AP Statistics Summer Assignment, 2016-17

AP Statistics students,

I hope this finds you all enjoying a wonderful summer vacation! I look forward to a great year of studying statistics with you. To allow us to complete the full curriculum with ample time for exam review in April and early May, we will use the summer assignment to get ahead on key basic concepts of statistics. But first, here are a few thoughts and guidelines regarding AP Statistics: Advanced Placement Statistics is a rigorous, college-level course intended to prepare students for the College Board's *AP Statistics Exam* in May. This course will no doubt be strikingly different from previous mathematics courses you have taken. **Careful reading and detailed writing are integral parts of the course.** Rest assured, we will work hard, and we will enjoy ourselves as we do so! **What will we learn?** The course is roughly comprised of four broad conceptual themes, as outlined below:

(a) Exploring Data: We will learn the methods necessary for exploring data and for describing our results graphically with histograms, stemplots, boxplots, ogives, and time-series plots. In addition, we will describe our results numerically through various "summary measures" such as mean, median, standard deviation, variance, and correlation.

(b) Sampling and Experimentation: We will learn the key differences between an *observational study* and an *experiment*, and we will design and implement both.

(c) Anticipating Patterns: We will explore random phenomena using probability and simulation. We will do extensive work with "Random Variables" which you will soon learn are markedly different from the "variables" you studied and used in algebra courses.

(d) Statistical Inference: We will learn how to estimate population parameters by constructing confidence intervals, and we will learn how to do a number of different *Hypothesis Tests*.

The following assignment will be due on the first full day of class. These problems are taken from the "Exploring Data" category above. You will be tested on this material within the first few weeks of school, so it is critical that you take this assignment very seriously and do your very best work. It is **very important that you do not attempt to do the assigned problems without having first carefully done the reading.**

Directions: Reading is critical in this course and our summer reading is just as essential as reading you will be assigned throughout the school year.

Your summer assignment is as follows:

Part A: Read the following:

Pages 2-6; Introduction: *Data Analysis: Making Sense of Data*

Pages 8-22; Section 1.1: *Analyzing Categorical Data*

Pages 27 -41: Section 1.2: *Displaying Quantitative Data with Graphs*

Part B: Do the following problems with your work clearly presented on notebook paper. Show all procedures clearly and make sure that all graphs are well-labeled (on both axes) and have a descriptive title. You must use a ruler for all axes and all bar graphs. Your work should be complete, concise, and well-documented. In short, it should represent your very best effort. The pages are attached.

Pages 7-8: # 3, 4, 6, 7, and 8

Pages 22-26: # 11, 14, 15, 18, 23, 24

Pages 42 – 49: # 37, 38, 39, 40, 43, 45, 46, 47, 51, 52

Questions? Email me at jroe@wcboe.org. I *look forward to a great year!*

Mr. Roe

AP Statistics Chapter 1 Notes - Exploring Data

1.1/2: Categorical Variables and Displaying Distributions with Graphs

Individuals and Variables

- **Individuals** are objects described by a set of data. Individuals may be people, but they may also be animals or things.
- A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

Categorical and Quantitative Variables

- A **categorical variable** places an individual into one of several groups or categories.
- A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense.

Distribution

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

Describing the Overall Pattern of a Distribution – Remember your SOCS

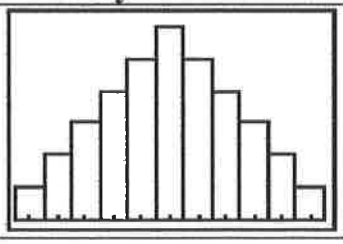
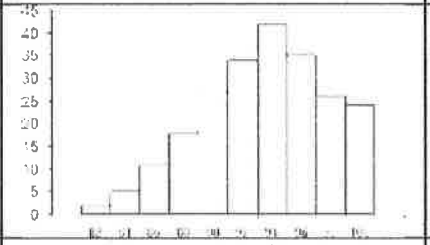
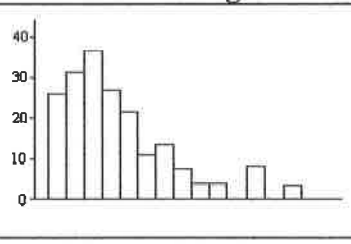
To describe the overall pattern of a distribution, address all of the following:

- Spread – give the lowest and highest value in the data set
- Outliers – are there any values that stand out as unusual?
- Center – what is the approximate average value of the data (only an estimation)
- Shape – does the graph show symmetry, or is it skewed in one direction (see below)

Outliers

An outlier in any graph of data is an individual observation that falls outside the overall pattern of the graph.

Describing the SHAPE of a distribution – Symmetric and Skewed Distributions

Symmetric	Skewed Left	Skewed Right
		
Mean = Median	Mean < Median	Mean > Median

Time Plot

- A **time plot** of a variable plots each observation against the time at which it was measured.
- Always mark the time scale on the horizontal axis and the variable of interest on the vertical axis. If there are not too many points, connecting the points by lines helps show the pattern of changes over time.

1.3: Describing Distributions with Numbers

The Mean ($\bar{x}$)

To find the **mean** of a set of observations, add their values and divide by the number of observations. If the n observations are x1, x2, …, xn, their mean is:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \quad \text{or simply,} \quad \bar{x} = \sum_{i=1}^{n} x_i$$

The Median (M)

- The median M is the midpoint of distribution, the number such that half the observations are smaller and the other half are larger. To find the median of distribution:
- Arrange all observation in order of size, from smallest to largest.
- If the number of observations n is odd, the median M is the center observation in the ordered list. The position of the center observation can be found at $(n + 1) / 2$
- If the number of observations n is even, the median M is the mean of the two center observations in the ordered list. The position of the two middle values are $n/2$ and $n/2 + 1$

The Five-Number Summary

The five-number summary of a data set consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is:

Minimum – Q_1 – M – Q_3 – Maximum

The Quartiles (Q_1 and Q_3)

- To calculate the quartiles, arrange the observations in increasing order and locate the median M in the ordered list of observations.
- The 1st quartile (Q1) is middle number of the values that are less than the median.
- The 3rd quartile (Q3) is the middle number of the values that are greater than the median.

Example

2	14	28	29	30	32	33	34	40	42	52
Min		**Q1**			**Med**			**Q3**		**Max**

The Interquartile Range (IQR)

The IQR is the distance between the first and third quartiles, IQR = Q3 - Q1

Outliers: The 1.5 x IQR Criterion

Call an observation an outlier if it falls more than 1.5 x IQR below the first quartile or above the third quartile. Using the 5-number summary from above as an example (IQR = 40-28=12)

- Low outlier cutoff: $Q_1 - 1.5 \times IQR$ (example: 28 – 1.5(12) = 28 – 18 = 10) Therefore, the 2 is an outlier.
- High outlier cutoff: $Q_3 + 1.5 \times IQR$ (example: 40 + 1.5(12) = 40 + 18 = 58) no outlier

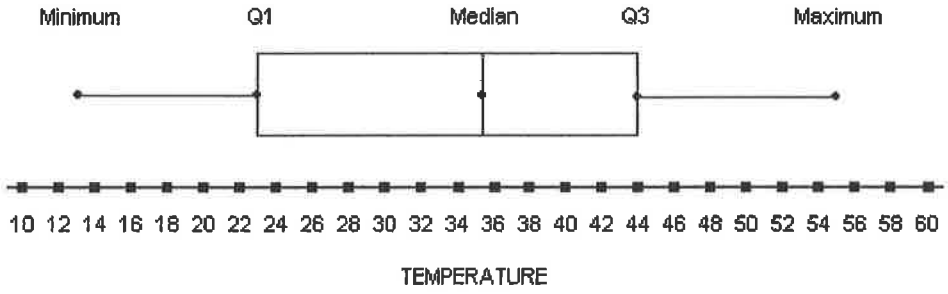
1.3: Describing Distributions with Numbers

Boxplot

A boxplot is a graph of the five-number summary, with outliers plotted individually.

- A central box spans the quartiles.
- A line in the box marks the median.
- Observations more than 1.5 x IQR outside the central box are plotted individually.
- Lines extend from the box out to the smallest and largest observations, not the outliers.

Example:



The Standard Deviation (S or Sx)

The standard deviation of a set of observations is the average of the squares of the deviations of the observations from their mean. The formula for the standard deviation of n observations $x_1, x_2, \ldots, x_n$ is:

$$s = \sqrt{\frac{\sum (x_i - x)^2}{n-1}}$$

Calculation of the Standard Deviation

Consider the data below which has a mean of 4.8:

x_i	x_i – mean	$(x_i\text{-mean})^2$
6	$6 - 4.8 = 1.2$	$(1.2)^2 = 1.44$
3	$3 - 4.8 = -1.8$	$(-1.8)^2 = 3.24$
8	$8 - 4.8 = 3.2$	$(3.2)^2 = 10.24$
5	$5 - 4.8 = 0.2$	$(0.2)^2 = 0.04$
2	$2 - 4.8 = -2.8$	$(-2.8)^2 = 7.84$
Sum	**0**	**22.8**

So the standard deviation is $\sqrt{22.8/(5-1)} = \sqrt{22.8/4} = \sqrt{5.7} = 2.387$

Exploring Data

Do Rewards Promote Creativity?

What motivates people to be creative? Is it the possibility of receiving an external reward—like money, praise, fame, or a good grade? Or is the personal satisfaction gained from doing creative work its own reward? Researcher Teresa Amabile designed a study to find out. Her specific research question was: Will competing for a prize improve children's artistic creativity?

Amabile gathered some elementary school students to take part in her study. The children were divided into two groups and instructed to make a "silly" collage using materials that were provided. Before they started, the children in one group were told that their collages would be judged by experts and that the winners would receive prizes. The children in the other group were told that they would share their collages at an art party. In fact, expert judges rated the creativity of all the collages.[1]

Want to know what happened? By the end of this chapter, you'll have your answer.

Introduction

Data Analysis: Making Sense of Data

In the Introduction, you'll learn about:

- Individuals and variables
- From data analysis to inference

Statistics is the science of data. The volume of data available to us is overwhelming. For example, the Census Bureau's American Community Survey collects data from 3,000,000 housing units each year. Astronomers work with data on tens of millions of galaxies. The checkout scanners at Walmart's 6500 stores in 15 countries record hundreds of millions of transactions every week. In all these cases, the data are trying to tell us a story—about U.S. households, objects in space, or Walmart shoppers. To hear what the data are saying, we need to help them speak by organizing, displaying, summarizing, and asking questions. That's **data analysis.**

Data analysis

Individuals and Variables

Any set of data contains information about some group of **individuals.** The characteristics we measure on each individual are called **variables.**

DEFINITION: Individuals and variables

Individuals are the objects described by a set of data. Individuals may be people, animals, or things.

A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

A high school's student data base, for example, includes data about every currently enrolled student. The students are the *individuals* described by the data set. For each individual, the data contain the values of *variables* such as age, gender, grade point average, homeroom, and grade level. In practice, any set of data is accompanied by background information that helps us understand the data. When you first meet a new data set, ask yourself the following questions:

1. *Who* are the individuals described by the data? How many individuals are there?
2. *What* are the variables? In what *units* is each variable recorded? Weights, for example, might be recorded in grams, pounds, thousands of pounds, or kilograms.

We could follow a newspaper reporter's lead and extend our list of questions to include *Why, When, Where,* and *How* were the data produced? For now, we'll focus on the first two questions.

Some variables, like gender and grade level, simply place individuals into categories. Others, like age and grade point average (GPA), take numerical values for which we can do arithmetic. It makes sense to give an average GPA for a group of students, but it doesn't make sense to give an "average" gender.

AP EXAM TIP If you learn to distinguish categorical from quantitative variables now, it will pay big rewards later. The type of data determines what kinds of graphs and which numerical summaries are appropriate. You will be expected to analyze categorical and quantitative data effectively on the AP exam.

DEFINITION: Categorical variable and quantitative variable

A **categorical variable** places an individual into one of several groups or categories.

A **quantitative variable** takes numerical values for which it makes sense to find an average.

CAUTION

Not every variable that takes number values is quantitative. Zip code is one example. Although zip codes are numbers, it doesn't make sense to talk about the average zip code. In fact, zip codes place individuals (people or dwellings) into categories based on location. Some variables—such as gender, race, and occupation—are categorical by nature. Other categorical variables are created by grouping values of a quantitative variable into classes. For instance, we could classify people in a data set by age: 0–9, 10–19, 20–29, and so on.

The proper method of analysis for a variable depends on whether it is categorical or quantitative. As a result, it is important to be able to distinguish these two types of variables.

EXAMPLE

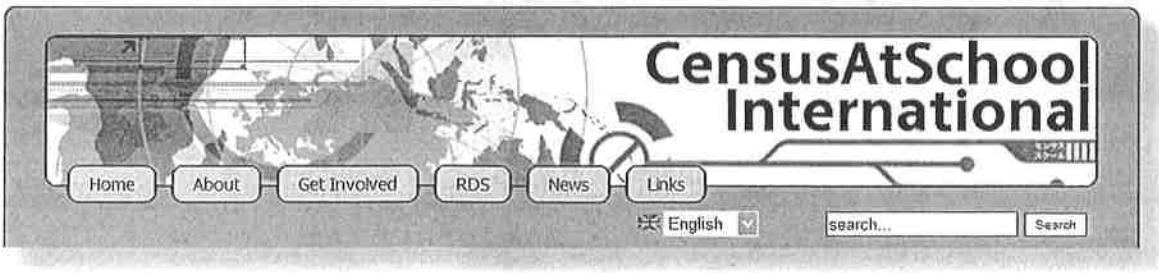
Census at School

Data, individuals, and variables

www

CensusAtSchool is an international project that collects data about primary and secondary school students using surveys. Hundreds of thousands of students from Australia, Canada, New Zealand, South Africa, and the United Kingdom have taken part in the project since 2000. Data from the surveys are available at the project's Web site (www.censusatschool.com). We used the site's "Random Data Selector" to choose 10 Canadian students who completed the survey in a recent year. The table below displays the data.

Province	Gender	Languages spoken	Handed	Height (cm)	Wrist circum. (mm)	Preferred communication	Travel to school (min)
Ontario	Male	1	Right	175	175	Internet chat or MSN	25
Alberta	Female	3	Right	147	140	MySpace/Facebook	20
Ontario	Male	1	Right	165	170	Internet chat	4
British Columbia	Female	1	Right	155	145	In person	10
New Brunswick	Male	9	Left	130.5	130	Other	40
Ontario	Male	2	Right	170	165	In person	7
Ontario	Male	3	Left	150	100	Internet chat	10
New Brunswick	Male	2	Both	167.5	220	Internet chat	30
Ontario	Female	1	Right	161	104	Text messaging	10
Ontario	Male	6	Right	190.5	180	Internet chat	10



PROBLEM:

(a) Who are the individuals in this data set?

(b) What variables were measured? Identify each as categorical or quantitative. In what units were the quantitative variables measured?

(c) Describe the individual in the highlighted row.

SOLUTION:

(a) The individuals are the 10 randomly selected Canadian students.

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

(b) The eight variables measured are province where student lives (categorical), gender (categorical), number of languages spoken (quantitative, in whole numbers), dominant hand (categorical), height (quantitative, in centimeters), wrist circumference (quantitative, in millimeters), preferred communication method (categorical), and travel time to school (quantitative, in minutes).

(c) This student lives in Ontario, is male, speaks three languages, is left-handed, is 150 cm tall (about 59 inches), has a wrist circumference of 100 mm (about 4 inches), prefers to communicate via Internet chat, and travels 10 minutes to get to school.

For Practice *Try Exercise 3*

Most data tables follow the format shown in the example—each row is an individual, and each column is a variable. Sometimes the individuals are called *cases*.

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

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

DEFINITION: Distribution

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

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

How to Explore Data

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

CHECK YOUR UNDERSTANDING

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

1. Who are the individuals in Jake's study?
2. What variables did Jake measure? Identify each as categorical or quantitative.

From Data Analysis to Inference

Inference

Sometimes, we're interested in drawing conclusions that go beyond the data at hand. That's the idea of **inference**. In the CensusAtSchool example, 7 of the 10 randomly selected Canadian students are right-handed. That's 70% of the *sample*. Can we conclude that 70% of the *population* of Canadian students who participated in CensusAtSchool are right-handed? No. If another random sample of 10 students was selected, the percent who are right-handed would probably not be exactly 70%. Can we at least say that the actual population value is "close" to 70%? That depends on what we mean by "close."

The following Activity gives you an idea of how statistical inference works.

ACTIVITY *Hiring discrimination—it just won't fly!*

MATERIALS: Deck of cards for each student or pair of students

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

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

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

1. Separate the cards into piles by suit. Use ten cards from one suit to represent the female pilots. To represent the 15 male pilots, you'll need all 13 cards of another suit plus two extra cards from a third suit. A second student or group can use the leftover cards from the deck to set up their simulation in a similar way.

2. Shuffle your stack of 25 cards thoroughly and deal 8 cards. Count the number of female pilots selected. Record this value in a table like the one shown below.

Trial:	1	2	3	4	5
No. of females:					

3. Return the 8 cards to your stack. Shuffle and deal four more times so that you have a total of five simulated lottery results.

4. Your teacher will draw and label axes for a class dotplot. Each student should plot the number of females obtained in each of the five simulation trials on the graph.

5. Discuss the results with your classmates. Does it seem believable that airline managers carried out a fair lottery? What advice would you give the male pilot who contacted you?

6. Would your advice change if the lottery had chosen 6 female (and 2 male) pilots? Explain.

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

Summary

- A data set contains information on a number of **individuals.** Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables.** A variable describes some characteristic of an individual, such as a person's height, gender, or salary.
- Some variables are **categorical** and others are **quantitative.** A categorical variable places each individual into a category, such as male or female. A quantitative variable has numerical values that measure some characteristic of each individual, such as height in centimeters or salary in dollars.
- The **distribution** of a variable describes what values the variable takes and how often it takes them.

INTRODUCTION Exercises

1. **Protecting wood** How can we help wood surfaces resist weathering, especially when restoring historic wooden buildings? In a study of this question, researchers prepared wooden panels and then exposed them to the weather. Here are some of the variables recorded: type of wood (yellow poplar, pine, cedar); type of water repellent (solvent-based, water-based); paint thickness (millimeters); paint color (white, gray, light blue); weathering time (months). Identify each variable as categorical or quantitative.

2. **Medical study variables** Data from a medical study contain values of many variables for each of the people who were the subjects of the study. Here are some of the variables recorded: gender (female or male); age (years); race (Asian, black, white, or other); smoker (yes or no); systolic blood pressure (millimeters of mercury); level of calcium in the blood (micrograms per milliliter). Identify each as categorical or quantitative.

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

pg 3

Gender	Hand	Height (in)	Homework time (min)	Favorite music	Pocket change (cents)
F	L	65	200	Hip-hop	50
M	L	72	30	Country	35
M	R	62	95	Rock	35
F	L	64	120	Alternative	0
M	R	63	220	Hip-hop	0
F	R	58	60	Alternative	76
F	R	67	150	Rock	215

(a) What individuals does this data set describe?

(b) Clearly identify each of the variables. Which are quantitative? In what units are they measured?

(c) Describe the individual in the highlighted row.

4. **Coaster craze** Many people like to ride roller coasters. Amusement parks try to increase attendance by building exciting new coasters. The table below displays data on several roller coasters that were opened in 2009.[2]

Roller coaster	Type	Height (ft)	Design	Speed (mph)	Duration (s)
Wild mouse	Steel	49.3	Sit down	28	70
Terminator	Wood	95	Sit down	50.1	180
Manta	Steel	140	Flying	56	155
Prowler	Wood	102.3	Sit down	51.2	150
Diamondback	Steel	230	Sit down	80	180

(a) What individuals does this data set describe?

(b) Clearly identify each of the variables. Which are quantitative? In what units are they measured?

(c) Describe the individual in the highlighted row.

5. **Ranking colleges** Popular magazines rank colleges and universities on their "academic quality" in serving undergraduate students. Describe two categorical variables and two quantitative variables that you might record for each institution. Give the units of measurement for the quantitative variables.

6. **Students and TV** You are preparing to study the television-viewing habits of high school students. Describe two categorical variables and two quantitative variables that you might record for each student. Give the units of measurement for the quantitative variables.

Multiple choice: Select the best answer.

Exercises 7 and 8 refer to the following setting. At the Census Bureau Web site, you can view detailed data collected by the American Community Survey. The table below includes data for 10 people chosen at random from the more than one million people in households contacted by the survey. "School" gives the highest level of education completed.

Weight (lb)	Age (yr)	Travel to work (min)	School	Gender	Income last year ($)
187	66	0	Ninth grade	1	24,000
158	66	n/a	High school grad	2	0
176	54	10	Assoc. degree	2	11,900
339	37	10	Assoc. degree	1	6,000
91	27	10	Some college	2	30,000
155	18	n/a	High school grad	2	0
213	38	15	Master's degree	2	125,000
194	40	0	High school grad	1	800
221	18	20	High school grad	1	2,500
193	11	n/a	Fifth grade	1	0

7. The individuals in this data set are
   (a) households.
   (b) people.
   (c) adults.
   (d) 120 variables.
   (e) columns.

8. This data set contains
   (a) 7 variables, 2 of which are categorical.
   (b) 7 variables, 1 of which is categorical.
   (c) 6 variables, 2 of which are categorical.
   (d) 6 variables, 1 of which is categorical.
   (e) None of these.

1.1

Analyzing Categorical Data

In Section 1.1, you'll learn about:

- Bar graphs and pie charts
- Graphs: Good and bad
- Two-way tables and marginal distributions
- Relationships between categorical variables: Conditional distributions
- Organizing a statistical problem
- Simpson's paradox*

The values of a categorical variable are labels for the categories, such as "male" and "female." The distribution of a categorical variable lists the categories and gives either the *count* or the *percent* of individuals who fall in each category. Here's an example.

EXAMPLE

Radio Station Formats

Distribution of a categorical variable

The radio audience rating service Arbitron places the country's 13,838 radio stations into categories that describe the kinds of programs they broadcast. Here are two different tables showing the distribution of station formats:[3]

Frequency table

Format	Count of stations
Adult contemporary	1,556
Adult standards	1,196
Contemporary hit	569
Country	2,066
News/Talk/Information	2,179
Oldies	1,060
Religious	2,014
Rock	869
Spanish language	750
Other formats	1,579
Total	13,838

Relative frequency table

Format	Percent of stations
Adult contemporary	11.2
Adult standards	8.6
Contemporary hit	4.1
Country	14.9
News/Talk/Information	15.7
Oldies	7.7
Religious	14.6
Rock	6.3
Spanish language	5.4
Other formats	11.4
Total	99.9

Frequency table

Relative frequency table

In this case, the *individuals* are the radio stations and the *variable* being measured is the kind of programming that each station broadcasts. The table on the left, which we call a **frequency table**, displays the counts (*frequencies*) of stations in each format category. On the right, we see a **relative frequency table** of the data that shows the percents (*relative frequencies*) of stations in each format category.

*This is an interesting topic, but it is not required for the AP Statistics exam.

It's a good idea to check data for consistency. The counts should add to 13,838, the total number of stations. They do. The percents should add to 100%. In fact, they add to 99.9%. What happened? Each percent is rounded to the nearest tenth. The exact percents would add to 100, but the rounded percents only come close. This is **roundoff error.** Roundoff errors don't point to mistakes in our work, just to the effect of rounding off results.

Roundoff error

Bar Graphs and Pie Charts

Pie chart
Bar graph

Columns of numbers take time to read. You can use a **pie chart** or a **bar graph** to display the distribution of a categorical variable more vividly. Figure 1.1 illustrates both displays for the distribution of radio stations by format.

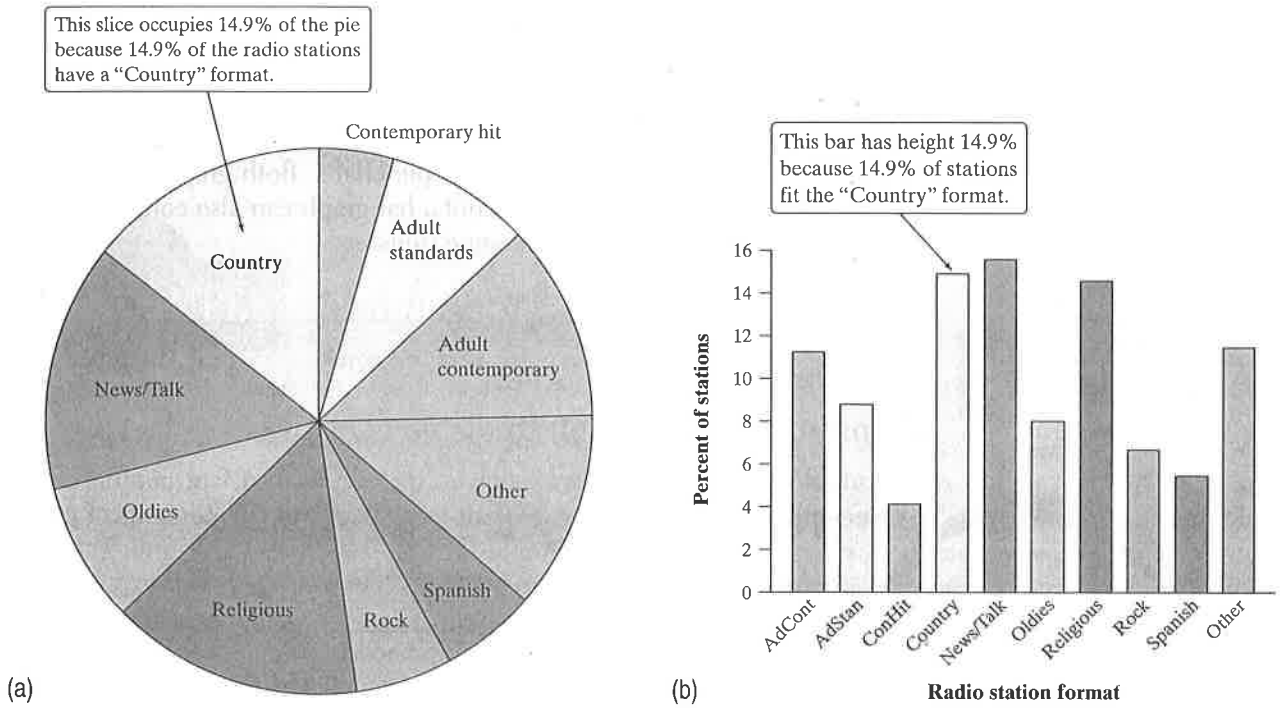
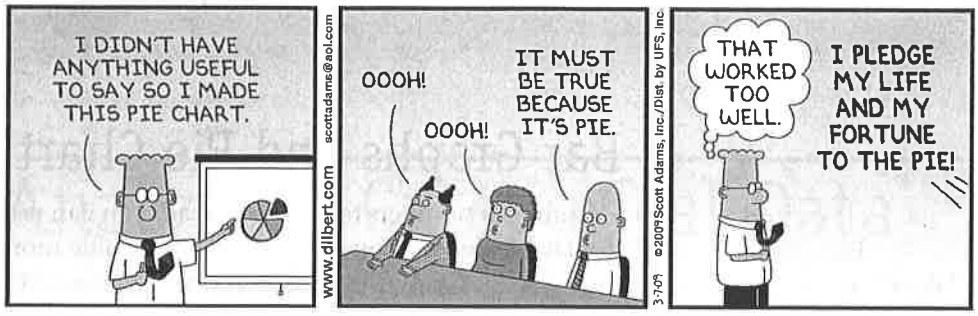


FIGURE 1.1 (a) Pie chart and (b) bar graph of U.S. radio stations by format.

THINK ABOUT IT

Do the data tell you what you want to know? Let's say that you plan to buy radio time to advertise your Web site for downloading MP3 music files. How helpful are the data in Figure 1.1? Not very. You are not interested in counting *stations*, but in counting *listeners*. For example, 14.6% of all stations are religious, but they have only a 5.5% share of the radio audience, according to Arbitron. In fact, you aren't even interested in the entire radio audience, because MP3 users are mostly young people. You really want to know what kinds of radio stations reach the largest numbers of young people. *Always think about whether the data you have help answer your questions.*

Pie charts show the distribution of a categorical variable as a "pie" whose slices are sized by the counts or percents for the categories. A pie chart must include all the categories that make up a whole. In the radio station example, we needed the "Other formats" category to complete the whole (all radio stations) and allow us to make a pie chart. Use a pie chart only when you want to emphasize each category's relation to the whole. Pie charts are awkward to make by hand, but technology will do the job for you.



Bar graphs are also called *bar charts.*

Bar graphs represent each category as a bar. The bar heights show the category counts or percents. Bar graphs are easier to make than pie charts and are also easier to read. To convince yourself, try to use the pie chart in Figure 1.1 to estimate the percent of radio stations that have an "Oldies" format. Now look at the bar graph—it's easy to see that the answer is about 8%.

Bar graphs are also more flexible than pie charts. Both 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.

EXAMPLE

Who Owns an MP3 Player?

Choosing the best graph to display the data

Portable MP3 music players, such as the Apple iPod, are popular—but not equally popular with people of all ages. Here are the percents of people in various age groups who own a portable MP3 player, according to an Arbitron survey of 1112 randomly selected people.[4]

Age group (years)	Percent owning an MP3 player
12 to 17	54
18 to 24	30
25 to 34	30
35 to 54	13
55 and older	5

PROBLEM:

(a) Make a well-labeled bar graph to display the data. Describe what you see.

(b) Would it be appropriate to make a pie chart for these data? Why or why not?

SOLUTION:

(a) We start by labeling the axes: age group goes on the horizontal axis, and percent who own an MP3 player goes on the vertical axis. For the vertical scale, which is measured in percents, we'll start at 0

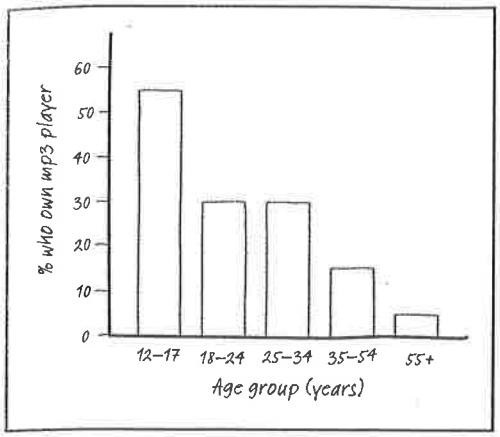


FIGURE 1.2 Bar graph comparing the percents of several age groups who own portable MP3 players.

and go up to 60, with tick marks for every 10. Then for each age category, we draw a bar with height corresponding to the percent of survey respondents who said they have an MP3 player. Figure 1.2 shows the completed bar graph. It appears that MP3 players are more popular among young people and that their popularity generally decreases as the age category increases.

(b) Making a pie chart to display these data is not appropriate because each percent in the table refers to a different age group, not to parts of a single whole.

For Practice *Try Exercise 15*

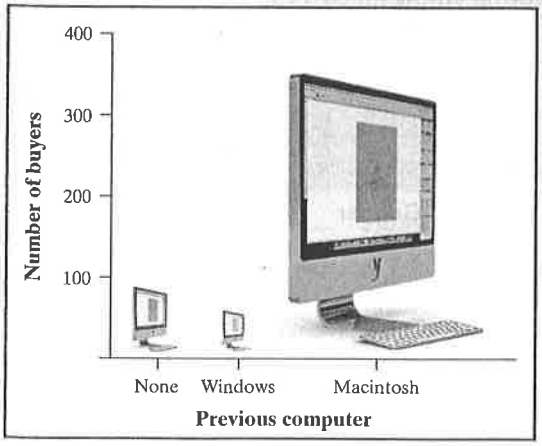
Graphs: Good and Bad

Bar graphs compare several quantities by comparing the heights of bars that represent the quantities. Our eyes, however, 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. When you draw a bar graph, make the bars equally wide. Artistically speaking, bar graphs are a bit dull. It is tempting to replace the bars with pictures for greater eye appeal. Don't do it! The following example shows why.

Who Buys iMacs?

Beware the pictograph!

When Apple, Inc., introduced the iMac, the company wanted to know whether this new computer was expanding Apple's market share. Was the iMac mainly being bought by previous Macintosh owners, or was it being purchased by first-time computer buyers and by previous PC users who were switching over? To find out, Apple hired a firm to conduct a survey of 500 iMac customers. Each customer was categorized as a new computer purchaser, a previous PC owner, or a previous Macintosh owner. The table summarizes the survey results.[5]

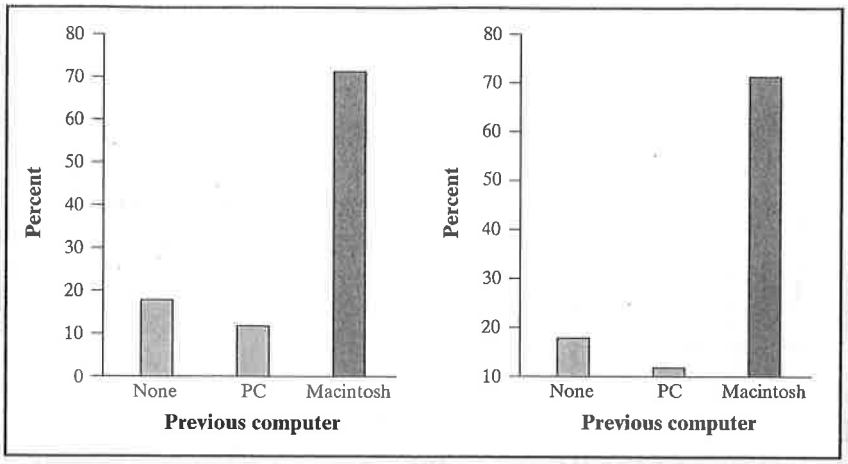


Previous ownership	Count	Percent
None	85	17.0
PC	60	12.0
Macintosh	355	71.0
Total	500	100.0

PROBLEM:

(a) Here's a clever graph of the data that uses pictures instead of the more traditional bars. How is this graph misleading?

(b) Two possible bar graphs of the data are shown on the next page. Which one could be considered deceptive? Why?



SOLUTION:

(a) Although the heights of the pictures are accurate, our eyes respond to the area of the pictures. The pictograph makes it seem like the percent of iMac buyers who are former Mac owners is at least ten times higher than either of the other two categories, which isn't the case.

(b) The bar graph on the right is misleading. By starting the vertical scale at 10 instead of 0, it looks like the percent of iMac buyers who previously owned a PC is less than half the percent who are first-time computer buyers. We get a distorted impression of the relative percents in the three categories.

For Practice *Try Exercise 17*

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

Two-Way Tables and Marginal Distributions

We have learned some techniques for analyzing the distribution of a single categorical variable. What do we do when a data set involves two categorical variables? We begin by examining the counts or percents in various categories for one of the variables. Here's an example to show what we mean.

EXAMPLE

I'm Gonna Be Rich!

Relationship between two categorical variables

A survey of 4826 randomly selected young adults (aged 19 to 25) asked, "What do you think are the chances you will have much more than a middle-class income at age 30?" The table below shows the responses, omitting a few people who refused to respond or who said they were already rich.[6]

Young adults by gender and chance of getting rich

	Gender		
Opinion	**Female**	**Male**	**Total**
Almost no chance	96	98	194
Some chance but probably not	426	286	712
A 50-50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

Two-way table

This is a **two-way table** because it describes two categorical variables, gender and opinion about becoming rich. Opinion is the *row variable* because each row in the table describes young adults who held one of the five opinions about their chances. Because the opinions have a natural order from "Almost no chance" to "Almost certain," the rows are also in this order. Gender is the *column variable*. The entries in the table are the counts of individuals in each opinion-by-gender class.

How can we best grasp the information contained in the two-way table above? First, *look at the distribution of each variable separately*. The distribution of a categorical variable says how often each outcome occurred. The "Total" column at the right of the table contains the totals for each of the rows. These row totals give the distribution of opinions about becoming rich in the entire group of 4826 young adults: 194 felt that they had almost no chance, 712 thought they had just some chance, and so on. (If the row and column totals are missing, the first thing to do in studying a two-way table is to calculate them.) The distributions of opinion alone and gender alone are called **marginal distributions** because they appear at the right and bottom margins of the two-way table.

DEFINITION: Marginal distribution

The **marginal distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among all individuals described by the table.

Percents are often more informative than counts, especially when we are comparing groups of different sizes. We can display the marginal distribution of opinions in percents by dividing each row total by the table total and converting to a percent. For instance, the percent of these young adults who think they are almost certain to be rich by age 30 is

$$\frac{\text{almost certain total}}{\text{table total}} = \frac{1083}{4826} = 0.224 = 22.4\%$$

EXAMPLE

I'm Gonna Be Rich!

Examining a marginal distribution

PROBLEM:

(a) Use the data in the two-way table to calculate the marginal distribution (in percents) of opinions.

(b) Make a graph to display the marginal distribution. Describe what you see.

SOLUTION:

(a) We can do four more calculations like the one shown above to obtain the marginal distribution of opinions in percents. Here is the complete distribution.

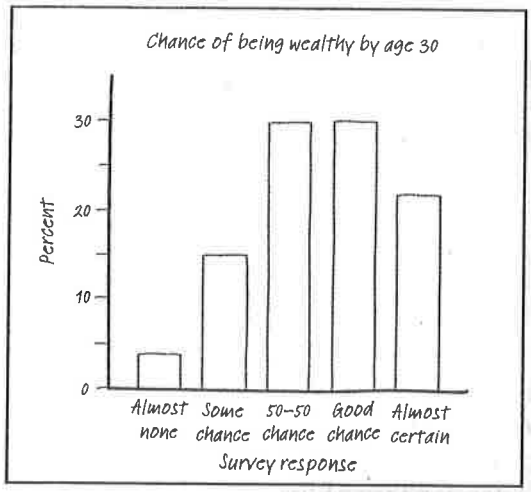


FIGURE 1.3 Bar graph showing the marginal distribution of opinion about chance of being rich by age 30.

Response	Percent
Almost no chance	$\frac{194}{4826} = 4.0\%$
Some chance	$\frac{712}{4826} = 14.8\%$
A 50–50 chance	$\frac{1416}{4826} = 29.3\%$
A good chance	$\frac{1421}{4826} = 29.4\%$
Almost certain	$\frac{1083}{4826} = 22.4\%$

(b) Figure 1.3 is a bar graph of the distribution of opinion among these young adults. It seems that many young adults are optimistic about their future income. Over 50% of those who responded to the survey felt that they had "a good chance" or were "almost certain" to be rich by age 30.

For Practice *Try Exercise 19*

Each marginal distribution from a two-way table is a distribution for a single categorical variable. As we saw earlier, we can use a bar graph or a pie chart to display such a distribution.

CHECK YOUR UNDERSTANDING

1. Use the data in the two-way table on page 12 to calculate the marginal distribution (in percents) of gender.
2. Make a graph to display the marginal distribution. Describe what you see.

Relationships between Categorical Variables: Conditional Distributions

The two-way table contains much more information than the two marginal distributions of opinion alone and gender alone. *Marginal distributions tell us nothing about the relationship between two variables.* To describe a relationship between two categorical variables, we must calculate some well-chosen percents from the counts given in the body of the table.

Young adults by gender and chance of getting rich

	Gender		
Opinion	**Female**	**Male**	**Total**
Almost no chance	96	98	194
Some chance but probably not	426	286	712
A 50-50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

Conditional distribution of opinion among women	
Response	**Female**
Almost no chance	$\frac{96}{2367} = 4.1\%$
Some chance	$\frac{426}{2367} = 18.0\%$
A 50-50 chance	$\frac{696}{2367} = 29.4\%$
A good chance	$\frac{663}{2367} = 28.0\%$
Almost certain	$\frac{486}{2367} = 20.5\%$

We can study the opinions of women alone by looking only at the "Female" column in the two-way table. To find the percent of *young women* who think they are almost certain to be rich by age 30, divide the count of such women by the total number of women, the column total:

$$\frac{\text{women who are almost certain}}{\text{column total}} = \frac{486}{2367} = 0.205 = 20.5\%$$

Doing this for all five entries in the "Female" column gives the **conditional distribution** of opinion among women. See the table in the margin. We use the term "conditional" because this distribution describes only young adults who satisfy the condition that they are female.

DEFINITION: Conditional distribution

A **conditional distribution** of a variable describes the values of that variable among individuals who have a specific value of another variable. There is a separate conditional distribution for each value of the other variable.

Now let's examine the men's opinions.

EXAMPLE

I'm Gonna Be Rich!

Calculating a conditional distribution

PROBLEM: Calculate the conditional distribution of opinion among the men.

SOLUTION: To find the percent of *men* who think they are almost certain to be rich by age 30, divide the count of such men by the total number of men, the column total:

$$\frac{\text{men who are almost certain}}{\text{column total}} = \frac{597}{2459} = 24.3\%$$

If we do this for all five entries in the "Male" column, we get the conditional distribution shown in the table.

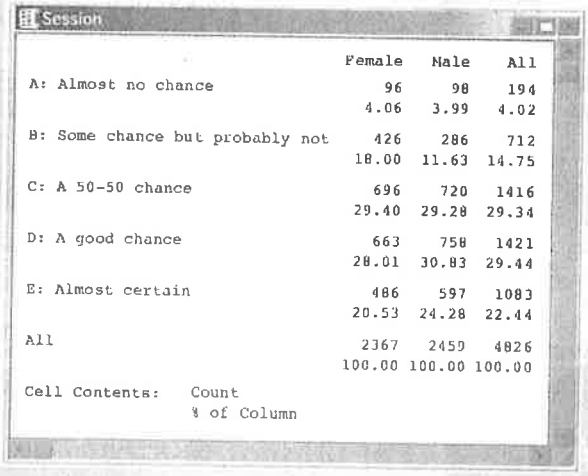
Conditional distribution of opinion among men	
Response	**Male**
Almost no chance	$\frac{98}{2459} = 4.0\%$
Some chance	$\frac{286}{2459} = 11.6\%$
A 50-50 chance	$\frac{720}{2459} = 29.3\%$
A good chance	$\frac{758}{2459} = 30.8\%$
Almost certain	$\frac{597}{2459} = 24.3\%$

For Practice *Try Exercise 21*

Software will calculate conditional distributions for you. Most programs allow you to choose which conditional distributions you want to compute.

TECHNOLOGY CORNER Analyzing two-way tables

Figure 1.4 presents the two conditional distributions of opinion, for women and for men, and also the marginal distribution of opinion for all of the young adults. The distributions agree (up to rounding) with the results in the last two examples.


Session

	Female	Male	All
A: Almost no chance	96	98	194
	4.06	3.99	4.02
B: Some chance but probably not	426	286	712
	18.00	11.63	14.75
C: A 50-50 chance	696	720	1416
	29.40	29.28	29.34
D: A good chance	663	758	1421
	28.01	30.83	29.44
E: Almost certain	486	597	1083
	20.53	24.28	22.44
All	2367	2459	4826
	100.00	100.00	100.00

Cell Contents: Count
% of Column

FIGURE 1.4 Minitab output for the two-way table of young adults by gender and chance of being rich, along with each entry as a percent of its column total. The "Female" and "Male" columns give the conditional distributions of opinion for women and men, and the "All" column shows the marginal distribution of opinion for all these young adults.

There are *two sets* of conditional distributions for any two-way table. So far, we have looked at the conditional distributions of opinion for the two genders. We could also examine the five conditional distributions of gender, one for each of the five opinions, by looking separately at the rows in the original two-way table. For instance, the conditional distribution of gender among those who responded "Almost certain" is

Female	Male
$\frac{486}{1083} = 44.9\%$	$\frac{597}{1083} = 55.1\%$

That is, of the young adults who said they were almost certain to be rich by age 30, 44.9% were female and 55.1% were male.

Because the variable "gender" has only two categories, comparing the five conditional distributions amounts to comparing the percents of women among young adults who hold each opinion. Figure 1.5 makes this comparison in a bar graph. The bar heights do *not* add to 100%, because each bar represents a different group of people.

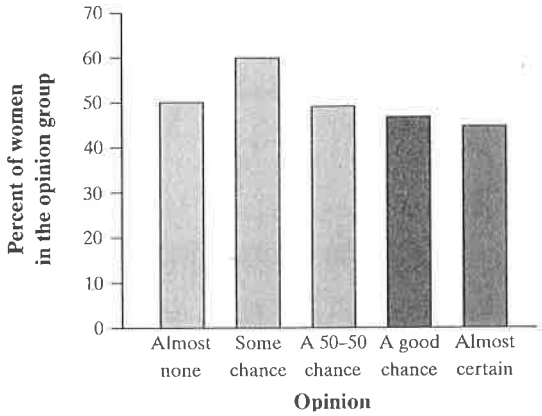


FIGURE 1.5 Bar graph comparing the percents of females among those who hold each opinion about their chance of being rich by age 30.

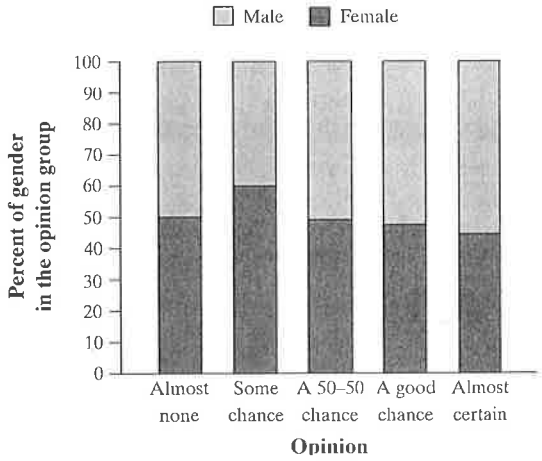


FIGURE 1.6 Segmented bar graph showing the conditional distribution of gender for each opinion category.

Segmented bar graph

An alternative to the bar graph in Figure 1.5 is a **segmented bar graph**, like the one shown in Figure 1.6. For each opinion category, there is a single bar with "segments" that correspond to the different genders. The height of each segment is determined by the percent of young adults having that opinion who were of each gender. We can see the two percents we calculated earlier displayed in the "Almost certain" bar—female 44.9% and male 55.1%. Notice that each bar has a total height of 100%.

THINK ABOUT IT

Which conditional distributions should we compare? Our goal all along has been to analyze the relationship between gender and opinion about chances of becoming rich for these young adults. We started by examining the conditional distributions of opinion for males and females. Then we looked at the conditional distributions of gender for each of the five opinion categories. Which of these two gives us the information we want? Here's a hint: think about whether changes in one variable might help explain changes in the other. In this case, it seems reasonable to think that gender might influence young adults' opinions about their chances of getting rich. To see whether the data support this idea, we should compare the conditional distributions of opinion for women and men.

CHECK YOUR UNDERSTANDING

1. Find the conditional distributions of gender among each of the other four opinion categories (we did "Almost certain" earlier). Use Figure 1.5 or Figure 1.6 to check that your answers are approximately correct.
2. Make a revised version of Figure 1.4 that includes your results from Question 1.

Organizing a Statistical Problem

As you learn more about statistics, you will be asked to solve more complex problems. Although no single strategy will work on every problem, it might be helpful to have a general framework for organizing your thinking. Here is a four-step process you can follow.

How to Organize a Statistical Problem: A Four-Step Process

STEP 4

State: What's the question that you're trying to answer?

Plan: How will you go about answering the question? What statistical techniques does this problem call for?

Do: Make graphs and carry out needed calculations.

Conclude: Give your practical conclusion in the setting of the real-world problem.

To keep the four steps straight, just remember: **S**tatistics **P**roblems **D**emand **C**onsistency!

Many examples and exercises in this book will tell you what to do—construct a graph, perform a calculation, interpret a result, and so on. Real statistics problems don't come with such detailed instructions, however. From now on, you will encounter some examples and exercises that are more realistic. They are marked with the four-step icon. Use the four-step process as a guide to solving these problems, as the following example illustrates.

EXAMPLE

Women's and Men's Opinions

Conditional distributions and relationships

STEP 4

Based on the survey data, can we conclude that young men and women differ in their opinions about the likelihood of future wealth? Give appropriate evidence to support your answer. Follow the four-step process.

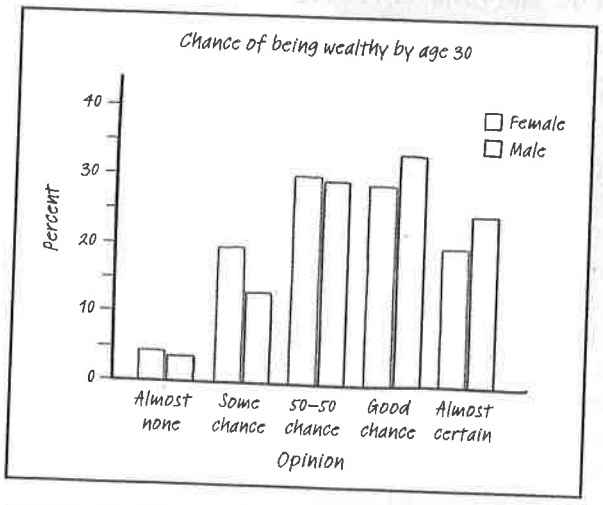


FIGURE 1.7 Side-by-side bar graph comparing the opinions of males and females.

STATE: What is the relationship between gender and responses to the question "What do you think are the chances you will have much more than a middle-class income at age 30?"

PLAN: We suspect that gender might influence a young adult's opinion about the chance of getting rich. So we'll compare the conditional distributions of response for men alone and for women alone.

Response	Female	Male
Almost no chance	$\frac{96}{2367} = 4.1\%$	$\frac{98}{2459} = 4.0\%$
Some chance	$\frac{426}{2367} = 18.0\%$	$\frac{286}{2459} = 11.6\%$
A 50-50 chance	$\frac{696}{2367} = 29.4\%$	$\frac{720}{2459} = 29.3\%$
A good chance	$\frac{663}{2367} = 28.0\%$	$\frac{758}{2459} = 30.8\%$
Almost certain	$\frac{486}{2367} = 20.5\%$	$\frac{597}{2459} = 24.3\%$

Side-by-side bar graph

DO: We'll make a side-by-side bar graph to compare the opinions of males and females. Figure 1.7 displays the completed graph.

CONCLUDE: Based on the sample data, men seem somewhat more optimistic about their future income than women. Men were less likely to say that they have "some chance but probably not" than women (11.6% vs. 18.0%). Men were more likely to say that they have "a good chance" (30.8% vs. 28.0%) or are "almost certain" (24.3% vs. 20.5%) to have much more than a middle-class income by age 30 than women were.

For Practice Try Exercise 25

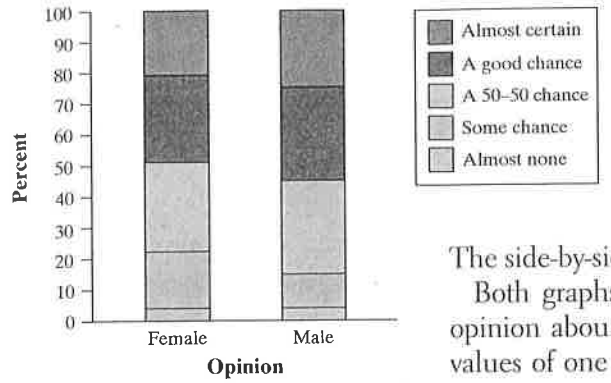


FIGURE 1.8 Segmented bar graph comparing the opinions of males and females.

We could have used a segmented bar graph to compare the distributions of male and female responses in the previous example. Figure 1.8 shows the completed graph. Each bar has five segments—one for each of the opinion categories. It's fairly difficult to compare the percents of males and females in each category because the "middle" segments in the two bars start at different locations on the vertical axis. The side-by-side bar graph in Figure 1.7 makes comparison easier.

Both graphs provide evidence of an **association** between gender and opinion about future wealth in this sample of young adults. That is, the values of one variable (opinion) tend to occur more or less frequently in combination with specific values of the other variable (gender). Men more often rated their chances of becoming rich in the two highest categories; women said "some chance but probably not" much more frequently. Can we say that there is an association between gender and opinion in the *population* of young adults? Making this determination requires formal inference, which will have to wait a few chapters.

DEFINITION: Association

We say that there is an **association** between two variables if specific values of one variable tend to occur in common with specific values of the other.

There's one caution that we need to offer: *even a strong association between two categorical variables can be influenced by other variables lurking in the background.* The Data Exploration that follows gives you a chance to explore this idea using a famous (or infamous) data set.

DATA EXPLORATION *A* Titanic *disaster*

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

	Survival Status	
Class of Travel	**Survived**	**Died**
First class	197	122
Second class	94	167
Third class	151	476

Here's another table that displays data on survival status by gender and class of travel.

	Gender			
	Female		Male	
Class of Travel	**Survived**	**Died**	**Survived**	**Died**
First class	140	4	57	118
Second class	80	13	14	154
Third class	76	89	75	387

The movie *Titanic*, starring Leonardo DiCaprio and Kate Winslet, suggested the following:

- First-class passengers received special treatment in boarding the lifeboats, while some other passengers were prevented from doing so (especially third-class passengers).
- Women and children boarded the lifeboats first, followed by the men.

1. What do the data tell us about these two suggestions? Give appropriate graphical and numerical evidence to support your answer.
2. How does gender affect the relationship between class of travel and survival status? Explain.

Simpson's Paradox*

In the most extreme cases, it is possible for an association between two categorical variables to be "reversed" when we consider a third variable. Here is an example that demonstrates the surprises that can await the unsuspecting user of data.

EXAMPLE

Do Medical Helicopters Save Lives?

Reversing an association

Accident victims are sometimes taken by helicopter from the accident scene to a hospital. Helicopters save time. Do they also save lives? Let's compare the percents of accident victims who die with helicopter evacuation and with the usual transport to a hospital by road. Here are hypothetical data that illustrate a practical difficulty:[7]

	Helicopter	Road
Victim died	64	260
Victim survived	136	840
Total	200	1100

We see that 32% (64 out of 200) of helicopter patients died, but only 24% (260 out of 1100) of the others did. That seems discouraging.

The explanation is that the helicopter is sent mostly to serious accidents, so that the victims transported by helicopter are more often seriously injured. They are more likely to die with or without helicopter evacuation. Here are the same data broken down by the seriousness of the accident:

Serious Accidents		
	Helicopter	**Road**
Died	48	60
Survived	52	40
Total	100	100

Less Serious Accidents		
	Helicopter	**Road**
Died	16	200
Survived	84	800
Total	100	1000

Inspect these tables to convince yourself that they describe the same 1300 accident victims as the original two-way table. For example, 200 (100 + 100) were moved by helicopter, and 64 (48 + 16) of these died.

*This is an interesting topic, but it is not required for the AP Statistics exam.

Among victims of serious accidents, the helicopter saves 52% (52 out of 100) compared with 40% for road transport. If we look only at less serious accidents, 84% of those transported by helicopter survive, versus 80% of those transported by road. Both groups of victims have a higher survival rate when evacuated by helicopter.

How can it happen that the helicopter does better for both groups of victims but worse when all victims are lumped together? Examining the data makes the explanation clear. Half the helicopter transport patients are from serious accidents, compared with only 100 of the 1100 road transport patients. So the helicopter carries patients who are more likely to die. The seriousness of the accident was a "lurking variable" that, until we uncovered it, hid the true relationship between survival and mode of transport to a hospital. This example illustrates **Simpson's paradox.**

DEFINITION: Simpson's paradox

An association between two variables that holds for each individual value of a third variable can be changed or even reversed when the data for all values of the third variable are combined. This reversal is called **Simpson's paradox.**

SECTION 1.1 Summary

- The distribution of a categorical variable lists the categories and gives the count (**frequency table**) or percent (**relative frequency table**) of individuals that fall in each category.
- **Pie charts** and **bar graphs** display the distribution of a categorical variable. Bar graphs can also compare any set of quantities measured in the same units. When examining any graph, ask yourself, "What do I see?"
- A **two-way table** of counts organizes data about two categorical variables. Two-way tables are often used to summarize large amounts of information by grouping outcomes into categories.
- The row totals and column totals in a two-way table give the **marginal distributions** of the two individual variables. It is clearer to present these distributions as percents of the table total. Marginal distributions tell us nothing about the relationship between the variables.
- There are two sets of **conditional distributions** for a two-way table: the distributions of the row variable for each value of the column variable, and the distributions of the column variable for each value of the row variable. You may want to use a **side-by-side bar graph** (or possibly a **segmented bar graph**) to display conditional distributions.

- A statistical problem has a real-world setting. You can organize many problems using the four steps **state, plan, do,** and **conclude.**
- To describe the **association** between the row and column variables, compare an appropriate set of conditional distributions. Remember that even a strong association between two categorical variables can be influenced by other variables lurking in the background.
- An association between two variables that holds for each individual value of a third variable can be changed or even reversed when the data for all values of the third variable are combined. This is **Simpson's paradox.**

1.1 TECHNOLOGY CORNER

Analyzing two-way tables ..... page 16

SECTION 1.1 Exercises

9. **Cool car colors** The most popular colors for cars and light trucks change over time. Silver passed green in 2000 to become the most popular color worldwide, then gave way to shades of white in 2007. Here is the distribution of colors for vehicles sold in North America in 2008.[8]

Color	Percent of vehicles
White	20
Black	17
Silver	17
Blue	13
Gray	12
Red	11
Beige/brown	5
Green	3
Yellow/gold	2

(a) What percent of vehicles had colors other than those listed?

(b) Display these data in a bar graph. Be sure to label your axes and title your graph.

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

10. **Spam** Email spam is the curse of the Internet. Here is a compilation of the most common types of spam:[9]

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

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

(b) Display these data in a bar graph. Be sure to label your axes and title your graph.

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

11. **Birth days** Births are not evenly distributed across the days of the week. Here are the average numbers of babies born on each day of the week in the United States in a recent year:[10]

Day	Births
Sunday	7,374
Monday	11,704
Tuesday	13,169
Wednesday	13,038
Thursday	13,013
Friday	12,664
Saturday	8,459

(a) Present these data in a well-labeled bar graph. Would it also be correct to make a pie chart?

(b) Suggest some possible reasons why there are fewer births on weekends.

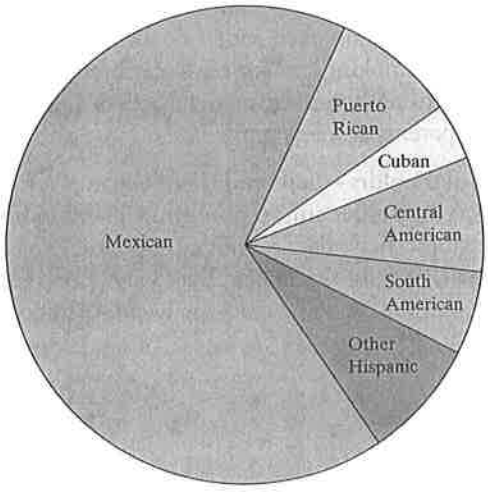
12. **Deaths among young people** Among persons aged 15 to 24 years in the United States, the leading causes of death and number of deaths in a recent year were as follows: accidents, 15,567; homicide, 5359; suicide, 4139; cancer, 1717; heart disease, 1067; congenital defects, 483.[11]

(a) Make a bar graph to display these data.

(b) To make a pie chart, you need one additional piece of information. What is it?

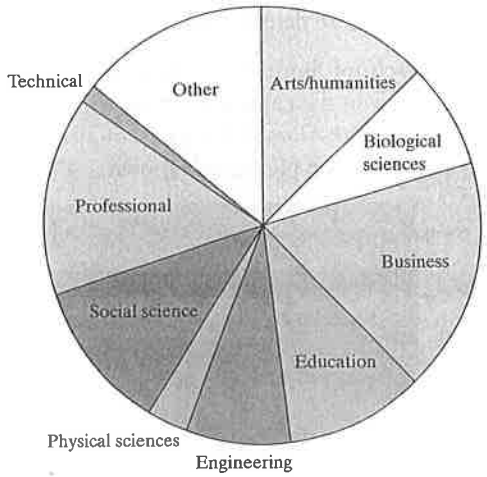
13. **Hispanic origins** Below is a pie chart prepared by the Census Bureau to show the origin of the more than 43 million Hispanics in the United States in 2006.[12] About what percent of Hispanics are Mexican? Puerto Rican?

Percent Distribution of Hispanics by Type: 2006



Comment: You see that it is hard to determine numbers from a pie chart. Bar graphs are much easier to use. (The Census Bureau did include the percents in its pie chart.)

14. **Which major?** About 1.6 million first-year students enroll in colleges and universities each year. What do they plan to study? The pie chart displays data on the percents of first-year students who plan to major in several discipline areas.[13] About what percent of first-year students plan to major in business? In social science?



15. **Buying music online** Young people are more likely than older folk to buy music online. Here are the percents of people in several age groups who bought music online in 2006:[14]

pg 10

Age group	Bought music online
12 to 17 years	24%
18 to 24 years	21%
25 to 34 years	20%
35 to 44 years	16%
45 to 54 years	10%
55 to 64 years	3%
65 years and over	1%

(a) Explain why it is *not* correct to use a pie chart to display these data.

(b) Make a bar graph of the data. Be sure to label your axes and title your graph.

16. **The audience for movies** Here are data on the percent of people in several age groups who attended a movie in the past 12 months:[15]

Age group	Movie attendance
18 to 24 years	83%
25 to 34 years	73%
35 to 44 years	68%
45 to 54 years	60%
55 to 64 years	47%
65 to 74 years	32%
75 years and over	20%

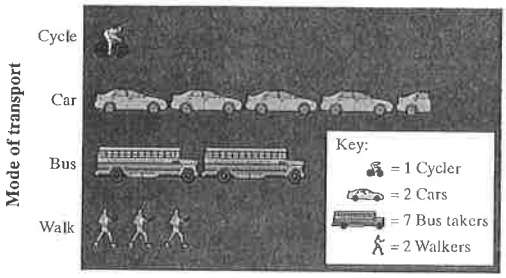
(a) Display these data in a bar graph. Describe what you see.

(b) Would it be correct to make a pie chart of these data? Why or why not?

(c) A movie studio wants to know what percent of the total audience for movies is 18 to 24 years old. Explain why these data do not answer this question.

17. **Going to school** Students in a high school statistics class were given data about the primary method of transportation to school for a group of 30 students. They produced the pictograph shown.

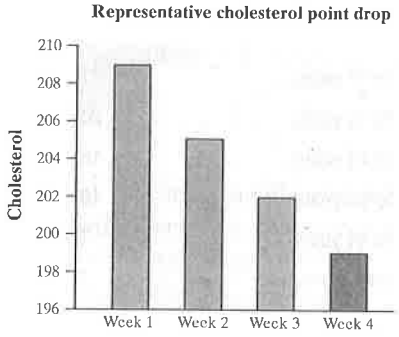
pg 11



(a) How is this graph misleading?

(b) Make a new graph that isn't misleading.

18. **Oatmeal and cholesterol** Does eating oatmeal reduce cholesterol? An advertisement included the following graph as evidence that the answer is "Yes."



(a) How is this graph misleading?

(b) Make a new graph that isn't misleading. What do you conclude about the effect of eating oatmeal on cholesterol reduction?

19. **Attitudes toward recycled products** Recycling is supposed to save resources. Some people think recycled products are lower in quality than other products, a fact that makes recycling less practical. People who actually use a recycled product may have different opinions from those who don't use it. Here are data on attitudes toward coffee filters made of recycled paper among people who do and don't buy these filters:[16]

pg 13

	Think the quality of the recycled product is:		
	Higher	The same	Lower
Buyers	20	7	9
Nonbuyers	29	25	43

(a) How many people does this table describe? How many of these were buyers of coffee filters made of recycled paper?

(b) Give the marginal distribution of opinion about the quality of recycled filters. What percent think the quality of the recycled product is the same or higher than the quality of other filters?

20. **Smoking by students and parents** Here are data from a survey conducted at eight high schools on smoking among students and their parents:[17]

	Neither parent smokes	One parent smokes	Both parents smoke
Student does not smoke	1168	1823	1380
Student smokes	188	416	400

(a) How many students are described in the two-way table? What percent of these students smoke?

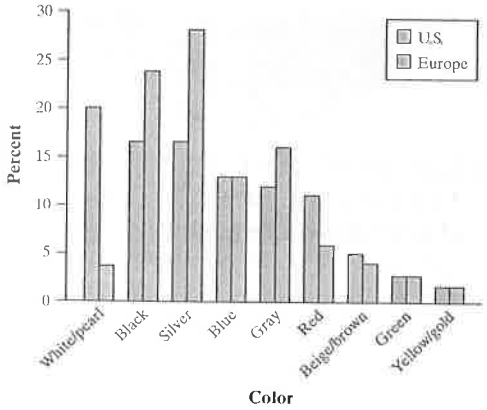
(b) Give the marginal distribution of parents' smoking behavior, both in counts and in percents.

21. **Attitudes toward recycled products** Exercise 19 gives data on the opinions of people who have and have not bought coffee filters made from recycled paper. To see the relationship between opinion and experience with the product, find the conditional distributions of opinion (the response variable) for buyers and nonbuyers. What do you conclude?

pg 15

22. **Smoking by students and parents** Refer to Exercise 20. Calculate three conditional distributions of students' smoking behavior: one for each of the three parental smoking categories. Describe the relationship between the smoking behaviors of students and their parents in a few sentences.

23. **Popular colors—here and there** Favorite vehicle colors may differ among countries. The side-by-side bar graph shows data on the most popular colors of cars in 2008 for the United States and Europe. Write a few sentences comparing the two distributions.



24. **Comparing car colors** Favorite vehicle colors may differ among types of vehicle. Here are data on the most popular colors in 2008 for luxury cars and for SUVs, trucks, and vans.

Color	Luxury cars (%)	SUVs, trucks, vans (%)
Black	22	13
Silver	16	16
White pearl	14	1
Gray	12	13
White	11	25
Blue	7	10
Red	7	11
Yellow/gold	6	1
Green	3	4
Beige/brown	2	6

(a) Make a graph to compare colors by vehicle type.

(b) Write a few sentences describing what you see.

25. **Snowmobiles in the park** Yellowstone National Park surveyed a random sample of 1526 winter visitors to the park. They asked each person whether they owned, rented, or had never used a snowmobile. Respondents were also asked whether they belonged to an environmental organization (like the Sierra Club). The two-way table summarizes the survey responses.

pg 18

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

Do these data provide convincing evidence of an association between environmental club membership and snowmobile use for the population of visitors to Yellowstone National Park? Follow the four-step process.

26. **Angry people and heart disease** People who get angry easily tend to have more heart disease. That's the conclusion of a study that followed a random sample of 12,986 people from three locations for about four years. All subjects were free of heart disease at the beginning of the study. The subjects took the Spielberger Trait Anger Scale test, which measures how prone a person is to sudden anger. Here are data for the 8474 people in the sample who had normal blood pressure. CHD stands for "coronary heart disease." This includes people who had heart attacks and those who needed medical treatment for heart disease.

	Low anger	Moderate anger	High anger	Total
CHD	53	110	27	190
No CHD	3057	4621	606	8284
Total	3110	4731	633	8474

Do these data support the study's conclusion about the relationship between anger and heart disease? Follow the four-step process.

Multiple choice: Select the best answer.

Exercises 27 to 32 refer to the following setting. The National Survey of Adolescent Health interviewed several thousand teens (grades 7 to 12). One question asked was "What do you think are the chances you will be married in the next ten years?" Here is a two-way table of the responses by gender:[18]

	Female	Male
Almost no chance	119	103
Some chance, but probably not	150	171
A 50-50 chance	447	512
A good chance	735	710
Almost certain	1174	756

27. The percent of females among the respondents was

(a) 2625. (c) about 46%. (e) None of these.

(b) 4877. (d) about 54%.

28. Your percent from the previous exercise is part of

(a) the marginal distribution of females.

(b) the marginal distribution of gender.

(c) the marginal distribution of opinion about marriage.

(d) the conditional distribution of gender among adolescents with a given opinion.

(e) the conditional distribution of opinion among adolescents of a given gender.

29. What percent of females thought that they were almost certain to be married in the next ten years?

(a) About 16% (c) About 40% (e) About 61%

(b) About 24% (d) About 45%

30. Your percent from the previous exercise is part of

(a) the marginal distribution of gender.

(b) the marginal distribution of opinion about marriage.

(c) the conditional distribution of gender among adolescents with a given opinion.

(d) the conditional distribution of opinion among adolescents of a given gender.

(e) the conditional distribution of "Almost certain" among females.

31. What percent of those who thought they were almost certain to be married were female?

(a) About 16% (c) About 40% (e) About 61%

(b) About 24% (d) About 45%

32. Your percent from the previous exercise is part of

(a) the marginal distribution of gender.

(b) the marginal distribution of opinion about marriage.

(c) the conditional distribution of gender among adolescents with a given opinion.

(d) the conditional distribution of opinion among adolescents of a given gender.

(e) the conditional distribution of females among those who said "Almost certain."

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

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

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

34.* **Baseball paradox** Most baseball hitters perform differently against right-handed and left-handed pitching. Consider two players, Joe and Moe, both of whom bat right-handed. The table below records their performance against right-handed and left-handed pitchers:

Player	Pitcher	Hits	At-bats
Joe	Right	40	100
	Left	80	400
Moe	Right	120	400
	Left	10	100

(a) Use these data to make a two-way table of player (Joe or Moe) versus outcome (hit or no hit).

(b) Show that Simpson's paradox holds: one player has a higher overall batting average, but the other player hits better against both left-handed and right-handed pitching.

(c) The manager doesn't believe that one player can hit better against both left-handers and right-handers yet have a lower overall batting average. Explain in simple language why this happens to Joe and Moe.

35.* **Race and the death penalty** Whether a convicted murderer gets the death penalty seems to be influenced by the race of the victim. Here are data on 326 cases in which the defendant was convicted of murder:[19]

	White Defendant		Black Defendant		
	White victim	Black victim		White victim	Black victim
Death	19	0	Death	11	6
Not	132	9	Not	52	97

(a) Use these data to make a two-way table of defendant's race (white or black) versus death penalty (yes or no).

(b) Show that Simpson's paradox holds: a higher percent of white defendants are sentenced to death overall, but for both black and white victims a higher percent of black defendants are sentenced to death.

(c) Use the data to explain why the paradox holds in language that a judge could understand..

36. **Fuel economy** (Introduction) Here is a small part of a data set that describes the fuel economy (in miles per gallon) of model year 2009 motor vehicles:

Make and model	Vehicle type	Transmission type	Number of cylinders	City mpg	Highway mpg
Aston Martin Vantage	Two-seater	Manual	8	12	19
Honda Civic	Subcompact	Automatic	4	25	36
Toyota Prius	Midsize	Automatic	4	48	45
Chevrolet Impala	Large	Automatic	6	18	29

(a) What are the individuals in this data set?

(b) What variables were measured? Identify each as categorical or quantitative.

*These exercises relate to the optional content on Simpson's paradox.

1.2

Displaying Quantitative Data with Graphs

In Section 1.2, you'll learn about:

- Dotplots
- Describing shape
- Comparing distributions
- Stemplots
- Histograms
- Using histograms wisely

To display the distribution of a categorical variable, use a bar graph or a pie chart. How can we picture the distribution of a quantitative variable? In this section, we present several types of graphs that can be used to display quantitative data.

Dotplots

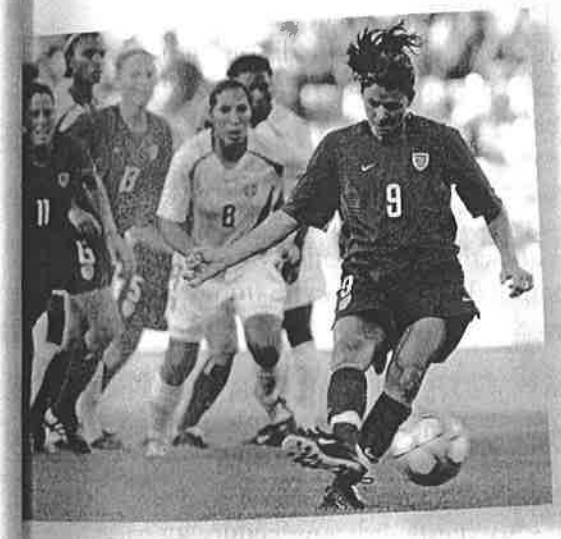
Dotplot

One of the simplest graphs to construct and interpret is a **dotplot.** Each data value is shown as a dot above its location on a number line. We'll show how to make a dotplot using some sports data.

EXAMPLE

Goooooaaaaalllllll!

How to make a dotplot



How good was the 2004 U.S. women's soccer team? With players like Brandi Chastain, Mia Hamm, and Briana Scurry, the team put on an impressive showing en route to winning the gold medal at the 2004 Olympics in Athens. Here are data on the number of goals scored by the team in 34 games played during the 2004 season:[20]

3 0 2 7 8 2 4 3 5 1 1 4 5 3 1 1 3
3 3 2 1 2 2 2 4 3 5 6 1 5 5 1 1 5

Here are the steps in making a dotplot:

- *Draw a horizontal axis (a number line) and label it with the variable name.* In this case, the variable is number of goals scored.
- *Scale the axis.* Start by looking at the minimum and maximum values of the variable. For these data, the minimum number of goals scored was 0, and the maximum was 8. So we mark our scale from 0 to 8, with tick marks at every whole-number value.
- *Mark a dot above the location on the horizontal axis corresponding to each data value.* Figure 1.9 displays a completed dotplot for the soccer data.

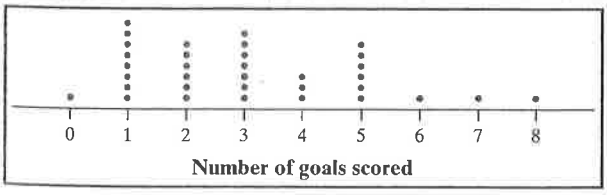


FIGURE 1.9 A dotplot of goals scored by the U.S. women's soccer team in 2004.

Making a graph is not an end in itself. The purpose of a graph is to help us understand the data. After you make a graph, always ask, "What do I see?" Here is a general strategy for interpreting graphs of quantitative data.

How to Examine the Distribution of a Quantitative Variable

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

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

We'll learn more formal ways of describing shape, center, and spread and identifying outliers shortly. For now, let's use our informal understanding of these ideas to examine the graph in Figure 1.9.

Mode

Shape: The dotplot has a peak at 1. This indicates that the team's most frequent number of goals scored in games that season (known as the **mode**) was 1. In most of its games, the U.S. women's soccer team scored between 1 and 5 goals. However, the distribution has a long tail to the right. (Later, we will describe the shape of Figure 1.9 as *skewed to the right.*)

Center: We can describe the center by finding a value that divides the observations so that about half take larger values and about half take smaller values. This value is called the *median* of the distribution. In Figure 1.9, the median is 3. That is, in a typical game during the 2004 season, the U.S. women's soccer team scored about 3 goals. Of course, we could also summarize the center of the distribution by calculating the average (*mean*) number of goals scored per game. For the 2004 season, the team's mean was 3.06 goals.

Range

Spread: The spread of a distribution tells us how much *variability* there is in the data. One way to describe the variability is to give the smallest and largest values. The spread in Figure 1.9 is from 0 goals to 8 goals scored. Alternatively, we can compute the **range** of the distribution by subtracting the smallest value from the largest value. For these data, the range is $8 - 0 = 8$ goals.

When describing a distribution of quantitative data, don't forget your SOCS (shape, outliers, center, spread)!

Outliers: Was the game in which the women's team scored 8 goals an outlier? How about the team's 7-goal game? These values differ somewhat from the overall pattern. However, they don't clearly stand apart from the rest of the distribution. For now, let's agree to call attention only to potential outliers that suggest something special about an observation. In Section 1.3, we'll establish a procedure for determining whether a particular data value is an outlier.

EXAMPLE

Are You Driving a Gas Guzzler?

Interpreting a dotplot

The Environmental Protection Agency (EPA) is in charge of determining and reporting fuel economy ratings for cars (think of those large window stickers on a new car). For years, consumers complained that their actual gas mileages were noticeably lower than the values reported by the EPA. It seems that the EPA's tests—all of which are done on computerized devices to ensure consistency—did not consider things like outdoor temperature, use of the air conditioner, or realistic accel-

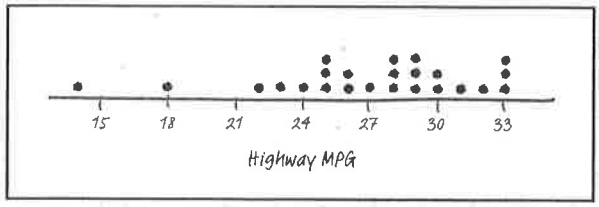
eration and braking by drivers. In 2008, the EPA changed the method for measuring a vehicle's fuel economy to try to give more accurate estimates.

The table below displays the EPA estimates of highway gas mileage in miles per gallon (mpg) for a sample of 24 model year 2009 midsize cars.

Model	Mpg	Model	Mpg	Model	Mpg
Acura RL	22	Dodge Avenger	30	Mercury Milan	29
Audi A6 Quattro	23	Hyundai Elantra	33	Mitsubishi Galant	27
Bentley Arnage	14	Jaguar XF	25	Nissan Maxima	26
BMW 528I	28	Kia Optima	32	Rolls Royce Phantom	18
Buick Lacrosse	28	Lexus GS 350	26	Saturn Aura	33
Cadillac CTS	25	Lincoln MKZ	28	Toyota Camry	31
Chevrolet Malibu	33	Mazda 6	29	Volkswagen Passat	29
Chrysler Sebring	30	Mercedes-Benz E350	24	Volvo S80	25

Source: 2009 Fuel Economy Guide, from the U.S. Environmental Protection Agency's Web site at www.fueleconomy.gov.

Here is a dotplot of the data:



PROBLEM: Describe the shape, center, and spread of the distribution. Are there any outliers?

SOLUTION: Don't forget your SOCS (shape, outliers, center, spread)! **Shape:** In the dotplot, we can see three *clusters* of values: cars that get around 25 mpg, cars that get about 28 to 30 mpg, and cars that get around 33 mpg. We can also see large *gaps* between the Acura RL at 22 mpg, the Rolls Royce Phantom at 18 mpg, and the Bentley Arnage at 14 mpg. **Center:** The median is 28. So a "typical" model year 2009 midsize car got about 28 miles per gallon on the highway. **Spread:** The highest value is 33 mpg and the lowest value is 14 mpg. The range is $33 - 14 = 19$ mpg. **Outliers:** We see two midsize cars with unusually low gas mileage ratings—the Bentley Arnage (14 mpg) and the Rolls Royce Phantom (18 mpg). These cars are potential outliers.

For Practice *Try Exercise 39*

Describing Shape

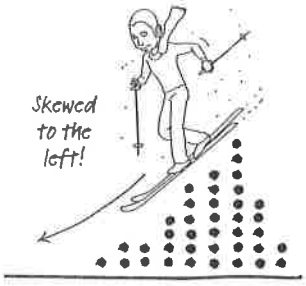
When you describe a distribution's shape, concentrate on the main features. Look for major peaks, not for minor ups and downs in the graph. Look for clusters of values and obvious gaps. Look for potential outliers, not just for the smallest and largest observations. Look for rough **symmetry** or clear **skewness.**

For brevity, we sometimes say "left-skewed" instead of "skewed to the left" and "right-skewed" instead of "skewed to the right." 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."

DEFINITION: Symmetric and skewed distributions

A distribution is roughly **symmetric** if the right and left sides of the graph are approximately mirror images of each other.

A distribution is **skewed to the right** if the right side of the graph (containing the half of the observations with larger values) is much longer than the left side. It is **skewed to the left** if the left side of the graph is much longer than the right side.



For his own safety, which way should Mr. Starnes go "skewing"?

The direction of skewness is the direction of the long tail, not the direction where most observations are clustered. See the drawing in the margin for a cute but corny way to help you keep this straight.

CAUTION

EXAMPLE

Die Rolls and Quiz Scores

Describing shape

Figure 1.10 displays dotplots for two different sets of quantitative data. Let's practice describing the shapes of these distributions. Figure 1.10(a) shows the results of rolling a pair of fair, six-sided dice and finding the sum of the up-faces 100 times. This distribution is roughly symmetric. The dotplot in Figure 1.10(b) shows the scores on an AP Statistics class's first quiz. This distribution is skewed to the left.

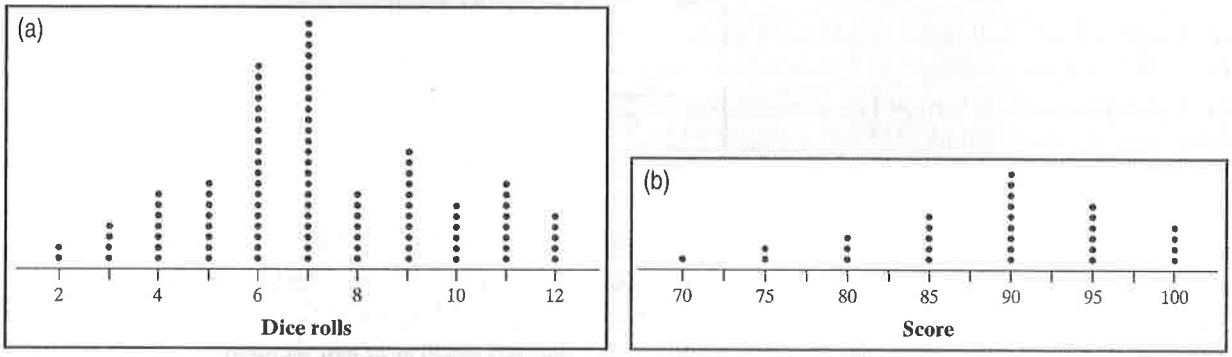


FIGURE 1.10 Dotplots displaying different shapes: (a) roughly symmetric; (b) skewed to the left.

Unimodal

Although the dotplots in the previous example have different shapes, they do have something in common. Both are **unimodal,** that is, they have a single peak: the graph of dice rolls at 7 and the graph of quiz scores at 90. (We don't count minor ups and downs in a graph, like the "bumps" at 9 and 11 in the dice rolls dotplot, as "peaks.") Figure 1.11 is a dotplot of the duration (in minutes) of 220 eruptions of the Old Faithful geyser. We would describe this distribution's shape as **bimodal** since it has two clear peaks: one near 2 minutes and the other near 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.**)

Bimodal

Multimodal

THINK ABOUT IT

What shape will the graph have? Some variables have distributions with predictable shapes. Many biological measurements on individuals from the same species and gender—lengths of bird bills, heights of young women—have symmetric distributions. Salaries and home prices, on the other hand, usually have right-skewed distributions. There are many moderately priced houses, for example, but the few very expensive mansions give the distribution of house prices a

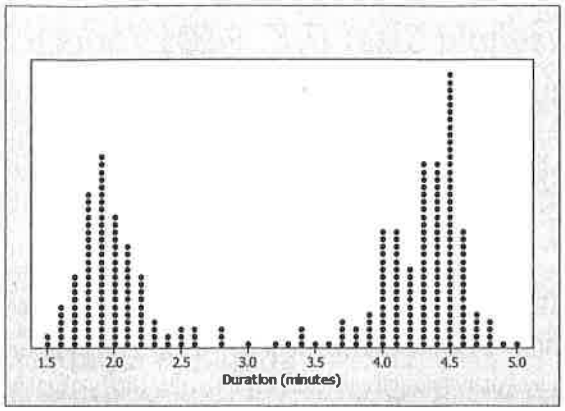
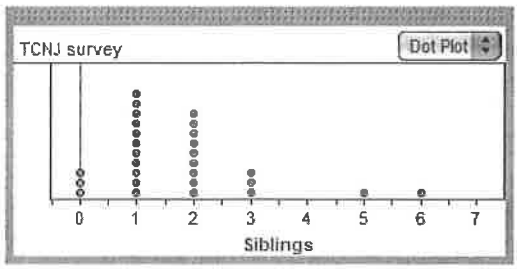


FIGURE 1.11 Dotplot displaying duration (in minutes) of Old Faithful eruptions. This graph has a bimodal shape.

strong right-skew. Many distributions have irregular shapes that are neither symmetric nor skewed. Some data show other patterns, such as the two peaks in Figure 1.11. Use your eyes, describe the pattern you see, and then try to explain the pattern.

CHECK YOUR UNDERSTANDING

The Fathom dotplot displays data on the number of siblings reported by each student in a statistics class.



1. Describe the shape of the distribution.
2. Describe the center of the distribution.
3. Describe the spread of the distribution.
4. Identify any potential outliers.

Comparing Distributions

Some of the most interesting statistics questions involve comparing two or more groups. Which of two popular diets leads to greater long-term weight loss? Who texts more—males or females? Does the number of people living in a household differ among countries? As the following example suggests, you should always discuss shape, center, spread, and possible outliers whenever you compare distributions of a quantitative variable.

EXAMPLE

Household Size: U.K. versus South Africa

Comparing distributions

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 used CensusAtSchool's "Random Data Selector" to choose 50 students from each country. Figure 1.12 is a dotplot of the household sizes reported by the survey respondents.

PROBLEM: Compare the distributions of household size for these two countries.

SOLUTION: Don't forget your SOCS! **Shape:** The distribution of household size for the U.K. sample is roughly symmetric and unimodal, while the distribution for the South Africa sample is skewed to the right and unimodal. **Center:** Household sizes for the South African students tended to be larger than for the U.K. students. The median household sizes for the two groups are 6 people and 4 people, respectively. **Spread:** There is more variability (greater spread) in the household sizes for the South African students than for the U.K. students. The range for the South African data is $26 - 3 = 23$ people, while the range for the U.K. data is $6 - 2 = 4$ people. **Outliers:** There don't appear to be any potential outliers in the U.K. distribution. The South African distribution has two potential outliers in the right tail of the distribution—students who reported living in households with 15 and 26 people. (The U.K. households with 2 people actually *will* be classified as outliers when we introduce a procedure in the next section.)

AP EXAM TIP When comparing distributions of quantitative data, it's not enough just to list values for the center and spread of each distribution. You have to explicitly *compare* these values, using words like "greater than," "less than," or "about the same as."

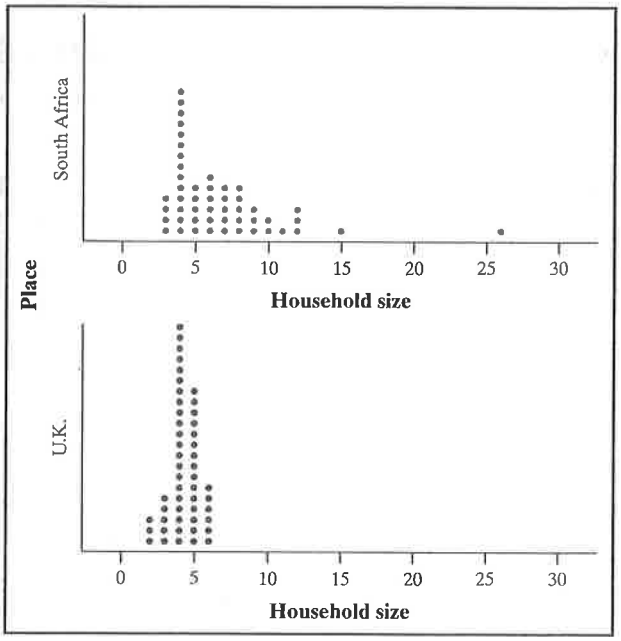


FIGURE 1.12 Dotplot of household size for random samples of 50 students from the United Kingdom and South Africa.

For Practice *Try Exercise 43*

Notice that we discussed the distributions of household size only for the two *samples* of 50 students in the previous example. 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 a few chapters 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 students in our samples were chosen at random.

Stemplots

Another simple graphical display for fairly small data sets is a **stemplot** (also called a stem-and-leaf plot). Stemplots give us a quick picture of the shape of a distribution while including the actual numerical values in the graph. Here's an example that shows how to make a stemplot.

EXAMPLE

How Many Shoes?

Making a stemplot

How many pairs of shoes does a typical teenager have? To find out, a group of AP Statistics students conducted a survey. They selected a random sample of 20 female students from their school. Then they recorded the number of pairs of shoes that each respondent reported having. Here are the data:

50	26	26	31	57	19	24	22	23	38
13	50	13	34	23	30	49	13	15	51

Here are the steps in making a stemplot. Figure 1.13 displays the process.

- *Separate each observation into a stem, consisting of all but the final digit, and a leaf, the final digit. Write the stems in a vertical column with the smallest at the top, and 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.* For these data, the tens digits are the stems, and the ones digits are the leaves. The stems run from 1 to 5.
- *Write each leaf in the row to the right of its stem.* For example, the female student with 50 pairs of shoes would have stem 5 and leaf 0, while the student with 31 pairs of shoes would have stem 3 and leaf 1.
- *Arrange the leaves in increasing order out from the stem.*
- *Provide a key that explains in context what the stems and leaves represent.*

1	1	93335	1	33359	Key: 4|9 represents a female student who reported having 49 pairs of shoes.
2	2	664233	2	233466	
3	3	1840	3	0148	
4	4	9	4	9	
5	5	0701	5	0017	
Stems		*Add leaves*		*Order leaves*	*Add a key*

FIGURE 1.13 Making a stemplot of the shoe data. (1) Write the stems. (2) Go through the data and write each leaf on the proper stem. (3) Arrange the leaves on each stem in order out from the stem. (4) Add a key.

The AP Statistics students in the previous example also collected data from a random sample of 20 male students at their school. Here are the numbers of pairs of shoes reported by each male in the sample:

14	7	6	5	12	38	8	7	10	10
10	11	4	5	22	7	5	10	35	7

What would happen if we tried the same approach as before: using the first digits as stems and the last digits as leaves? The completed stemplot is shown in Figure 1.14(a). What shape does this distribution have? It is difficult to tell with so few stems. We can get a better picture of male shoe ownership by **splitting stems.**

Splitting stems

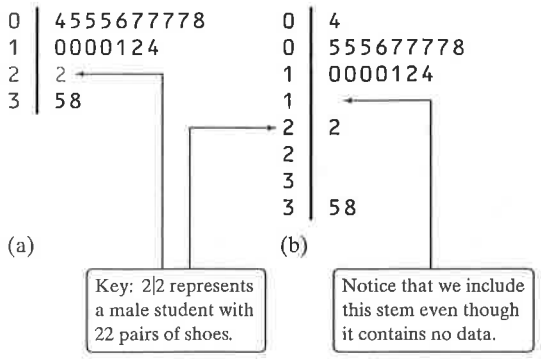


FIGURE 1.14 Two stemplots showing the male shoe data. Figure 1.14(b) improves on the stemplot of Figure 1.14(a) by splitting stems.

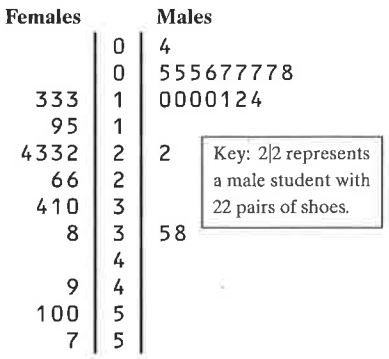


FIGURE 1.15 Back-to-back stemplot comparing numbers of pairs of shoes for male and female students at a school.

In Figure 1.14(a), the values from 0 to 9 are placed on the "0" stem. Figure 1.14(b) shows another stemplot of the same data. This time, values having leaves 0 through 4 are placed on one stem, while values ending in 5 through 9 are placed on another stem. Now we can see the single peak, the cluster of values between 4 and 14, and the large gap between 22 and 35 more clearly.

Back-to-back stemplot

What if we want to compare the number of pairs of shoes that males and females have? That calls for a **back-to-back stemplot** with common stems. The leaves on each side are ordered out from the common stem. Figure 1.15 is a back-to-back stemplot for the male and female shoe data. Note that we have used the split stems from Figure 1.14(b) as the common stems. The values on the right are the male data from Figure 1.14(b). The values on the left are the female data, ordered out from the stem from right to left. We'll ask you to compare these two distributions shortly.

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

- Stemplots do not work well for large data sets, where each stem must hold a large number of leaves.
- There is no magic number of stems to use, but five is a good minimum. Too few or too many stems will make it difficult to see the distribution's shape.
- If you split stems, be sure that each stem is assigned an equal number of possible leaf digits (two stems, each with five possible leaves; or five stems, each with two possible leaves).
- You can get more flexibility by rounding the data so that the final digit after rounding is suitable as a leaf. Do this when the data have too many digits. For example, in reporting teachers' salaries, using all five digits (for example, $42,549) would be unreasonable. It would be better to round to the nearest thousand and use 4 as a stem and 3 as a leaf.

Instead of rounding, you can also *truncate* (remove one or more digits) when data have too many digits. The teacher's salary of $42,549 would truncate to $42,000.

CHECK YOUR UNDERSTANDING

1. Use the back-to-back stemplot in Figure 1.15 to write a few sentences comparing the number of pairs of shoes owned by males and females. Be sure to address shape, center, spread, and outliers.

Multiple choice: Select the best answer for Questions 2 through 4.
Here is a stemplot of the percents of residents aged 65 and older in the 50 states and the District of Columbia. The stems are whole percents and the leaves are tenths of a percent.

```
 6 | 8
 7 |
 8 | 8
 9 | 79
10 | 08
11 | 15566
12 | 012223444457888999
13 | 01233333444899
14 | 02666
15 | 23
16 | 8
```

Key: 8|8 represents a state in which 8.8% of residents are 65 and older.

2. The low outlier is Alaska. What percent of Alaska residents are 65 or older?
(a) 0.68 (b) 6.8 (c) 8.8 (d) 16.8 (e) 68
3. Ignoring the outlier, the shape of the distribution is
(a) skewed to the right (c) skewed to the left. (e) skewed to the middle.
(b) roughly symmetric (d) bimodal.
4. The center of the distribution is close to
(a) 13.3%. (b) 12.8%. (c) 12.0%. (d) 11.6%. (e) 6.8% to 16.8%.

Histograms

Quantitative variables often take many values. A graph of the distribution is clearer if nearby values are grouped together. The most common graph of the distribution of one quantitative variable is a **histogram.** Let's look at how to make a histogram using data on foreign-born residents in the United States.

Histogram

EXAMPLE

Foreign-Born Residents

Making a histogram

What percent of your home state's residents were born outside the United States? The country as a whole has 12.5% foreign-born residents, but the states vary from 1.2% in West Virginia to 27.2% in California. The table below presents the data for all 50 states.[21] The *individuals* in this data set are the states. The *variable* is the percent of a state's residents who are foreign-born. It's much easier to see from a graph than from the table how your state compares with other states.

State	Percent	State	Percent	State	Percent
Alabama	2.8	Louisiana	2.9	Ohio	3.6
Alaska	7.0	Maine	3.2	Oklahoma	4.9
Arizona	15.1	Maryland	12.2	Oregon	9.7
Arkansas	3.8	Massachusetts	14.1	Pennsylvania	5.1
California	27.2	Michigan	5.9	Rhode Island	12.6
Colorado	10.3	Minnesota	6.6	South Carolina	4.1
Connecticut	12.9	Mississippi	1.8	South Dakota	2.2
Delaware	8.1	Missouri	3.3	Tennessee	3.9
Florida	18.9	Montana	1.9	Texas	15.9
Georgia	9.2	Nebraska	5.6	Utah	8.3
Hawaii	16.3	Nevada	19.1	Vermont	3.9
Idaho	5.6	New Hampshire	5.4	Virginia	10.1
Illinois	13.8	New Jersey	20.1	Washington	12.4
Indiana	4.2	New Mexico	10.1	West Virginia	1.2
Iowa	3.8	New York	21.6	Wisconsin	4.4
Kansas	6.3	North Carolina	6.9	Wyoming	2.7
Kentucky	2.7	North Dakota	2.1		

Here are the steps in making a histogram:

• *Divide the range of the data into classes of equal width.* The data in the table vary from 1.2 to 27.2, so we might choose to use classes of width 5, beginning at 0:

0–5 5–10 10–15 15–20 20–25 25–30

But we need to specify the classes so that each individual falls into exactly one class. For instance, what if a state had exactly 5.0% of its residents born outside the United States? Since a value of 0.0% would go in the 0–5 class, we'll agree to place a value of 5.0% in the 5–10 class, a value of 10.0% in the 10–15 class, and so on. In reality, then, our classes for the percent of foreign-born residents in the states are

0 to <5 5 to <10 10 to <15 15 to <20 20 to <25 25 to <30

• *Find the count (frequency) or percent (relative frequency) of individuals in each class.* Here is a frequency table and a relative frequency table for these data:

Notice that the frequencies add to 50, the number of individuals (states) in the data, and that the relative frequencies add to 100%.

Frequency table	
Class	**Count**
0 to < 5	20
5 to < 10	13
10 to < 15	9
15 to < 20	5
20 to < 25	2
25 to < 30	1
Total	50

Relative frequency table	
Class	**Percent**
0 to < 5	40
5 to < 10	26
10 to < 15	18
15 to < 20	10
20 to < 25	4
25 to < 30	2
Total	100

• *Label and scale your axes and draw the histogram.* Label the horizontal axis with the variable whose distribution you are displaying. That's the percent of a state's residents who are foreign-born. The scale on the horizontal axis runs from 0 to 30 because that is the span of the classes we chose. The vertical axis contains the scale of counts or percents. Each bar represents a class. The base of the bar covers the class, and the bar height is the class frequency or relative frequency. Draw the bars with no horizontal space between them unless a class is empty, so that its bar has height zero.

Figure 1.16(a) shows a completed frequency histogram; Figure 1.16(b) shows a completed relative frequency histogram. The two graphs look identical except for the vertical scales.

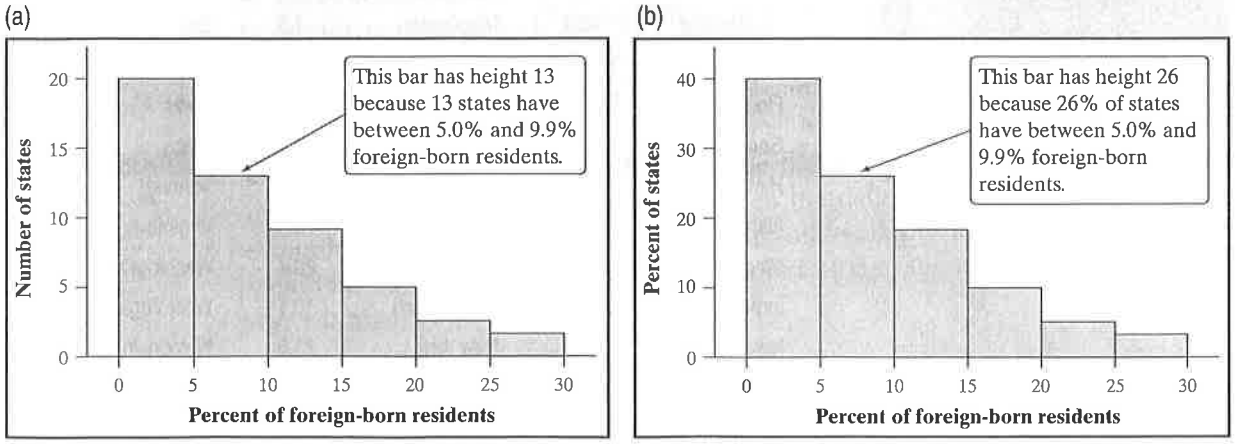


FIGURE 1.16 (a) Frequency histogram and (b) relative frequency histogram of the distribution of the percent of foreign-born residents in the 50 states.

What do the histograms in Figure 1.16 tell us about the percent of foreign-born residents in the states? To find out, we follow our familiar routine: describe the pattern and look for any departures from the pattern.

Shape: The distribution is skewed to the right. A majority of states have fewer than 10% foreign-born residents, but several states have much higher percents, so that the graph extends quite far to the right of its peak. The distribution has a *single peak* at the left, which represents states in which between 0% and 4.9% of residents are foreign-born.

Center: From the graph, we see that the midpoint (median) would fall somewhere in the 5.0% to 9.9% class. Remember that we're looking for the value having 25 states with smaller percents foreign-born and 25 with larger. (Arranging the observations from the table in order of size shows that the median is 6.1%.)

Spread: The histogram shows that the percent of foreign-born residents in the states varies from less than 5% to over 25%. (Using the data in the table, we see that the range is 27.2% − 1.2% = 26.0%.)

Outliers: We don't see any observations outside the overall single-peaked, right-skewed pattern of the distribution.

Figure 1.17 shows (a) a frequency histogram and (b) a relative frequency histogram of the same distribution, with classes half as wide. The new classes are 0–2.4, 2.5–4.9, etc. Now California, at 27.2%, stands out as a potential outlier in the right tail. The choice of classes in a histogram can influence the appearance of a distribution. Histograms with more classes show more detail but may have a less clear pattern.

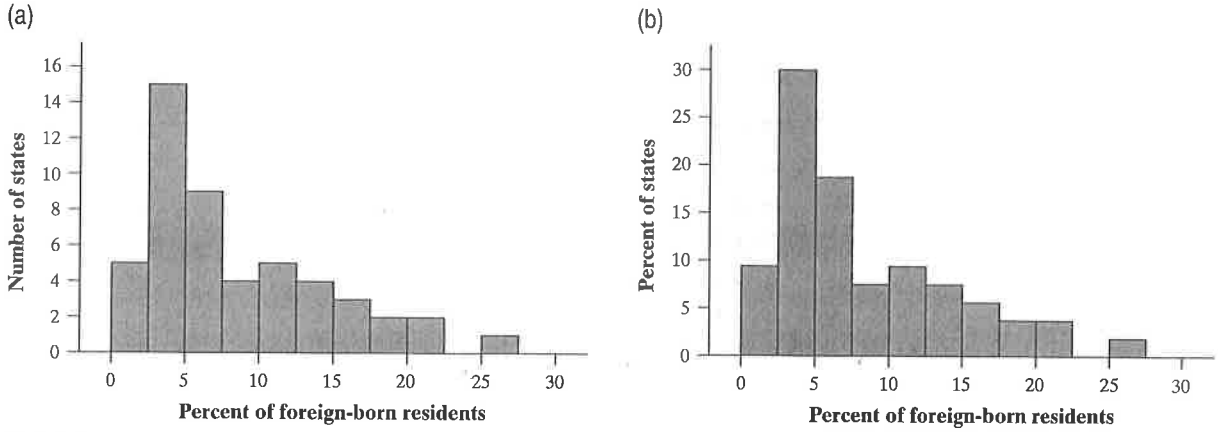


FIGURE 1.17 (a) Frequency histogram and (b) relative frequency histogram of the distribution of the percent of foreign-born residents in the 50 states, with classes half as wide as in Figure 1.16.

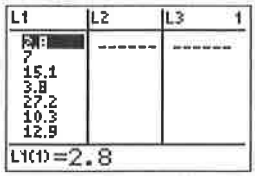
Statistical software and graphing calculators will choose the classes for you. The default choice is a good starting point, but you should adjust the classes to suit your needs. To see what we're talking about, launch the *One-Variable Statistical Calculator* applet at the book's Web site, www.whfreeman.com/tps4e. Select the "Percent of foreign-born residents" data set, and then click on the "Histogram" tab. You can change the number of classes by dragging the horizontal axis with your mouse or pointing device. By doing so, it's easy to see how the choice of classes affects the histogram. *Bottom line: Use your judgment in choosing classes to display the shape.*

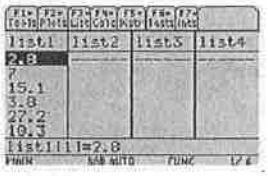
TECHNOLOGY CORNER Histograms on the calculator

TI-83/84 TI-89

1. Enter the data for the percent of state residents born outside the United States in your Statistics/List Editor.

- Press STAT and choose 1:Edit... / Press APPS and select Stats/List Editor.
- Type the values into list L1 / Type the values into list1.





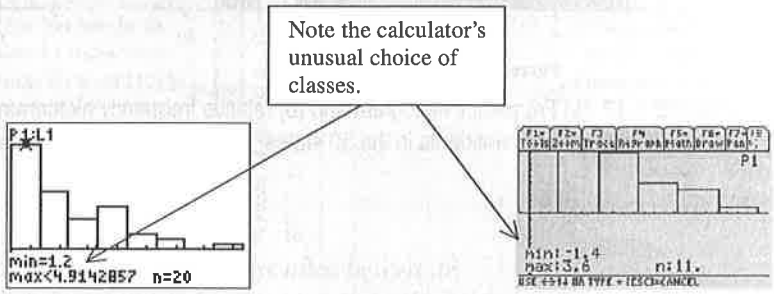
2. Set up a histogram in the Statistics Plots menu.

- Press 2nd Y = (STAT PLOT) / Press F2 and choose 1:Plot Setup...
- Press ENTER or 1 to go into Plot1 / With Plot1 highlighted, press F1 to define.

- Adjust the settings as shown. / Adjust the settings as shown.

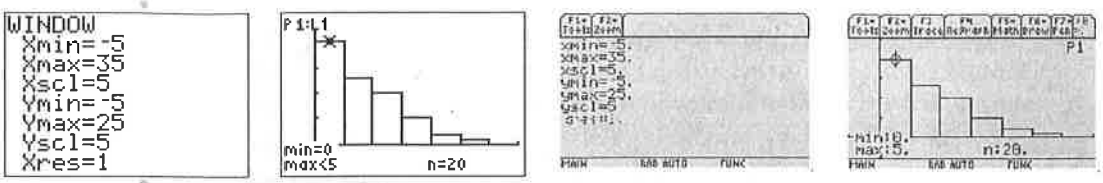
3. Use ZoomStat (ZoomData on the TI-89) to let the calculator choose classes and make a histogram.

- Press ZOOM and choose 9:ZoomStat. / Press F5 (ZoomData).
- Press TRACE and ◄ ► to examine the classes. / Press F3 (Trace) and ◄ ► to examine the classes.



4. Adjust the classes to match those in Figure 1.16, and then graph the histogram.

- Press WINDOW and enter the values shown. / Press ♦ F2 (WINDOW) and enter the values shown.
- Press GRAPH / Press ♦ F3 (GRAPH)
- Press TRACE and ◄ ► to examine the classes. / Press F3 (trace and ◄ ► to examine the classes)



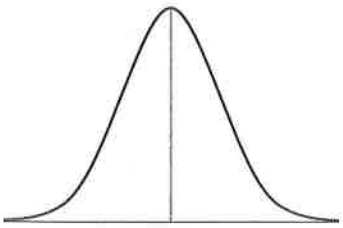
5. See if you can match the histogram in Figure 1.17.

TI-Nspire instructions in Appendix B

Here are some important things to consider when you are constructing a histogram:

AP EXAM TIP If you're asked to make a graph on a free-response question, be sure to label and scale your axes. Unless your calculator shows labels and scaling, don't just transfer a calculator screen shot to your paper.

- Our eyes respond to the area of the bars in a histogram, so *be sure to choose classes that are all the same width.* Then area is determined by height, and all classes are fairly represented.
- There is no one right choice of the classes in a histogram. Too few classes will give a "skyscraper" graph, with all values in a few classes with tall bars. Too many will produce a "pancake" graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution. Five classes is a good minimum.



CHECK YOUR UNDERSTANDING

Many people believe that the distribution of IQ scores follows a "bell curve," like the one shown in the margin. But is this really how such scores are distributed? The IQ scores of 60 fifth-grade students chosen at random from one school are shown below.[22]

145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

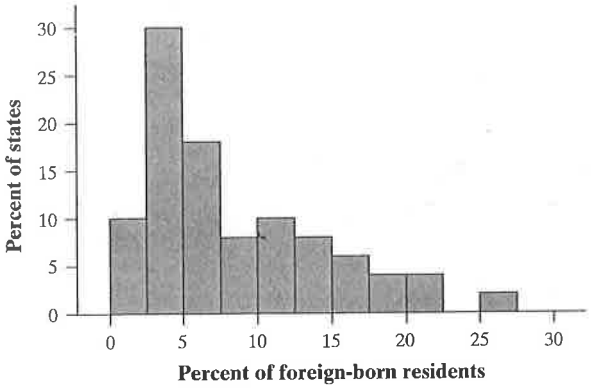
1. Construct a histogram that displays the distribution of IQ scores effectively.
2. Describe what you see. Is the distribution bell-shaped?

Using Histograms Wisely

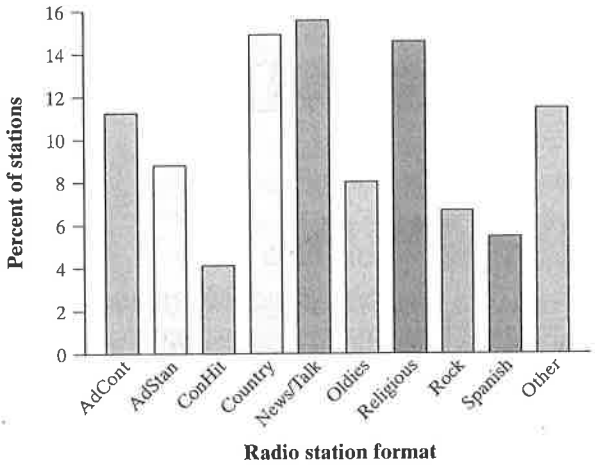
We offer several cautions based on common mistakes students make when using histograms.

1. *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. The horizontal axis of a histogram is marked in the units of measurement for the variable. A bar graph is used to display the distribution of a categorical variable or to compare the sizes of different quantities. The horizontal axis of a bar graph identifies the categories or quantities being compared. Draw bar graphs with blank space between the bars to separate the items being compared. Draw histograms with no space, to show the equal-width classes. For comparison, here is one of each type of graph from previous examples.

Histogram

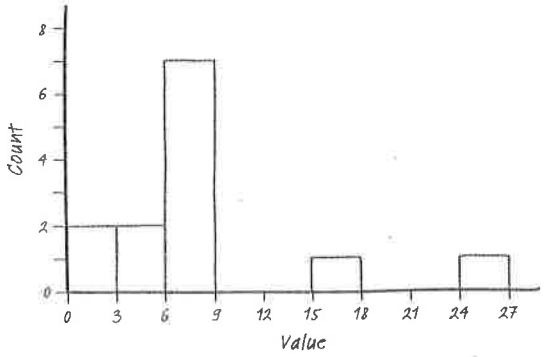


Bar graph



2. *Don't use counts (in a frequency table) or percents (in a relative frequency table) as data.* Below is a frequency table displaying the lengths (number of letters) of the first 100 words in a journal article.

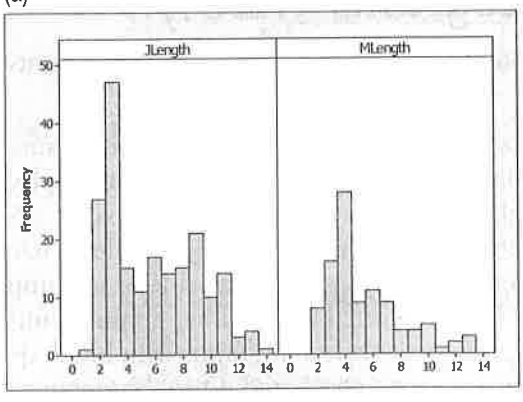
Length:	1	2	3	4	5	6	7	8	9	10	11	12	13
Count:	1	15	25	7	5	7	8	7	7	6	8	3	1



Billy made the histogram shown to display these data. Can you see what Billy did wrong? (He used the counts as data when drawing the histogram—so there were two counts of 1, two counts between 3 and 5, and so on.) Question 1 in the Check Your Understanding below asks you to make a correct graph.

3. *Use percents instead of counts on the vertical axis when comparing distributions with different numbers of observations.* Mary was interested in comparing the reading levels of a medical journal and an airline's in-flight magazine. She counted the number of letters in the first 200 words of an article in the medical journal and of the first 100 words of an article in the airline magazine. Mary then used Minitab statistical software to produce the histograms shown in Figure 1.18(a). This figure is misleading—it compares frequencies, but the two samples were of very different

(a)



(b)

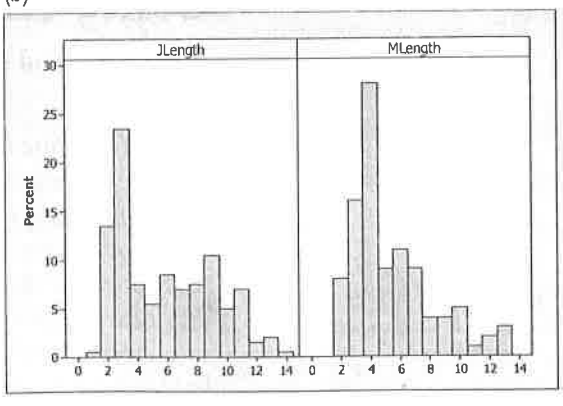
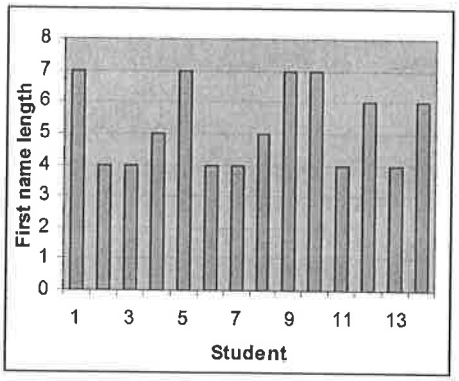


FIGURE 1.18 Two sets of histograms comparing word lengths in articles from a journal and from an airline magazine. In (a), the vertical scale uses frequencies. The graph in (b) fixes this problem by using percents on the vertical scale.

samples were of very different sizes (100 and 200). Using the same data, Mary's teacher produced the histograms in Figure 1.18(b). By using relative frequencies, this figure provides an accurate comparison of word lengths in the two samples.

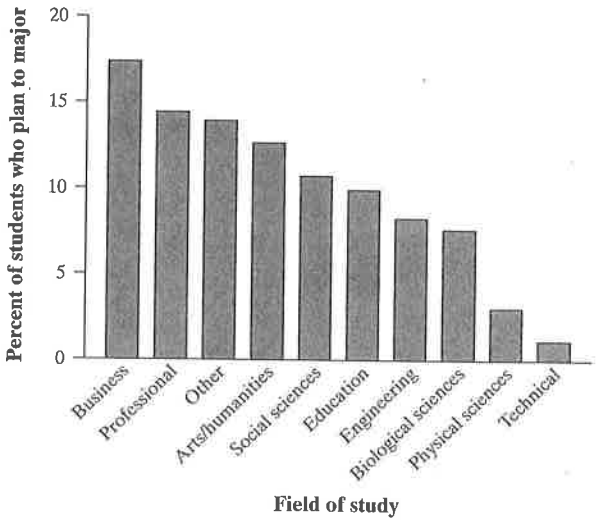
4. *Just because a graph looks nice, it's not necessarily a meaningful display of data.* The students in a small statistics class recorded the number of letters in their first names. One student entered the data into an Excel spreadsheet and then used Excel's "chart maker" to produce the graph shown. What kind of graph is this? It's neither a bar graph nor a histogram. Both of these types of graphs display the number or percent of individuals in a given category or class. This graph shows the individual data values, in the order that they were entered into the spreadsheet. It is not a very meaningful display of the data.



CHECK YOUR UNDERSTANDING

1. Draw a correct histogram to replace Billy's graph of the word length data from Caution 2.
2. Draw a more meaningful graph of the first-name length data from Caution 4.

Questions 3 and 4 relate to the following setting. About 1.6 million first-year students enroll in colleges and universities each year. What do they plan to study? The graph displays data on the percents of first-year students who plan to major in several discipline areas.[23]



3. Is this a bar graph or a histogram? Explain.
4. Would it be correct to describe this distribution as right-skewed? Why or why not?

SECTION 1.2 Summary

- You can use a **dotplot, stemplot,** or **histogram** to show the distribution of a quantitative variable. A dotplot displays individual values on a number line. Stemplots separate each observation into a stem and a one-digit leaf. Histograms plot the counts (frequencies) or percents (relative frequencies) of values in equal-width classes.
- When examining any graph, look for an **overall pattern** and for notable **departures** from that pattern. **Shape, center, and spread** describe the overall pattern of the distribution of a quantitative variable. **Outliers** are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them. Don't forget your SOCS!
- Some distributions have simple shapes, such as **symmetric** or **skewed.** The number of **modes** (major peaks) is another aspect of overall shape. Not all distributions have a simple overall shape, especially when there are few observations.
- When comparing distributions of quantitative data, be sure to discuss shape, center, spread, and possible outliers.
- Remember: histograms are for quantitative data; bar graphs are for categorical data. Also, be sure to use relative frequency histograms when comparing data sets of different sizes.

1.2 TECHNOLOGY CORNER

Histograms on the calculator ..... page 38

TI-Nspire instructions in Appendix B

SECTION 1.2 Exercises

37. **Feeling sleepy?** Students in a college statistics class responded to a survey designed by their teacher. One of the survey questions was "How much sleep did you get last night?" Here are the data (in hours):

9	6	8	6	8	8	6	6.5	6	7	9	4	3	4
5	6	11	6	3	6	6	10	7	8	4.5	9	7	7

(a) Make a dotplot to display the data.

(b) Describe the overall pattern of the distribution and any deviations from that pattern.

38. **Olympic gold!** The following table displays the total number of gold medals won by a sample of countries in the 2008 Summer Olympic Games in China.

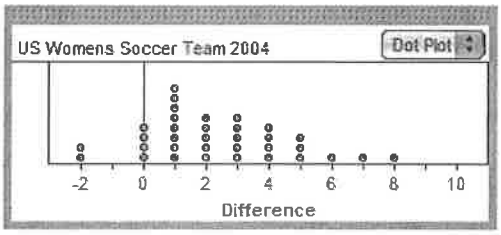
Country	Gold medals	Country	Gold medals
Sri Lanka	0	Thailand	2
China	51	Kuwait	0
Vietnam	0	Bahamas	0
Great Britain	19	Kenya	5
Norway	3	Trinidad and Tobago	0
Romania	4	Greece	0
Switzerland	2	Mozambique	0
Armenia	0	Kazakhstan	2
Netherlands	7	Denmark	2
India	0	Latvia	1
Georgia	3	Czech Republic	3
Kyrgyzstan	0	Hungary	3
Costa Rica	0	Sweden	0
Brazil	3	Uruguay	0
Uzbekistan	1	United States	36

(a) Make a dotplot to display these data. Describe the overall pattern of the distribution and any deviations from that pattern.

(b) Overall, 204 countries participated in the 2008 Summer Olympics, of which 55 won at least one gold medal. Do you believe that the sample of countries listed in the table is representative of this larger population? Why or why not?

39. **U.S. women's soccer—2004** Earlier, we examined data on the number of goals scored by the U.S. women's soccer team in games during the 2004 season. The dotplot below displays the goal differential for those same games, computed as U.S. score minus opponent's score.

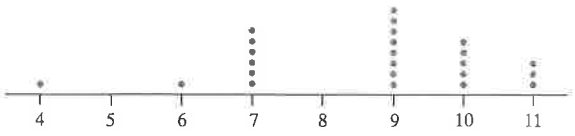
pg 28



(a) Explain what the two dots above −2 represent.

(b) What does the graph tell us about how well the team did in 2004? Be specific.

40. **Fuel efficiency** In an earlier example, we examined data on highway gas mileages of model year 2009 midsize cars. The dotplot below shows the difference (highway – city) in EPA mileage ratings for each of the 24 car models from the earlier example.



(a) Explain what the dot above 6 represents.

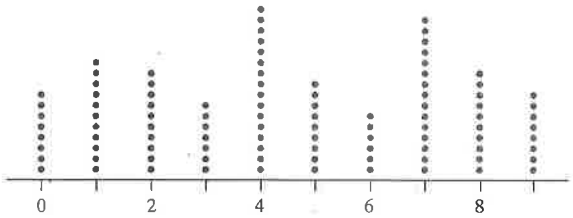
(b) What does the graph tell us about fuel economy in the city versus on the highway for these car models? Be specific.

41. **Dates on coins**

(a) Sketch a dotplot for a distribution that is skewed to the left.

(b) Suppose that you and your friends emptied your pockets of coins and recorded the year marked on each coin. The distribution of dates would be skewed to the left. Explain why.

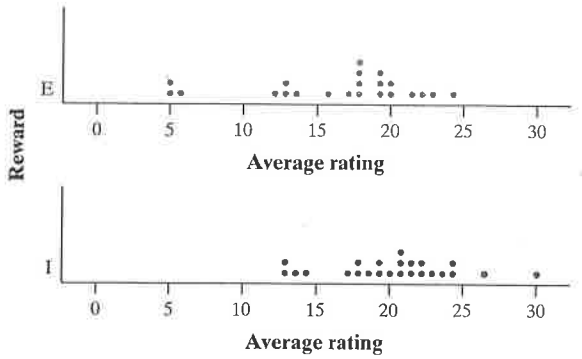
42. **Phone numbers** The dotplot below displays the last digit of 100 phone numbers chosen at random from a phone book. Describe the shape of the distribution. Does this shape make sense to you? Explain.



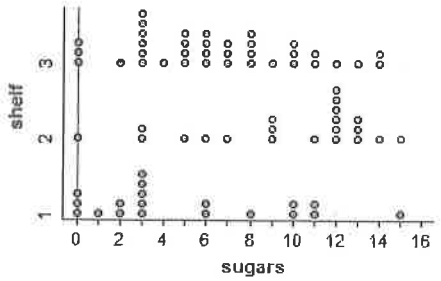
43. **Creative writing** The chapter-opening Case Study described research by Teresa Amabile investigating whether external rewards would promote creativity in children's artwork. Dr. Amabile conducted another study involving college students, who were divided into two groups using a chance process (like drawing names from a hat). 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.[24] The 12 poets' ratings of each student's poem were averaged to obtain an overall creativity score.

pg 32

A dotplot of the two groups' creativity scores is shown below. Compare the two distributions. What do you conclude about whether external rewards promote creativity?



44. **Healthy cereal?** Researchers collected data on 77 brands of cereal at a local supermarket.[25] 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. A dotplot of the data is shown below. Compare the three distributions. Critics claim that supermarkets tend to put sugary kids' cereals on lower shelves, where the kids can see them. Do the data from this study support this claim?



45. **Where do the young live?** Below is a stemplot of the percent of residents aged 25 to 34 in each of the 50 states. As in the stemplot for older residents (page 35), the stems are whole percents, and the leaves are tenths of a percent. This time, each stem has been split in two, with values having leaves 0 through 4 placed on one stem, and values ending in 5 through 9 placed on another stem.

11	44
11	66778
12	0134
12	666778888
13	0000001111444
13	7788999
14	0044
14	567
15	11
15	
16	0

Key: 12|1 means that 12.1% of that state's residents are aged 25 to 34.

(a) Why did we split stems?

(b) Utah has the highest percent of residents aged 25 to 34. What is that percent? Why do you think Utah has an unusually high percent of residents in this age group?

(c) Describe the shape, center, and spread of the distribution, ignoring Utah.

46. **Watch that caffeine!** The U.S. Food and Drug Administration (USFDA) 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 below.

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

(a) Why did we split stems?

(b) Give an appropriate key for this graph.

(c) Describe the shape, center, and spread of the distribution. Compare the caffeine content of these drinks with the USFDA's limit.

47. **El Niño and the monsoon** It appears that El Niño, the periodic warming of the Pacific Ocean west of South America, affects the monsoon rains that are essential for agriculture in India. Here are the monsoon rains (in millimeters) for the 23 strong El Niño years between 1871 and 2004:[26]

628	669	740	651	710	736	717	698	653	604	781	784
790	811	830	858	858	896	806	790	792	957	872	

(a) To make a stemplot of these rainfall amounts, round the data to the nearest 10, so that stems are hundreds of millimeters and leaves are tens of millimeters. Make two stemplots, with and without splitting the stems. Which plot do you prefer?

(b) Describe the shape, center, and spread of the distribution.

(c) The average monsoon rainfall for all years from 1871 to 2004 is about 850 millimeters. What effect does El Niño appear to have on monsoon rains?

48. **Shopping spree** A marketing consultant observed 50 consecutive shoppers at a supermarket. One variable of interest was how much each shopper spent in the store. Here are the data (in dollars), arranged in increasing order:

3.11	8.88	9.26	10.81	12.69	13.78	15.23	15.62	17.00	17.39
18.36	18.43	19.27	19.50	19.54	20.16	20.59	22.22	23.04	24.47
24.58	25.13	26.24	26.26	27.65	28.06	28.08	28.38	32.03	34.98
36.37	38.64	39.16	41.02	42.97	44.08	44.67	45.40	46.69	48.65
50.39	52.75	54.80	59.07	61.22	70.32	82.70	85.76	86.37	93.34

(a) Round each amount to the nearest dollar. Then make a stemplot using tens of dollars as the stems and dollars as the leaves.

(b) Make another stemplot of the data by splitting stems. Which of the plots shows the shape of the distribution better?

(c) Write a few sentences describing the amount of money spent by shoppers at this supermarket.

49. **Do women study more than men?** We asked the students in a large first-year college class how many minutes they studied on a typical weeknight. Here are the responses of random samples of 30 women and 30 men from the class:

Women					Men				
180	120	180	360	240	90	120	30	90	200
120	180	120	240	170	90	45	30	120	75
150	120	180	180	150	150	120	60	240	300
200	150	180	150	180	240	60	120	60	30
120	60	120	180	180	30	230	120	95	150
90	240	180	115	120	0	200	120	120	180

(a) Examine the data. Why are you not surprised that most responses are multiples of 10 minutes? Are there any responses you consider suspicious?

(b) Make a back-to-back stemplot to compare the two samples. Does it appear that women study more than men (or at least claim that they do)? Justify your answer.

50. **Basketball playoffs** Here are the scores of games played in the California Division I-AAA high school basketball playoffs:[27]

71-38	52-47	55-53	76-65	77-63	65-63	68-54	64-62
87-47	64-56	78-64	58-51	91-74	71-41	67-62	106-46

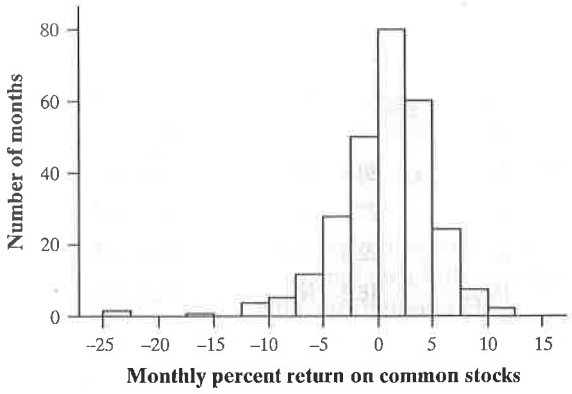
On the same day, the final scores of games in Division V-AA were

98-45	67-44	74-60	96-54	92-72	93-46
98-67	62-37	37-36	69-44	86-66	66-58

(a) Construct a back-to-back stemplot to compare the points scored by the 32 teams in the Division I-AAA playoffs and the 24 teams in the Division V-AA playoffs.

(b) Write a few sentences comparing the two distributions.

51. **Returns on common stocks** The return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percent of the beginning price. The figure below shows a histogram of the distribution of the monthly returns for all common stocks listed on U.S. markets from January 1985 to September 2007 (273 months).[28] The extreme low outlier represents the market crash of October 1987, when stocks lost 23% of their value in one month.



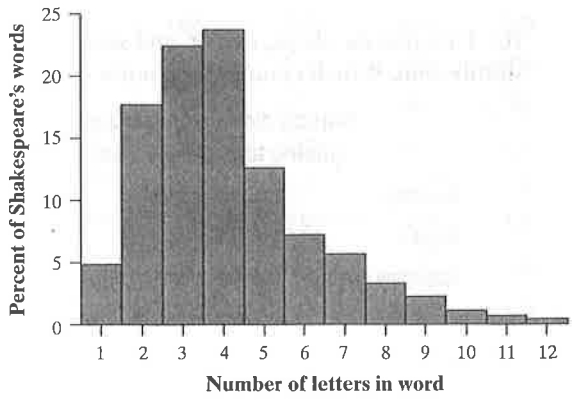
(a) Ignoring the outliers, describe the overall shape of the distribution of monthly returns.

(b) What is the approximate center of this distribution?

(c) Approximately what were the smallest and largest monthly returns, leaving out the outliers?

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

52. **Shakespeare** The histogram below shows the distribution of lengths of words used in Shakespeare's plays.[29] Describe the shape, center, and spread of this distribution.



53. **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) for workers in each state and the District of Columbia who are at least 16 years old and don't work at home.[30]

AL	23.6	LA	25.1	OH	22.1
AK	17.7	ME	22.3	OK	20.0
AZ	25.0	MD	30.6	OR	21.8
AR	20.7	MA	26.6	PA	25.0
CA	26.8	MI	23.4	RI	22.3
CO	23.9	MN	22.0	SC	22.9
CT	24.1	MS	24.0	SD	15.9
DE	23.6	MO	22.9	TN	23.5
FL	25.9	MT	17.6	TX	24.6
GA	27.3	NE	17.7	UT	20.8
HI	25.5	NV	24.2	VT	21.2
ID	20.1	NH	24.6	VA	26.9
IL	27.9	NJ	29.1	WA	25.2
IN	22.3	NM	20.9	WV	25.6
IA	18.2	NY	30.9	WI	20.8
KS	18.5	NC	23.4	WY	17.9
KY	22.4	ND	15.5	DC	29.2

(a) Make a histogram of the travel times using classes of width 2 minutes, starting at 14 minutes. That is, the first class is 14 to 16 minutes, the second is 16 to 18 minutes, and so on.

(b) The shape of the distribution is a bit irregular. Is it closer to symmetric or skewed? About where is the center of the data? What is the spread in terms of the smallest and largest values? Are there any outliers?

54. **Carbon dioxide emissions** Burning fuels in power plants and motor vehicles emits carbon dioxide (CO_2), which contributes to global warming. The table below displays CO_2 emissions per person from countries with populations of at least 20 million.[31]

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

(b) Describe the shape, center, and spread of the distribution. Which countries are outliers?

Carbon dioxide emissions (metric tons per person)

Country	CO_2	Country	CO_2
Algeria	2.6	Egypt	2.0
Argentina	3.6	Ethiopia	0.1
Australia	18.4	France	6.2
Bangladesh	0.3	Germany	9.9
Brazil	1.8	Ghana	0.3
Canada	17.0	India	1.1
China	3.9	Indonesia	1.6
Colombia	1.3	Iran	6.0
Congo	0.2	Iraq	2.9
Italy	7.8	Romania	4.2
Japan	9.5	Russia	10.8
Kenya	0.3	Saudi Arabia	13.8
Korea, North	3.3	South Africa	7.0
Korea, South	9.3	Spain	7.9
Malaysia	5.5	Sudan	0.3
Mexico	3.7	Tanzania	0.1
Morocco	1.4	Thailand	3.3
Myanmar	0.2	Turkey	3.0
Nepal	0.1	Ukraine	6.3
Nigeria	0.4	United Kingdom	8.8
Pakistan	0.8	United States	19.6
Peru	1.0	Uzbekistan	4.2
Philippines	0.9	Venezuela	5.4
Poland	7.8	Vietnam	1.0

55. **DRP test scores** There are many ways to measure the reading ability of children. One frequently used test is the Degree of Reading Power (DRP). In a research study on third-grade students, the DRP was administered to 44 students.[32] Their scores were:

40	26	39	14	42	18	25	43	46	27	19
47	19	26	35	34	15	44	40	38	31	46
52	25	35	35	33	29	34	41	49	28	52
47	35	48	22	33	41	51	27	14	54	45

Make a histogram to display the data. Write a paragraph describing the distribution of DRP scores.

56. **Drive time** Professor Moore, who lives a few miles outside a college town, records the time he takes to drive to the college each morning. Here are the times (in minutes) for 42 consecutive weekdays:

8.25	7.83	8.30	8.42	8.50	8.67	8.17	9.00	9.00	8.17	7.92
9.00	8.50	9.00	7.75	7.92	8.00	8.08	8.42	8.75	8.08	9.75
8.33	7.83	7.92	8.58	7.83	8.42	7.75	7.42	6.75	7.42	8.50
8.67	10.17	8.75	8.58	8.67	9.17	9.08	8.83	8.67		

Make a histogram of these drive times. Is the distribution roughly symmetric, clearly skewed, or neither? Are there any clear outliers?

57. **The statistics of writing style** Numerical data can distinguish different types of writing, and sometimes even individual authors. Here are data on the percent of words of 1 to 15 letters used in articles in *Popular Science* magazine:[33]

Length:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Percent:	3.6	14.8	18.7	16.0	12.5	8.2	8.1	5.9	4.4	3.6	2.1	0.9	0.6	0.4	0.2

(a) Make a histogram of this distribution. Describe its shape, center, and spread.

(b) How does the distribution of lengths of words used in *Popular Science* compare with the similar distribution for Shakespeare's plays in Exercise 52? Look in particular at short words (2, 3, and 4 letters) and very long words (more than 10 letters).

58. **Chest out, Soldier!** In 1846, a published paper provided chest measurements (in inches) of 5738 Scottish militiamen. The table below summarizes the data.

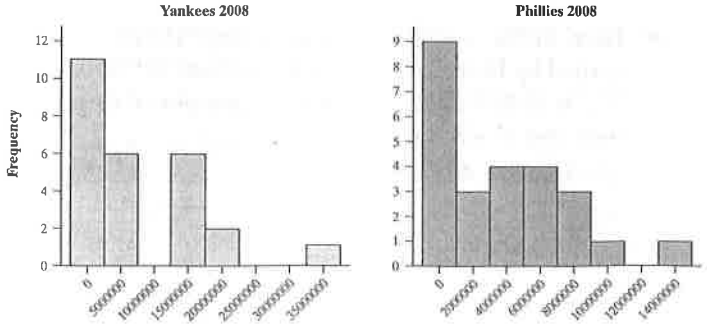
Chest size	Count	Chest size	Count
33	3	41	934
34	18	42	658
35	81	43	370
36	185	44	92
37	420	45	50
38	749	46	21
39	1073	47	4
40	1079	48	1

Source: Online Data and Story Library (DASL).

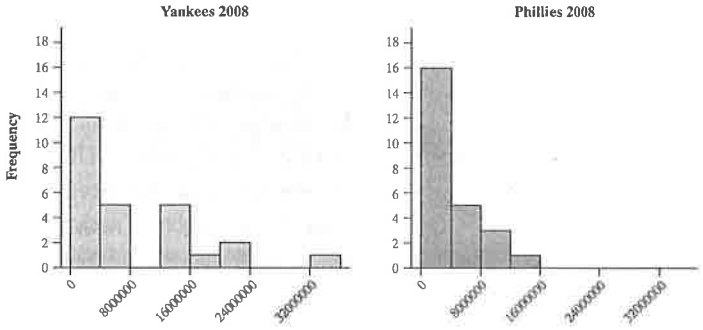
(a) Make a histogram.

(b) Describe the shape, center, and spread of the chest measurements distribution. Why might this information be useful?

59. **Paying for championships** Does paying high salaries lead to more victories in professional sports? The New York Yankees have long been known for having Major League Baseball's highest team payroll. And over the years, the team has won many championships. This strategy didn't pay off in 2008, when the Philadelphia Phillies won the World Series. Maybe the Yankees didn't spend enough money that year. The graph below shows histograms of the salary distributions for the two teams during the 2008 season. Why can't you use this graph to effectively compare the team payrolls?



60. **Paying for championships** Refer to Exercise 59. Here is another graph of the 2008 salary distributions for the Yankees and the Phillies. Write a few sentences comparing these two distributions.



61. **Birth months** Imagine asking a random sample of 60 students from your school about their birth months. Draw a plausible graph of the distribution of birth months. (*Hint:* Should you use a bar graph or a histogram?)

62. **Die rolls** Imagine rolling a fair, six-sided die 60 times. Draw a plausible graph of the distribution of die rolls. (*Hint:* Should you use a bar graph or a histogram?)

63. **Who makes more?** A manufacturing company is reviewing the salaries of its full-time employees below the executive level at a large plant. The clerical staff is almost entirely female, while a majority of the production workers and technical staff are male. As a result, the distributions of salaries for male and female employees may be quite different. The table below gives the frequencies and relative frequencies for women and men.

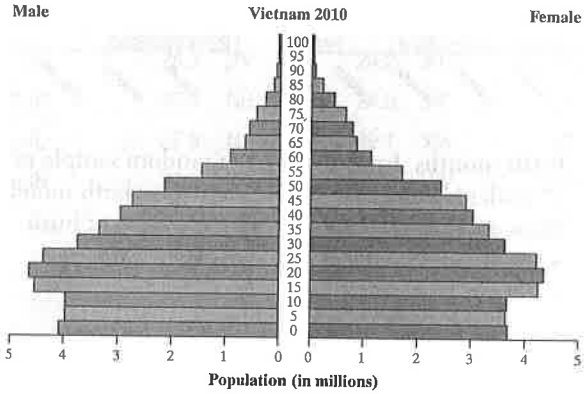
	Women		Men	
Salary ($1000)	Number	%	Number	%
10–15	89	11.8	26	1.1
15–20	192	25.4	221	9.0
20–25	236	31.2	677	27.6
25–30	111	14.7	823	33.6
30–35	86	11.4	365	14.9
35–40	25	3.3	182	7.4
40–45	11	1.5	91	3.7
45–50	3	0.4	33	1.3
50–55	2	0.3	19	0.8
55–60	0	0.0	11	0.4
60–65	0	0.0	0	0.0
65–70	1	0.1	3	0.1
Total	**756**	**100.1**	**2451**	**99.9**

(a) Explain why the total for women is greater than 100%.

(b) Make histograms for these data, choosing the vertical scale that is most appropriate for comparing the two distributions.

(c) Write a few sentences comparing the salary distributions for men and women.

64. **Population pyramids** A population pyramid is a helpful graph for examining the distribution of a country's population. Here is a population pyramid for Vietnam in the year 2010. Describe what the graph tells you about Vietnam's population that year. Be specific.



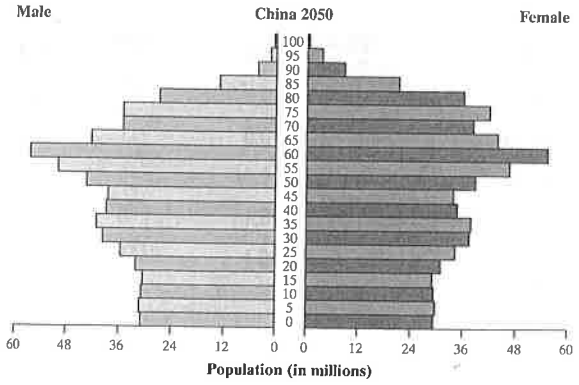
65. **Comparing AP scores** The table below gives the distribution of grades earned by students taking the AP Calculus AB and AP Statistics exams in 2009.[34]

		Grade				
	No. of exams	5	4	3	2	1
Calculus AB	230,588	52,921	43,140	41,204	35,843	57,480
Statistics	116,876	14,353	26,050	28,276	22,283	25,914

(a) Make an appropriate graphical display to compare the grade distributions for AP Calculus AB and AP Statistics.

(b) Write a few sentences comparing the two distributions of exam grades.

66. **Population pyramids** Refer to Exercise 64. Here is a graph of the projected population distribution for China in the year 2050. Describe what the graph suggests about China's future population. Be specific.



67. **Student survey** A survey of a large high school class asked the following questions:

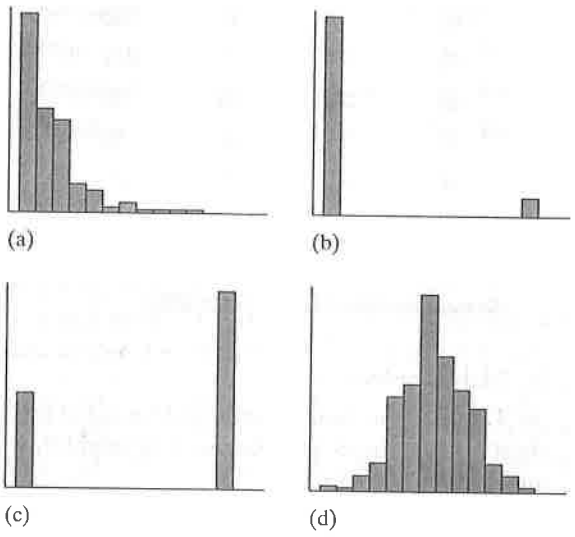
(i) Are you female or male? (In the data, male = 0, female = 1.)

(ii) Are you right-handed or left-handed? (In the data, right = 0, left = 1.)

(iii) What is your height in inches?

(iv) How many minutes do you study on a typical weeknight?

The figure below shows histograms of the student responses, in scrambled order and without scale markings. Which histogram goes with each variable? Explain your reasoning.



68. **Choose a graph** What type of graph or graphs would you plan to make in a study of each of the following issues at your school? Explain your choices.

(a) Which radio stations are most popular with students?

(b) How many hours per week do students study?

(c) How many calories do students consume per day?

Multiple choice: Select the best answer for Exercises 69 to 74.

69. Here are the amounts of money (cents) in coins carried by 10 students in a statistics class: 50, 35, 0, 97, 76, 0, 0, 87, 23, 65. To make a stemplot of these data, you would use stems

(a) 0, 1, 2, 3, 4, 5, 6, 7, 8, 9.

(b) 0, 2, 3, 5, 6, 7, 8, 9.

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

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

(e) None of these.

70. One of the following 12 scores was omitted from the stemplot below:

84 76 92 92 88 96 68 80 92 88 76 96

```
6 | 8
7 | 66
8 | 0488
9 | 2266
```

The missing number is

(a) 76. (b) 88. (c) 90. (d) 92. (e) 96.

71. You look at real estate ads for houses in Naples, Florida. There are many houses ranging from $200,000 to $500,000 in price. The few houses on the water, however, have prices up to $15 million. The distribution of house prices will be

(a) skewed to the left.

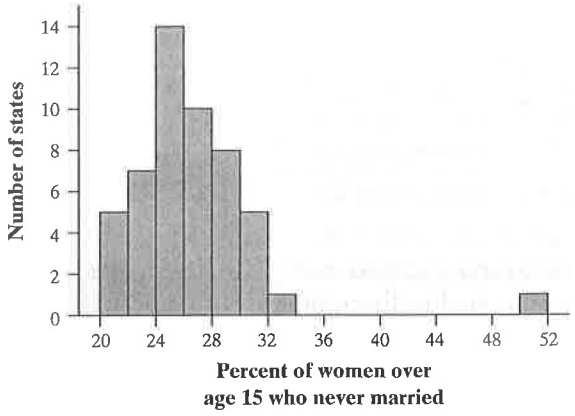
(b) roughly symmetric.

(c) skewed to the right.

(d) unimodal.

(e) too high.

Exercises 72 to 74 refer to the following setting. The histogram below shows the distribution of the percents of women aged 15 and over who have never married in each of the 50 states and the District of Columbia.



72. The leftmost bar in the histogram covers percents of never-married women ranging from about

(a) 20% to 24%. (d) 0% to 5%.

(b) 20% to 22%. (e) None of these.

(c) 0% to 20%.

73. The center of this distribution is in the interval

(a) 22% to 24%. (d) 28% to 30%.

(b) 24% to 26%. (e) 36% to 38%.

(c) 26% to 28%.

74. In about what percent of states have at least 30% of women aged 15 and over never married?

(a) 4% (b) 7% (c) 10% (d) 14% (e) 32%

75. **Baseball players** (Introduction) Here is a small part of a data set that describes Major League Baseball players as of opening day of the 2009 season:

Player	Team	Position	Age	Height	Weight	Salary
Rodriguez, Alex	Yankees	Infielder	33	6-3	230	33,000,000
Ramirez, Manny	Dodgers	Outfielder	36	6-0	200	23,854,494
Santana, Johan	Mets	Pitcher	30	6-0	210	18,876,139
Zambrano, Carlos	Cubs	Pitcher	27	6-5	255	18,750,000
Suzuki, Ichiro	Mariners	Outfielder	35	5-11	170	18,000,000

(a) What individuals does this data set describe?

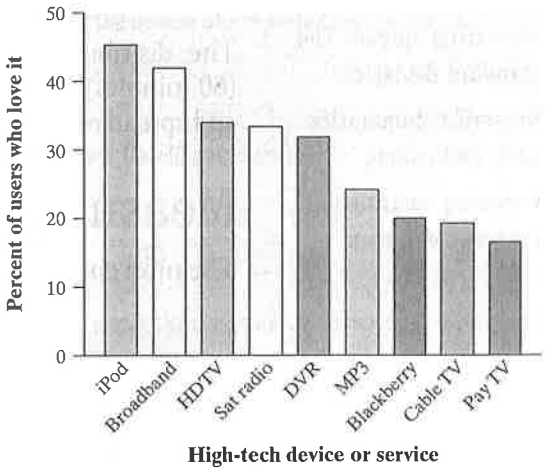
(b) In addition to the player's name, how many variables does the data set contain? Which of these variables are categorical and which are quantitative?

(c) What do you think are the units of measurement for each of the quantitative variables?

76. **I love my iPod!** (1.1) The rating service Arbitron asked adults who used several high-tech devices and services whether they "loved" using them. Below is a graph of the percents who said they did.[35]

(a) Summarize what this graph tells you in a sentence or two.

(b) Would it be appropriate to make a pie chart of these data? Why or why not?



77. **Risks of playing soccer** (1.1) A study in Sweden looked at former elite soccer players, people who had played soccer but not at the elite level, and

people of the same age who did not play soccer. Here is a two-way table that classifies these individuals by whether or not they had arthritis of the hip or knee by their mid-fifties:[36]

	Elite	Non-Elite	Did not play
Arthritis	10	9	24
No arthritis	61	206	548

(a) What percent of the people in this study were elite soccer players? What percent had arthritis?

(b) What percent of the elite soccer players had arthritis? What percent of those who had arthritis were elite soccer players?

78. **Risks of playing soccer** (1.1) Refer to Exercise 77. We suspect that the more serious soccer players have more arthritis later in life. Do the data confirm this suspicion? Give graphical and numerical evidence to support your answer.

1.3

Describing Quantitative Data with Numbers

In Section 1.3, you'll learn about:

- Measuring center: The mean
- Measuring center: The median
- Comparing the mean and the median
- Measuring spread: The interquartile range (*IQR*)
- Identifying outliers
- The five-number summary and boxplots
- Measuring spread: The standard deviation
- Numerical summaries with technology
- Choosing measures of center and spread

How long do people spend traveling to work? The answer may depend on where they live. Here are the travel times in minutes for 15 workers in North Carolina, chosen at random by the Census Bureau:[37]

30 20 10 40 25 20 10 60 15 40 5 30 12 10 10

We aren't surprised that most people estimate their travel time in multiples of 5 minutes. Here is a stemplot of these data:

0	5
1	000025
2	005
3	00
4	00
5	
6	0

Key: 2|5 is a NC worker who travels 25 minutes to work.

The distribution is single-peaked and right-skewed. The longest travel time (60 minutes) may be an outlier. Our goal in this section is to describe the center and spread of this and other distributions of quantitative data with numbers.

Measuring Center: The Mean

The most common measure of center is the ordinary arithmetic average, or **mean.**

DEFINITION: The mean $\bar{x}$

To find the **mean** $\bar{x}$ (pronounced "x-bar") of a set of observations, add their values and divide by the number of observations. If the n observations are $x_1, x_2, \ldots, x_n$, their mean is

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$