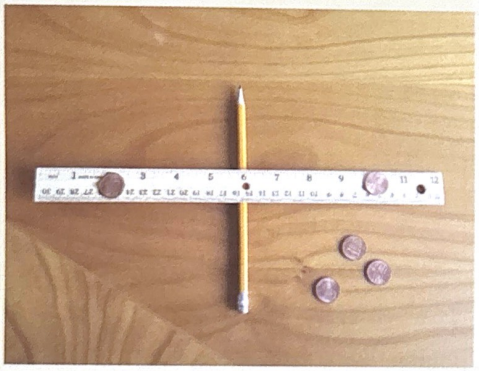
DEFINITION Resistant

A statistical measure is **resistant** if it is not affected much by extreme data values.

The median *is* a resistant measure of center. In the preceding example, the median percent air in all 14 bags of chips is 45.5. If we remove the possible outlier bag of Fritos, the median percent air in the remaining 13 bags is nearly the same (46).

Why is the mean so sensitive to extreme values? The following activity provides some insight.

ACTIVITY Interpreting the mean

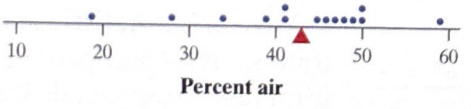


Ann Heath

In this activity, you will investigate a physical interpretation of the mean of a distribution.

1. Stack 5 pennies on top of the 6-inch mark on a 12-inch ruler. Place a pencil under the ruler to make a "seesaw" on a desk or table. Move the pencil until the ruler balances. What is the relationship between the location of the pencil and the mean of the five data values 6, 6, 6, 6, and 6?
2. Move one penny off the stack to the 8-inch mark on your ruler. Now move one other penny so that the ruler balances again without moving the pencil. Where did you put the other penny? What is the mean of the five data values represented by the pennies now?
3. Move one more penny off the stack to the 2-inch mark on your ruler. Now move both remaining pennies from the 6-inch mark so that the ruler still balances with the pencil in the same location. Is the mean of the data values still 6?
4. Discuss with your classmates: Why is the mean called the "balance point" of a distribution?

The activity gives a physical interpretation of the mean as the balance point of a distribution. For the data on percent air in each of 14 brands of chips, the dotplot balances at $\bar{x} = 42.57\%$.



COMPARING THE MEAN AND MEDIAN

Which measure — the mean or the median — should we report as the center of a distribution? That depends on both the shape of the distribution and whether there are any outliers.

- **Shape:** Figure 1.10 shows the mean and median for dotplots with three different shapes. Notice how these two measures of center compare in each case. The mean is pulled in the direction of the long tail in a skewed distribution.

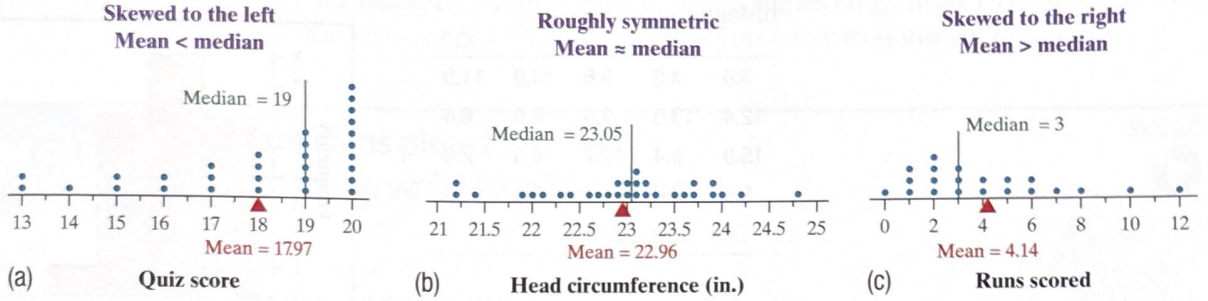


FIGURE 1.10 Dotplots that show the relationship between the mean and the median in distributions with different shapes: (a) scores on a 20-point quiz for each of the 30 students in a statistics class, (b) head circumference (in inches) for 30 players on a youth football team, and (c) runs scored by a softball team in 21 games played.

- **Outliers:** We noted earlier that the median is a resistant measure of center, but the mean is not. If we remove the two possible outliers (10 and 12) in Figure 1.10(c), the mean number of runs scored per game decreases from 4.14 to 3.42, but the median number of runs scored is still 3.

EFFECT OF SKEWNESS AND OUTLIERS ON MEASURES OF CENTER

- If a distribution of quantitative data is roughly symmetric and has no outliers, the mean and the median will be similar.
- If the distribution is strongly skewed, the mean will be pulled in the direction of the skewness but the median won't. For a right-skewed distribution, we expect the mean to be greater than the median. For a left-skewed distribution, we expect the mean to be less than the median.
- The median is resistant to outliers but the mean isn't.

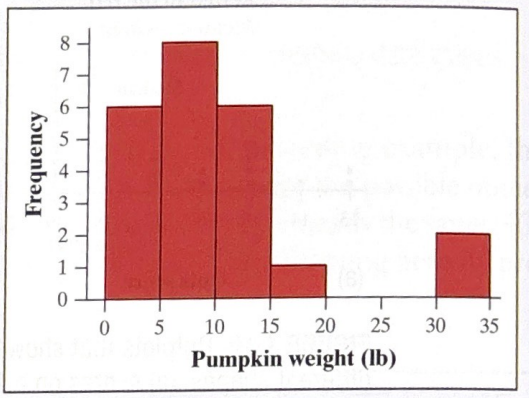
The mean and the median measure center in different ways, and both are useful. In Major League Baseball (MLB), the distribution of player salaries is strongly skewed to the right. Most players earn close to the minimum salary, which was $700,000 in 2022, while a few earn more than $20 million. The median salary for MLB players in 2022 was about $1.2 million — but the mean salary was about $4.4 million. Max Scherzer, Mike Trout, Anthony Rendon, Gerrit Cole, Carlos Correa, Manny Machado, and several other highly paid superstars pulled the mean up but their huge paychecks did not affect the median. The median gives us a good idea of what a "typical" MLB salary is. If we want to know the total salary paid to MLB players in 2022, however, we would multiply the mean salary by the total number of players: ($4.4 million)(975) ≈ $4.3 billion!



CHECK YOUR UNDERSTANDING

Some students purchased pumpkins for a carving contest. Before the contest began, they weighed the pumpkins. The weights (in pounds) are shown here, along with a histogram of the data.

3.6	4.0	9.6	14.0	11.0
12.4	13.0	2.0	6.0	6.6
15.0	3.4	12.7	6.0	2.8
9.6	4.0	6.1	5.4	11.9
5.4	31.0	33.0		



1. Find the median weight of the pumpkins.
2. Calculate the mean weight of the pumpkins.
3. Explain why the mean of the distribution would be larger than the median before doing any calculations.

Measuring Variability: The Range

Being able to describe the shape and center of a distribution of quantitative data is a great start. However, two distributions can have the same shape and center, but still look quite different.

Figure 1.11 shows comparative dotplots of the length (in millimeters, mm) of separate random samples of PVC pipe from two suppliers, A and B.[67] Both distributions are roughly symmetric and single-peaked (unimodal), with centers at about 600 mm, but the variability of these two distributions is quite different. The sample of pipes from Supplier A has a much more consistent length (less variability) than the sample from Supplier B.

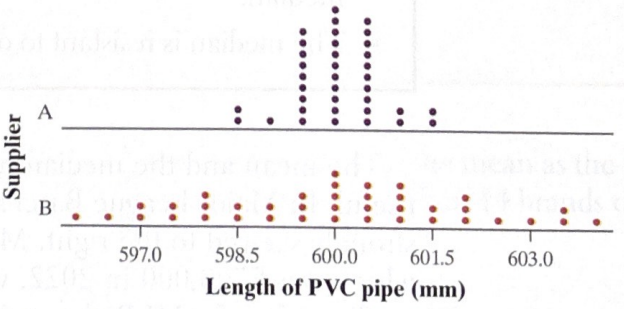


FIGURE 1.11 Comparative dotplots of the length of PVC pipes in separate random samples from Supplier A and Supplier B.

There are several ways to measure the variability of a distribution. The simplest is the **range**.

DEFINITION Range

The **range** of a distribution is the distance between the minimum value and the maximum value. That is,

$$\text{range} = \text{maximum} - \text{minimum}$$

Refer to the parallel dotplots of the length of PVC pipe from the two suppliers in Figure 1.11. For Supplier A, the range of the distribution is 601.5 − 598.5 = 3.0 mm. That's quite a bit smaller than the range of Supplier B's distribution: 604.0 − 596.0 = 8.0 mm.

Note that **the range of a data set is a single number**. In everyday language, people sometimes say things like "The data values range from 17 to 308." Be sure to use the term *range* correctly, now that you know its statistical definition.

EXAMPLE More chips please
Measuring variability: The range

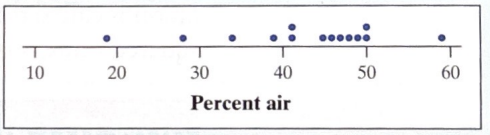
Skill 2.C

PROBLEM: Here are the data on percent air in the 14 bags of chips from the preceding two examples, along with a dotplot:

Brand	Percent air	Brand	Percent air
Cape Cod	46	Popchips	45
Cheetos	59	Pringles	28
Doritos	48	Ruffles	50
Fritos	19	Stacy's Pita Chips	50
Kettle Brand	47	Sun Chips	41
Lays	41	Terra	49
Lays Baked	39	Tostitos Scoops	34



PackStock/Alamy Stock Photo



Find the range of the distribution.

SOLUTION:

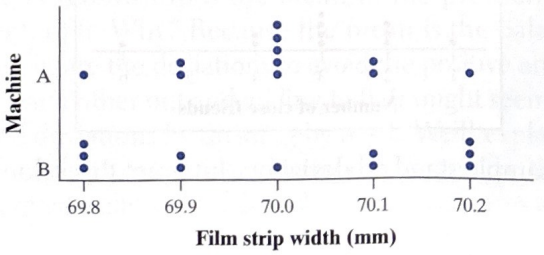
range = 59 − 19 = 40% air

range = maximum − minimum

FOR PRACTICE, TRY EXERCISE 11

The range is *not* a resistant measure of variability. It depends on only the maximum and minimum values, which may be outliers. Look again at the data on the percent air in bags of chips. Without the possible outlier at 19%, the range of the distribution would decrease from 40% air to 59% − 28% = 31% air.

The following graph illustrates another problem with using the range as a measure of variability. Many older movies were recorded on high-resolution film with a width of 70 mm, much larger than the standard 35-mm film used for modern movies. The parallel dotplots show the widths (in millimeters) of a sample of 11 strips of 70-mm movie film produced by each of two machines.[68] Both distributions are symmetric, centered at 70 mm, and have a range of 70.2 − 69.8 = 0.4 mm. But the widths of the film strips made by Machine B clearly vary more from the center of 70 mm than the film strips made by Machine A.



Measuring Variability: The Standard Deviation

To avoid the problem illustrated by the film strip example, we prefer a measure of variability that uses more than just the minimum and maximum values of a quantitative data set. If we summarize the center of a distribution with the mean, then we should use the **standard deviation** to describe the variation of data values around the mean.

DEFINITION Standard deviation

The **standard deviation** measures the typical distance of the values in a distribution from the mean. It is calculated by finding an "average" of the squared deviations of the individual data values from the mean, and then taking the square root.

If the values in a quantitative data set are $x_1, x_2, \ldots, x_n$, the *sample standard deviation* s_x is given by the formula

$$s_x = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots+(x_n-\bar{x})^2}{n-1}} = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$$

AP® EXAM TIP
Formula sheet

The formula for the sample standard deviation s_x is included on the formula sheet provided on both sections of the AP® Statistics exam.

The value obtained before taking the square root in the standard deviation calculation is called the *sample variance*, denoted by s_x^2. Because variance is measured in squared units, it is not a very helpful way to describe the variability of a distribution.

HOW TO CALCULATE THE SAMPLE STANDARD DEVIATION s_x

1. Find the mean of the distribution.
2. Calculate the *deviation* of each value from the mean:

$$\text{deviation} = \text{value} - \text{mean}$$

3. Square each of the deviations.
4. Add all of the squared deviations, then divide by $n-1$ to get the sample variance s_x^2.
5. Take the square root to return to the original units.

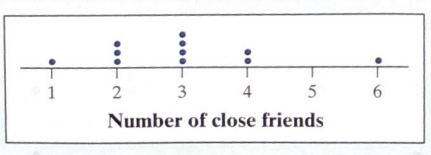
EXAMPLE

How many friends?
Measuring variability: The standard deviation

Skill 2.C

PROBLEM: A random sample of 11 high school students are asked how many "close" friends they have. Here are their responses, along with a dotplot:

1 2 2 2 3 3 3 3 4 4 6





Ian Allenden/Alamy Stock Photo

Calculate the sample standard deviation. Interpret this value.

SOLUTION:

$$\bar{x} = \frac{1+2+2+2+3+3+3+3+4+4+6}{11} = 3$$

1. Find the mean of the distribution.

Value x_i	Deviation from mean $x_i - \bar{x}$	Squared deviation $(x_i - \bar{x})^2$
1	$1-3=-2$	$(-2)^2=4$
2	$2-3=-1$	$(-1)^2=1$
2	$2-3=-1$	$(-1)^2=1$
2	$2-3=-1$	$(-1)^2=1$
3	$3-3=0$	$0^2=0$
3	$3-3=0$	$0^2=0$
3	$3-3=0$	$0^2=0$
3	$3-3=0$	$0^2=0$
4	$4-3=1$	$1^2=1$
4	$4-3=1$	$1^2=1$
6	$6-3=3$	$3^2=9$
		Sum = 18

2. Calculate the *deviation* of each value from the mean: deviation = value − mean.

3. Square each deviation.

$$s_x^2 = \frac{18}{11-1} = 1.80$$

4. Add all the squared deviations and divide by $n-1$. This gives the sample *variance*.

$$s_x = \sqrt{1.80} = 1.34 \text{ close friends}$$

5. Take the square root to return to the original units.

Interpretation: The number of close friends these students have typically varies from the mean by about 1.34 close friends.

FOR PRACTICE, TRY EXERCISE 13

The notation s_x refers to the standard deviation of a *sample*. When we need to refer to the standard deviation of a population, we'll use the symbol σ (lowercase Greek letter sigma). We often use the sample statistic s_x to estimate the population parameter σ. The population standard deviation σ is calculated by dividing the sum of squared deviations from the population mean μ by the population size N (not $N-1$) before taking the square root.

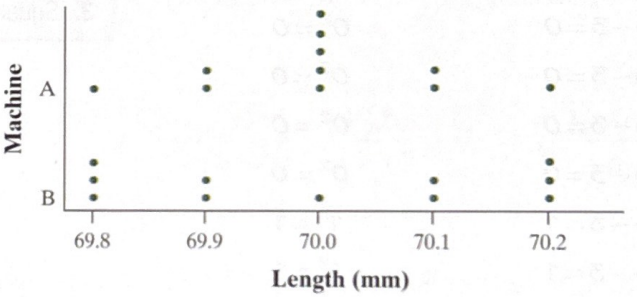
Think About It

WHY IS THE STANDARD DEVIATION CALCULATED IN SUCH A COMPLEX WAY? Add the deviations from the mean in the preceding example. You should get a sum of 0. Why? Because the mean is the balance point of the distribution. We square the deviations to avoid the positive and negative deviations balancing each other out and adding to 0. It might seem strange to "average" the squared deviations by dividing by $n-1$. We'll explain the reason for doing this in Unit 5. It's easier to understand why we take the square root: to return to the original units (close friends).

PROPERTIES OF THE STANDARD DEVIATION

More important than the details of calculating s_x are the properties of the standard deviation as a measure of variability:

- **s_x is always greater than or equal to 0.** $s_x = 0$ only when there is no variability — that is, when all values in a distribution are the same.
- **Greater variation from the mean results in larger values of s_x.** For instance, the widths of 70-mm strips of film produced by Machine A have a standard deviation of 0.110 mm, while the widths of 70-mm strips of film produced by Machine B have a standard deviation of about 0.167 mm. That's about 52% more variability in the widths of film strips produced by Machine B!



- **s_x is not a resistant measure of variability.** The use of squared deviations makes s_x even more sensitive than $\bar{x}$ to extreme values in a distribution. In the preceding example, the distribution of number of close friends has standard deviation $s_x = 1.34$ close friends. If we omit the student with 6 close friends, the standard deviation decreases to $s_x = 0.949$ close friends.
- **s_x measures variation about the mean.** It should be used only when the mean is chosen as the measure of center.

In the preceding example, 11 high school students had an average of $\bar{x} = 3$ close friends with a standard deviation of $s_x = 1.34$ close friends. How would the sample standard deviation be affected if a 12th high school student was added to the sample who had 3 close friends? The mean number of close friends in the sample would still be $\bar{x} = 3$. Because the standard deviation measures the typical distance of the values in a distribution from the mean, s_x would *decrease* because this 12th value is at a distance of 0 from the mean. In fact, the new standard deviation would be

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{18}{12-1}} = 1.28 \text{ close friends}$$

Measuring Variability: The Interquartile Range (*IQR*)

We can avoid the impact of extreme values on our measure of variability by focusing on the middle of the distribution. Here's the basic strategy: Order the data values from

smallest to largest. Then find the **quartiles,** the values that divide the distribution into four groups of roughly equal size. The **first quartile** Q_1 lies one-fourth of the way through the ordered list. The second quartile is the median, which is halfway through the list. The **third quartile** Q_3 lies three-fourths of the way through the list. The first and third quartiles mark out the middle half of the distribution.

DEFINITION Quartiles, First quartile Q_1, Third quartile Q_3

The **quartiles** of a distribution divide an ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values left to right from smallest to largest and find the median.

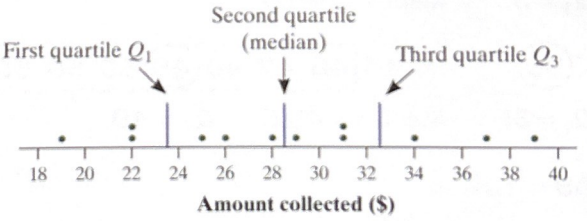
The **first quartile Q_1** is the median of the data values that are to the left of the median in the ordered list.

The **third quartile Q_3** is the median of the data values that are to the right of the median in the ordered list.

For example, here are the amounts collected each hour by a charity at a local store:

$19 $26 $25 $37 $31 $28 $22 $22 $29 $34 $39 $31

The dotplot displays the data. Because there are 12 data values, the quartiles divide the distribution into 4 groups of 3 values each.



The quartiles can be used to help us describe the *position* of an individual data value in a distribution of quantitative data. We'll discuss ways to measure position/location in Section 1E.

The **interquartile range** (*IQR*) measures variability using the quartiles. It's simply the range of the middle half of the distribution.

DEFINITION Interquartile range (*IQR*)

The **interquartile range (*IQR*)** is the distance between the first and third quartiles of a distribution. In symbols:

$$IQR = Q_3 - Q_1$$

For the distribution of amount collected each hour at the charity store, you can check that $Q_1 = \$23.50$, Median $= \$28.50$, and $Q_3 = \$32.50$. The interquartile range of the distribution is

$$IQR = \$32.50 - \$23.50 = \$9.00$$

EXAMPLE More chips please

Measuring variability: The interquartile range (*IQR*)

Skill 2.C

PROBLEM: Here again are the data on the percent air in a sample of 14 bags of chips.

Brand	Percent air	Brand	Percent air
Cape Cod	46	Popchips	45
Cheetos	59	Pringles	28
Doritos	48	Ruffles	50
Fritos	19	Stacy's Pita Chips	50
Kettle Brand	47	Sun Chips	41
Lays	41	Terra	49
Lays Baked	39	Tostitos Scoops	34

Find the interquartile range.

SOLUTION:

19 28 34 39 41 41 (45 | 46) 47 48 49 50 50 59

Median = 45.5

19 28 34 (39) 41 41 45 | 46 47 48 49 50 50 59

$Q_1 = 39$ Median = 45.5

19 28 34 (39) 41 41 45 | 46 47 48 (49) 50 50 59

$Q_1 = 39$ Median = 45.5 $Q_3 = 49$

$IQR = 49 - 39 = 10\%$ air

Sort the data values left to right from smallest to largest and find the median.

Find the first quartile Q_1, which is the median of the data values to the left of the median in the ordered list.

Find the third quartile Q_3, which is the median of the data values to the right of the median in the ordered list.

$IQR = Q_3 - Q_1$

FOR PRACTICE, TRY EXERCISE 19

How should we interpret the interquartile range from the example, $IQR = 10\%$ air? The middle half of the distribution of percent air in these 14 bags of chips has a range of 10%.

The quartiles and the interquartile range are *resistant* because they are not affected by a few extreme values. For the air in chips data, Q_3 would still be 49 and the *IQR* would still be 10 if the maximum were 69 rather than 59.

Be sure to leave out the median when you locate the quartiles. In the preceding example, the median was not one of the data values. For the earlier Close Friends example data set, we ignore the circled median of 3 when finding Q_1 and Q_3.

1 2 (2) 2 3 | (3) | 3 3 (4) 4 6

Q_1 Median Q_3

Choosing Summary Statistics

Which measure of center and variability should we use to summarize the distribution of a quantitative variable? That depends on the shape of the distribution

and whether there are any outliers. As you learned earlier, the median is a resistant measure of center but the mean is not. Among measures of variability, the interquartile range (*IQR*) is resistant to extreme values but the range and standard deviation are not. For now, follow this advice:

- If a distribution of quantitative data is roughly symmetric with no outliers, the mean $\bar{x}$ and standard deviation s_x are the preferred measures of center and variability.
- If the distribution is clearly skewed or has outliers, use the median as the measure of center and the *IQR* as the measure of variability.

CHOOSING MEASURES OF CENTER AND VARIABILITY

The median and *IQR* are usually better choices than the mean and standard deviation for describing a skewed distribution or a distribution with outliers. Use $\bar{x}$ and s_x for roughly symmetric distributions that don't have outliers.

We recommend using the range to measure variability only as a last resort because it gives so little information about how the individual data values are distributed.

EXAMPLE

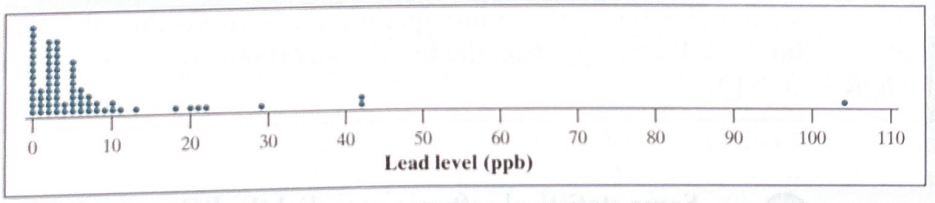
Lead in the water
Choosing summary statistics

Skill 4.B



Jim West/Alamy Stock Photo

PROBLEM: Here once again is a dotplot of the lead levels (in parts per billion, ppb) in 71 water samples taken from randomly selected Flint, Michigan, dwellings after the city switched its water supply from Lake Huron to the Flint River. Summary statistics for the data set are also provided.[69]



n	Mean	SD	Min	Q_1	Med	Q_3	Max
71	7.31	14.347	0	2	3	7	104

Which measures of center and variability should we choose to summarize this distribution? Explain your answer.

SOLUTION:

The distribution of lead level is right-skewed and has an obvious outlier at 104 ppb. Consequently, we should choose measures of center and variability that are resistant: the median of 3 ppb and the interquartile range of $IQR = 7 - 2 = 5$ ppb.

FOR PRACTICE, TRY EXERCISE 23

For other statistical reasons, the mean and standard deviation will become the preferred measures of center and variability for the distribution of a quantitative variable in future sections. You'll see why when we get there!

Graphing calculators and computer software will calculate summary statistics for you. Using technology to perform calculations will allow you to focus on choosing the right methods and interpreting your results.

2. Tech Corner CALCULATING SUMMARY STATISTICS

TI-Nspire and other technology instructions are on the book's website at bfwpub.com/tps7e.

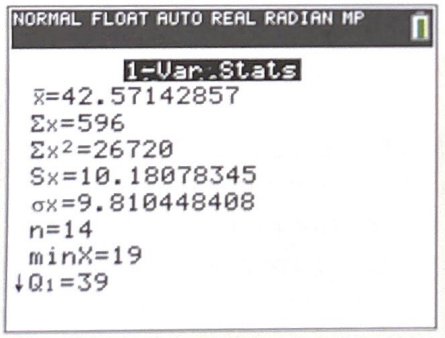
You can use technology to calculate measures of center and variability for a distribution of quantitative data. We'll illustrate using data on the percent air in a sample of 14 bags of chips from previous examples.

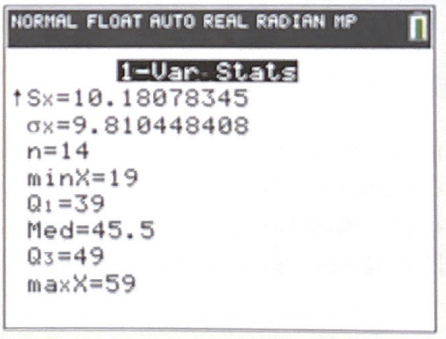
1. Type the values into list L1.
2. Press STAT, arrow over to the CALC menu, and choose 1-Var Stats.

   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 the down arrow to see the rest of the one-variable statistics.





Note: The TI-83/84 does not give the range or interquartile range (*IQR*) directly when you calculate one-variable statistics. But the calculator provides the five-number summary, so you can calculate range = max − min and $IQR = Q_3 - Q_1$.

Some statistical software uses slightly different rules for calculating quartiles. We used Minitab statistical software to calculate summary statistics for the percent air in chips data. In the output, Minitab reports that $Q_1 = 37.75$ and $Q_3 = 49.25$. The TI-83/84 gives the first and third quartiles as $Q_1 = 39$ and $Q_3 = 49$, which match the values we calculated in an earlier example. Results from the various rules are usually close to each other, especially for large data sets.

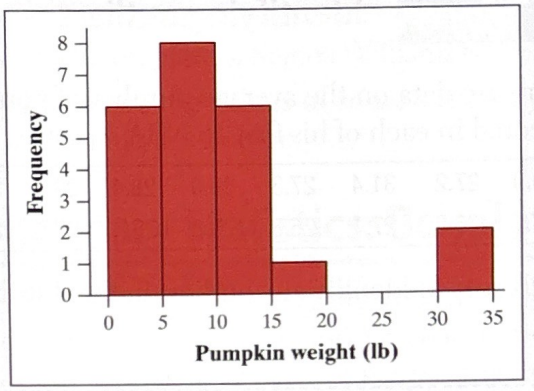
Descriptive Statistics

Variable	N	Mean	StDev	Minimum	Q_1	Median	Q_3	Maximum
Percent air	14	42.57	10.18	19.00	37.75	45.50	49.25	59.00

CHECK YOUR UNDERSTANDING

Some students purchased pumpkins for a carving contest. Before the contest began, they weighed the pumpkins. The weights in pounds are shown here, along with a histogram of the data.

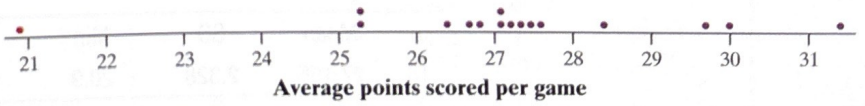
3.6	4.0	9.6	14.0	11.0	12.4	13.0	2.0	6.0	6.6	15.0	3.4
12.7	6.0	2.8	9.6	4.0	6.1	5.4	11.9	5.4	31.0	33.0	



1. Explain why you cannot calculate the range exactly from the histogram. Then use the data to calculate the range of the distribution.
2. The mean and standard deviation of the distribution are 9.93 lb and 8.01 lb, respectively. Interpret the standard deviation.
3. Calculate the interquartile range of the distribution.
4. Which measures of center and variability would you choose to describe the distribution? Explain your answer.

Identifying Outliers

LeBron James emerged as a superstar in the National Basketball Association (NBA) during his rookie season (2003–2004). He maintained a consistent level of excellence over the first 16 years of his professional career, reaching the NBA Finals eight consecutive times and winning three NBA championships. The dotplot shows the average number of points per game that LeBron scored in each of these 16 seasons.[70]



LeBron's 20.9 points per game average in his rookie season stands out (in red) from the rest of the distribution. Should this value be classified as an *outlier*?

The most common method for identifying outliers in a distribution of quantitative data uses the interquartile range (*IQR*). Besides being a resistant measure of variability, the *IQR* serves as a kind of "ruler" for determining how extreme an individual data value must be to be classified as an outlier.

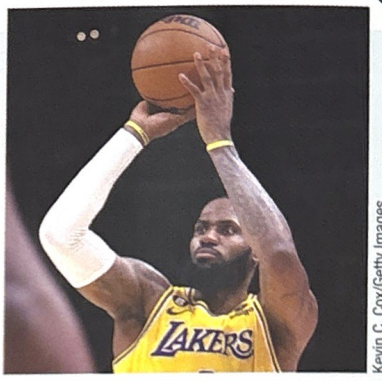
HOW TO IDENTIFY OUTLIERS: THE 1.5 × *IQR* RULE

Call an observation an outlier if it falls more than $1.5 \times IQR$ above the third quartile or more than $1.5 \times IQR$ below the first quartile. That is,

Low outliers $< Q_1 - 1.5 \times IQR$ High outliers $> Q_3 + 1.5 \times IQR$

EXAMPLE An NBA legend Identifying outliers

Skill 4.B



Kevin C. Cox/Getty Images

PROBLEM: Here are data on the average number of points per game that LeBron James scored in each of his first 16 NBA seasons:

20.9	27.2	31.4	27.3	30.0	28.4	29.7	26.7
27.1	26.8	27.1	25.3	25.3	26.4	27.5	27.4

Use the $1.5 \times IQR$ rule to identify any outliers in the distribution.

SOLUTION:

20.9 25.3 25.3 26.4 | 26.7 26.8 27.1 27.1 | 27.2 27.3 27.4 27.5 | 28.4 29.7 30.0 31.4

$Q_1 = 26.55$ Median $= 27.15$ $Q_3 = 27.95$

$IQR = 27.95 - 26.55 = 1.40$

Low outliers $< 26.55 - 1.5 \times 1.40 = 24.45$
High outliers $> 27.95 + 1.5 \times 1.40 = 30.05$

- Find the interquartile range: $IQR = Q_3 - Q_1$.
- Then calculate the upper and lower cutoff values for outliers:
  Low outliers $< Q_1 - 1.5 \times IQR$
  High outliers $> Q_3 + 1.5 \times IQR$

LeBron James's rookie-season average of 20.9 points per game is a low outlier because it is less than 24.45. The season when LeBron averaged 31.4 points per game is a high outlier because it is greater than 30.55.

FOR PRACTICE, TRY EXERCISE 27

There are other methods for determining outliers, such as "any value that is more than 2 standard deviations from the mean." Let's apply this $2 \times SD$ *rule* to the data from the preceding example. Here are summary statistics on the average number of points scored by LeBron James in each of his first 16 NBA seasons:

n	Mean	SD	Min	Q_1	Med	Q_3	Max
16	27.156	2.328	20.9	26.55	27.15	27.95	31.4

Low outliers $<$ Mean $- 2 \times$ SD $= 27.156 - 2 \times 2.328 = 22.50$

High outliers $>$ Mean $+ 2 \times$ SD $= 27.156 + 2 \times 2.328 = 31.812$

By the $2 \times$ SD rule, the season in which LeBron James averaged 20.9 points per game is an outlier because it is less than 22.50. However, the season in which LeBron averaged 31.4 points per game is *not* an outlier by this rule because 31.4 is not greater than the upper cutoff of 31.812.

Unless otherwise indicated, we always use the $1.5 \times IQR$ rule in this book to identify outliers because it is based on statistics that are resistant to extreme data values,

AP® EXAM TIP

You may be asked to determine whether a quantitative data set has any outliers. Be prepared to state and use the $1.5 \times IQR$ rule and the $2 \times$ SD rule for identifying outliers.

unlike the mean and standard deviation. As we will explain in Section 1F, there is some justification for using the $2 \times$ SD rule in the special case of roughly symmetric, single-peaked, mound-shaped distributions called *normal distributions*.

It is important to identify outliers in a distribution for several reasons:

1. **They might be inaccurate data values.** Maybe someone recorded a value as 10.1 instead of 101. Perhaps a measuring device broke down. Or maybe someone gave a silly response, like the student in a class survey who claimed to study 30,000 minutes per night! Try to correct errors like these if possible. If you can't, give summary statistics with and without the outlier.
2. **They can indicate a remarkable occurrence.** For example, in a graph of career earnings of professional tennis players, Serena Williams is likely to be an outlier.
3. **They can heavily influence the values of some summary statistics,** such as the mean, range, and standard deviation.

Displaying Summary Statistics: Boxplots

You can use a dotplot, stemplot, or histogram to display the distribution of a quantitative variable. Another graphical option for quantitative data is a **boxplot** (sometimes called a *box-and-whisker* plot). A boxplot summarizes a distribution by displaying the location of five important values within the distribution, known as its **five-number summary.**

DEFINITION Five-number summary, Boxplot

The **five-number summary** of a distribution of quantitative data consists of the minimum, the first quartile Q_1, the median, the third quartile Q_3, and the maximum.

A **boxplot** is a visual representation of the five-number summary.

Figure 1.12 illustrates the process of making a boxplot. The dotplot in Figure 1.12(a) shows LeBron James's average points per game for each of 16 seasons.

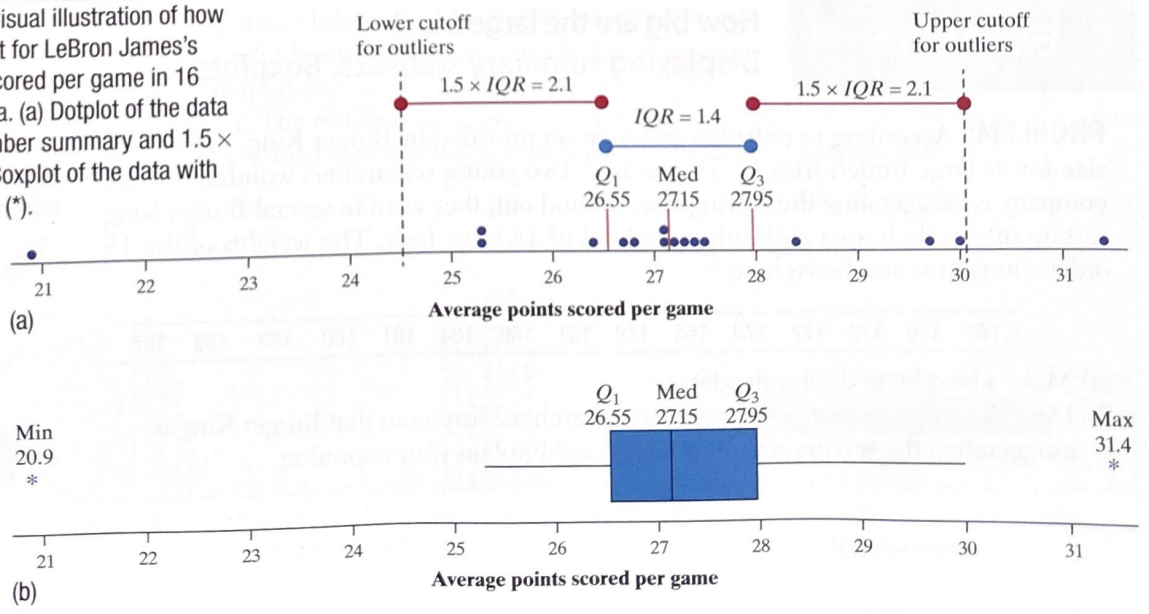


FIGURE 1.12 A visual illustration of how to make a boxplot for LeBron James's average points scored per game in 16 NBA seasons data. (a) Dotplot of the data with the five-number summary and $1.5 \times IQR$ marked. (b) Boxplot of the data with outliers identified (*).

We have labeled the first quartile, the median, and the third quartile and marked them with orange line segments. The process of testing for outliers with the $1.5 \times IQR$ rule is illustrated above the dotplot. Because the values of 20.9 and 31.4 are outliers, we mark these separately. To get the finished boxplot in Figure 1.12(b), we make a box spanning from Q_1 to Q_3 and then draw whiskers to the smallest and largest data values that are not outliers.

As you can see, it is fairly easy to make a boxplot by hand for small sets of quantitative data. Here's a summary of the steps.

HOW TO MAKE A BOXPLOT

1. **Find the five-number summary** for the distribution.
2. **Identify outliers** using the $1.5 \times IQR$ rule.
3. **Draw and label the axis.** Draw a horizontal axis and put the name of the quantitative variable underneath it, including units if applicable.
4. **Scale the axis.** Look at the minimum and maximum values in the data set. Start the horizontal axis at a convenient number less than or equal to the minimum and place tick marks at equal intervals until you equal or exceed the maximum.
5. **Draw a box** that spans from the first quartile (Q_1) to the third quartile (Q_3).
6. **Mark the median** with a vertical line segment that's the same height as the box.
7. **Mark any outliers** with a special symbol such as an asterisk (*).
8. **Draw whiskers** — lines that extend from the ends of the box to the smallest and largest data values that are *not* outliers.

We can see from the boxplot in Figure 1.12 that the middle 50% of LeBron James's season scoring averages fall between 26.55 and 27.95 points per game.

EXAMPLE

How big are the large fries?

Skills 2.A, 2.B

Displaying summary statistics: Boxplots



Michael Neelon/Alamy Stock Photo

PROBLEM: According to nutrition information provided by Burger King, the serving size for its large french fries is 173 grams.[71] Two young researchers wondered if the company is exaggerating the serving size. To find out, they went to several Burger King restaurants in their area and ordered a total of 15 large fries. The weights of the 15 orders (in grams) are shown here.[72]

178	176	173	172	179	165	179	181	186	184	181	180	183	183	187

(a) Make a boxplot to display the data.

(b) Does the graph in part (a) support the researchers' suspicion that Burger King is exaggerating the serving size of its large fries? Explain your reasoning.

SOLUTION:

(a) 165 172 173 (176) 178 179 179 (180) 181 181 183 (183) 184 186 187

Min Q_1 Med Q_3 Max

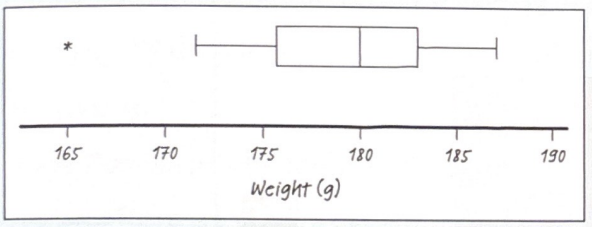
$IQR = Q_3 - Q_1 = 183 - 176 = 7$

Low outliers $< Q_1 - 1.5 \times IQR = 176 - 1.5 \times 7 = 165.5$

High outliers $> Q_3 + 1.5 \times IQR = 183 + 1.5 \times 7 = 193.5$

The order of large fries that weighed 165 grams is an outlier.

1. Find the five-number summary.
2. Identify outliers.
3. Draw and label the axis.
4. Scale the axis.
5. Draw a box from Q_1 to Q_3.
6. Mark the median.
7. Mark any outliers.
8. Draw whiskers.

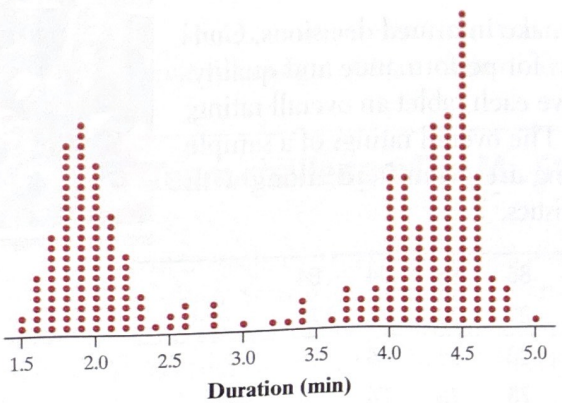


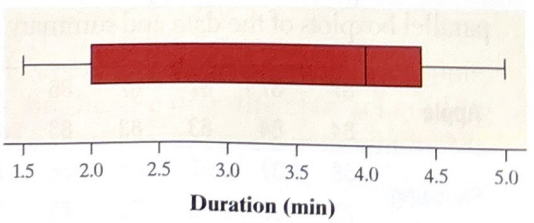
(b) No. From the boxplot, $Q_1 = 176$, so at least 75% of the orders of large fries that the researchers bought from local Burger King restaurants weighed 176 grams or more. Only the outlier (165 grams) and one other order (172 grams) of large fries weighed less than the advertised 173-gram serving size.

FOR PRACTICE, TRY EXERCISE 31

Boxplots provide a quick summary of the center and variability of a distribution. The median is displayed as a vertical line segment in the central box, the interquartile range is the length of the box, and the range is the length of the entire plot, including outliers. *Note that some statistical software orients boxplots vertically.*

Boxplots don't give a complete picture of the shape of a distribution because they do not display each individual data value. From the boxplot in the example, we can see that the distribution of weight for orders of Burger King's large fries is skewed to the left because the left half of the distribution varies from 165 to 180 grams, while the right half of the distribution varies from 180 to 187 grams. But we can't identify peaks, gaps, or clusters from a boxplot. For instance, the following dotplot displays the duration, in minutes, of 263 eruptions of the Old Faithful geyser. The distribution of eruption durations is clearly bimodal (two-peaked). However, the boxplot of the data hides this important information about shape.



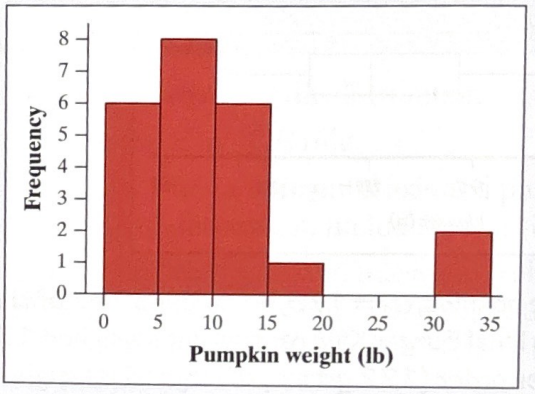




CHECK YOUR UNDERSTANDING

Some students purchased pumpkins for a carving contest. Before the contest began, they weighed the pumpkins. The weights in pounds are shown here, along with a histogram of the data.

3.6	4.0	9.6	14.0	11.0	12.4	13.0	2.0	6.0	6.6	15.0	3.4
12.7	6.0	2.8	9.6	4.0	6.1	5.4	11.9	5.4	31.0	33.0	



1. Identify any outliers in the distribution.
2. Make a boxplot to display the data.
3. Explain why the boxplot does not completely display the shape of the distribution.

Comparing Distributions with Boxplots and Summary Statistics

Boxplots are especially effective for comparing the distribution of a quantitative variable in two or more groups. Just remember to discuss shape, outliers, center, and variability as you did with comparative dotplots, stemplots, and histograms in Section 1C.

EXAMPLE

Which company makes better tablets?
Comparing distributions with boxplots and summary statistics

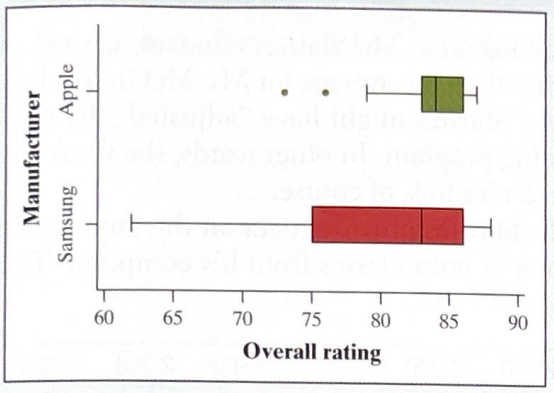
Skill 2.D



Scott Olson/Getty Images

PROBLEM: To help potential purchasers make informed decisions, *Consumer Reports* rated many tablet computers for performance and quality. Based on several variables, the magazine gave each tablet an overall rating, where higher scores indicate better ratings. The overall ratings of a sample of tablets produced by Apple and Samsung are given here, along with parallel boxplots of the data and summary statistics.[73]

Apple	87	87	87	87	86	86	86	86	84	84		
	84	84	83	83	83	83	81	79	76	73		
Samsung	88	87	87	86	86	86	86	86	84	84	83	83
	77	76	76	75	75	75	75	75	74	71	62	



	$\bar{x}$	s_x	Min	Q_1	Median	Q_3	Max	*IQR*
Apple	83.45	3.762	73	83	84	86	87	3
Samsung	79.87	6.74	62	75	83	86	88	11

Compare the distributions of overall rating for Apple and Samsung.

SOLUTION:

Shape: Both distributions of overall rating are skewed to the left.

Outliers: There are two low outliers in the Apple tablet distribution with overall ratings of 73 and 76. The Samsung tablet distribution has no outliers.

Center: The Apple tablets had a slightly higher median overall rating (84) than the Samsung tablets (83). More importantly, about 75% of the Apple tablets had overall ratings that were greater than or equal to the median for the Samsung tablets.

Variability: There is much more variation in overall rating among the Samsung tablets than among the Apple tablets. The *IQR* for Samsung tablets (11) is almost four times larger than the *IQR* for Apple tablets (3).

Be sure to discuss shape, outliers, center, and variability, and to include context (variable name)!

Don't forget to explicitly *compare* measures of center and variability, using words like "greater than," "less than," or "about the same as." Due to the strong skewness and outliers, use the median and *IQR* instead of the mean and standard deviation when comparing center and variability.

FOR PRACTICE, TRY EXERCISE 35

AP® EXAM TIP

Use statistical terms carefully and correctly on the AP® Statistics exam. Don't say "mean" if you really mean "median." Range is a single number; so are Q_1, Q_3, and *IQR*. Avoid sloppy use of language, like "the outlier *skews* the mean" or "the median is in the middle of the *IQR*." Skewed is a shape and the *IQR* is a single number, not a region. If you misuse a term, expect to lose some credit.

Here's an activity that gives you a chance to use the skills you have learned in this unit: selecting statistical methods, data analysis, and making statistical arguments.

ACTIVITY Team challenge: Did Mr. Starnes stack his class?

In this activity, you will work in a team of three or four students to resolve a dispute.

Mr. Starnes teaches AP® Statistics, but he also does the class scheduling for the high school. There are two AP® Statistics classes — one taught by Mr. Starnes and one taught by Ms. McGrail. The two teachers give the same first

test to their classes and grade the test together. Mr. Starnes's students earned an average score that was 8 points higher than the average for Ms. McGrail's class. Ms. McGrail wonders whether Mr. Starnes might have "adjusted" the class rosters from the computer scheduling program. In other words, she thinks he might have "stacked" his class. He denies this, of course.

To help resolve the dispute, Mr. Starnes provides data on the cumulative grade point averages of the students in both classes from his computer. The following table displays the data.

McGrail	2.900	3.300	3.980	2.900	3.200	3.500	2.800	2.900	3.950
	3.100	2.850	2.900	3.245	3.000	3.000	2.800	2.900	3.200
Starnes	2.900	2.860	2.600	3.600	3.200	2.700	3.100	3.085	3.750
	3.400	3.338	3.560	3.800	3.200	3.100			

Based on these data, did Mr. Starnes stack his class? Give appropriate graphical and numerical evidence to support your conclusion.

You can use technology to make boxplots, as the following Tech Corner illustrates.

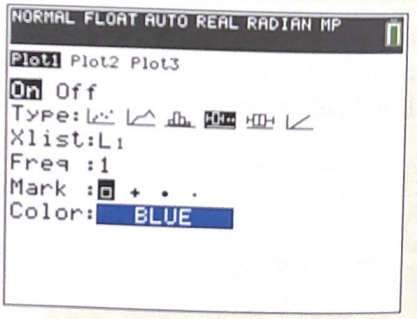
3. Tech Corner MAKING BOXPLOTS

TI-Nspire and other technology instructions are on the book's website at bfwpub.com/tps7e.

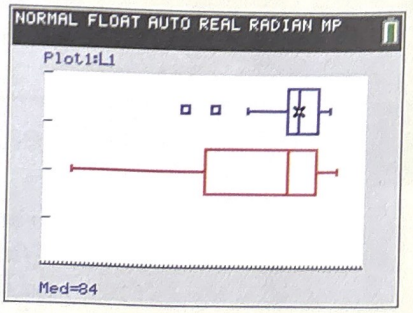
The TI-83/84 can plot up to three boxplots in the same viewing window. Let's use the calculator to make parallel boxplots of the overall rating data for Apple and Samsung tablets.

1. Enter the ratings for the Apple tablets in list L1 and those for the Samsung tablets in list L2.
2. Set up two statistics plots: Plot1 to show a boxplot of the Apple data and Plot2 to show a boxplot of the Samsung data. The setup for Plot1 is shown. When you define Plot2, be sure to change L1 to L2.

   Note: The calculator offers two types of boxplots: one that shows outliers and one that doesn't. We'll always use the type that identifies outliers.



3. Press ZOOM and select ZoomStat to display the parallel boxplots. Then press TRACE and use the arrow keys to view the five-number summary.



SECTION 1D Summary

- A numerical summary of a distribution of quantitative data should include measures of *center* and *variability*.
- The **mean** and the **median** measure the center of a distribution in different ways. The median is the midpoint of the distribution, the number such that about half the observations are smaller and half are larger. The mean is the average of the observations. In symbols, the sample mean is given by $\bar{x} = \frac{\sum x_i}{n}$.
- A **statistic** is a number that describes a sample. A **parameter** is a number that describes a population. We often use statistics (like the sample mean $\bar{x}$) to estimate parameters (like the population mean μ).
- The simplest measure of variability for a distribution of quantitative data is the **range**, which is the distance from the minimum value to the maximum value.
- When you use the mean to describe the center of a distribution, use the **standard deviation** to measure variability. The standard deviation gives the typical distance of the values in a distribution from the mean. In symbols, the sample standard deviation is given by $s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$. The value obtained before taking the square root is known as the *sample variance*, denoted by s_x^2. The standard deviation s_x is 0 when there is no variability and gets larger as variability from the mean increases.
- When you use the median to describe the center of a distribution, use the **interquartile range** (*IQR*) to describe the distribution's variability. The **first quartile** Q_1 has about one-fourth of the individual data values at or below it, and the **third quartile** Q_3 has about three-fourths of the individual data values at or below it. The interquartile range measures variability in the middle half of the distribution and is found by calculating $IQR = Q_3 - Q_1$.
- The median is a **resistant** measure of center because it is relatively unaffected by extreme values. The mean is not resistant. Among measures of variability, the *IQR* is resistant, but the range and standard deviation are not.
- The mean and standard deviation are good descriptions for roughly symmetric distributions with no outliers. The median and *IQR* are a better description for skewed distributions or distributions with outliers.
- The most common method of identifying outliers in a distribution of quantitative data is the **1.5 × *IQR* rule**. According to this rule, an individual data value is an outlier if it is less than $Q_1 - 1.5 \times IQR$ or greater than $Q_3 + 1.5 \times IQR$. Another method for identifying outliers is the 2 × SD rule, which says that any data value more than 2 standard deviations from the mean of the distribution is an outlier.
- The **five-number summary** of a distribution consists of the minimum, Q_1, the median, Q_3, and the maximum. **A boxplot** displays the five-number summary, marking outliers with a special symbol. The box shows the variability in the middle half of the distribution. The median is marked within the box. Lines extend from the box to the smallest and largest observations that are not outliers. Boxplots are helpful for comparing the center (median) and variability (range, *IQR*) of multiple distributions. Boxplots aren't as useful for identifying the shape of a distribution because they do not display peaks, clusters, gaps, and other interesting features.

AP® EXAM TIP
AP® Daily Videos

Review the content of this section and get extra help by watching the AP® Daily Videos for Topics 1.7–1.9, which are available in AP® Classroom.

1D Tech Corners

TI-Nspire and other technology instructions are on the book's website at bfwpub.com/tps7e.

2. Calculating summary statistics Page 62
3. Making boxplots Page 70

SECTION 1D Exercises

Measuring Center: The Median

1. pg 49 **Soccer stars** How good was the 2019 U.S. women's soccer team? With players like Carli Lloyd, Alex Morgan, and Megan Rapinoe, the team put on an impressive showing en route to winning the 2019 Women's World Cup. Here are data on the number of goals scored by the team in 24 games played in the 2019 season:[74]

1	1	2	2	1	5	6	3	5	3	13	3
2	2	2	2	2	3	4	3	2	1	3	6

Find the median.

2. **How fuel efficient?** The EPA is in charge of determining and reporting fuel economy ratings for cars. Here are the EPA estimates of highway gas mileage in miles per gallon (mpg) for a sample of 21 model-year 2022 midsize cars:[75]

25	30	27	31	38	26	28	40	25	28	30
31	30	30	34	30	31	31	32	48	31	

Find the median.

Measuring Center: The Mean

3. pg 51 **Goooaaaalll!** Refer to Exercise 1.

(a) Calculate the mean number of goals scored by the U.S. women's soccer team in the 24 games it played during the 2019 season.

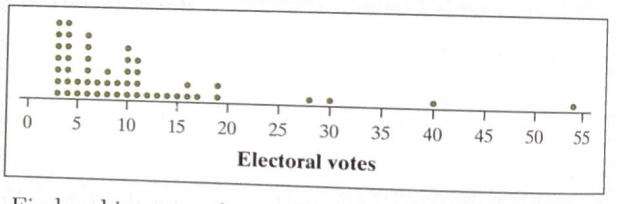
(b) The one game when the team scored 13 goals is a possible outlier in the distribution. Recalculate the mean number of goals scored by the team in the remaining 23 games. What do you notice?

4. **Average fuel efficiency** Refer to Exercise 2.

(a) Calculate the mean fuel efficiency in the sample of 21 model-year 2020 midsize cars.

(b) The Toyota Prius, with its 48 mpg fuel efficiency, is a possible outlier in the distribution. Recalculate the mean fuel efficiency for the other 20 car models. What do you notice?

5. **Electing the president** To become president of the United States, a candidate does not have to receive a majority of the popular vote, but that candidate does have to win a majority of the 538 votes that are cast in the Electoral College. Here is a dotplot of the number of electoral votes in 2024 for each of the 50 states and the District of Columbia:

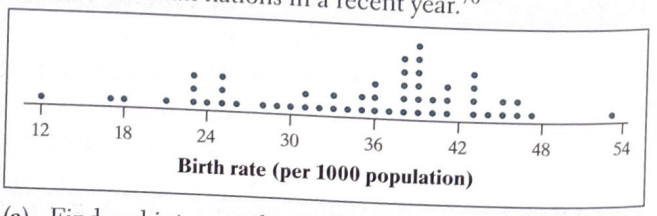


(a) Find and interpret the median.

(b) Without doing any calculations, explain how the mean and the median of this distribution compare.

(c) Is the value you found in part (a) a statistic or a parameter? Justify your answer.

6. **Birth rates in Africa** One of the important factors in determining population growth rates is the birth rate in a country. The dotplot shows the birth rates per 1000 individuals (rounded to the nearest whole number) for all 54 African nations in a recent year.[76]



(a) Find and interpret the median.

(b) Without doing any calculations, explain how the mean and the median compare.

(c) Is the value you found in part (a) a statistic or a parameter? Justify your answer.

7. **College loans** In 2021, college student loan debt in the United States reached $1.71 trillion. At the time, there were 45.3 million U.S. borrowers with student loan debt. More than 3 million had loan debts greater than $100,000; about 900,000 had loan debts exceeding $200,000.[77]

(a) Find the mean amount of student loan debt among U.S. borrowers in 2021.

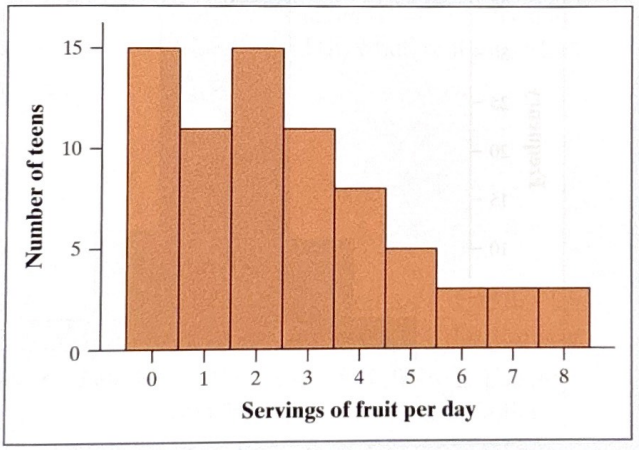
(b) Do you think the median amount of student loan debt was greater than or less than the mean amount? Justify your answer.

8. **House prices** The mean and median selling prices of existing single-family homes sold in the United States in the first 3 months of 2022 were $428,700 and $507,800.[78]

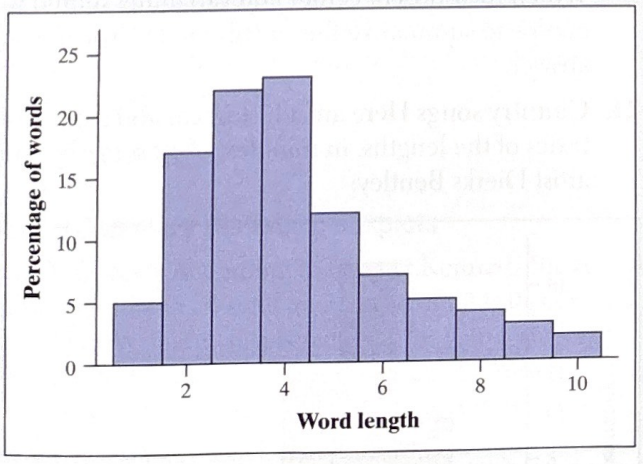
(a) Which of these numbers is the mean and which is the median? Explain your reasoning.

(b) What shape would you expect the distribution of home prices to have? Justify your answer.

9. **Teens and fruit** We all know that fruit is good for us. Here is a histogram of the number of servings of fruit per day that 74 seventeen-year-olds claimed they consumed in a study in Pennsylvania.[79] Find the mean and the median. Show your method clearly.



10. **Shakespeare** The following histogram shows the distribution of lengths of words used in Shakespeare's plays.[80] Find the mean and the median. Show your method clearly.



Measuring Variability: The Range

11. **Soccer range** Refer to Exercise 1.

pg 55

(a) Find the range of the distribution of goals scored by the U.S. women's soccer team in all 24 games played that season.

(b) The one game when the team scored 13 goals is a possible outlier in the distribution. Recalculate the range excluding this game. Explain why the result is so different than the answer to part (a).

12. **Mileage range** Refer to Exercise 2.

(a) Find the range of the distribution of highway gas mileage for the sample of 21 model-year 2022 midsize cars.

(b) The Toyota Prius, with its 48 mpg fuel efficiency, is a possible outlier in the distribution. Recalculate the range excluding this car model. Explain why the result is so different than the answer to part (a).

Measuring Variability: The Standard Deviation

13. **Foot length** Here are the foot lengths (in centimeters) for a random sample of seven 14-year-olds from the United Kingdom:

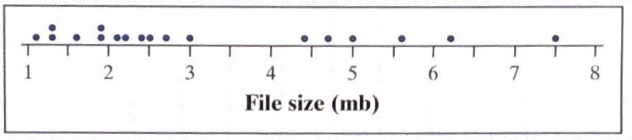
pg 56

25	22	20	25	24	24	28

Calculate the sample standard deviation. Interpret this value.

14. **Well rested?** A random sample of 6 students in a first-period statistics class was asked how much sleep (to the nearest hour) they got last night. Their responses were 6, 7, 7, 8, 10, and 10. Calculate the sample standard deviation. Interpret this value.

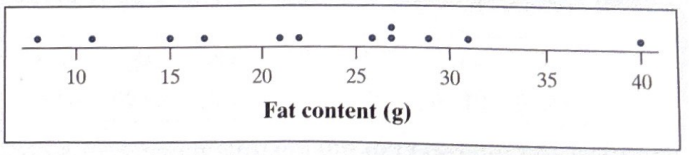
15. **Digital photos** How much storage space do digital photos use? Here is a dotplot of the file sizes (to the nearest tenth of a megabyte) for 18 randomly selected photos in Noemi's cloud storage:[81]



(a) The distribution of file size has a mean of $\bar{x} = 3.2$ megabytes and a standard deviation of $s_x = 1.9$ megabytes. Interpret the standard deviation.

(b) Suppose the music file that takes up 7.5 megabytes of storage space is replaced with another version of the file that takes up only 4 megabytes. How would this affect the mean and the standard deviation? Justify your answer.

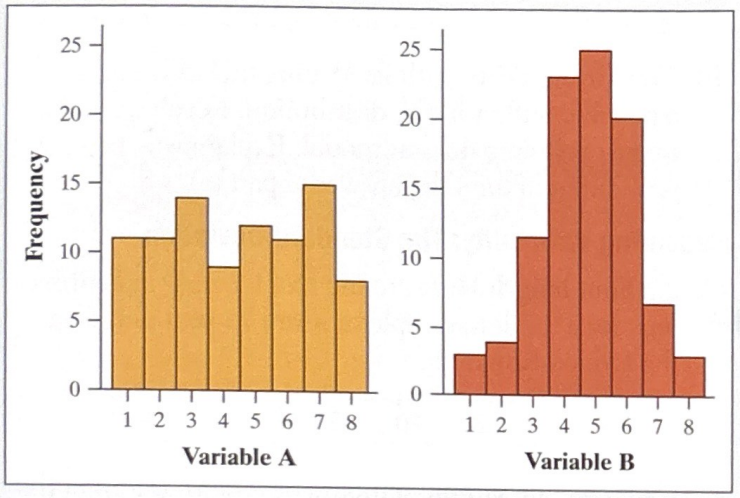
16. **Healthy fast food?** Here is a dotplot of the amount of fat (to the nearest gram) in 12 different hamburgers served at a fast-food restaurant:



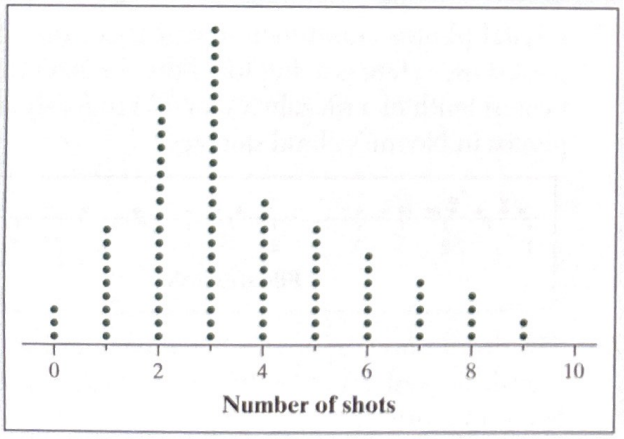
(a) The distribution of fat content has a mean of $\bar{x} = 22.83$ grams and a standard deviation of $s_x = 9.06$ grams. Interpret the standard deviation.

(b) Suppose the restaurant replaces the burger that has 22 grams of fat with a new burger that has 35 grams of fat. How would this affect the mean and the standard deviation? Justify your answer.

17. **Comparing SD** Which of the following distributions has a smaller standard deviation? Justify your answer.



18. **Estimating SD** The dotplot shows the number of shots per game taken by National Hockey League (NHL) player Sidney Crosby in 81 regular season games in a recent season.[82] Is the standard deviation of this distribution closest to 2, 5, or 10? Explain your answer without doing any calculations.



Measuring Variability: The Interquartile Range (*IQR*)

19. **Soccer *IQR*** Refer to Exercise 1. Find the interquartile range of the distribution.

pg 60

20. **Fuel efficiency *IQR*** Refer to Exercise 2. Find the interquartile range of the distribution.

21. **Shopping spree** The table displays summary statistics for data on the amount (in dollars) spent by a sample of 50 grocery shoppers.

$\bar{x}$	s_x	Min	Q_1	Med	Q_3	Max
34.70	21.70	3.11	19.27	27.86	45.40	93.34

(a) Find and interpret the interquartile range.

(b) What would you guess is the shape of the distribution based only on the summary statistics? Explain your reasoning.

22. **C-sections** A study in Switzerland examined the number of cesarean sections (surgical deliveries of babies) performed in a year by a sample of doctors. Here are summary statistics for the distribution:

$\bar{x}$	s_x	Min	Q_1	Med	Q_3	Max
19.1	10.126	5	10	18.5	29	33

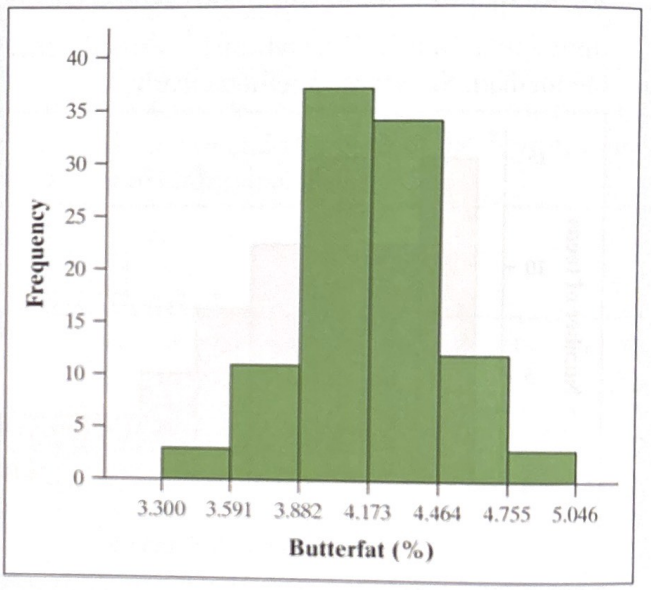
(a) Find and interpret the interquartile range.

(b) What would you guess is the shape of the distribution based only on the summary statistics? Explain your reasoning.

Choosing Summary Statistics

23. **Cow's milk** Here are a histogram and summary statistics of data on the percentage of butterfat in milk from a random sample of 100 three-year-old Ayershire cows:[83]

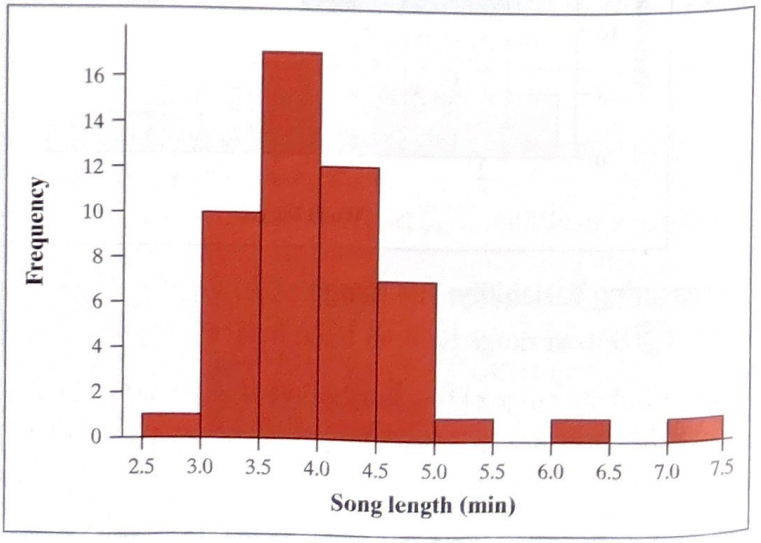
pg 61



n	Mean	SD	Min	Q_1	Med	Q_3	Max
100	4.173	0.291	3.52	3.97	4.17	4.37	4.91

Which measures of center and variability should we choose to summarize this distribution? Explain your answer.

24. **Country songs** Here are a histogram and summary statistics of the lengths, in minutes, of 50 songs by country artist Dierks Bentley:



n	Mean	SD	Min	Q_1	Med	Q_3	Max
50	3.998	0.739	2.9	3.5	3.85	4.4	7

Which measures of center and variability should we choose to summarize this distribution? Explain your answer.

25. SD contest This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 10, with repeats allowed.

(a) Choose four numbers that have the smallest possible standard deviation.

(b) Choose four numbers that have the largest possible standard deviation.

(c) Is more than one choice possible in either part (a) or (b)? Explain your answer.

26. What do they measure? For each of the following summary statistics, decide (i) whether it could be used to measure center or variability and (ii) whether it is resistant.

(a) $\dfrac{Q_1 + Q_3}{2}$

(b) $\dfrac{\text{Max} - \text{Min}}{2}$

Identifying Outliers

27. Soccer outliers Refer to Exercise 1. Use the $1.5 \times IQR$ rule to identify any outliers in the distribution.
pg 64

28. Fuel efficiency outliers Refer to Exercise 2. Use the $1.5 \times IQR$ rule to identify any outliers in the distribution.

29. Cow's milk outliers Refer to Exercise 23. Determine whether the distribution has low or high outliers using

(a) The $1.5 \times IQR$ rule

(b) The $2 \times$ SD rule

30. Country music outliers Refer to Exercise 24. Determine whether the distribution has low or high outliers using

(a) The $1.5 \times IQR$ rule

(b) The $2 \times$ SD rule

Displaying Summary Statistics: Boxplots

31. Daily texts According to blogger Kenneth Burke, adults ages 18 to 24 send an average of 64 texts per day.[84] Dr. Williams suspected that this value was exaggerated, and collected data from a class of Introductory Statistics students on the number of texts they had sent in the past 24 hours. Here are the data:
pg 66

0	7	1	29	25	8	5	1	25	98	9	0
268	118	72	0	92	52	14	3	3	44	5	42

(a) Make a boxplot to display the data.

(b) Based on the boxplot in part (a), what percentage of Dr. Williams's students sent more than 64 texts in the previous day: less than 25%, between 25% and 50%, between 50% and 75%, or more than 75%? Do the data confirm Dr. Williams's suspicion? Explain your answer.

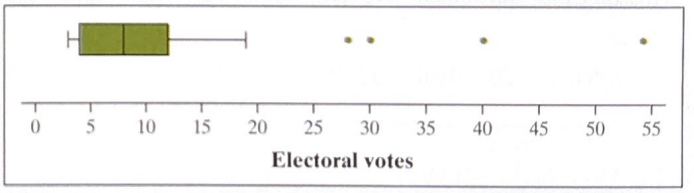
32. Copper mining In March 2021, a Canadian company found substantial deposits of copper at shallow depths in Arizona's Copper World region, which previously yielded about 440,000 tons of high-quality copper from 1874 to 1969. The company drilled test holes in several randomly selected locations at each of the old mine sites, and measured the amount of copper (as a percentage) in the rock extracted from each hole. Here are the data from the 28 test holes drilled at the Broad Top Butte site:[85]

0.00	0.75	1.43	0.43	0.19	0.52	0.00
0.00	0.20	1.38	0.00	0.28	0.71	0.30
0.30	0.44	0.59	0.91	0.33	0.39	0.00
0.70	0.37	0.38	0.67	0.00	0.38	0.30

(a) Make a boxplot to display the data.

(b) An economically viable copper mine has about 0.6% copper in its rock. Based on the boxplot in part (a), what percentage of the rock at Broad Top Butte should the company estimate is more than 0.6% copper: less than 25%, between 25% and 50%, between 50% and 75%, or more than 75%? Explain your answer.

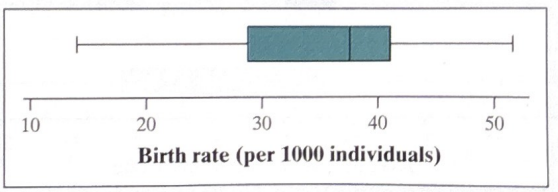
33. Electoral boxplot Refer to Exercise 5. Here is a boxplot of the electoral vote data:



(a) Use the boxplot to estimate the IQR of the distribution.

(b) Identify an aspect of the distribution that the dotplot in Exercise 5 reveals that the boxplot does not.

34. Birth rate boxplot Refer to Exercise 6. Here is a boxplot of the birth rate data:



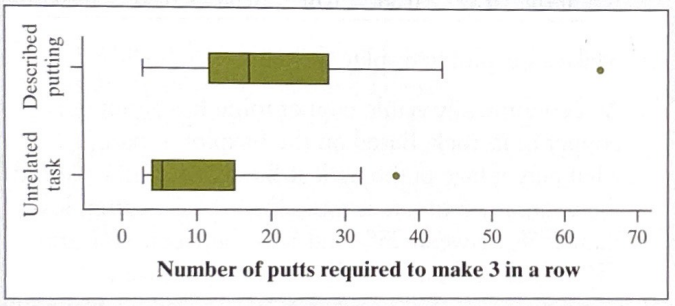
(a) Use the boxplot to estimate the IQR of the distribution.

(b) Identify an aspect of the distribution that the dotplot in Exercise 6 reveals that the boxplot does not.

Comparing Distributions with Boxplots and Summary Statistics

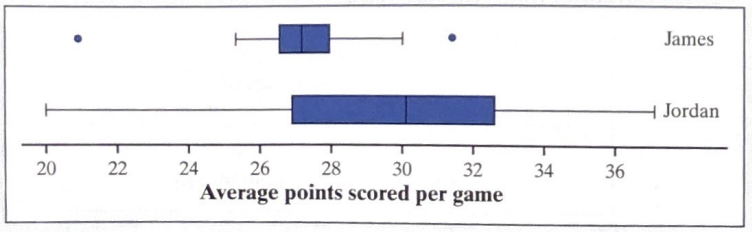
35. **Overthinking it?** Athletes often comment that they try not to "overthink it" when competing in their sport. Is it possible to "overthink it"? To investigate, researchers put some golfers to the test. They recruited 40 experienced golfers and allowed them some time to practice their putting. After practicing, they randomly assigned the golfers in equal numbers to two groups. Golfers in one group were asked to write a detailed description of their putting technique (which could lead to "overthinking it"). Golfers in the other group were asked to do an unrelated verbal task for the same amount of time. After completing their tasks, each golfer attempted putts from a fixed distance until they made three putts in a row. Here are parallel boxplots displaying the number of putts required for the golfers in each group to make three putts in a row, along with summary statistics.[86] Compare the distributions.

pg 68



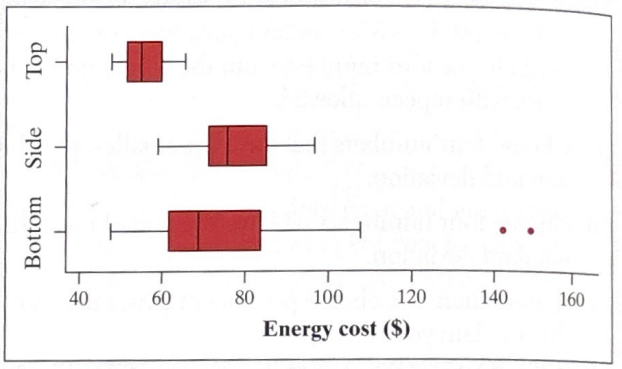
Group Name	n	Mean	SD	Min	Q_1	Med	Q_3	Max
Described putting	20	21.2	15.07	3	11.5	17	28	65
Unrelated task	20	10.6	9.832	3	4	5.5	15.5	37

36. **Who is the GOAT?** Which NBA player is the Greatest of All Time: LeBron James or Michael Jordan? Here are parallel boxplots displaying the average points scored per game by each player in the seasons they played through 2019, along with summary statistics.[87] Compare the distributions.



Player	$\bar{x}$	s_x	Min	Q_1	Med	Q_3	Max	IQR
James	27.16	2.33	20.90	26.55	27.15	27.95	31.40	1.40
Jordan	29.45	4.76	20.00	26.90	30.10	32.60	37.10	5.70

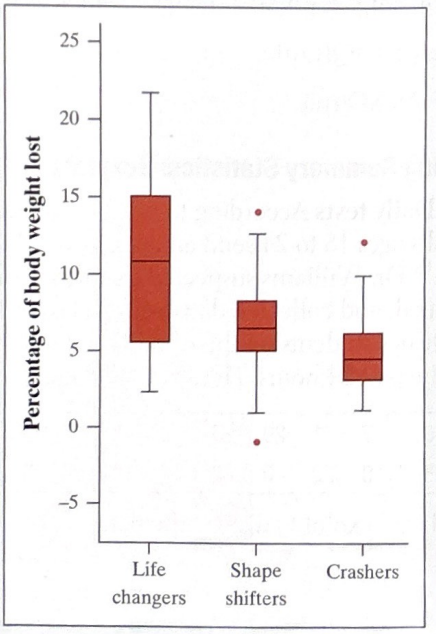
37. **Energetic refrigerators** *Consumer Reports* magazine rated different types of refrigerators, including those with bottom freezers, those with top freezers, and those with side freezers. One of the variables it measured was annual energy cost (in dollars). The following boxplots show the energy cost distributions for each of these types.[88]



(a) From the boxplots, what can you say about the percentage of each type of freezer that costs more than $60 per year to operate?

(b) Compare the energy cost distributions for the three types of refrigerators.

38. **Weight-loss strategies** Which dieting strategy works best? Researchers used data from the U.K. television show *How to Lose Weight Well* to investigate. In each episode, six participants are assigned in pairs to adopt three dieting strategies: crashers, shape shifters, and life changers. The crashers follow extreme diets for 1–2 weeks, such as replacing meals with bone broth or baby food. The shape shifters follow moderately extreme diets for 4–6 weeks, like the blood sugar diet that involves consuming at most 800 calories per day. The life changers follow less-extreme diets for 12–16 weeks, like low-carbohydrate or Mediterranean diets. Here are boxplots that summarize the percentage of body weight lost by 137 pairs of participants in the TV show over a 5-year period:[89]



(a) From the boxplots, what can you say about the percentage of people in each of the three dieting groups who lost less than 5% of their body weight?

(b) Compare the distributions of percentage of body weight lost for the three dieting groups.

39. **On-site worker commutes** How long do on-site workers typically spend traveling to their workplace? The answer may depend on where they live. Here are the travel times (in minutes) of 20 randomly chosen on-site workers in New York state and 15 randomly chosen on-site workers in North Carolina:[90]

New York	10	30	5	25	40	20	10	15	30	20	15
	20	85	15	65	15	60	60	40	45		
North Carolina	30	20	10	40	25	20	10	60	15	40	5
	30	10	12	10							

(a) Make parallel boxplots to display the data.

(b) Do these data provide strong evidence that travel times to work differ for workers in these two states? Give appropriate evidence to support your answer.

40. **SSHA scores** Higher scores on the Survey of Study Habits and Attitudes (SSHA) indicate good study habits and attitudes toward learning. Here are scores for 18 first-year college students and 20 second-year college students:

First-year students	154	109	137	115	152	140	154	178	101	
	103	126	126	137	165	165	129	200	148	
Second-year students	108	140	114	91	180	115	126	92	169	146
	109	132	75	88	113	151	70	115	187	104

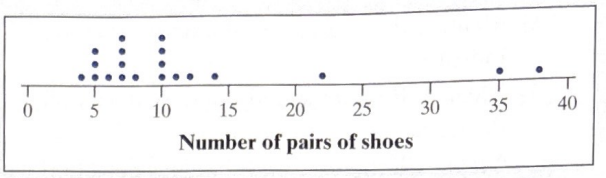
(a) Make parallel boxplots to compare the distributions.

(b) Do these data support the belief that first-year and second-year college students differ in their study habits and attitudes toward learning? Give appropriate evidence to support your answer.

For Investigation *Apply the skills from the section in a new context or nonroutine way.*

41. **Trimmed mean** Another measure of center for a quantitative data set is the *trimmed mean*. To calculate the trimmed mean, order the data set from lowest to highest, remove the same number of data values from each end, and calculate the mean of the remaining values. For example, to calculate the 10% trimmed mean, start by removing the smallest 10% and the largest 10% of values in the data set. For a data set with 50 values, you would remove 5 values from each "end" of the data set because $5/50 = 0.10$ or 10%.

Researchers asked a random sample of 20 students from a large high school how many pairs of shoes they had. Here is a dotplot of the data:

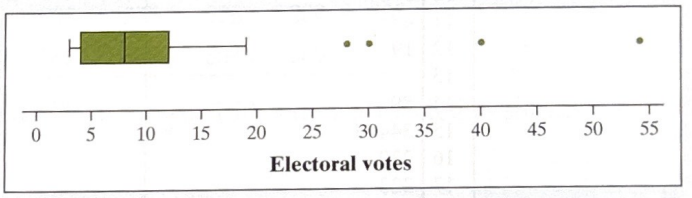


(a) Calculate the mean of the distribution.

(b) Calculate the 10% trimmed mean.

(c) Why is the trimmed mean a better summary of the center of this distribution than the mean?

42. **Measuring skewness** Here is a boxplot of the number of electoral votes in 2024 for each of the 50 states and the District of Columbia, along with summary statistics. You can see that the distribution is skewed to the right with four high outliers. How might we compute a numerical measure of skewness?



n	Mean	SD	Min	Q_1	Med	Q_3	Max
51	10.549	9.653	3	4	8	12	54

(a) One simple formula for calculating skewness is $\frac{\text{maximum} - \text{median}}{\text{median} - \text{minimum}}$. Compute this value for the electoral vote data. Explain why this formula should yield a value greater than 1 for a right-skewed distribution.

(b) Based solely on the summary statistics provided, define a formula for a different statistic that measures skewness. Compute the value of this statistic for the electoral vote data. What values of the statistic might indicate that a distribution is skewed to the right? Explain your reasoning.

Multiple Choice *Select the best answer for each question.*

43. If a distribution is strongly skewed to the right with no outliers, which of the following relationships is most likely correct?

(A) mean < median
(B) mean ≈ median
(C) mean = median
(D) mean > median
(E) We can't tell without examining the data.

44. The scores on a statistics test had a mean of 81 and a standard deviation of 9. One student was absent on the test day, and their score wasn't included in the calculation. If this student's score of 84 was added to the distribution of scores, what would happen to the mean and standard deviation?

(A) Mean will increase and standard deviation will increase.
(B) Mean will increase and standard deviation will decrease.
(C) Mean will increase and standard deviation will stay the same.
(D) Mean will decrease and standard deviation will increase.
(E) Mean will decrease and standard deviation will decrease.

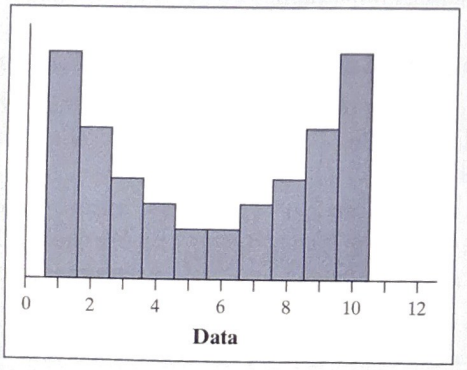
45. The stemplot shows the number of home runs hit by each of the 30 Major League Baseball teams in a single season. Home run totals greater than what value should be considered outliers?

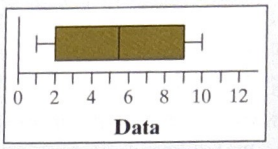
09	15
10	3789
11	47
12	19
13	
14	89
15	34445
16	239
17	223
18	356
19	1
20	3
21	0
22	2

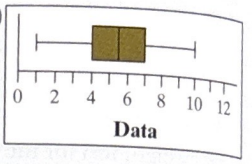
Key: 14|8 is a team with 148 home runs.

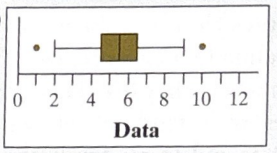
(A) 173
(B) 210
(C) 222
(D) 229
(E) 257

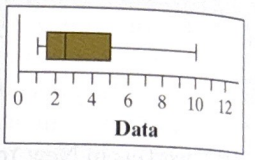
46. Which of the following boxplots best matches the distribution shown in the histogram?

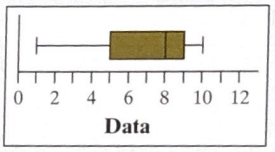


(A)


(D)


(B)


(E)


(C)


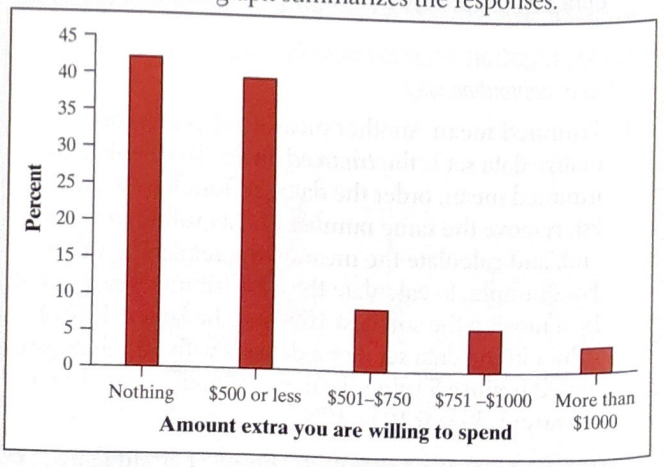
Recycle and Review *Practice what you learned in previous sections.*

47. **How tall are you? (1C)** We chose a random sample of 50 Canadian students who completed an online survey that included the question, "How tall are you without your shoes on? Answer to the nearest half-centimeter." Here are the students' responses:

166.5	170.0	178.0	163.0	150.5	169.0	173.0	169.0	171.0	166.0
190.0	183.0	178.0	161.0	171.0	170.0	191.0	168.5	178.5	173.0
175.0	160.5	166.0	164.0	163.0	174.0	160.0	174.0	182.0	167.0
166.0	170.0	170.0	181.0	171.5	160.0	178.0	157.0	165.0	187.0
168.0	157.5	145.5	156.0	182.0	168.5	177.0	162.5	160.5	185.5

Make an appropriate graph to display the data. Describe the shape, center, and variability of the distribution. Are there any outliers?

48. **Climate change insurance? (1A, 1B)** Qualtrics conducted a survey of 1070 randomly selected U.S. homeowners. Respondents were asked, "How much extra would you be willing to spend per year on insurance policies that focus on coverage for climate change–induced risks?" The bar graph summarizes the responses.[91]



(a) Identify the individuals and the variable displayed in the bar graph. Is the variable categorical or quantitative? Justify your answer.
(b) Describe what you see.

UNIT 1 Wrap-Up

PART I

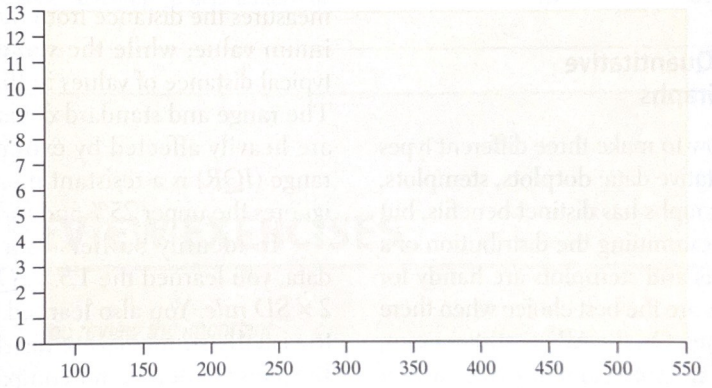
FRAPPY! Free Response AP® Problem, Yay!

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.

A realtor collects data about the sales price (in $1000s) of 43 homes that sold in their city in the past several months. Here are the data, sorted from lowest to highest sales price:[92]

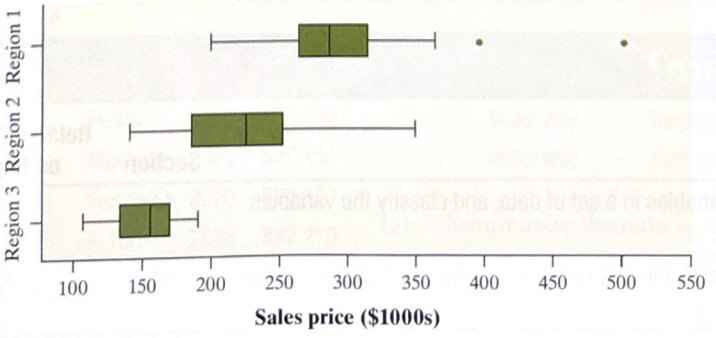
107	133	135	141	147	165	165	170	173	175	182
190	191	195	199	205	210	215	220	226	228	242
249	251	255	255	260	265	265	275	275	280	285
289	295	301	310	315	350	365	365	397	503	

(a) Make a histogram of the data using intervals of width 50, starting at 100, with axes scaled as shown.



(b) Determine whether or not the maximum sales price of $503,000 is an outlier in the distribution of sales price for these 43 homes. Show your method clearly.

Here are parallel boxplots of the sales prices for the homes in each of three regions of the city:



(c) Compare the distributions of sales price in the three regions of the city.

After you finish the FRAPPY!, you can view two example solutions on the book's website (**bfwpub.com/tps7e**). 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 a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score.

UNIT 1, PART I REVIEW

SECTION 1A Statistics: The Language of Variation

In this brief section, you learned how to identify the **individuals** and **variables** in a data set. You also learned how to distinguish between **categorical variables** and **quantitative variables**, as well as how to summarize the **distribution** of a variable with a **frequency table** or **relative frequency table.**

SECTION 1B Displaying and Describing Categorical Data

In this section, you learned how to display the distribution of a categorical variable with a **bar graph** or a **pie chart**, and what to look for when describing these graphs. Next, you learned how to compare the distribution of a categorical variable in two or more groups using **side-by-side bar graphs.** Remember to properly label your graphs! Poor labeling is an easy way to lose credit on the AP® Statistics exam. You should also be able to recognize misleading graphs and be careful to avoid making misleading graphs yourself.

SECTION 1C Displaying Quantitative Data with Graphs

In this section, you learned how to make three different types of graphs for displaying quantitative data: **dotplots, stemplots,** and **histograms.** Each of the graphs has distinct benefits, but all of them are good tools for examining the distribution of a quantitative variable. Dotplots and stemplots are handy for smaller sets of data. Histograms are the best choice when there are a large number of data values. On the AP® Statistics exam, you will be expected to make, interpret, and describe each of these types of graphs.

When you are describing the distribution of a quantitative variable, you should look at a graph of the data to determine the overall pattern (**shape, center, variability**) and striking departures from that pattern (**outliers**). When comparing distributions of quantitative data, you should include explicit comparison words for center and variability such as "is greater than" or "is approximately the same as." When asked to compare distributions, a very common mistake on the AP® Statistics exam is describing the characteristics of each distribution separately without making these explicit comparisons.

SECTION 1D Describing Quantitative Data with Numbers

To measure the *center* of a distribution of quantitative data you learned how to calculate the **median** and the **mean** of a distribution. You also learned that the median is a **resistant** measure of center, but the mean isn't resistant because it can be greatly affected by skewness or outliers.

To measure the *variability* of a distribution of quantitative data, you learned how to calculate the **range, standard deviation,** and **interquartile range** (*IQR*). The range measures the distance from the minimum value to the maximum value, while the standard deviation measures the typical distance of values in the distribution from the mean. The range and standard deviation are not resistant — both are heavily affected by extreme values. The interquartile range (*IQR*) is a resistant measure of variability because it ignores the upper 25% and lower 25% of the distribution.

To identify outliers in a distribution of quantitative data, you learned the **1.5 × *IQR* rule** and the less common *2 × SD rule*. You also learned that **boxplots** are a great way to visually summarize a distribution of quantitative data. Boxplots are helpful for comparing the center (median) and the variability (range, *IQR*) for multiple distributions. Boxplots aren't as useful for identifying the shape of a distribution because they do not display peaks, clusters, gaps, and other interesting features.

What Did You Learn?

Learning Target	Section	Related Example on Page(s)	Relevant Review Exercise(s)
Identify the individuals and variables in a set of data, and classify the variables as categorical or quantitative.	1A	4	R1
Make and interpret a frequency table or a relative frequency table for a distribution of data.	1A	6	R2
Make and interpret bar graphs of categorical data.	1B	12	R2
Compare distributions of categorical data.	1B	13	R4
Identify what makes some graphs of categorical data misleading.	1B	15	R3
Make and interpret dotplots of quantitative data.	1C	23	R5